Supplements for "Crowdsourcing Hypothesis Tests"

Supplements

Table of Contents:

SUPPLEMENT 1 - Pre-registration of materials and analyses for the Main Studies

Research Teams	
Team	**Members**
1	Jay Van Bavel (New York University); Jennifer Ray (New York University); Diego Reinero (New York University); William Brady (New York University); Julian Wills (New York University)
2	Chris Bauman (University of California, Irvine); Elizabeth Mullen (San Jose State University)
3	Adam Hahn (University of Cologne); Simone Dohle (University of Cologne)
4	Kai Chi Yam (National University of Singapore); Jared Koh (National University of Singapore); Runkun Su (Sun Yat-sen University)
5	Miaolei Jia (National University of Singapore); Isabel Ding (National University of Singapore)
6	Jun Pang (Renmin University of China)
7	Michael Hall (University of Michigan); Walter Sowden (University of Michigan)
8	Benoit Monin (Stanford University); Jesse Reynolds (Stanford University)
9	William Jiménez-Real (Universidad de los Andes); Andres Montealegre (Universidad de los Andes)
10	Xiaobing Xu (Tsinghua University); Xiaoyu Yang (Tsinghua University)
11	Justin Landy (University of Chicago); Daniel Walco (University of Chicago); Daniel Bartels (University of Chicago)
12	Andrei Cimpian (University of Illinois); Christina Tworek (University of Illinois); Daniel Storage (University of Illinois)
13	M. Brent Donnellan (Texas A&M University); Richard Lucas (Michigan State University); Felix Cheung (Michigan State University); David Johnson (Michigan State University)

Research Questions

1. When directly asked, do people explicitly self-report an awareness of harboring negative automatic associations with members of negatively stereotyped social groups?

2. Are negotiators who make extreme first offers trusted more, less, or the same relative to negotiators who make moderate first offers?

3. What are the effects of continuing to work despite having no material/financial need to work on moral judgments of that individual-- beneficial, detrimental, or no effect?

4. Part of why people are opposed to the use of performance enhancing drugs in sports is because they are "against the rules". But which contributes more to this judgment - whether the performance enhancer is against the law, or whether it is against the rules established by a more proximal authority (e.g., the league)?

5. Is a utilitarian vs. deontological moral orientation related to personal happiness?

General Notes on Materials

Condition Names and Presentation Orders

In naming each team's conditions, we have adhered as closely to the language the teams originally used in their materials as possible. This means that there is some variation in how analogous conditions are named, between teams. For instance, for Research Question 3, Team 2's "Continue to Work" condition and Team 7's "Intrinsically Motivated" condition are conceptually very similar. To make it easier to see which conditions are analogous to one another, we present the materials below in the same orders for each team:

- Hypothesis 1: Most teams have only one condition for this hypothesis, but if they have two, we present the condition with the direct question(s) about stigmatized group(s) first, followed by the comparison/control condition second.

- Hypothesis 2: We present the "extreme offer" condition (or whatever the team called it) first, then the "moderate offer" condition second.

- Hypothesis 3: We present the "continues to work despite having no financial need" condition first, and the comparison condition second.

- Hypothesis 4: We present the "banned but legal" condition first, and the "illegal but not banned" condition second.

- Hypothesis 5: We present the measure(s) of moral orientation first, and the measure(s) of happiness second.

Question Names

We have labeled all materials below with question names from the Qualtrics survey. All data from this project will be made publicly available upon publication, so these question names serve as a key for researchers working with this data. For the sake of consistency, we have adhered to the following format in naming all questions.

All question names begin with the team number, then the hypothesis number, then the condition. So, they all begin with alphanumeric strings like "*11_1_Ctrl*" or "*7_3_Work*". The condition is indicated by a short word or abbreviation related to the condition names. For Hypothesis 5, the measure(s) of moral orientation are always called "Moral" (e.g., for Team 11, they begin with "*11_5_Moral*") and the measure(s) of happiness are always called "Happy" (e.g., for Team 11, they begin with "*11_5_Happy*").

Each type of question has another descriptor following this string, explaining what kind of question it is. These are standardized:

- Introductory questions explaining the task, presenting a scenario, etc., have "*_Intro*" at the end. For example, Team 11's scenario for Hypothesis 3 is named "*11_3_Ret_Intro*" in the "Retire" condition, and "*11_3_Work_Intro*" in the "Work" condition. If there is more than one, they are numbered, starting with "*_Intro1*", then "*_Intro2*", etc.

- Dependent variables have "*_DV*" at the end of the question name. If there is more than one, they are numbered, starting with "*_DV1*", then "*_DV2*", etc. For instance, Team 11 has two dependent variable for Hypothesis 3, which are named (in the "Work" condition) "*11_3_Work_DV1*" and "*11_3_Work_DV2*".

- Filler questions have "*_F*" at the end of the question name. If there is more than one, they are numbered as above, starting with "*_F1*", then "*_F2*", and so forth.

- Manipulation checks have "*_MC*" at the end. Again, if there is more than one, they are numbered starting with "*_MC1*", then "*_MC2*", and so on. Some manipulation check questions are used for excluding participants, while others are not, depending on what a team requested. We detail this below in the Analysis Plans.

General Notes on Data Collection and Analyses

Sample Size and Stopping Rule

We will collect as many subjects for the project as are available on Amazon Mechanical Turk, which we estimate to be approximately 3900 to 7800 participants in total or 300-600 participants per study version. In the large online data collection, each Mechanical Turk worker will complete five sets of materials, each corresponding to one of the five key hypotheses being tested in the project. For each of the five hypotheses, they will be randomly assigned to complete one of the twelve or more study versions testing that hypothesis created by the independent research teams. Thus, a total sample size of 3900-7800 corresponds to approximately 300-600 participants per study version: 150-300 participants per cell for a between-subjects experiment and 300-600 observations for a within-subjects experiment or correlational design. There will also be a stand-alone satellite condition testing Hypothesis 5 with a long set of study materials that takes too much time to complete to be paired with the materials for Hypotheses 1-4, again with 300-600 participants collected.

Data collection will begin on Monday, May 15, 2017. We will cease data collection when recruitment rates drop below 15 participants in a day, or when the study has run for 25 consecutive weekdays (i.e., five weeks, Friday June 16, 2017) or we reach 7,500 total participants (the maximum our budget permits), whichever comes first.

For all study versions as well as for the project as a whole, the data will only be analyzed once, after we have exhausted the supply of available research participants on Mechanical Turk and downloaded the final set of data.

Overview of Analyses

The primary analyses for this project are meta-analytic in nature. We will examine the level of support obtained for each of the five unpublished findings listed above. The key variables in this analysis will be the effect sizes, which will be meta-analyzed. For Research Questions 1-4, all effect sizes will be converted to uncorrected Cohen's *d*s (i.e., single-sample and independent-groups *d*s), and for Question 5, all derived effect sizes will be Pearson *r*s. For each research question, a mean effect size will be calculated, weighting each effect size by the inverse of its sampling variance. This mean will be compared against a null hypothesis of zero (i.e., no effect). We will also examine the median effect size and range of observed effect sizes for each unpublished finding.

We will also use more simplistic (but conceptually simpler) "count" methods to assess whether the crowdsourced materials replicated the original findings. The relevant variables are again the effect sizes described above. We will count two outcomes: how many teams' materials produced effect sizes in the same direction as the original results, and how many produced statistically significant effect sizes in the same direction as the original results? Effect sizes will simply be counted as “successful” or “failed” replications in these analyses.

For each of the five research questions, we will test for heterogeneity of effect sizes, using the *Q* statistic. Once again, the variables in this analysis are the effect sizes. This analysis asks, essentially, “do the decisions made by researchers significantly impact the observed effect sizes?” A significant *Q* statistic means that there is heterogeneity among the effect sizes, and we can reject the null hypothesis that none of the observed effect sizes for a given research question differ from each other. Because all participants were drawn from the same large sample and randomly assigned to conditions, heterogeneity in effect sizes cannot be attributed to “hidden moderators” and must be due to differences in the materials designed by different teams. We will further quantify the amount of heterogeneity observed using the I^2 statistic, which indicates the percentage of variance among effect sizes attributable to heterogeneity, rather than sampling variance. By convention, I^2 values of 25%, 50%, and 75% indicate low, moderate, and high levels of heterogeneity, respectively (Higgins, Thompson, Deeks, & Altman, 2003).

Further, we will compare the amount of variance in effect sizes attributable to the hypothesis being tested, versus the team that developed the materials. Specifically, we will compute intraclass correlation coefficients of hypotheses across teams, and teams across hypotheses (see Klein et al., 2014 for a similar analysis). If the former is substantially larger than the latter, it would suggest that much of the variance in effect sizes observed in scientific research is attributable to the truth or falsehood of the underlying hypothesis, as is typically assumed. If the latter is substantially larger, it would suggest that much of the variance in effect sizes in

scientific research is not due to the truth or falsehood of hypotheses *per se,* but the skill of the researcher designing the study.

Additionally, we will conduct a survey in which independent scientists rate the quality of the materials developed to test each hypothesis (to be pre-registered separately). What is important for our purposes here is that this study will involve fellow scientists rating the quality of each set of materials on a 1-10 scale (1= not at all informative, 10= extremely informative). We will repeat the analyses above, including only effect sizes estimated from materials rated as a 6 or above, on average, by the independent scientists. Further, we will correlate the quality ratings from independent scientists with observed effect sizes, to examine whether higher-quality materials (as assessed via peer evaluations) show greater support for the hypotheses.

Calculating Effect Size Estimates and Sampling Variances

There is no agreed-upon method for calculating the sampling variance of a single-sample Cohen's *d* effect size, and there are multiple proposed methods for calculating the sampling variance of a repeated-measures *d* (e.g., Lipsey & Wilson, 2001; Morris & DeShon, 2002). We will therefore use a bootstrapping approach to calculating effect size estimates and sampling variances, for consistency across designs. Responses to each team's materials for each research question will be resampled 10,000 times with replacement and the effect size for each resample will be calculated. The mean of these 10,000 effect size estimates will be taken as the point estimate for the true effect size, and the variance of the distribution of 10,000 effect sizes will be taken to be a distribution-free estimate of the sampling variance of the true effect size (Landy & Montoya, in progress). R code for these calculations is appended to this preregistration. The formulas that will be used to calculate each effect size are:

- Independent-groups *d*: $d_{IG} = (M_1 - M_2) / SD_{Pool}$, where M_1 and M_2 are the observed means of the two conditions, and SD_{Pool} is the pooled standard deviation (see Lipsey & Wilson, 2001; Morris & DeShon, 2002).

- Single-sample *d*: $d_{SS} = (M - \mu) / SD$, where M is the observed mean, μ is the mean under the null hypothesis, and SD is the observed standard deviation (see Lipsey & Wilson, 2001).

- Repeated-measures *d:* $d_{RM} = M_{Diff} / SD_{Diff}$, where M_{Diff} is the mean of the difference scores between the two variables of interest, and SD_{Diff} is the standard deviation of these difference scores. To make these repeated-measures *d*s comparable to the independentgroups and single-sample *d*s above, they will be converted to independent-groups *d*s, correcting for within-subjects correlations, using Equation 11 from Morris & DeShon (2002).

For Hypothesis 5, the effect size will be the Pearson correlation between the measure of moral orientation and the measure of happiness.

Study Materials and Detailed Analysis Plans

Research Question 1: When directly asked, do people explicitly self-report an awareness of harboring negative automatic associations with members of negatively stereotyped social groups?

Team 1 Materials

1_1_Mat_DV Regardless of my explicit (i.e. conscious) beliefs about social equality, I believe I possess automatic (i.e. unconscious) negative associations towards members of stigmatized social groups.

- ❍ Strongly Disagree
- ❍ Somewhat Disagree
- ❍ Neither Agree nor Disagree
- ❍ Somewhat Agree
- ❍ Strongly Agree

Team 1 Analysis Plan

The DV will be responses to question *1_1_Mat_DV*. We will compare the mean of these responses to a null hypothesis of $\mu = 3$ (Neither Agree nor Disagree) using a one-sample t-test. The effect size estimate will be a single-sample Cohen's *d*.

Team 2 Materials

2_1_Mat_Intro Automatic associations (also known as implicit attitudes) are evaluations that people make about others without effort or conscious awareness. Automatic associations often are consistent with stereotypes in society about a particular group. For example, many people automatically associate youth with good and old-age with bad, and others automatically associate women with family and men with careers. Importantly, a person who has an automatic association does not necessarily agree with the stereotype or act in a discriminatory way toward others. Automatic associations are just mental connections, and people's conscious values and ideals can override them. In other words, automatic associations are like initial, default settings that you might override upon further consideration. Do you think you have some automatic associations for certain groups? Please answer the questions as honestly as possible.

2_1_Mat_DV1 In your initial reaction, how much do you tend to associate youth with good and old-age with bad?

- ❍ not at all
- ❍ slightly
- ❍ moderately
- ❍ much
- ❍ very much

2_1_Mat_DV2 In your initial reaction, how much do you tend to associate women with family-based roles and men with career-based roles?

- ❍ not at all
- ❍ slightly
- ❍ moderately
- ❍ much
- ❍ very much

2_1_Mat_DV3 In your initial reaction, how much do you tend to associate fat people with bad and thin people with good?

- ❍ not at all
- ❍ slightly
- ❍ moderately
- ❍ much
- ❍ very much

2_1_Mat_DV4 In your initial reaction, how much do you tend to associate straight/heterosexual people with good and gay/homosexual people with bad?

- ❍ not at all
- ❍ slightly
- ❍ moderately
- ❍ much
- ❍ very much

2_1_Mat_MC How strongly do you agree or disagree with the following statement: It is possible for people to have some negative automatic associations for certain groups of people that they are not consciously aware of?

- ❍ strongly disagree
- ❍ moderately disagree
- ❍ slightly disagree
- ❍ neither agree nor disagree
- ❍ slightly agree
- ❍ moderately agree

❍ strongly agree

Team 2 Analysis Plan
The DV will be the mean of responses to questions *2_1_Mat_DV1*, *2_1_Mat_DV2*, *2_1_Mat_DV3*, and *2_1_Mat_DV4*. We will compare the mean of this composite DV to a null hypothesis of $\mu = 3$ (Moderately) using a one-sample t-test. The effect size estimate will be a single-sample Cohen's *d*. If this composite DV shows poor internal reliability ($\alpha < .70$), we will use *2_1_Mat_DV3* as the DV instead of the composite.

Team 3 Materials
3_1_Mat_Intro Over recent decades psychological research has pointed to the fact that there might be different kinds of attitudes. Specifically, it has been claimed that there are attitudes people express when they are asked about them - let's call them "explicit attitudes". However, additionally, people may sometimes hold positive or negative associations with people that are triggered automatically when they think about them or encounter them. And these automatic associations may or may not be at odds with people's consciously held beliefs. For example, you may consciously know that a fashion model's body is not a healthy or realistic ideal one should attain, and yet on some level you may still harbor positive automatic associations when you think of supermodels and think they look beautiful. In this study we are interested in such automatically triggered associations people may harbor towards different people or groups. These automatic associations may or may not conflict with consciously held beliefs.

3_1_Mat_DV Think about the following groups. What is your first automatic reaction when you think about them? We are interested in whether you think you harbor more positive or more negative automatic associations with these groups. Please rate your associations with the following groups.

	1 More negative automatic associations	2	3	4 Neither negative nor positive automatic associations	5	6	7 More positive automatic associations
African Americans	❍	❍	❍	❍	❍	❍	❍
Latino Americans	❍	❍	❍	❍	❍	❍	❍
White Americans	❍	❍	❍	❍	❍	❍	❍
Asian Americans	❍	❍	❍	❍	❍	❍	❍

Gay people	❍	❍	❍	❍	❍	❍	❍
Straight people	❍	❍	❍	❍	❍	❍	❍

Team 3 Analysis Plan

Difference scores between responses to racial minorities (African Americans, Latino Americans, and Asian Americans) and responses to White Americans will be computed for each participant. We will then average these together to form a composite measure of acknowledged implicit racial bias. Similarly, we will compute difference scores between responses Gay people and Straight people to measure acknowledged implicit sexual-orientation bias. The primary DV will be the average of these two measures. However, we will exclude racial bias measures from participants who identify as non-White, and sexual-orientation bias measures from participants who do not identify as straight. If the primary DV shows poor internal reliability ($\alpha < .70$), we will use only the composite measure of acknowledged implicit racial bias. If this measure also shows poor reliability, ($\alpha < .70$), we will use only the difference score between responses to African Americans and White Americans. Whichever DV is ultimately used, it will be compared to a null hypothesis of $\mu = 0$ (no difference between implicit reactions to minority and majority groups) using a one-sample t-test. The effect size estimate will be a single-sample Cohen's *d*.

Team 4 Materials

4_1_Mat_Intro In the following space, please write down a paragraph describing an individual holding a morally tainted job (e.g., exotic dancers, collection agents, abortion clinic medical staff, etc.). In your writing, please describe what you think of his/her job and his/her daily schedule. When you are done, please respond to the following survey questions.

4_1_Mat_DV

	1 strongly disagree	2	3	4	5	6	7 strongly agree
This person is conscientious	❍	❍	❍	❍	❍	❍	❍
This person is the life of the party	❍	❍	❍	❍	❍	❍	❍
Without too much deliberation, I think this person is bad	❍	❍	❍	❍	❍	❍	❍

This person gets chores done right away	❍	❍	❍	❍	❍	❍	❍
This person has frequent mood swings	❍	❍	❍	❍	❍	❍	❍
This person has a vivid imagination.	❍	❍	❍	❍	❍	❍	❍
I intuitively think badly about this person	❍	❍	❍	❍	❍	❍	❍
This person gets upset easily	❍	❍	❍	❍	❍	❍	❍
This person makes a mess of things	❍	❍	❍	❍	❍	❍	❍
I am aware of my automatic bias against this individual	❍	❍	❍	❍	❍	❍	❍

Team 4 Analysis Plan

The DV will be the mean of responses to "without too much deliberation, I think this person is bad, "I intuitively think badly about this person", and "I am aware of my automatic bias against this person". We will compare the mean of this composite DV to a null hypothesis of $\mu = 4$ using a one-sample t-test. The effect size estimate will be a single-sample Cohen's *d*. If this composite DV shows poor internal reliability ($\alpha < .70$), we will use responses to "I intuitively think badly about this person" as the DV instead of the composite.

Team 5 Materials

5_1_Mat_Intro Please write down a negatively stereotyped social group that you can think of:

5_1_Mat_DV Are you aware that you harbor negative automatic associations with the social group that you just named?

❍ Yes

❍ No

Team 5 Analysis Plan

The DV will be the proportion of "yes" responses to *5_1_Mat_DV*. We will compare the mean of this DV ("yes" = 1, "no" = 0) to a null hypothesis of $\mu = 0.5$ using a one-sample t-test. The effect size estimate will be a single-sample Cohen's *d*.

Team 6 Materials

6_1_Mat_Intro In this study, we are interested in people's opinions of different things. Please read the questions carefully and then provide your answers. There are no right or wrong answers. We are only interested in your real opinions.

6_1_Mat_DV1 Are you aware that you have automatic negative associations with Lesbian, Gay, Bisexual and Transgender (LGBT) persons?

❍ Yes, I am.

❍ No, I am not.

6_1_Mat_DV2 Are you aware that you have automatic negative associations with ethnic minorities?

❍ Yes, I am.

❍ No, I am not.

Team 6 Analysis Plan

The DV will be the average of the proportion of "yes" responses to *6_1_Mat_DV1* and *6_1_Mat_DV2*. We will compare the mean of this composite DV ("yes" = 1, "no" = 0) to a null hypothesis of $\mu = 0.5$ using a one-sample t-test. The effect size estimate will be a single-sample Cohen's *d*.

Team 7 Materials

General Note: The design of this team's method is somewhat hard to understand from the materials presented below. In short:

- Participants indicate their political party affiliation
- Participants who indicate that they are "weakly affiliated" with either party are excused from participation
- Participants who indicate that they are "strongly affiliated" with a party read that Charles is affiliated with the opposition party (i.e., strongly affiliated Democrats read that Charles is a Republican, and strongly affiliated Republicans read that Charles is a Democrat), a group about whom they likely have negative stereotypes. They then indicate how their reaction when thinking about Charles.

- Participants who indicate that they are "not affiliated" with either party are randomly assigned to read the Charles is a Democrat or a Republican, because neither constitutes a stereotyped out-group. They then indicate how their reaction when thinking about Charles.

7_1_ExpA_Intro1 Please mark the following that best represents your political party affiliation:
❍ strongly affiliated with the Republican Party
❍ weakly affiliated with the Republican Party
❍ not affiliated with either the Republican or the Democratic Party
❍ weakly affiliated with the Democratic Party
❍ strongly affiliated with the Democratic Party

7_1_ExpA_Intro2 Charles is a Democrat and voted for Hillary Clinton.

7_1_ExpA_Intro3 Charles is a Democrat and voted for Hillary Clinton.

7_1_ExpA_Intro4 Charles is a Republican and voted for Donald Trump.

7_1_ExpA_DV Mark the emoji below that best represents your initial reaction when thinking about Charles.

7_1_ExpA_F1 Thank you for completing this survey. Please move to the next study by clicking the button below.

7_1_ExpB_Intro1 Please mark the following that best represents your political party affiliation:
❍ strongly affiliated with the Republican Party
❍ weakly affiliated with the Republican Party
❍ not affiliated with either the Republican or the Democratic Party
❍ weakly affiliated with the Democratic Party
❍ strongly affiliated with the Democratic Party

7_1_ExpB_Intro2 Charles is a Republican and voted for Donald Trump.

7_1_ExpB_Intro3 Charles is a Democrat and voted for Hillary Clinton.

7_1_ExpB_Intro4 Charles is a Republican and voted for Donald Trump.

7_1_ExpB_DV Mark the emoji below that best represents your initial reaction when thinking about Charles.

7_1_ExpB_F1 Thank you for completing this survey. Please move to the next study by clicking the button below.

Team 7 Analysis Plan

Participants who identify as "weakly affiliated" with either the Republican or Democratic party will not be analyzed. "Strongly affiliated" partisans (who will read that Charles supports the party they oppose) will be compared with "politically neutral" participants (who will be randomly assigned to read that Charles is a Democrat or a Republican).

The DV will be responses to *7_1_ExpA_DV* or *7_1_ExpB_DV*. We will compare strong partisans to politically neutral participants using an independent-samples t-test. The effect size will be an independent-groups Cohen's *d* comparing strong partisans to politically neutral participants.

Team 8 Materials

8_1_Mat_DV Some people argue that we all internalize some of the negative social stereotypes propagated by the media and the society we grew up in, and that we therefore automatically base much of our first impressions of the people we meet on those stereotypes, whether we like it or not.

Think about your reactions to **obese people**. Do you think you may harbor automatic negative associations towards members of that group (regardless of what you <u>want</u> to think about them)? Take a second to look at the picture and notice what automatic reactions it elicits in you. We ask you to be as honest and candid as you can in your responses. How much would do you agree with the following statements using the following scale?

	Strongly Disagree	Disagree	Slightly Disagree	Don't Know	Slightly Agree	Agree	Strongly Agree
I do not harbor any automatic negative associations towards obese people.	❍	❍	❍	❍	❍	❍	❍
When I meet obese people, I sometimes catch myself having negative thoughts about them just because of their weight.	❍	❍	❍	❍	❍	❍	❍
I could definitely be sexually attracted to an obese person.	❍	❍	❍	❍	❍	❍	❍
I do not ever assume that obese people have lower self-control.	❍	❍	❍	❍	❍	❍	❍

To be honest I find the naked bodies of obese people rather disgusting.	❍	❍	❍	❍	❍	❍	❍
I am confident that seeing obese people does not ever bring negative thoughts to my mind.	❍	❍	❍	❍	❍	❍	❍
If a new coworker was obese, I would immediately expect that they are less intelligent than if they were average-size.	❍	❍	❍	❍	❍	❍	❍

Team 8 Analysis Plan

Participant's height and weight will be asked in the demographics section, and used to compute BMI (BMI = [[Weight in pounds]/[Height in inches]^2] X 703). Obese participants (BMI > 30) will be excluded from analysis.

The DV will be the mean of responses to the items in *8_1_Mat_DV* (items 1, 3, 4, and 6 will be reverse-scored). We will compare the mean of these responses to a null hypothesis of $\mu = 4$ ("Don't know") using a one-sample t-test. The effect size estimate will be a single-sample Cohen's *d*. If this composite DV shows poor internal reliability ($\alpha < .70$), we will use responses to item 1 ("I do not harbor any automatic negative associations towards obese people" [reverse-scored]).

Team 9 Materials

9_1_Bia_DV In the present task you will be asked some questions about your social attitudes. For each of the following statements please indicate your level of agreement from 1 (strongly disagree) to 7 (strongly agree).

	1	2	3	4	5	6	7
Even though I know it's not appropriate, I sometimes feel that I hold unconscious negative attitudes toward Blacks	❍	❍	❍	❍	❍	❍	❍
When talking to Black people, I sometimes worry that I am unintentionally acting in a prejudiced way	❍	❍	❍	❍	❍	❍	❍
Even though I like Black people, I still worry that I have unconscious biases toward Blacks	❍	❍	❍	❍	❍	❍	❍
I never worry that I may be acting in a subtly prejudiced way toward Blacks	❍	❍	❍	❍	❍	❍	❍

Team 9 Analysis Plan

The DV will be the mean of responses to the items in *9_1_Mat_DV* (item 4 will be reverse-scored). We will compare the mean of these responses to a null hypothesis of $\mu = 4$ using a one-sample t-test. The effect size estimate will be a single-sample Cohen's *d*. If this composite DV shows poor internal reliability ($\alpha < .70$), we will use responses to item 3 ("Even though I like Black people, I still worry that I have unconscious biases toward Blacks").

Team 10 Materials

10_1_Mod_Intro The following are examples of some negative stereotypes. Please rate your agreement with the sentences, and the extent to which you are aware that you harbor similar associations with the sentences.

10_1_Mod_MC1 1. Athletes are terrible at managing money.
(1) To what extent do you agree with this sentence?

- ❍ 1 Not at all
- ❍ 2
- ❍ 3
- ❍ 4
- ❍ 5
- ❍ 6
- ❍ 7 Very much

10_1_Mod_DV1 (2) To what extent are you aware that you harbor associations consistent with the belief expressed in this sentence?

- ❍ 1 Not at all
- ❍ 2
- ❍ 3
- ❍ 4
- ❍ 5
- ❍ 6
- ❍ 7 Very much

10_1_Mod_MC2 2. Men who like pink are effeminate.
(1) To what extent do you agree with this sentence?

- ❍ 1 Not at all
- ❍ 2
- ❍ 3
- ❍ 4
- ❍ 5
- ❍ 6
- ❍ 7 Very much

10_1_Mod_DV2 (2) To what extent are you aware that you harbor associations consistent with the belief expressed in this sentence?

- ❍ 1 Not at all
- ❍ 2
- ❍ 3
- ❍ 4
- ❍ 5
- ❍ 6
- ❍ 7 Very much

10_1_Mod_MC3 3. Asian women are submissive, and really bad drivers.
(1) To what extent do you agree with this sentence?

- ❍ 1 Not at all
- ❍ 2
- ❍ 3
- ❍ 4
- ❍ 5
- ❍ 6
- ❍ 7 Very much

10_1_Mod_DV3 (2) To what extent are you aware that you harbor associations consistent with the belief expressed in this sentence?

- ❍ 1 Not at all
- ❍ 2
- ❍ 3
- ❍ 4
- ❍ 5
- ❍ 6
- ❍ 7 Very much

10_1_Mod_MC4 4. All politicians are philanderers and think only of personal gain and benefit.
(1) To what extent do you agree with this sentence?

- ❍ 1 Not at all

❍ 2

❍ 3

❍ 4

❍ 5

❍ 6

❍ 7 Very much

10_1_Mod_DV4 (2) To what extent are you aware that you harbor associations consistent with the belief expressed in this sentence?

❍ 1 Not at all

❍ 2

❍ 3

❍ 4

❍ 5

❍ 6

❍ 7 Very much

Team 10 Analysis Plan

Difference scores will be calculated by subtracting each *MC* question from its corresponding *DV* (e.g., *10_1_Mat_DV1 – 10_1_Mat_MC1*). Positive difference scores indicate awareness of biases in the *DV* questions not explicitly endorsed in the *MC* questions. The DV will be the mean of these four difference scores. We will compare this DV to a null hypothesis of $\mu = 0$ using a one-sample t-test. The effect size will be a single-sample Cohen's *d*. If this composite DV shows poor internal reliability ($\alpha < .70$), we will use difference scores computed by subtracting *10_1_Mat_MC3* from *10_1_Mat_DV3*.

Team 11 Materials

11_1_Exp_DV Over the past several decades, research has demonstrated that humans can "think" in two very different ways. One is what we usually mean by "thinking" - it is conscious and intentional. The other is something quite different –very fast, automatic responses to things in the world – what we sometimes call "gut reactions." For instance, if you see a wild bear, you do not need to put in effort to figure out that it is dangerous - you just immediately associate the bear with danger, and react. This kind of thinking is very efficient, and usually serves us well. However, it can also produce reactions that we may not agree with, upon reflection. For instance, many people have automatic negative reactions to certain social groups, even though they do not want to, and even though they do not consciously endorse these reactions. This study is about those kinds of reactions. Specifically, we want to know how negative or positive your *immediate, gut reaction* is to members of each of the groups below, regardless of what your more reasoned,

reflective reaction is. You may not consciously agree with your *immediate, gut reaction*, but we want you to tell us what it is, as honestly as you can.
How negative is your *immediate, gut reaction* to each of the following groups?

	Not at all negative				Somewhat negative				Extremely negative
Gay Men	❍	❍	❍	❍	❍	❍	❍	❍	❍
African Americans	❍	❍	❍	❍	❍	❍	❍	❍	❍
Muslims	❍	❍	❍	❍	❍	❍	❍	❍	❍
Jewish Americans	❍	❍	❍	❍	❍	❍	❍	❍	❍
College Students	❍	❍	❍	❍	❍	❍	❍	❍	❍
Women	❍	❍	❍	❍	❍	❍	❍	❍	❍

11_1_Ctrl_DV Over the past several decades, research has demonstrated that humans can "think" in two very different ways. One is what we usually mean by "thinking" - it is conscious and intentional. The other is something quite different - very fast, automatic responses to things in the world – what we sometimes call "gut reactions." For instance, if you see a wild bear, you do not need to put in effort to figure out that it is dangerous - you just immediately associate the bear with danger, and react. This kind of thinking is very efficient, and usually serves us well. However, it can also produce reactions that we may not agree with, upon reflection. For instance, many people have automatic negative reactions to certain social groups, even though they do not want to, and even though they do not consciously endorse these reactions. This study is about those kinds of reactions. Specifically, we want to know how negative or positive your *immediate, gut reaction* is to members of each of the groups below, regardless of what your more reasoned, reflective reaction is. You may not consciously agree with your *immediate, gut reaction*, but we want you to tell us what it is, as honestly as you can.
How negative is your *immediate, gut reaction* to each of the following groups?

	Not at all negative				Somewhat negative				Extremely negative
Straight Men	❍	❍	❍	❍	❍	❍	❍	❍	❍

White Americans	❍	❍	❍	❍	❍	❍	❍	❍	❍
Christians	❍	❍	❍	❍	❍	❍	❍	❍	❍
Jewish Americans	❍	❍	❍	❍	❍	❍	❍	❍	❍
College Students	❍	❍	❍	❍	❍	❍	❍	❍	❍
Women	❍	❍	❍	❍	❍	❍	❍	❍	❍

Team 11 Analysis Plan
The filler groups (Jewish Americans, college students, women) will not be analyzed. The DV will be the mean of responses to the critical groups (*11_1_Exp_DV:* gay men, African Americans, Muslims; *11_1_Ctrl_DV:* straight men, White Americans, Christians). We will compare the experimental and control conditions using an independent-samples t-test. The effect size will be an independent-groups Cohen's *d*. If this composite DV shows poor internal reliability ($\alpha < .70$), we will use responses to "Muslims" and "Christians".

Team 12 did not develop materials for this research question.

Team 13 Materials
13_1_Mat_Intro Recent psychological research suggests that people have immediate 'gut level' reactions to other people and objects. Sometimes those reactions are negative and other times they are positive. Of course, people can ignore these unintentional and automatic reactions but they serve as a starting point for interactions.
In this task we will show you pictures of exemplars from different groups. We are interested in whether you are AWARE that you have such automatic reactions to members of this group.

13_1_Mat_DV1

Are you aware that you have an immediate “gut level” reaction towards this person?

- ❍ Yes. I have a Strong Negative Reaction
- ❍ Yes. I have a Negative Reaction
- ❍ Yes. I have a Slight Negative Reaction
- ❍ No. I am unaware of Any Reaction
- ❍ Yes. I have a Slight Positive Reaction
- ❍ Yes. I have a Positive Reaction
- ❍ Yes. I have a Strong Positive Reaction

13_1_Mat_DV2

Are you aware that you have an immediate “gut level” reaction towards this person?

- ❍ Yes. I have a Strong Negative Reaction
- ❍ Yes. I have a Negative Reaction
- ❍ Yes. I have a Slight Negative Reaction
- ❍ No. I am unaware of Any Reaction
- ❍ Yes. I have a Slight Positive Reaction
- ❍ Yes. I have a Positive Reaction
- ❍ Yes. I have a Strong Positive Reaction

13_1_Mat_F1

Are you aware that you have an immediate “gut level” reaction towards this person?

- ❍ Yes. I have a Strong Negative Reaction
- ❍ Yes. I have a Negative Reaction
- ❍ Yes. I have a Slight Negative Reaction
- ❍ No. I am unaware of Any Reaction
- ❍ Yes. I have a Slight Positive Reaction
- ❍ Yes. I have a Positive Reaction
- ❍ Yes. I have a Strong Positive Reaction

13_1_Mat_DV3

Are you aware that you have an immediate “gut level” reaction towards this person?

- ❍ Yes. I have a Strong Negative Reaction
- ❍ Yes. I have a Negative Reaction
- ❍ Yes. I have a Slight Negative Reaction
- ❍ No. I am unaware of Any Reaction
- ❍ Yes. I have a Slight Positive Reaction
- ❍ Yes. I have a Positive Reaction
- ❍ Yes. I have a Strong Positive Reaction

13_1_Mat_DV4

Are you aware that you have an immediate “gut level” reaction towards this person?

- ❍ Yes. I have a Strong Negative Reaction
- ❍ Yes. I have a Negative Reaction
- ❍ Yes. I have a Slight Negative Reaction
- ❍ No. I am unaware of Any Reaction
- ❍ Yes. I have a Slight Positive Reaction
- ❍ Yes. I have a Positive Reaction
- ❍ Yes. I have a Strong Positive Reaction

13_1_Mat_F2

Are you aware that you have an immediate “gut level” reaction towards this person?

- ❍ Yes. I have a Strong Negative Reaction
- ❍ Yes. I have a Negative Reaction
- ❍ Yes. I have a Slight Negative Reaction
- ❍ No. I am unaware of Any Reaction
- ❍ Yes. I have a Slight Positive Reaction
- ❍ Yes. I have a Positive Reaction
- ❍ Yes. I have a Strong Positive Reaction

13_1_Mat_F3

Are you aware that you have an immediate “gut level” reaction towards this animal?

- ❍ Yes. I have a Strong Negative Reaction
- ❍ Yes. I have a Negative Reaction
- ❍ Yes. I have a Slight Negative Reaction
- ❍ No. I am unaware of Any Reaction
- ❍ Yes. I have a Slight Positive Reaction
- ❍ Yes. I have a Positive Reaction
- ❍ Yes. I have a Strong Positive Reaction

13_1_Mat_F4

Are you aware that you have an immediate "gut level" reaction towards this animal?

- ❍ Yes. I have a Strong Negative Reaction
- ❍ Yes. I have a Negative Reaction
- ❍ Yes. I have a Slight Negative Reaction
- ❍ No. I am unaware of Any Reaction
- ❍ Yes. I have a Slight Positive Reaction
- ❍ Yes. I have a Positive Reaction
- ❍ Yes. I have a Strong Positive Reaction

Team 13 Analysis Plan

The DV will be the mean of responses to *13_1_Mat_1, 13_1_Mat_2, 13_1_Mat_3,* and *13_1_Mat_4* (i.e., the four members of stigmatized groups [Blacks and overweight individuals]). We will compare this DV to a null hypothesis of $\mu = 4$ (No. I am unaware of any reaction) using a one-sample t-test. The effect size will be a single-sample Cohen's *d*. The effect size will be an independent-groups Cohen's *d*. If this composite DV shows poor internal reliability ($\alpha < .70$), we will use responses to *13_1_Mat_DV4.*

Original Materials

14_1_Mat_DV1 Although I don't necessarily agree with them, I sometimes have prejudiced feelings (like gut reactions or spontaneous thoughts) that I don't feel I can prevent.

- ❍ 1 strongly disagree
- ❍ 2
- ❍ 3

❍ 4

❍ 5

❍ 6

❍ 7 strongly agree

14_1_Mat_DV2 At times stereotypical thoughts about minorities coming into my head without my necessarily intending them to.

❍ 1 strongly disagree

❍ 2

❍ 3

❍ 4

❍ 5

❍ 6

❍ 7 strongly agree

Original Materials Analysis Plan

The DV will be the mean of responses to *14_1_Mat_1* and *14_1_Mat_2*. We will compare this DV to a null hypothesis of $\mu = 4$ using a one-sample t-test. The effect size will be a single-sample Cohen's *d*. If this composite DV shows poor internal reliability ($\alpha < .70$), we will use responses to *14_1_Mat_DV1*.

Research Question 2: Are negotiators who make extreme first offers trusted more, less, or the same relative to negotiators who make moderate first offers?

Team 1 Materials

Extreme Low condition

1_2_ELow_Intro Imagine you are about to enter a business negotiation where you are the buyer and the other person is the seller. Although you would like to buy their product for as little as possible, your advisors have informed you that the seller's product is likely valued around $100,000. You have both agreed that they will make the first offer. The seller goes on to propose $130,000 as their initial offer. Please answer the following question honestly.

1_2_ELow_DV To what extent do you feel you can trust your partner?

- ❍ Strongly distrust
- ❍ Moderately distrust
- ❍ Somewhat distrust
- ❍ Neither trust nor distrust
- ❍ Somewhat trust
- ❍ Moderately trust
- ❍ Strongly trust

Extreme High condition

1_2_EHigh_Intro Imagine you are about to enter a business negotiation where you are the seller and the other person is the buyer. Although you would like to sell your product for as much as possible, your advisors have informed you that your product is likely valued around $100,000. You have both agreed that they will make the first offer. The buyer goes on to propose $70,000 as their initial offer. Please answer the following question honestly.

1_2_EHigh_DV To what extent do you feel you can trust your partner?

- ❍ Strongly distrust
- ❍ Moderately distrust
- ❍ Somewhat distrust
- ❍ Neither trust nor distrust
- ❍ Somewhat trust
- ❍ Moderately trust
- ❍ Strongly trust

Moderate Low Condition
1_2_MLow_Intro Imagine you are about to enter a business negotiation where you are the buyer and the other person is the seller. Although you would like to buy their product for as little as possible, your advisors have informed you that the seller's product is likely valued around $100,000. You have both agreed that they will make the first offer. The seller goes on to propose $110,000 as their initial offer. Please answer the following question honestly.

1_2_MLow_DV To what extent do you feel you can trust your partner?
- ❍ Strongly distrust
- ❍ Moderately distrust
- ❍ Somewhat distrust
- ❍ Neither trust nor distrust
- ❍ Somewhat trust
- ❍ Moderately trust
- ❍ Strongly trust

Moderate High condition
1_2_MHigh_Intro Imagine you are about to enter a business negotiation where you are the seller and the other person is the buyer. Although you would like to sell your product for as much as possible, your advisors have informed you that your product is likely valued around $100,000. You have both agreed that they will make the first offer. The buyer goes on to propose $90,000 as their initial offer. Please answer the following question honestly.

1_2_MHigh_DV To what extent do you feel you can trust your partner?
- ❍ Strongly distrust
- ❍ Moderately distrust
- ❍ Somewhat distrust
- ❍ Neither trust nor distrust
- ❍ Somewhat trust
- ❍ Moderately trust
- ❍ Strongly trust

Team 1 Analysis Plan
We will collapse across the low and high conditions (full data will be made public, so a more complete analysis can be undertaken later). Responses from the *DV* questions in the extreme and moderate conditions will be compared using an independent-samples t-test. The effect size will be an independent-groups Cohen's *d*.

Team 2 Materials

Extreme Condition

2_2_Ext_Intro1 Imagine you owned a car that was recently destroyed beyond repair in an accident. No one was hurt and your insurance company gave you a check for damages in the amount of $7,300. You're now shopping for a used car and $7,300 is your budget limit. You've thought carefully about your options and concluded that your ideal car would be a Nissan Altima that is about 8–10 years old. You've done some research and found that this kind of car, in good shape with low mileage, typically sells for $6,500 to $7,500.

You recently read an ad for a 2008 Nissan Altima. Everything looked promising: low mileage (about 50,000), in good shape, nice color. It's being sold by someone from out of town who inherited it from a relative. You arranged to get the keys from a neighbor of the relative and take it for a test drive. Everything looked good. You'd like to get this car if possible. You can't go over the limit from your insurance payout ($7,300). You're hoping to pay well under that amount, which would allow you to use that much needed cash for other things.

You have exchanged emails with the seller and inquired about their asking price. They replied by saying that they were looking to get $8,400 for the car.

2_2_Ext_MC1 How reasonable was the seller's asking price?

- ❍ Very unreasonable
- ❍ moderately unreasonable
- ❍ slightly unreasonable
- ❍ neither reasonable or unreasonable
- ❍ slightly reasonable
- ❍ moderately reasonable
- ❍ Very reasonable

2_2_Ext_Intro2 Now imagine that the negotiation is over. You were able to reach a deal with the seller; you agreed to purchase the car for $7,000, which is $300 under the limit that you were given above.

Now imagine that the researchers running this study were going to give you the remaining $300. However, they also gave you two options about what to do with the money:

OPTION 1: The researchers would send you a check for the $300, and that would be the end.

OPTION 2: You could send some or all of your $300 to the person with whom you just negotiated. Whatever amount you keep would be yours, just as in Option 1. But, whatever amount you sent to the seller in the negotiation you just completed would triple, and then that person would have the option of returning any portion of the tripled amount back to you. In

other words, you could send the seller $0, $5, $100, the entire $300—any portion of the $300 you wanted. Anything you sent will multiply by a factor of three. For example, if you sent $300, then the seller would receive $900, or if you sent $5, then the seller would receive $15. Then, the seller would decide how much money (if any) to return you.

The more you sent to the seller, the more money the seller could potentially return to you. However, the seller would not be required to give anything back to you. Therefore, you could end up with more or less than the $300 you started with, depending on what the seller decides.

2_2_Ext_DV How much of the $300 would you send to the seller? Enter a number from $0 to $300:

Moderate Condition

2_2_Mod_Intro1 Imagine you owned a car that was recently destroyed beyond repair in an accident. No one was hurt and your insurance company gave you a check for damages in the amount of $7,300. You're now shopping for a used car and $7,300 is your budget limit. You've thought carefully about your options and concluded that your ideal car would be a Nissan Altima that is about 8–10 years old. You've done some research and found that this kind of car, in good shape with low mileage, typically sells for $6,500 to $7,500.

You recently read an ad for a 2008 Nissan Altima. Everything looked promising: low mileage (about 50,000), in good shape, nice color. It's being sold by someone from out of town who inherited it from a relative. You arranged to get the keys from a neighbor of the relative and take it for a test drive. Everything looked good. You'd like to get this car if possible. You can't go over the limit from your insurance payout ($7,300). You're hoping to pay well under that amount, which would allow you to use that much needed cash for other things.

You have exchanged emails with the seller and inquired about their asking price. They replied by saying that they were looking to get $7,400 for the car.

2_2_Mod_MC1 How reasonable was the seller's asking price?

- ❍ Very unreasonable
- ❍ moderately unreasonable
- ❍ slightly unreasonable
- ❍ neither reasonable or unreasonable
- ❍ slightly reasonable
- ❍ moderately reasonable
- ❍ Very reasonable

2_2_Mod_Intro2 Now imagine that the negotiation is over. You were able to reach a deal with the seller; you agreed to purchase the car for $7,000, which is $300 under the limit that you were given above.

Now imagine that the researchers running this study were going to give you the remaining $300. However, they also gave you two options about what to do with the money:

OPTION 1: The researchers would send you a check for the $300, and that would be the end.

OPTION 2: You could send some or all of your $300 to the person with whom you just negotiated. Whatever amount you keep would be yours, just as in Option 1. But, whatever amount you sent to the seller in the negotiation you just completed would triple, and then that person would have the option of returning any portion of the tripled amount back to you. In other words, you could send the seller $0, $5, $100, the entire $300—any portion of the $300 you wanted. Anything you sent will multiply by a factor of three. For example, if you sent $300, then the seller would receive $900, or if you sent $5, then the seller would receive $15. Then, the seller would decide how much money (if any) to return you.

The more you sent to the seller, the more money the seller could potentially return to you. However, the seller would not be required to give anything back to you. Therefore, you could end up with more or less than the $300 you started with, depending on what the seller decides.

2_2_Mod_DV How much of the $300 would you send to the seller? Enter a number from $0 to $300:

Team 2 Analysis Plan

Responses to *2_2_Ext_DV* and *2_2_Mod_DV* will be compared using an independent-samples t-test. The effect size will be an independent-groups *Cohen's* d.

Team 3 Materials

Extreme Condition

3_2_Ext_Intro Imagine you moved to a new city. You would like to rent a 1-bedroom apartment in the city centre. According to your friends and information you find on websites, the typical rent for a 1-bedroom apartment in that city ranges between $1000 – $1600 per month, but that it's quite common to negotiate, such that landlords may sometimes, though not always, start with a higher price than they actually expect to get in the end. You visit the first 1-bedroom apartment. George, the landlord, shows you around. You like the apartment - it is modern and in good condition. George starts negotiating with you about the rent. He tells you that the monthly rent for this apartment is $3200. What is your impression of George at this point? Please answer the following question:

3_2_Ext_DV How trustworthy do you think George is?

❍ 1 Not at all trustworthy

❍ 2

❍ 3

❍ 4

❍ 5

❍ 6

❍ 7 Very trustworthy

Moderate Condition

3_2_Mod_Intro Imagine you moved to a new city. You would like to rent a 1-bedroom apartment in the city centre. According to your friends and information you find on websites, the typical rent for a 1-bedroom apartment in that city ranges between $1000 – $1600 per month, but that it's quite common to negotiate, such that landlords may sometimes, though not always, start with a higher price than they actually expect to get in the end. You visit the first 1-bedroom apartment. George, the landlord, shows you around. You like the apartment - it is modern and in good condition. George starts negotiating with you about the rent. He tells you that the monthly rent for this apartment is $1600. What is your impression of George at this point? Please answer the following question:

3_2_Mod_DV How trustworthy do you think George is?

❍ 1 Not at all trustworthy

❍ 2

❍ 3

❍ 4

❍ 5

❍ 6

❍ 7 Very trustworthy

Team 3 Analysis Plan

Responses to *3_2_Ext_DV* and 3*_2_Mod_DV* will be compared using an independent-samples t-test. The effect size will be an independent-groups Cohen's *d*.

Team 4 Materials

Extreme Condition

4_2_Ext_Intro **In this study, you are asked to put yourself in the shoes of a buyer negotiating the price of the items you intend to purchase. Please read the short passage and answer the questions indicated below.**

Imagine that you and a friend are in Istanbul on vacation. You want to bring home gifts for friends and family members and so you decide to go to the Grand Bazaar for the afternoon. The Grand Bazaar is the world's largest covered market. You are positive that you will be able to find

a few Turkish trinkets to bring home with you. After perusing the goods for 30 minutes, you happen upon a set of bangles that you think would make the perfect present for your sister. In the bazaar, the norm is to try to negotiate with the shopkeeper over the price. You know that other stalls are selling similar bangles for $20 USD. You approach the shopkeeper and ask him the price of the bangles. He tells you that they are $80 USD for the set.

4_2_Ext_DV Based on the scenario, I feel that

	1 strongly disagree	2	3	4	5	6	7 strongly agree
The shopkeeper's offer is extreme	❍	❍	❍	❍	❍	❍	❍
I can trust the shopkeeper	❍	❍	❍	❍	❍	❍	❍
There are times when the shopkeeper cannot be trusted	❍	❍	❍	❍	❍	❍	❍
The shopkeeper is truly sincere	❍	❍	❍	❍	❍	❍	❍

Moderate Condition

4_4_Mod_Intro **In this study, you are asked to put yourself in the shoes of a buyer negotiating the price of the items you intend to purchase. Please read the short passage and answer the questions indicated below.**

Imagine that you and a friend are in Istanbul on vacation. You want to bring home gifts for friends and family members and so you decide to go to the Grand Bazaar for the afternoon. The Grand Bazaar is the world's largest covered market. You are positive that you will be able to find a few Turkish trinkets to bring home with you. After perusing the goods for 30 minutes, you happen upon a set of bangles that you think would make the perfect present for your sister. In the bazaar, the norm is to try to negotiate with the shopkeeper over the price. You know that other stalls are selling similar bangles for $20 USD. You approach the shopkeeper and ask him the price of the bangles. He tells you that they are $25 USD for the set.

4_2_Mod_DV Based on the scenario, I feel that

	1 strongly disagree	2	3	4	5	6	7 strongly agree
The shopkeeper's offer is extreme	❍	❍	❍	❍	❍	❍	❍
I can trust the shopkeeper	❍	❍	❍	❍	❍	❍	❍
There are times when the shopkeeper cannot be trusted	❍	❍	❍	❍	❍	❍	❍
The shopkeeper is truly sincere	❍	❍	❍	❍	❍	❍	❍

Team 4 Analysis Plan

The DV will be the mean of responses to items 2-4 in *4_2_Ext_DV* and *4_2_Mod_DV* (item 3 will be reverse-scored). We will compare this composite DV in the extreme and moderate conditions using an independent-samples t-test. The effect size will be an independent-groups Cohen's *d*. If this composite DV shows poor internal reliability ($\alpha < .70$), we will use responses to item 2 (I can trust the shopkeeper).

Team 5 Materials

Extreme Condition

5_2_Ext_Intro Imagine that you met a local businessman, Johnson, who is looking for a partner to start a business to sell household products at a local mall. After chatting with Johnson, you realize that this business might be potentially profitable, since there are many young families in the area who are interested in purchasing household products. You are considering partnering with Johnson in this new business venture. Both of you will work on the new business together. During your meeting, Johnson offers a 95/5 profit split, with him getting 95% of the profit and you getting the remaining 5% of the profit.

5_2_Ext_DV1 How trustworthy do you think Johnson is?
❍ 1 Not at all trustworthy

❍ 2

❍ 3

❍ 4

❍ 5

❍ 6

❍ 7 Very trustworthy

5_2_Ext_DV2 How credible do you think Johnson is?
❍ 1 Not at all credible

❍ 2

❍ 3

❍ 4

❍ 5

❍ 6

❍ 7 Very credible

Moderate Condition
5_2_Mod_Intro Imagine that you met a local businessman, Johnson, who is looking for a partner to start a business to sell household products at a local mall. After chatting with Johnson, you realize that this business might be potentially profitable, since there are many young families in the area who are interested in purchasing household products. You are considering partnering with Johnson in this new business venture. Both of you will work on the new business together. During your meeting, Johnson offers a 55/45 profit split, with him getting 55% of the profit and you getting the remaining 45% of the profit.

5_2_Mod_DV1 How trustworthy do you think Johnson is?
❍ 1 Not at all trustworthy

❍ 2

❍ 3

❍ 4

❍ 5

❍ 6

❍ 7 Very trustworthy

5_2_Mod_DV2 How credible do you think Johnson is?
❍ 1 Not at all credible

❍ 2

❍ 3

❍ 4

❍ 5

❍ 6

❍ 7 Very credible

Team 5 Analysis Plan
The DV will be the mean of the two *DV* questions in each condition (i.e., the mean of *5_2_Ext_DV1* and *5_2_Ext_DV2* or the mean of *5_2_Mod_DV1* and *5_2_Mod_DV2*, depending on condition). We will compare the extreme and moderate conditions using an independent-samples t-test. The effect size will be an independent-groups Cohen's *d*. If this composite DV shows poor internal reliability ($\alpha < .70$), we will use responses to *5_2_Ext_DV1* and *5_2_Mod_DV1*.

Team 6 Materials
Extreme Condition
6_2_Ext_Intro1 In this study, we are interested in people's responses to different scenarios. Please read the scenario on the next page carefully and then answer questions.

6_2_Ext_Intro2 Imagine that you are going to work in London for two years and you want to rent an apartment there. Local colleagues tell you the range of weekly rent for a one-bedroom apartment in London is between £ 85 and £140. After searching at a local renting website, you find a nice apartment that suits you well. The landlord charges £280 a week for this apartment.

6_2_Ext_DV1 How reliable do you think the landlord is?
❍ 1 not at all

❍ 2

❍ 3

❍ 4

❍ 5

❍ 6

❍ 7 very much

6_2_Ext_DV2 How trustworthy do you think the landlord is?
❍ 1 not at all

❍ 2

❍ 3

❍ 4

❍ 5

❍ 6

❍ 7 very much

6_2_Ext_DV3 How likely would you be to contact and bargain with this landlord?
❍ 1 highly unlikely

❍ 2

❍ 3

❍ 4

❍ 5

❍ 6

❍ 7 highly likely

6_2_Ext_MC1 Have you ever lived in London before?
❍ Yes

❍ No

Moderate Condition

6_2_Mod_Intro1 In this study, we are interested in people's responses to different scenarios. Please read the scenario on the next page carefully and then answer questions.

6_2_Mod_Intro2 Imagine that you are going to work in London for two years and you want to rent an apartment there. Local colleagues tell you the range of weekly rent for a one-bedroom apartment in London is between £ 85 and £140. After searching at a local renting website, you find a nice apartment that suits you well. The landlord charges £140 a week for this apartment.

6_2_Mod_DV1 How reliable do you think the landlord is?
❍ 1 not at all

❍ 2

❍ 3

❍ 4

❍ 5

❍ 6

❍ 7 very much

6_2_Mod_DV2 How trustworthy do you think the landlord is?
❍ 1 not at all

❍ 2

❍ 3

❍ 4

❍ 5

❍ 6

❍ 7 very much

6_2_Mod_DV3 How likely would you be to contact and bargain with this landlord?
❍ 1 highly unlikely

❍ 2

❍ 3

❍ 4

❍ 5

❍ 6

❍ 7 highly likely

6_2_Mod_MC Have you ever lived in London before?
❍ Yes

❍ No

Team 6 Analysis Plan
Participants who answer "yes" to *6_2_Ext_MC* or *6_2_Mod_MC* will be excluded from analysis because their knowledge of London might affect their perceptions of the landlord's offer. The

DV will be the mean of the three *DV* questions in each condition (i.e., the mean of 6*_2_Ext_DV1, 6_2_Ext_DV2,* and *6_2_Ext_DV3* or the mean of *6_2_Mod_DV1, 6_2_Mod_DV2* and *6_2_Mod_DV3,* depending on condition). We will compare the extreme and moderate conditions using an independent-samples t-test. The effect size will be an independent-groups Cohen's *d*. If this composite DV shows poor internal reliability ($\alpha < .70$), we will use responses to *6_2_Ext_DV2* and *6_2_Mod_DV2*.

Team 7 Materials

Extreme Condition

7_2_Ext_Intro Jim needs a new car, and after much research, he decides he'd like to purchase a new Subaru Impreza. He drives to the local Subaru dealership on a hot Saturday at the end of July, and his timing is no coincidence. Jim knows that because it's both the end of the month and the end of the model year, the Subaru salesmen will probably be eager to make a good deal with him; they need to hit their monthly sales quotas, and they're probably trying to sell as many of the old model year's vehicles before the new batch of vehicles arrives.

Jim arrives at the dealership, and he's enthusiastically greeted by a salesman named Scott. As expected, Scott is very friendly and is thrilled to hear that Jim wants to buy a car that day. Jim describes the characteristics of the ideal car that he'd like to buy: he wants the hatchback version, he'd like the slightly less expensive "Limited" version of the car (rather than the "Premium" version), and he likes the charcoal color best in the Impreza. "However," Jim said, "I would also consider silver or blue if there are no charcoal ones available."

Scott went over to his computer to check on the dealership's inventory of vehicles. "Jim, we have pretty much exactly what you're looking for," he said. "We have an Impreza in the back lot in the hatchback version, and it's the 'Limited' version."

"That sounds great," said Jim. "What color is it?"

"Unfortunately, charcoal is a popular color," said Scott. "So we're out of those. This one is in blue, which you said you also liked."

"Ok, let's have a look," said Jim.

Scott heads to the back lot to retrieve the car while Jim waits in front of the dealership for him to bring the car around. Scott pulls up in the car, and Jim generally likes what he sees. He definitely preferred charcoal, but the blue still looks pretty good.

"The price is $22,549, and that's after taking off $1,000 for our summer savings discount," said Scott. "Great deals right now! And this is such a fantastic car. So, what do you say?"

"It's a nice car, even though blue was not my first choice," replied Jim. "So, I'll give you $19,000 for it."

7_2_Ext_DV How trustworthy of a person is Jim?

❍ 1 not at all trustworthy

- ❍ 2
- ❍ 3
- ❍ 4
- ❍ 5
- ❍ 6
- ❍ 7 very trustworthy

Moderate Condition
7_2_Mod_Intro Jim needs a new car, and after much research, he decides he'd like to purchase a new Subaru Impreza. He drives to the local Subaru dealership on a hot Saturday at the end of July, and his timing is no coincidence. Jim knows that because it's both the end of the month and the end of the model year, the Subaru salesmen will probably be eager to make a good deal with him; they need to hit their monthly sales quotas, and they're probably trying to sell as many of the old model year's vehicles before the new batch of vehicles arrives.

Jim arrives at the dealership, and he's enthusiastically greeted by a salesman named Scott. As expected, Scott is very friendly and is thrilled to hear that Jim wants to buy a car that day. Jim describes the characteristics of the ideal car that he'd like to buy: he wants the hatchback version, he'd like the slightly less expensive "Limited" version of the car (rather than the "Premium" version), and he likes the charcoal color best in the Impreza. "However," Jim said, "I would also consider silver or blue if there are no charcoal ones available."

Scott went over to his computer to check on the dealership's inventory of vehicles. "Jim, we have pretty much exactly what you're looking for," he said. "We have an Impreza in the back lot in the hatchback version, and it's the 'Limited' version."

"That sounds great," said Jim. "What color is it?" "Unfortunately, charcoal is a popular color," said Scott. "So we're out of those. This one is in blue, which you said you also liked." "Ok, let's have a look," said Jim.

Scott heads to the back lot to retrieve the car while Jim waits in front of the dealership for him to bring the car around. Scott pulls up in the car, and Jim generally likes what he sees. He definitely preferred charcoal, but the blue still looks pretty good.

"The price is $22,549, and that's after taking off $1,000 for our summer savings discount," said Scott. "Great deals right now! And this is such a fantastic car. So, what do you say?"

"It's a nice car, even though blue was not my first choice," replied Jim. "So, I'll give you $21,500 for it."

7_2_Mod_DV How trustworthy of a person is Jim?
- ❍ 1 not at all trustworthy

❍ 2

❍ 3

❍ 4

❍ 5

❍ 6

❍ 7 very trustworthy

Team 7 Analysis Plan

The DV will be the responses to *7_2_Ext_DV* and *7_2_Mod_DV,* depending on condition. We will compare the extreme and moderate conditions using an independent-samples t-test. The effect size will be an independent-groups Cohen's *d.*

Team 8 Materials

Extreme Condition

8_2_Ext_Intro Imagine you are looking for a new car and see the ad below posted on Craigslist. Before calling about the car, you consult local retailers, online price estimating services, and your friend who is an auto dealer and find that the car is worth between $1,500 and $3000. When you call about the car the next morning, a man named Mark answers and says that he is looking for $3800 for the vehicle but is open to negotiating the price.

2000 HONDA ACCORD LX, 37K, 4 Door, White w/tan leather interior, Cruise Control, Dual Airbag, A/C, PWR Steering, PWR Mirrors and Windows, FM/CD, Alarm System - Call XXX-XXXX to make an offer!

8_2_Ext_MC1 Would you accept Mark's initial offer price?

❍ Yes

❍ No

8_2_Ext_MC2 How much would you offer Mark as a counterproposal?

8_2_Ext_DV1 Mark assures you that the car has had no major maintenance problems in the past five years. He says that he has changed the car's oil every three months, replaced the tires after 25,000 miles, and generally kept up perfect maintenance on the vehicle. How likely do you think it is that Mark is telling the truth about his car's maintenance history? Mark is...

- ❍ definitely lying
- ❍ probably lying
- ❍ possibly lying
- ❍ possibly telling the truth
- ❍ probably telling the truth
- ❍ definitely telling the truth

8_2_Ext_DV2 Mark also tells you that he has 6 other potential buyers who have left him voicemails to inquire about the price, and that because there is so much interest in his vehicle, you should decide quickly. How likely do you think it is that Mark is telling the truth about getting so many calls inquiring about the car? Mark is...

- ❍ definitely lying
- ❍ probably lying
- ❍ possibly lying
- ❍ possibly telling the truth
- ❍ probably telling the truth
- ❍ definitely telling the truth

8_2_Ext_DV3 Suppose you do decide to buy the car. The next day when Mark drops off the car at your apartment, he tells you that he forgot the title to the car. He says he really needs the money right now, but promises to bring you the title the following day. How much do you believe that he is telling the truth about bringing the car title when he says he will? Mark is...

- ❍ definitely lying
- ❍ probably lying
- ❍ possibly lying
- ❍ possibly telling the truth
- ❍ probably telling the truth
- ❍ definitely telling the truth

Moderate Condition

8_2_Mod_Intro Imagine you are looking for a new car and see the ad below posted on Craigslist. Before calling about the car, you consult local retailers, online price estimating services, and your friend who is an auto dealer and find that the car is worth between $1,500 and $3000. When you call about the car the next morning, a man named Mark answers and says that he is looking for $2800 for the vehicle but is open to negotiating the price.

2000 HONDA ACCORD LX, 37K, 4 Door, White w/tan leather interior, Cruise Control, Dual Airbag, A/C, PWR Steering, PWR Mirrors and Windows, FM/CD, Alarm System - Call XXX-XXXX to make an offer!

8_2_Mod_MC1 Would you accept Mark's initial offer price?

- ❍ Yes
- ❍ No

8_2_Mod_MC2 How much would you offer Mark as a counterproposal?

8_2_Mod_DV1 Mark assures you that the car has had no major maintenance problems in the past five years. He says that he has changed the car's oil every three months, replaced the tires after 25,000 miles, and generally kept up perfect maintenance on the vehicle. How likely do you think it is that Mark is telling the truth about his car's maintenance history? Mark is...

- ❍ definitely lying
- ❍ probably lying
- ❍ possibly lying
- ❍ possibly telling the truth
- ❍ probably telling the truth
- ❍ definitely telling the truth

8_2_Mod_DV2 Mark also tells you that he has 6 other potential buyers who have left him voicemails to inquire about the price, and that because there is so much interest in his vehicle,

you should decide quickly. How likely do you think it is that Mark is telling the truth about getting so many calls inquiring about the car? Mark is...

- ❍ definitely lying
- ❍ probably lying
- ❍ possibly lying
- ❍ possibly telling the truth
- ❍ probably telling the truth
- ❍ definitely telling the truth

8_2_Mod_DV3 Suppose you do decide to buy the car. The next day when Mark drops off the car at your apartment, he tells you that he forgot the title to the car. He says he really needs the money right now, but promises to bring you the title the following day. How much do you believe that he is telling the truth about bringing the car title when he says he will? Mark is...

- ❍ definitely lying
- ❍ probably lying
- ❍ possibly lying
- ❍ possibly telling the truth
- ❍ probably telling the truth
- ❍ definitely telling the truth

Team 8 Analysis Plan

The DV will be the mean of the three *DV* questions in each condition (i.e., the mean of *8_2_Ext_DV1*, *8_2_Ext_DV2*, and *8_2_Ext_DV3* or the mean of *8_2_Mod_DV1, 8_2_Mod_DV2* and *8_2_Mod_DV3,* depending on condition). We will compare the extreme and moderate conditions using an independent-samples t-test. The effect size will be an independent-groups Cohen's *d.* If this composite DV shows poor internal reliability ($\alpha < .70$), we will use responses to *6_2_Ext_DV1* and *6_2_Mod_DV1*.

Team 9 Materials

Extreme Condition

9_2_Ext_Intro1 In this task you will play against a hypothetical player.

This game involves two players: one is called the proposer and the other is called the responder. The game rules are as follows: At the beginning of each round, the proposer is endowed with $10 by the experimenters. The proposer will then decide how to split the $10 endowment between him/herself and the responder. For example, the proposer can decide to split the money 90/10: 90% to him/herself ($9) and 10% ($1) to the responder. After that, the responder can decide whether to accept the proposer's split or not. If the responder accepted this example split, he/she would get $1 and the proposer would get $9. If the responder thinks an offer is unfair and rejects it, both the responder and the proposer get $0.

Imagine that you play the role of the responder and we have paired you with a proposer. You will need to decide whether to accept or reject their offer. It is important to remember that your decision today would affect both how much you and the proposer would get paid. This is because if you accept their offer, we would pay you your share, and we would also send the proposer their share. If you reject the split offer on a given trial, neither you nor the proposer would get anything.

9_2_Ext_MC1 You will now see the proposed split of the $10 endowment. You will need to come to a decision (e.g. accept or reject). The proposer offered to split the money 90/10: 90% to him/herself ($9) and 10% ($1) to you.

- ❍ Accept
- ❍ Reject

9_2_Ext_Intro2 Imagine now that you will play another game with **the same hypothetical individual** with whom you played before. This individual is subject to the same rules. In this decision-making game there are two roles, A and B. In today's game, you will play the role of Player A and the other player the role of Player B.

Here are the rules of the game:
1) Player A and Player B will both receive $10.
2) Player A must decide whether to send some, none or all of the experimenter dollars to Player B.
3) Any money that Player A sends to Player B is tripled by the experimenter. For example, if Player A sends $3, Player B receives $9. If Player A sends $10, Player B receives $30. If Player A sends nothing, Player B receives nothing.
4) If Player A sends money to Player B, Player B must decide whether to send some, none, or all of the money received back to Player A. For example, if Player A sends $3, Player B receives $9 and can send any amount between $0 and $9 back to Player A.

There is only one round in this game. The game is over after Player A decides whether to send money to Player B and after Player B decides whether to send money back to Player A. The other participant has read the same game instructions as you.

At the end of the experiment, you would receive whatever amount of money out of the original $10 that you keep for yourself in addition to any amount the other participant would send back to you (out of the tripled amount that you choose to send). Your payment at the end of the experiment could be more than, less than, or equal to $10.

9_2_Ext_DV Please enter the amount of money (out of $10) that you would send to the other participant. Remember, the amount you send would be tripled. The other participant would have the option of giving some, none, or all of the tripled amount back to you.

Moderate Condition
9_2_Mod_Intro1 In this task you will play against a hypothetical player.
In this task you will play against a hypothetical player.
This game involves two players: one is called the proposer and the other is called the responder. The game rules are as follows: At the beginning of each round, the proposer is endowed with $10 by the experimenters. The proposer will then decide how to split the $10 endowment between him/herself and the responder. For example, the proposer can decide to split the money 90/10: 90% to him/herself ($9) and 10% ($1) to the responder. After that, the responder can decide whether to accept the proposer's split or not. If the responder accepted this example split, he/she would get $1 and the proposer would get $9. If the responder thinks an offer is unfair and rejects it, both the responder and the proposer get $0.
Imagine that you play the role of the responder and we have paired you with a proposer. You will need to decide whether to accept or reject their offer. It is important to remember that your decision today would affect both how much you and the proposer would get paid. This is because if you accept their offer, we would pay you your share, and we would also send the proposer their share. If you reject the split offer on a given trial, neither you nor the proposer would get anything.
9_2_Mod_MC1 You will now see the proposed split of the $10 endowment. You will need to come to a decision (e.g. accept or reject). The proposer offered to split the money 50/50: 50% to him/herself ($5) and 50% ($5) to you.
❍ Accept

❍ Reject

9_2_Mod_Intro2 Imagine now that you will play another game with **the same hypothetical individual** with whom you played before. This individual is subject to the same rules. In this decision-making game there are two roles, A and B. In today's game, you will play the role of Player A and the other player the role of Player B.

Here are the rules of the game:
1) Player A and Player B will both receive $10.
2) Player A must decide whether to send some, none or all of the experimenter dollars to Player B.
3) Any money that Player A sends to Player B is tripled by the experimenter. For example, if Player A sends $3, Player B receives $9. If Player A sends $10, Player B receives $30. If Player A sends nothing, Player B receives nothing.
4) If Player A sends money to Player B, Player B must decide whether to send some, none, or all of the money received back to Player A. For example, if Player A sends $3, Player B receives $9 and can send any amount between $0 and $9 back to Player A.

There is only one round in this game. The game is over after Player A decides whether to send money to Player B and after Player B decides whether to send money back to Player A. The other participant has read the same game instructions as you.

At the end of the experiment, you would receive whatever amount of money out of the original $10 that you keep for yourself in addition to any amount the other participant would send back to

you (out of the tripled amount that you choose to send). Your payment at the end of the experiment could be more than, less than, or equal to $10.

9_2_Mod_DV Please enter the amount of money (out of $10) that you would send to the other participant. Remember, the amount you send would be tripled. The other participant would have the option of giving some, none, or all of the tripled amount back to you.

Team 9 Analysis Plan

The DV will be responses to *9_2_Ext_DV* or *9_2_Mod_DV*, depending on condition. We will compare the extreme and moderate conditions using an independent-samples t-test. The effect size will be an independent-groups Cohen's *d*.

Team 10 Materials

Extreme Condition

10_2_Ext_Intro Suppose you need to buy a house. The size and architecture styling must match your preferences. The surrounding areas need to be safe and clean. Most importantly, it needs to be fairly priced.

Information about a house that fits your requirements has been passed to you but you don't know the price of the property. After contacting the salesperson, he meets you to discuss the specifics on price.

You two meet on the housing property to discuss the price. After some brief small talk and questions, the salesperson concludes: "This house is an excellent fit for you, and 900 thousand dollars is a fair and good price for this house."

You have looked at other prices before. Similar level house you've looked at in the past were all in the $400-$500 thousand dollars range.

10_2_Ext_DV How trustworthy do you think the salesperson is?

- ❍ 1 not at all
- ❍ 2
- ❍ 3
- ❍ 4
- ❍ 5
- ❍ 6
- ❍ 7 very much

Moderate Condition
10_2_Mod_Intro Suppose you need to buy a house. The size and architecture styling must match your preferences. The surrounding areas need to be safe and clean. Most importantly, it needs to be fairly priced.

Information about a house that fits your requirements has been passed to you but you don't know the price of the property. After contacting the salesperson, he meets you to discuss the specifics on price.

You two meet on the housing property to discuss the price. After some brief small talk and questions, the salesperson concludes: "This house is an excellent fit for you, and 450 thousand dollars is a fair and good price for this house."

You have looked at other prices before. Similar level house you've looked at in the past were all in the $400-$500 thousand dollars range.

10_2_Mod_DV How trustworthy do you think the salesperson is?
❍ 1 not at all

❍ 2

❍ 3

❍ 4

❍ 5

❍ 6

❍ 7 very much

Team 10 Analysis Plan
The DV will be responses to *10_2_Ext_DV* or *10_2_Mod_DV*, depending on condition. We will compare the extreme and moderate conditions using an independent-samples t-test. The effect size will be an independent-groups Cohen's *d*.

Team 11 Materials
Extreme Condition
11_2_Ext_Intro Imagine that you are looking to purchase a house. You have found a house that you are interested in buying, and are about to sit down with the seller, Sam, to negotiate on the price. Of course, you would like to pay as little as possible, and Sam would like you to pay as much as possible, so you anticipate that there will be some back-and-forth during the negotiation. Six months ago, three independent appraisers valued the house at $290,000, $300,000, and $320,000, but this does not account for changes in property values since the appraisals. After the two of you shake hands and sit down, Sam abruptly makes a first offer of $550,000.

11_2_Ext_DV How trustworthy is Sam?
❍ Not at all

❍

❍

❍

❍

❍

❍

❍

❍ Extremely

11_2_Ext_F1 How intelligent is Sam?
❍ Not at all

❍

❍

❍

❍

❍

❍

❍

❍ Extremely

11_2_Ext_F2 How friendly is Sam?
❍ Not at all

❍

❍

❍

❍

❍

❍

❍

❍ Extremely

11_2_Ext_F3 How positive is your overall impression of Sam?
- ❍ Not at all
- ❍
- ❍
- ❍
- ❍
- ❍
- ❍
- ❍ Extremely

11_2_Ext_MC How reasonable do you think Sam's initial offer was?
- ❍ Sam's offer was much too low
- ❍
- ❍
- ❍
- ❍ Sam's offer was reasonable
- ❍
- ❍
- ❍
- ❍ Sam's offer was much too high

11_2_Mod_Intro Imagine that you are looking to purchase a house. You have found a house that you are interested in buying, and are about to sit down with the seller, Sam, to negotiate on the price. Of course, you would like to pay as little as possible, and Sam would like you to pay as much as possible, so you anticipate that there will be some back-and-forth during the negotiation. Six months ago, three independent appraisers valued the house at $290,000, $300,000, and $320,000, but this does not account for changes in property values since the appraisals. After the two of you shake hands and sit down, Sam abruptly makes a first offer of $350,000.

11_2_Mod_DV How trustworthy is Sam?
- ❍ Not at all
- ❍

❍

❍

❍

❍

❍

❍

❍ Extremely

11_2_Mod_F1 How intelligent is Sam?
❍ Not at all

❍

❍

❍

❍

❍

❍

❍

❍ Extremely

11_2_Mod_F2 How friendly is Sam?
❍ Not at all

❍

❍

❍

❍

❍

❍

❍

❍ Extremely

11_2_Mod_F3 How positive is your overall impression of Sam?
❍ Not at all

❍

❍

❍

❍

❍

❍

❍

❍ Extremely

11_2_Mod_MC How reasonable do you think Sam's initial offer was?

❍ Sam's offer was much too low

❍

❍

❍

❍ Sam's offer was reasonable

❍

❍

❍

❍ Sam's offer was much too high

Team 11 Analysis Plan

The DV will be responses to *11_2_Ext_DV* or *11_2_Mod_DV*, depending on condition. We will compare the extreme and moderate conditions using an independent-samples t-test. The effect size will be an independent-groups Cohen's *d*.

Team 12 did not develop materials for this research question.

Team 13 Materials

General Note: These materials consist of an "extreme" version and a "moderate" version of six different scenarios. Participants will be randomly assigned to see one version of each scenario, for a total of six scenarios, three of which will be "extreme" and three of which will be "moderate". The order of presentation of the six scenarios will be randomized.

13_2_Intro We are interested in perceptions of negotiators. On following pages you will find several descriptions of negotiations. Do your best to imagine yourself in each situation and how you would feel about the other party in the negotiation.

13_2_Ext_DV1 You are trying to sell your car. It works fine but has some signs of age and wear. Your car was appraised by a dealer for roughly $1,500 so you decide to list the price as $1,700. After posting advertisements for your car, you find one serious buyer. This person proposes to buy the car for $600 as a first offer.

Based on this information, how trustworthy do you think the buyer is?
- ❍ Not at all trustworthy
- ❍ A little trustworthy
- ❍ Slightly trustworthy
- ❍ Neutral
- ❍ Moderately trustworthy
- ❍ Very trustworthy
- ❍ Extremely trustworthy

13_2_Mod_DV1 You are trying to sell your car. It works fine but has some signs of age and wear. Your car was appraised by a dealer for roughly $1,500 so you decide to list the price as $1,700. After posting advertisements for your car, you find one serious buyer. This person proposes to buy the car for $1400 as a first offer.

Based on this information, how trustworthy do you think the buyer is?
- ❍ Not at all trustworthy
- ❍ A little trustworthy
- ❍ Slightly trustworthy
- ❍ Neutral
- ❍ Moderately trustworthy
- ❍ Very trustworthy
- ❍ Extremely trustworthy

13_2_Ext_DV2 You are looking to buy a house and finally find the “right” one. You research the prices of other homes in the area and find that most homes in this neighborhood sell between $200,000 and $230,000. In addition, your realtor tells you that there is another home that is very similar on sale for $220,000. The seller lists an asking price of $275,000.

Based on this information, how trustworthy do you think the seller is?
- ❍ Not at all trustworthy
- ❍ A little trustworthy

- ❍ Slightly trustworthy
- ❍ Neutral
- ❍ Moderately trustworthy
- ❍ Very trustworthy
- ❍ Extremely trustworthy

13_2_Mod_DV2 You are looking to buy a house and finally find the "right" one. You research the prices of other homes in the area and find that most homes in this neighborhood sell between $200,000 and $230,000. In addition, your realtor tells you that there is another home that is very similar on sale for $220,000. The seller lists an asking price of $230,000.Based on this information, how trustworthy do you think the seller is?

- ❍ Not at all trustworthy
- ❍ A little trustworthy
- ❍ Slightly trustworthy
- ❍ Neutral
- ❍ Moderately trustworthy
- ❍ Very trustworthy
- ❍ Extremely trustworthy

13_2_Ext_DV3 You want to install flooring in your home. For a job of this size you should expect to pay at least $20,000 based on the prices your neighbors paid for the similarly sized homes. Unfortunately, their contractors are busy. You receive an initial bid from a new contractor for $40,000.

Based on this information, how trustworthy do you think the contractor is?

- ❍ Not at all trustworthy
- ❍ A little trustworthy
- ❍ Slightly trustworthy
- ❍ Neutral
- ❍ Moderately trustworthy
- ❍ Very trustworthy
- ❍ Extremely trustworthy

13_2_Mod_DV3 You want to install flooring in your home. For a job of this size you should expect to pay at least $20,000 based on the prices your neighbors paid for the similarly sized

homes. Unfortunately, their contractors are busy. You receive an initial bid from a new contractor for $21,000.

Based on this information, how trustworthy do you think the contractor is?
- ❍ Not at all trustworthy
- ❍ A little trustworthy
- ❍ Slightly trustworthy
- ❍ Neutral
- ❍ Moderately trustworthy
- ❍ Very trustworthy
- ❍ Extremely trustworthy

13_2_Ext_DV4 You are negotiating a 12-month extension contract to work as an engineer. Your current salary is $68,000. One of your colleagues with comparable experience got a 12-month contract for $80,000 at a new firm. You have a competing contract from a different firm for $70,000. Your current firm proposes to pay you $70,001 as a first offer to continue your contract.

Based on this information, how trustworthy do you think your current firm is?
Not at all trustworthy
- ❍ A little trustworthy
- ❍ Slightly trustworthy
- ❍ Neutral
- ❍ Moderately trustworthy
- ❍ Very trustworthy
- ❍ Extremely trustworthy

13_2_Mod_DV4 You are negotiating a 12-month extension contract to work as an engineer. Your current salary is $68,000. One of your colleagues with comparable experience got a 12-month contract for $80,000 at a new firm. You have a competing contract from a different firm for $70,000. Your current firm proposes to pay you $75,000 as a first offer to continue your contract.

Based on this information, how trustworthy do you think your current firm is?
- ❍ Not at all trustworthy
- ❍ A little trustworthy
- ❍ Slightly trustworthy
- ❍ Neutral

- ❍ Moderately trustworthy
- ❍ Very trustworthy
- ❍ Extremely trustworthy

13_2_Ext_DV5 You are working as an agent for a player in the NFL. Your client made the league minimum for a rookie last year ($435,000) and was only signed to a 1-year deal. However, he had a breakout season and was the second best running back in the league. The top running back just signed an $8 million deal and Top 10 running backs are typically making between $4 million and $5 million dollars this upcoming season. The team comes to you with initial salary offer of $1.5 million.

Based on this information, how trustworthy do you think this team is?

- ❍ Not at all trustworthy
- ❍ A little trustworthy
- ❍ Slightly trustworthy
- ❍ Neutral
- ❍ Moderately trustworthy
- ❍ Very trustworthy
- ❍ Extremely trustworthy

13_2_Mod_DV5 You are working as an agent for a player in the NFL. Your client made the league minimum for a rookie last year ($435,000) and was only signed to a 1-year deal. However, he had a breakout season and was the second best running back in the league. The top running back just signed an $8 million deal and Top 10 running backs are typically making between $4 million and $5 million dollars this upcoming season. The team comes to you with initial salary offer of $6 million.

Based on this information, how trustworthy do you think this team is?

- ❍ Not at all trustworthy
- ❍ A little trustworthy
- ❍ Slightly trustworthy
- ❍ Neutral
- ❍ Moderately trustworthy
- ❍ Very trustworthy
- ❍ Extremely trustworthy

13_2_Mod_DV6 You are looking to buy a new car. You research the typical amount paid for the car using the internet. You also find out prices from dealers all over the region but you would like to buy from your local dealer. You find the range of prices paid for this car in your region is between $26,000 and $31,000. The suggested retail price is $27,500. Your local dealer offers an initial asking price of $34,000.

Based on this information, how trustworthy do you think this car dealer is?
- ❍ Not at all trustworthy
- ❍ A little trustworthy
- ❍ Slightly trustworthy
- ❍ Neutral
- ❍ Moderately trustworthy
- ❍ Very trustworthy
- ❍ Extremely trustworthy

13_2_Ext_DV6 You are looking to buy a new car. You research the typical amount paid for the car using the internet. You also find out prices from dealers all over the region but you would like to buy from your local dealer. You find the range of prices paid for this car in your region is between $26,000 and $31,000. The suggested retail price is $27,500. Your local dealer offers an initial asking price of $28,000.

Based on this information, how trustworthy do you think this car dealer is?
- ❍ Not at all trustworthy
- ❍ A little trustworthy
- ❍ Slightly trustworthy
- ❍ Neutral
- ❍ Moderately trustworthy
- ❍ Very trustworthy
- ❍ Extremely trustworthy

Team 13 Analysis Plan

Participants will respond to three *DV* questions in the extreme condition, and three *DV* questions in the moderate condition. The DV will be the mean of responses to each of these questions. We will compare the extreme and moderate conditions using a paired-samples t-test. The effect size will be a repeated-measures Cohen's *d,* which will be converted to an independent-groups *d* for comparison to other effect sizes.

Original Materials
Extreme Condition

15_2_Ext_Intro1 Welcome to this study. Please take a few minutes to read the following scenario carefully. In this study you will be asked to assume the role of someone who is negotiating for a new mobile phone. Please read the following instructions carefully.

15_2_Ext_Intro2 With all the new mobile phones on the market, you are excited about buying a new mobile phone! You want to buy a mobile phone with a new service contract.

After doing some research, you have decided to buy a phone and contract from TeleCo. You go to the TeleCo store to negotiate the following four issues with the sales representative:

1. Price of phone
2. Price of accessories
3. Warranty period
4. Service contract

The table below shows you which outcomes are most favorable to you. Your goal in this negotiation is to **earn as many points as possible**.

15_2_Ext_Intro3

Price of phone		Price of Accessories		Warranty period		Service Contract	
$	*Points*	$	*Points*	Months	*Points*	Months	*Points*
$200	*0*	$50	*0*	6 months	*0*	6 months	*0*
$190	*50*	$45	*25*	9 months	*15*	9 months	*30*
$180	*100*	$40	*50*	12 months	*30*	12 months	*60*
$170	*150*	$35	*75*	15 months	*45*	15 months	*90*
$160	*200*	$30	*100*	18 months	*60*	18 months	*120*
$150	*250*	$25	*125*	21 months	*75*	21 months	*150*
$140	*300*	$20	*150*	24 months	*90*	24 months	*180*
$130	*350*	$15	*175*	27 months	*105*	27 months	*210*
$120	*400*	$10	*200*	30 months	*120*	30 months	*240*

15_2_Ext_Intro4 For example, if you negotiated a phone price of $200, you would earn **0 points**, but negotiating a phone price of $120 would earn you **400 points**.

Price of phone		Price of Accessories		Warranty period		Service Contract	
$	*Points*	$	*Points*	Months	*Points*	Months	*Points*
$200	*0*	$50	*0*	6 months	*0*	6 months	*0*
$190	*50*	$45	*25*	9 months	*15*	9 months	*30*
$180	*100*	$40	*50*	12 months	*30*	12 months	*60*
$170	*150*	$35	*75*	15 months	*45*	15 months	*90*
$160	*200*	$30	*100*	18 months	*60*	18 months	*120*
$150	*250*	$25	*125*	21 months	*75*	21 months	*150*
$140	*300*	$20	*150*	24 months	*90*	24 months	*180*
$130	*350*	$15	*175*	27 months	*105*	27 months	*210*
$120	*400*	$10	*200*	30 months	*120*	30 months	*240*

15_2_Ext_Intro6 The information in the table is private information, so the sales representative does not know your preferences. Similarly, you don't know the sales rep's preferences and priorities, which might differ from yours. Your goal is to **earn as many points as possible.**

15_2_Ext_Intro7 After talking about each of these issues in more detail with the sales representative, the sales rep hands you a note with his first offer for each of the four issues.

The note says that he offers you:

1.	Price of phone:	**$ 200**	(**0 POINTS** for you)
2.	Price of accessories:	**$ 50**	(**0 POINTS** for you)
3.	Warranty period:	**6 months**	(**0 POINTS** for you)
4.	Service contract:	**6 months**	(**0 POINTS** for you)

15_2_Ext_Intro8

Price of phone		Price of Accessories		Warranty period		Service Contract	
$	Points	$	Points	Months	Points	Months	Points
$200	0	$50	0	6 months	0	6 months	0
$190	50	$45	25	9 months	15	9 months	30
$180	100	$40	50	12 months	30	12 months	60
$170	150	$35	75	15 months	45	15 months	90
$160	200	$30	100	18 months	60	18 months	120
$150	250	$25	125	21 months	75	21 months	150
$140	300	$20	150	24 months	90	24 months	180
$130	350	$15	175	27 months	105	27 months	210
$120	400	$10	200	30 months	120	30 months	240

15_2_Ext_Intro9 You need to come to an agreement on **all four issues** to buy the new mobile phone. If you fail to come to an agreement on all four issues, you can walk away without a deal, at which point, you would look for another store from which to buy a phone.

15_2_Ext_Intro10 **[For your reference, please find the sales rep's offer to you below:]**

Price of phone		Price of Accessories		Warranty period		Service Contract	
$	Points	$	Points	Months	Points	Months	Points
$200	0	$50	0	6 months	0	6 months	0
$190	50	$45	25	9 months	15	9 months	30
$180	100	$40	50	12 months	30	12 months	60
$170	150	$35	75	15 months	45	15 months	90
$160	200	$30	100	18 months	60	18 months	120
$150	250	$25	125	21 months	75	21 months	150
$140	300	$20	150	24 months	90	24 months	180
$130	350	$15	175	27 months	105	27 months	210
$120	400	$10	200	30 months	120	30 months	240

15_2_Ext_Intro11 Please answer the following questions about your reactions to the above mobile phone negotiation.

15_2_Ext_MC1 What is your counteroffer on each of the four issues? Price of phone (in points)

15_2_Ext_MC2 Price of accessories (in points)

15_2_Ext_MC3 Warranty period (in points)

15_2_Ext_MC4 Service contract (in points)

15_2_Ext_DV1 How trustworthy is this sales rep?
- ❍ 1 Not at all
- ❍ 2
- ❍ 3
- ❍ 4
- ❍ 5
- ❍ 6
- ❍ 7 Very much so

15_2_Ext_DV2 How credible do you think this sales rep is?
- ❍ 1 Not at all
- ❍ 2
- ❍ 3
- ❍ 4
- ❍ 5
- ❍ 6
- ❍ 7 Very much so

Moderate Condition

15_2_Mod_Intro1 Welcome to this study. Please take a few minutes to read the following scenario carefully. In this study you will be asked to assume the role of someone who is negotiating for a new mobile phone. Please read the following instructions carefully.

15_2_Mod_Intro2 With all the new mobile phones on the market, you are excited about buying a new mobile phone! You want to buy a mobile phone with a new service contract.

After doing some research, you have decided to buy a phone and contract from TeleCo. You go to the TeleCo store to negotiate the following four issues with the sales representative:

1. Price of phone
2. Price of accessories
3. Warranty period
4. Service contract The table below shows you which outcomes are most favorable to you. Your goal in this negotiation is to earn as many points as possible.

15_2_Mod_Intro3

Price of phone		Price of Accessories		Warranty period		Service Contract	
$	*Points*	$	*Points*	Months	*Points*	Months	*Points*
$200	*0*	$50	*0*	6 months	*0*	6 months	*0*
$190	*50*	$45	*25*	9 months	*15*	9 months	*30*
$180	*100*	$40	*50*	12 months	*30*	12 months	*60*
$170	*150*	$35	*75*	15 months	*45*	15 months	*90*
$160	*200*	$30	*100*	18 months	*60*	18 months	*120*
$150	*250*	$25	*125*	21 months	*75*	21 months	*150*
$140	*300*	$20	*150*	24 months	*90*	24 months	*180*
$130	*350*	$15	*175*	27 months	*105*	27 months	*210*
$120	*400*	$10	*200*	30 months	*120*	30 months	*240*

15_2_Mod_Intro4 For example, if you negotiated a phone price of $200, you would earn **0 points**, but negotiating a phone price of $120 would earn you **400 points**.

Price of phone		Price of Accessories		Warranty period		Service Contract	
$	*Points*	$	*Points*	Months	*Points*	Months	*Points*
$200	*0*	$50	*0*	6 months	*0*	6 months	*0*
$190	*50*	$45	*25*	9 months	*15*	9 months	*30*
$180	*100*	$40	*50*	12 months	*30*	12 months	*60*
$170	*150*	$35	*75*	15 months	*45*	15 months	*90*
$160	*200*	$30	*100*	18 months	*60*	18 months	*120*
$150	*250*	$25	*125*	21 months	*75*	21 months	*150*
$140	*300*	$20	*150*	24 months	*90*	24 months	*180*
$130	*350*	$15	*175*	27 months	*105*	27 months	*210*
$120	*400*	$10	*200*	30 months	*120*	30 months	*240*

15_2_Mod_Intro5 The information in the table is private information, so the sales representative does not know your preferences. Similarly, you don't know the sales rep's preferences and priorities, which might differ from yours. Your goal is to **earn as many points as possible**.

15_2_Mod_Intro6 After talking about each of these issues in more detail with the sales representative, the sales rep hands you a note with his first offer for each of the four issues.

The note says that he offers you:

1.	Price of phone:	**$ 160**	(**200 POINTS** for you)
2.	Price of accessories:	**$ 30**	(**100 POINTS** for you)
3.	Warranty period:	**18 months**	(**60 POINTS** for you)
4.	Service contract:	**18 months**	(**120 POINTS** for you)

15_2_Mod_Intro7

Price of phone		Price of Accessories		Warranty period		Service Contract	
$	Points	$	Points	Months	Points	Months	Points
$200	0	$50	0	6 months	0	6 months	0
$190	50	$45	25	9 months	15	9 months	30
$180	100	$40	50	12 months	30	12 months	60
$170	150	$35	75	15 months	45	15 months	90
$160	200	$30	100	18 months	60	18 months	120
$150	250	$25	125	21 months	75	21 months	150
$140	300	$20	150	24 months	90	24 months	180
$130	350	$15	175	27 months	105	27 months	210
$120	400	$10	200	30 months	120	30 months	240

15_2_Mod_Intro8 You need to come to an agreement on **all four issues** to buy the new mobile phone. If you fail to come to an agreement on all four issues, you can walk away without a deal, at which point, you would look for another store from which to buy a phone.

15_2_Mod_Intro9 **[For your reference, please find the sales rep's offer to you below:]**

Price of phone		Price of Accessories		Warranty period		Service Contract	
$	Points	$	Points	Months	Points	Months	Points
$200	0	$50	0	6 months	0	6 months	0
$190	50	$45	25	9 months	15	9 months	30
$180	100	$40	50	12 months	30	12 months	60
$170	150	$35	75	15 months	45	15 months	90
$160	200	$30	100	18 months	60	18 months	120
$150	250	$25	125	21 months	75	21 months	150
$140	300	$20	150	24 months	90	24 months	180
$130	350	$15	175	27 months	105	27 months	210
$120	400	$10	200	30 months	120	30 months	240

15_2_Mod_Intro10 Please answer the following questions about your reactions to the above mobile phone negotiation.

15_2_Mod_MC1 What is your counteroffer on each of the four issues? Price of phone (in points)

15_2_Mod_MC2 Price of accessories (in points)

15_2_Mod_MC3 Warranty period (in points)

15_2_Mod_MC4 Service contract (in points)

15_2_Mod_DV1 How trustworthy is this sales rep?
- ❍ 1 Not at all
- ❍ 2
- ❍ 3
- ❍ 4
- ❍ 5
- ❍ 6
- ❍ 7 Very much so

15_2_Mod_DV2 How credible do you think this sales rep is?
- ❍ 1 Not at all
- ❍ 2
- ❍ 3
- ❍ 4
- ❍ 5
- ❍ 6
- ❍ 7 Very much so

Original Materials Analysis Plan

The DV will be the mean of the two *DV* questions in each condition (i.e., the mean of *15_2_Ext_DV1* and *15_2_Ext_DV2*, or the mean of *15_2_Mod_DV1* and *15_2_Mod_DV2* depending on condition). We will compare the extreme and moderate conditions using an independent-samples t-test. The effect size will be an independent-groups Cohen's *d*. If this composite DV shows poor internal reliability ($\alpha < .70$), we will use responses to *15_2_Ext_DV1* and *15_2_Mod_DV1*.

Research Question 3: What are the effects of continuing to work despite having no material/financial need to work on moral judgments of that individual -- beneficial, detrimental, or no effect?

Team 1 Materials

Male Work Condition

1_3_MWork_Intro You will read a short vignette about an individual, and you will then be asked to answer a series of questions. The vignette is based on a real situation that was recently faced by another person. John has been working and he comes into a large sum of money. Upon receiving the money, he reviews his finances and realizes that he could stop working now. He has no material or financial need to work anymore. However, John decides to continue working.

1_3_MWork_DV Given this decision, how do you view John as a person?

- ❍ Much less favorably
- ❍ Somewhat less favorably
- ❍ Slightly less favorably
- ❍ Neither more or less favorably
- ❍ Slightly more favorably
- ❍ Somewhat more favorably
- ❍ Much more favorably

Female Work Condition

1_3_FWork_Intro You will read a short vignette about an individual, and you will then be asked to answer a series of questions. The vignette is based on a real situation that was recently faced by another person. Julia has been working and she comes into a large sum of money. Upon receiving the money, she reviews her finances and realizes that she could stop working now. She has no material or financial need to work anymore. However, Julia decides to continue working.

1_3_FWork_DV Given this decision, how do you view Julia as a person?

- ❍ Much less favorably
- ❍ Somewhat less favorably
- ❍ Slightly less favorably
- ❍ Neither more or less favorably
- ❍ Slightly more favorably
- ❍ Somewhat more favorably
- ❍ Much more favorably

Male No Work Condition
1_3_MNoWork_Intro You will read a short vignette about an individual, and you will then be asked to answer a series of questions. The vignette is based on a real situation that was recently faced by another person. John has been working and he comes into a large sum of money. Upon receiving the money, he reviews his finances and realizes that he could stop working now. He has no material or financial need to work anymore. As such, John decides to stop working.

1_3_MNoWork_DV Given this decision, how do you view John as a person?
- ❍ Much less favorably
- ❍ Somewhat less favorably
- ❍ Slightly less favorably
- ❍ Neither more or less favorably
- ❍ Slightly more favorably
- ❍ Somewhat more favorably
- ❍ Much more favorably

Female No Work Condition
1_3_FNoWork_Intro You will read a short vignette about an individual, and you will then be asked to answer a series of questions. The vignette is based on a real situation that was recently faced by another person. Julia has been working and she comes into a large sum of money. Upon receiving the money, she reviews her finances and realizes that she could stop working now. She has no material or financial need to work anymore. As such, Julia decides to stop working.

1_3_FNoWork_DV Given this decision, how do you view Julia as a person?
- ❍ Much less favorably
- ❍ Somewhat less favorably
- ❍ Slightly less favorably
- ❍ Neither more or less favorably
- ❍ Slightly more favorably
- ❍ Somewhat more favorably
- ❍ Much more favorably

Team 1 Analysis Plan
We will collapse across the male and female conditions (full data will be made public, so a more complete analysis can be undertaken later). Responses from the *DV* questions in the work and no work conditions will be compared using an independent-samples t-test. The effect size will be an independent-groups Cohen's *d*.

Team 2 Materials
Continue to Work Condition
2_3_Work_Intro Please read the article below.

B.C. Resident Wins $25M Jackpot
CBC
Posted: 11/10/2012 2:15 pm EST Updated: 01/10/2013 5:12 am EST
Bob Erb, a resident from Terrace, British Columbia., is $25 million richer after scoring a winning ticket in last week's Lotto Max jackpot. "I just went in, checked the lottery ticket — 25 and a whole bunch of zeroes," new millionaire Bob Erb told CBC News. "I pulled the ticket out and I said, 'Oh my God. I think I won $25 million.'"

Erb bought the ticket in New Hazelton while on his way to Calgary. Erb has been purchasing lottery tickets for 43 years, always buying the exact same amount — but this time, the clerk ran in more plays than he wanted.

"I said, 'No, I wanted a $6-ticket for this upcoming Friday and the following Friday so he said, 'Okay, I'll just cancel this one.' I said, 'No no, this just might be the big one. I'll keep that.'"

The 50-year-old construction worker intends to keep working and living in Terrace, saying "I like my life, I see no reason to change it now". He works for Beutle Masonry in Terrace. As for the windfall, Erb says some will go to family, friends, and homeless shelters.
2_3_Work_DV To what extent is Bob . . . ?

	not at all	slightly	moderately	very	extremely
moral	❍	❍	❍	❍	❍
sincere	❍	❍	❍	❍	❍
honest	❍	❍	❍	❍	❍
righteous	❍	❍	❍	❍	❍
trustworthy	❍	❍	❍	❍	❍
respectful	❍	❍	❍	❍	❍
kind	❍	❍	❍	❍	❍
friendly	❍	❍	❍	❍	❍
likeable	❍	❍	❍	❍	❍
warm	❍	❍	❍	❍	❍
helpful	❍	❍	❍	❍	❍

intelligent	❍	❍	❍	❍	❍
competent	❍	❍	❍	❍	❍
efficient	❍	❍	❍	❍	❍
skillful	❍	❍	❍	❍	❍
capable	❍	❍	❍	❍	❍

2_3_Work_MC1 Did the lottery winner give some of his winnings to others?
❍ Yes

❍ No

❍ I don't know

2_3_Work_MC2 Did the lottery winner quit his job or continue to work?
❍ Quit his job

❍ Continued working

❍ I don't know

Quit Job Condition
2_3_Quit_Intro Please read the article below.

B.C. Resident Wins $25M Jackpot
CBC
Posted: 11/10/2012 2:15 pm EST Updated: 01/10/2013 5:12 am EST
Bob Erb, a resident from Terrace, British Columbia., is $25 million richer after scoring a winning ticket in last week's Lotto Max jackpot. "I just went in, checked the lottery ticket — 25 and a whole bunch of zeroes," new millionaire Bob Erb told CBC News. "I pulled the ticket out and I said, 'Oh my God. I think I won $25 million.'"

Erb bought the ticket in New Hazelton while on his way to Calgary. Erb has been purchasing lottery tickets for 43 years, always buying the exact same amount — but this time, the clerk ran in more plays than he wanted.

"I said, 'No, I wanted a $6-ticket for this upcoming Friday and the following Friday so he said, 'Okay, I'll just cancel this one.' I said, 'No no, this just might be the big one. I'll keep that.'" The 50-year-old construction worker quit his job but plans to continue living in Terrace, saying "I like my life, I see no reason to change it now". He worked for Beutle Masonry in Terrace. As for the windfall, Erb says some will go to family, friends, and homeless shelters.

2_3_Quit_DV To what extent is Bob . . . ?

	not at all	slightly	moderately	very	extremely
moral	❍	❍	❍	❍	❍
sincere	❍	❍	❍	❍	❍
honest	❍	❍	❍	❍	❍
righteous	❍	❍	❍	❍	❍
trustworthy	❍	❍	❍	❍	❍
respectful	❍	❍	❍	❍	❍
kind	❍	❍	❍	❍	❍
friendly	❍	❍	❍	❍	❍
likeable	❍	❍	❍	❍	❍
warm	❍	❍	❍	❍	❍
helpful	❍	❍	❍	❍	❍
intelligent	❍	❍	❍	❍	❍
competent	❍	❍	❍	❍	❍
efficient	❍	❍	❍	❍	❍
skillful	❍	❍	❍	❍	❍
capable	❍	❍	❍	❍	❍

2_3_Quit_MC1 Did the lottery winner give some of his winnings to others?

❍ Yes

❍ No

❍ I don't know

2_3_Quit_MC2 Did the lottery winner quit his job or continue to work?

❍ Quit his job

❍ Continued working

❍ I don't know

Team 2 Analysis Plan

The DV will be the mean of the responses to *moral, sincere, honest, righteous, trustworthy,* and *respectful* in *2_3_Work_DV* or *2_3_Quit_DV*, depending on condition. We will compare the continue to work and quit job conditions using an independent-samples t-test. The effect size will be an independent-groups Cohen's *d*. If this composite DV shows poor internal reliability ($\alpha < .70$), we will use responses to *moral*.

Team 3 Materials

Continue to Work Condition

3_3_Work_Intro Bob is an accountant at a large investment firm. Although his salary is modest, it is enough to provide for his family. Bob works hard and likes his job. One day, Bob wins a lottery with a cash prize in the amount of 150 million US dollars – enough to move to a bigger house and never work again. After giving it some careful thought, Bob decides to continue working as an accountant in the investment firm.

3_3_Work_DV1 Please rate your impression of Bob on the following items:

Bob is…

❍ 1 …a bad person

❍ 2

❍ 3

❍ 4

❍ 5

❍ 6

❍ 7 …a good person

3_3_Work_DV2 Bob is…

❍ 1 …immoral

❍ 2

❍ 3

❍ 4

❍ 5

❍ 6

❍ 7 …moral

3_3_Work_DV3 Bob is…

❍ 1 …dishonorable

❍ 2

❍ 3

❍ 4

❍ 5

❍ 6

❍ 7 …honorable

Quit Job Condition
3_3_Quit_Intro Bob is an accountant at a large investment firm. Although his salary is modest, it is enough to provide for his family. Bob works hard and likes his job. One day, Bob wins a lottery with a cash prize in the amount of 150 million US dollars – enough to move to a bigger house and never work again. After giving it some careful thought, Bob decides to quit his job at the investment firm.

3_3_Quit_DV1 Please rate your impression of Bob on the following items:

Bob is…
❍ 1 …a bad person

❍ 2

❍ 3

❍ 4

❍ 5

❍ 6

❍ 7 …a good person

3_3_Quit_DV2 Bob is…
❍ 1 …immoral

❍ 2

❍ 3

❍ 4

❍ 5

❍ 6

❍ 7 …moral

3_3_Quit_DV3 Bob is…
❍ 1 …dishonorable

❍ 2

❍ 3

❍ 4

❍ 5

❍ 6

❍ 7 …honorable

Team 3 Analysis Plan
The DV will be the mean of the three *DV* questions in each condition (i.e., the mean of *3_3_Work_DV1*, *3_3_Work_DV2*, and *3_3_Work_DV3* or the mean of *3_3_Quit_DV1*, *3_3_Quit_DV2*, and *3_3_Quit_DV3*, depending on condition). We will compare the continue to work and quit job conditions using an independent-samples t-test. The effect size will be an independent-groups Cohen's *d*. If this composite DV shows poor internal reliability ($\alpha < .70$), we will use responses to *3_3_Work_DV2* and *3_3_Quit_DV3*.

Team 4 Materials
No Material/Financial Need Condition
4_3_Work_Intro Paul has just clinched a full-time position that pays $500,000 USD a year. Despite having no material or financial need, Paul continues to work full time and plans to do so in the near future.

Now, please take a moment to think about Paul's daily life. Think about what he may do and feel. Once you have visualized the kind of person Paul is, please complete the following survey.

4_3_Work_DV To what extent do you agree with the following statements (1 = strongly disagree to 7 = strongly agree)?

	1 strongly disagree	2	3	4	5	6	7 strongly agree
Paul is a moral person	❍	❍	❍	❍	❍	❍	❍
Paul is conscientiousness	❍	❍	❍	❍	❍	❍	❍
Paul is agreeable	❍	❍	❍	❍	❍	❍	❍

Paul is someone I would describe as humorous	❍	❍	❍	❍	❍	❍	❍
Paul is someone I would describe as an extrovert	❍	❍	❍	❍	❍	❍	❍
Paul is an ethical person	❍	❍	❍	❍	❍	❍	❍
Paul is likeable	❍	❍	❍	❍	❍	❍	❍
Paul is a prosocial person	❍	❍	❍	❍	❍	❍	❍

Control Condition

4_3_Ctrl_Intro Paul has just clinched a full-time position that pays $50,000 USD a year. Paul works full time and plans to do so in the near future. Now, please take a moment to think about Paul's daily life. Think about what he may do and feel. Once you have visualized the kind of person Paul is, please complete the following survey.

4_3_Ctrl_DV To what extent do you agree with the following statements (1 = strongly disagree to 7 = strongly agree)?

	1 strongly disagree	2	3	4	5	6	7 strongly agree
Paul is a moral person	❍	❍	❍	❍	❍	❍	❍
Paul is conscientiousness	❍	❍	❍	❍	❍	❍	❍
Paul is agreeable	❍	❍	❍	❍	❍	❍	❍
Paul is someone I would describe as humorous	❍	❍	❍	❍	❍	❍	❍
Paul is someone I would describe as an extrovert	❍	❍	❍	❍	❍	❍	❍

Paul is an ethical person	❍	❍	❍	❍	❍	❍	❍
Paul is likeable	❍	❍	❍	❍	❍	❍	❍
Paul is a prosocial person	❍	❍	❍	❍	❍	❍	❍

Team 4 Analysis Plan

The DV will be the mean of the responses to items 1, 6, and 8 in *4_3_Work_DV* or *4_3_Ctrl_DV*, depending on condition. We will compare the no material/financial need condition and control condition using an independent-samples t-test. The effect size will be an independent-groups Cohen's *d*. If this composite DV shows poor internal reliability ($\alpha < .70$), we will use responses to item 1 (Paul is a moral person).

Team 5 Materials

Continue Working Condition

5_3_Work_Intro Mr. Smith is 65 years old. He has reached the retirement age, and has sufficient savings, pension, and insurance to secure him for the rest of his life. However, he continues to work even though he has no material or financial need.

5_3_Work_DV How praiseworthy do you think Mr. Smith's moral character is?

❍ 1 Not at all praiseworthy

❍ 2

❍ 3

❍ 4

❍ 5

❍ 6

❍ 7 Very praiseworthy

Stop Working Condition

5_3_Quit_Intro Mr. Smith is 65 years old. He has reached the retirement age, and has sufficient savings, pension, and insurance to secure him for the rest of his life. Therefore, he stops working because he has no material or financial need.

5_3_Quit_DV How praiseworthy do you think Mr. Smith's moral character is?

❍ 1 Not at all praiseworthy

❍ 2

❍ 3

❍ 4

❍ 5

❍ 6

❍ 7 Very praiseworthy

Team 5 Analysis Plan
The DV will be the mean of the responses to *5_3_Work_DV* or *5_3_Quit_DV*, depending on condition. We will compare the continue working condition and stop working condition using an independent-samples t-test. The effect size will be an independent-groups Cohen's *d*.

Team 6 Materials
Continue to Work Condition
6_3_Work_Intro1 In this study, we are interested in how people form impressions. On the next page you will be provided with a description of a person. Please read the information carefully and then answer questions.

6_3_Work_Intro2 David was a senior product manager in a local company. He has been working there for more than 10 years. Recently, David's aunt passed away and left him a significant sum of money. David does not need to work for money any more. Nevertheless, he has decided to continue to work in the company.

6_3_Work_DV To what extent do you think David is characterized by the following features?

	1 not at all	2	3	4	5	6	7 very much
Moral	❍	❍	❍	❍	❍	❍	❍
Honest	❍	❍	❍	❍	❍	❍	❍
Respectable	❍	❍	❍	❍	❍	❍	❍
Reliable	❍	❍	❍	❍	❍	❍	❍
Noble	❍	❍	❍	❍	❍	❍	❍

Quit Work Condition
6_3_Quit_Intro1 In this study, we are interested in how people form impressions. On the next page you will be provided with a description of a person. Please read the information carefully and then answer questions.

6_3_Quit_Intro2 David was a senior product manager in a local company. He has been working there for more than 10 years. Recently, David's aunt passed away and left him a significant sum of money. David does not need to work for money any more. Now he has decided to quit the job and live a different life.

6_3_Quit_DV To what extent do you think David is characterized by the following features?

	1 not at all	2	3	4	5	6	7 very much
Moral	❍	❍	❍	❍	❍	❍	❍
Honest	❍	❍	❍	❍	❍	❍	❍
Respectable	❍	❍	❍	❍	❍	❍	❍
Reliable	❍	❍	❍	❍	❍	❍	❍
Noble	❍	❍	❍	❍	❍	❍	❍

Team 6 Analysis Plan

The DV will be the mean of the responses to the five items in *6_3_Work_DV* or *4_3_Quit_DV*, depending on condition. We will compare the continue to work condition and quit work condition using an independent-samples t-test. The effect size will be an independent-groups Cohen's *d*. If this composite DV shows poor internal reliability ($\alpha < .70$), we will use responses to item 1 (Moral).

Team 7 Materials

Intrinsically Motivated Condition

7_3_Intrin_Intro An hour ago, David finished his work for the day and was compensated accordingly. David continues to work and will not be compensated whatsoever for this extra work.

7_3_Intrin_DV To what extent do you agree with the following statement? David is a morally good person.

❍ 1 definitely yes

❍ 2 probably yes

❍ 3 might or might not

❍ 4 probably not

❍ 5 definitely not

Extrinsically Motivated Condition
7_3_Extrin_Intro An hour ago, David finished his work for the day and was compensated accordingly. David continues to work and is being compensated for this extra work.

7_3_Extrin_DV To what extent do you agree with the following statement? David is a morally good person.

- ❍ 1 definitely yes
- ❍ 2 probably yes
- ❍ 3 might or might not
- ❍ 4 probably not
- ❍ 5 definitely not

Team 7 Analysis Plan
The DV will be the mean of the responses to *7_3_Intrin_DV* or *7_3_Ctrl_DV,* depending on condition. We will compare the intrinsically motivated condition and extrinsically motivated condition using an independent-samples t-test. The effect size will be an independent-groups Cohen's *d.*

Team 8 Materials
Keeps Working Condition
8_3_Work_Intro

Local Man Wins Big

June 29, 2016 - Murphys, CA - Yahoo News

Exactly a year after it was announced that an undisclosed local resident had won the jackpot of the California MegaBall lottery estimated at USD 50,000,000, the identity of the mysterious winner was finally revealed. Raymond Gonzalez, 34, a long-time resident of Douglas Flat, announced on Tuesday night that he was the lucky recipient of the financial windfall. "I never thought I would win," Gonzalez said in an interview, "but I guess good things happen sometimes." Gonzalez, who works as an accountant for the law firm of Hendricks and Lew in Murphys, said the numbers he played were a combination of his parents' birthdates and the jersey numbers of his favorite basketball players. He did not disclose whether when he won last year, he chose to receive the prize as a lump sum or a yearly installment; but whichever he chooses, it's fair to assume the recently-married Gonzalez will never need to work another day in his life.

Despite his newfound wealth, Gonzalez says he has kept his accounting job, and has been showing up for work every day in the last year. His employer, Attorney Leila Hendricks said in an email that she was "extremely happy for Gonzalez" and confirmed that he had continued to work every day since the day he won the lottery. "I have no plan to ever quit working," said Gonzalez, "I imagine I'll still be working in ten years, even though I don't need the money anymore."

8_3_Work_DV We are interested in perceptions of people in the media. While we know you have very little information about Mr. Gonzalez, we ask you try to guess where he stands on the following personality dimensions relative to the average person. Again, please take your best guess based on your first impression based on what you know about him.

	1	2	3	4	5	6	7
Unlucky:Lucky	❍	❍	❍	❍	❍	❍	❍
Dumb:Intelligent	❍	❍	❍	❍	❍	❍	❍
Dishonest:Honest	❍	❍	❍	❍	❍	❍	❍
Disorganized:Organized	❍	❍	❍	❍	❍	❍	❍
Cruel:Kind	❍	❍	❍	❍	❍	❍	❍

Shy:Outgoing	❍	❍	❍	❍	❍	❍	❍
Selfish:Generous	❍	❍	❍	❍	❍	❍	❍
Boring:Interesting	❍	❍	❍	❍	❍	❍	❍
Shifty:Trustworthy	❍	❍	❍	❍	❍	❍	❍

Quits Working Condition
8_3_Quit_Intro

Local Man Wins Big

June 29, 2016 - Murphys, CA - Yahoo News

Exactly a year after it was announced that an undisclosed local resident had won the jackpot of the California MegaBall lottery estimated at USD 50,000,000, the identity of the mysterious winner was finally revealed. Raymond Gonzalez, 34, a long-time resident of Douglas Flat, announced on Tuesday night that he was the lucky recipient of the financial windfall. "I never thought I would win," Gonzalez said in an interview, "but I guess good things happen sometimes." Gonzalez, who works as an accountant for the law firm of Hendricks and Lew in Murphys, said the numbers he played were a combination of his parents' birthdates and the jersey numbers of his favorite basketball players. He did not disclose whether when he won last year, he chose to receive the prize as a lump sum or a yearly installment; but whichever he chooses, it's fair to assume the recently-married Gonzalez will never need to work another day in his life.

Given his newfound wealth, Gonzalez says he has quit his accounting job, and has not worked a single day in the last year. His former employer, Attorney Leila Hendricks said in an email that she was "extremely happy for Gonzalez" and confirmed that he quit working there on the day he won the lottery. "I have no plan to ever work again," said Gonzalez, "I imagine I'll still not be working in ten years, because I don't need the money anymore."

8_3_Quit_DV We are interested in perceptions of people in the media. While we know you have very little information about Mr. Gonzalez, we ask you try to guess where he stands on the following personality dimensions relative to the average person. Again, please take your best guess based on your first impression based on what you know about him.

	1	2	3	4	5	6	7
Unlucky:Lucky	❍	❍	❍	❍	❍	❍	❍
Dumb:Intelligent	❍	❍	❍	❍	❍	❍	❍
Dishonest:Honest	❍	❍	❍	❍	❍	❍	❍
Disorganized:Organized	❍	❍	❍	❍	❍	❍	❍
Cruel:Kind	❍	❍	❍	❍	❍	❍	❍
Shy:Outgoing	❍	❍	❍	❍	❍	❍	❍
Selfish:Generous	❍	❍	❍	❍	❍	❍	❍
Boring:Interesting	❍	❍	❍	❍	❍	❍	❍
Shifty:Trustworthy	❍	❍	❍	❍	❍	❍	❍

Team 8 Analysis Plan

The DV will be the mean of the responses to items 3, 5, 7, and 9 in *8_3_Work_DV* or *8_3_Quit_DV*, depending on condition. We will compare the keeps working condition and quits working condition using an independent-samples t-test. The effect size will be an independent-groups Cohen's *d*. If this composite DV shows poor internal reliability ($\alpha < .70$), we will use responses to item 3 (Dishonest:Honest).

Team 9 Materials

Not Needing to Work Condition

9_3_Not_Intro Tom has recently retired. He worked as technician in a brewery for nearly 30 years. During this time he managed to buy a house for him and his family and saved enough money to live comfortably off his savings and pension. He does not have any need to work but now he has decided to continue working producing brewery supplies.

9_3_Not_DV1 From this information, to what extent do you think Tom could be described as…? (Where 1 means "not at all" and 7 "extremely")

	1	2	3	4	5	6	7
Caring	❍	❍	❍	❍	❍	❍	❍
Warm	❍	❍	❍	❍	❍	❍	❍
Empathic	❍	❍	❍	❍	❍	❍	❍
Principled	❍	❍	❍	❍	❍	❍	❍

Honorable	❍	❍	❍	❍	❍	❍	❍
Bold	❍	❍	❍	❍	❍	❍	❍
Gutsy	❍	❍	❍	❍	❍	❍	❍
Fearless	❍	❍	❍	❍	❍	❍	❍
Reasonable	❍	❍	❍	❍	❍	❍	❍
Effective	❍	❍	❍	❍	❍	❍	❍
Realistic	❍	❍	❍	❍	❍	❍	❍
Rational	❍	❍	❍	❍	❍	❍	❍

Needing to Work Condition

9_3_Need_Intro Tom has recently retired. He worked as technician in a brewery for nearly 30 years. During this time he could not manage to buy a house for him and his family nor to save enough money to live off his savings and pension. He does have the need to keep working and he has decided to continue working producing brewery supplies.

9_3_Need_DV From this information, to what extent do you think Tom could be described as…? (Where 1 means "not at all" and 7 "extremely")

	1	2	3	4	5	6	7
Caring	❍	❍	❍	❍	❍	❍	❍
Warm	❍	❍	❍	❍	❍	❍	❍
Empathic	❍	❍	❍	❍	❍	❍	❍
Principled	❍	❍	❍	❍	❍	❍	❍
Honorable	❍	❍	❍	❍	❍	❍	❍
Bold	❍	❍	❍	❍	❍	❍	❍
Gutsy	❍	❍	❍	❍	❍	❍	❍
Fearless	❍	❍	❍	❍	❍	❍	❍
Reasonable	❍	❍	❍	❍	❍	❍	❍
Effective	❍	❍	❍	❍	❍	❍	❍

Realistic	❍	❍	❍	❍	❍	❍	❍
Rational	❍	❍	❍	❍	❍	❍	❍

Team 9 Analysis Plan
The DV will be the mean of the responses to items 1, 2, and 3 in *9_3_Need_DV* or *9_3_Ctrl_DV*, depending on condition. We will compare the not needing to work condition and needing to work condition using an independent-samples t-test. The effect size will be an independent-groups Cohen's *d*. If this composite DV shows poor internal reliability ($\alpha < .70$), we will use responses to item 3 (Empathic).

Team 10 did not develop materials for this research question.

Team 11 Materials
Work Condition
11_3_Work_Intro Pat took a job at a mid-sized insurance company immediately after graduating from college, and was given stock in the company upon being hired. Twenty years later, Pat met with a financial adviser, who explained that Pat's company stock had grown considerably in value, and that Pat had more than enough money to retire and live comfortably without ever working again. Upon learning this, Pat decided not to retire, and has continued to work for the company for the past ten years.

11_3_Work_DV1 How good or bad is Pat's moral character?
❍ Extremely bad
❍
❍
❍
❍
❍
❍
❍
❍ Extremely good

11_3_Work_DV2 How good or bad a person is Pat?
❍ Extremely bad
❍
❍
❍

❍

❍

❍

❍

❍ Extremely good

11_3_Work_MC What did Pat do upon finding out that he had enough money to retire?
❍ Pat decided to retire

❍ Pat decided NOT to retire

Retire Condition
11_3_Ret_Intro Pat took a job at a mid-sized insurance company immediately after graduating from college, and was given stock in the company upon being hired. Twenty years later, Pat met with a financial adviser, who explained that Pat's company stock had grown considerably in value, and that Pat had more than enough money to retire and live comfortably without ever working again. Upon learning this, Pat decided to retire immediately, and has been retired for the past ten years.

11_3_Ret_DV1 How good or bad is Pat's moral character?
❍ Extremely bad

❍

❍

❍

❍

❍

❍

❍ Extremely good

11_3_RetDV2 How good or bad a person is Pat?
❍ Extremely bad

❍

❍

❍

❍

❍

❍

❍

❍ Extremely good

11_3_Ret_MC What did Pat do upon finding out that he had enough money to retire?
❍ Pat decided to retire

❍ Pat decided NOT to retire

Team 11 Analysis Plan
The DV will be the mean of the responses to *11_3_Work_DV* and *11_3_Work_DV2*, or *11_3_Ret_DV1 and 11_3_Ret_DV2*, depending on condition. We will compare the work and retire conditions using an independent-samples t-test. The effect size will be an independent-groups Cohen's *d*. If this composite DV shows poor internal reliability ($\alpha < .70$), we will use responses to *11_3_Work_DV1* and *11_3_Ret_DV1*.

Team 12 Materials
No Financial Need Condition
12_3_No-Fin_Intro INSTRUCTIONS: Please read the information about the individual below and answer the questions that follow.

12_3_No-Fin_DV Recently, Pat inherited a large sum of money from a distant relative—enough money to live on for the foreseeable future. However, Pat plans to continue working.

	1 Definitely No	2	3	4	5	6	7	8	9 Definitely Yes
Is Pat a good person?	❍	❍	❍	❍	❍	❍	❍	❍	❍
Is Pat a moral person?	❍	❍	❍	❍	❍	❍	❍	❍	❍
Is Pat praiseworthy?	❍	❍	❍	❍	❍	❍	❍	❍	❍

Should other people look up to Pat?	❍	❍	❍	❍	❍	❍	❍	❍	❍

Financial Need Condition
12_3_Fin_Intro INSTRUCTIONS: Please read the information about the individual below and answer the questions that follow.

12_3_Fin_DV Recently, Pat inherited a small sum of money from a distant relative—enough money to live on for about a week. Thus, Pat plans to continue working.

	1 Definitely No	2	3	4	5	6	7	8	9 Definitely Yes
Is Pat a good person?	❍	❍	❍	❍	❍	❍	❍	❍	❍
Is Pat a moral person?	❍	❍	❍	❍	❍	❍	❍	❍	❍
Is Pat praiseworthy?	❍	❍	❍	❍	❍	❍	❍	❍	❍
Should other people look up to Pat?	❍	❍	❍	❍	❍	❍	❍	❍	❍

Team 12 Analysis Plan
The DV will be the mean of responses to the four items in *12_3_No-Fin_DV* or *12_3_Fin_DV*, depending on condition. We will compare the no financial need and financial need conditions using an independent-samples t-test. The effect size will be an independent-groups Cohen's *d*. If this composite DV shows poor internal reliability ($\alpha < .70$), we will use responses item 1 (Is Pat a good person?).

Team 13 Materials
General Note: These materials consist of a "no financial need" version and a "financial need" version of six different scenarios. Participants will be randomly assigned to see one version of each scenario, for a total of six scenarios, three of which will be "no financial need" and three of which will be "financial need". The order of presentation of the six scenarios will be randomized.

13_3_Intro Instructions: On following pages are different descriptions of workers. We are interested in your impressions about their moral character.

13_3_No-Fin_DV1 John has worked for the same company for 20 years. Recently, John's boss asked him whether he would like to renew his contract. As John considers this decision, he finds out that he has saved enough money to live comfortably for the rest of his life. Despite having no financial need, he decides to renew the contract and continue to work.

John is a moral person.
- ❍ Strongly Disagree
- ❍ Disagree
- ❍ Somewhat Disagree
- ❍ Neither Disagree nor Agree
- ❍ Somewhat Agree
- ❍ Agree
- ❍ Strongly Agree

13_3_Fin_DV1 John has worked for the same company for 20 years. Recently, John's boss asked him whether he would like to renew his contract. As John considers this decision, he finds out that he needs to make more money to meet his needs in retirement. Because of his financial need, he decides to renew the contract and continue to work.

John is a moral person.
- ❍ Strongly Disagree
- ❍ Disagree
- ❍ Somewhat Disagree
- ❍ Neither Disagree nor Agree
- ❍ Somewhat Agree
- ❍ Agree
- ❍ Strongly Agree

13_3_No-Fin_DV2 Sophia is an employee at a startup company. When she first joined the company, she was given stock options. The company grew rapidly and the stock options are now worth a great deal of money. With the stock options, Sophia does not have to worry about her financial needs if she retires today. She decides to keep working at the company despite having no financial needs.

Sophia is a moral person.
- ❍ Strongly Disagree
- ❍ Disagree

- ❍ Somewhat Disagree
- ❍ Neither Disagree nor Agree
- ❍ Somewhat Agree
- ❍ Agree
- ❍ Strongly Agree

13_3_Fin_DV2 Sophia is an employee at a startup company. When she first joined the company, she was given stock options. The company grew rapidly and the stock options are now worth a great deal of money. Even with the stock options, Sophia still needs to save more money for her retirement. She decides to keep working at the company because of her financial needs.

Sophia is a moral person.
- ❍ Strongly Disagree
- ❍ Disagree
- ❍ Somewhat Disagree
- ❍ Neither Disagree nor Agree
- ❍ Somewhat Agree
- ❍ Agree
- ❍ Strongly Agree

13_3_No-Fin_DV3 Celine is an hourly worker who is contracted to work from 9-6. She is working on a task until 8pm despite having no financial incentive.

Celine is a moral person.
- ❍ Strongly Disagree
- ❍ Disagree
- ❍ Somewhat Disagree
- ❍ Neither Disagree nor Agree
- ❍ Somewhat Agree
- ❍ Agree
- ❍ Strongly Agree

13_3_Fin_DV3 Celine is an hourly worker who is contracted to work from 9-6. Despite needing about 2 more hours to complete a task, she left at 6 because of her contract.

Celine is a moral person.
- ❍ Strongly Disagree
- ❍ Disagree
- ❍ Somewhat Disagree
- ❍ Neither Disagree nor Agree
- ❍ Somewhat Agree
- ❍ Agree
- ❍ Strongly Agree

13_3_No-Fin_DV4 Anna just won the lottery along with a group of colleagues. The jackpot was huge so she has enough money to live well for the rest of her life. Anna continues to work at the company despite have no financial need to work.

Anna is a moral person.
- ❍ Strongly Disagree
- ❍ Disagree
- ❍ Somewhat Disagree
- ❍ Neither Disagree nor Agree
- ❍ Somewhat Agree
- ❍ Agree
- ❍ Strongly Agree

13_3_Fin_DV4 Anna just won the lottery along with a group of colleagues. The jackpot was modest and split among many people. Anna continues to work at the company because the financial gain from the lottery was quite modest.

Anna is a moral person.
- ❍ Strongly Disagree
- ❍ Disagree
- ❍ Somewhat Disagree
- ❍ Neither Disagree nor Agree
- ❍ Somewhat Agree
- ❍ Agree
- ❍ Strongly Agree

13_3_No-Fin_DV5 Roger's parents just died and left him a large inheritance. Roger continues to work at his job even though he could retire comfortably for the rest of his life.

Roger is a moral person.
- ❍ Strongly Disagree
- ❍ Disagree
- ❍ Somewhat Disagree
- ❍ Neither Disagree nor Agree
- ❍ Somewhat Agree
- ❍ Agree
- ❍ Strongly Agree

13_3_Fin_DV5 Roger's parents just died and left him a modest inheritance. Roger continues to work at his job because he does not have the financial resources to retire.

Roger is a moral person.
- ❍ Strongly Disagree
- ❍ Disagree
- ❍ Somewhat Disagree
- ❍ Neither Disagree nor Agree
- ❍ Somewhat Agree
- ❍ Agree
- ❍ Strongly Agree

13_3_No-Fin_DV6 Joan is a professor at the local university. Joan also has a side business making wine. Joan makes a great deal of money selling her wines. Joan continues to work as a professor even though she does not need her salary.

Joan is a moral person.
- ❍ Strongly Disagree
- ❍ Disagree
- ❍ Somewhat Disagree
- ❍ Neither Disagree nor Agree
- ❍ Somewhat Agree

❍ Agree

❍ Strongly Agree

Q13_3_Fin_DV6 Joan is a professor at the local university. Joan also has a side business making wine. Joan makes little money selling her wines. Joan continues to work as a professor because she needs her salary to support herself.

Joan is a moral person.
❍ Strongly Disagree

❍ Disagree

❍ Somewhat Disagree

❍ Neither Disagree nor Agree

❍ Somewhat Agree

❍ Agree

❍ Strongly Agree

Team 13 Analysis Plan
Participants will respond to three *DV* questions in the no financial need condition, and three *DV* questions in the financial need condition. The DV will be the mean of responses to each of these questions. We will compare the no financial need and financial need conditions using a paired-samples t-test. The effect size will be a repeated-measures Cohen's *d*, which will be converted to an independent-groups *d* for comparison to other effect sizes.

Original Materials
John Retires Condition

14_3_John_DV Robert and John are both 25-year-olds who work as lawn mowers for a landscaping company. Every week they buy a lottery ticket together. One week their ticket turns out to be the winning one and they share a $10 million jackpot 50-50. After winning the lottery, John retires young and never works at a paid job again. Robert continues to work as a lawn mower for the rest of his life. Who do you think is a morally better person?
❍ 1 definitely John

❍ 2

❍ 3

❍ 4

❍ 5

❍ 6

❍ 7 definitely Robert

Robert Retires Condition

14_3_Robert_DV Robert and John are both 25-year-olds who work as lawn mowers for a landscaping company. Every week they buy a lottery ticket together. One week their ticket turns out to be the winning one and they share a $10 million jackpot 50-50. After winning the lottery, Robert retires young and never works at a paid job again. John continues to work as a lawn mower for the rest of his life. Who do you think is a morally better person?

- ❍ 1 definitely John
- ❍ 2
- ❍ 3
- ❍ 4
- ❍ 5
- ❍ 6
- ❍ 7 definitely Robert

Original Materials Analysis Plan

The DV will be responses to *14_3_John_DV* or *14_3_Robert_DV*, depending on random assignment. Responses to *14_3_Robert_DV* will be reverse-scored, such that higher scores always reflect greater approval of the person who continues working. We will compare these responses to a null hypothesis of $\mu = 4$, using a one-sample t-test. The effect size will be a single-sample Cohen's *d*.

Research Question 4: Part of why people are opposed to the use of performance enhancing drugs in sports is because they are "against the rules". But which contributes more to this judgment - whether the performance enhancer is against the law, or whether it is against the rules established by a more proximal authority (e.g., the league)?

Team 1 Materials

League Law Condition

1_4_Lea_Intro Keith is a track and field athlete about to compete in his Host Country's National Track and Field Competition which is regulated by the World Track Association (WTA). In his Host Country, the use of performance enhancing drugs is legal under country law, but it is illegal under WTA rule. Keith really wants the gold medal this year and decides to take performance enhancing drugs as a “boost” to his performance.

1_4_Lea_DV How wrong do you think Keith is for taking the performance enhancing drugs?

- ❍ Not wrong at all 1
- ❍ 2
- ❍ 3
- ❍ Somewhat wrong 4
- ❍ 5
- ❍ 6
- ❍ Very wrong 7

Country Law Condition

1_4_Cou_Intro Keith is a track and field athlete about to compete in his Host Country's National Track and Field Competition which is regulated by the World Track Association (WTA). In his Host Country, the use of performance enhancing drugs is illegal under country law, but it is legal under WTA rule. Keith really wants the gold medal this year and decides to take performance enhancing drugs as a “boost” to his performance.

1_4_Cou_DV How wrong do you think Keith is for taking the performance enhancing drugs?

- ❍ Not wrong at all 1
- ❍ 2
- ❍ 3
- ❍ Somewhat wrong 4
- ❍ 5
- ❍ 6
- ❍ Very wrong 7

Team 1 Analysis Plan

The DV will be responses to *1_4_Lea_DV* or *1_4_Cou_DV*, depending on condition. We will compare the league law and country law conditions using an independent-samples t-test. The effect size will be an independent-groups Cohen's *d*.

Team 2 Materials

Legal But Banned Condition

2_4_Ban_Intro The sport of Adventure Racing is sweeping the world, with popularity growing by leaps and bounds in recent years. Each race is a unique expedition on a course designed to test athletes for up to 10 days of non-stop racing in the disciplines of trekking, mountain biking, kayaking, navigation and more. Adventure racers may find themselves ripping down rapids in a canoe, rappelling off a 100 foot rock face, and then shredding through tight single track on a mountain bike, all in one day.

The United States Adventure Racing Association (USARA) is the official governing body of adventure racing in the USA. It originally formed as a means to bring interested athletes together and popularize the sport. As interest in Adventure Racing grew around the world, the USARA partnered with similar organizations around the world and created the Adventure Racing (AR) World Series. The AR World Series leads up to the AR World Championship, which has crowned the world's top adventure racers since 2001.

As popularity has grown, so too has the prize money and even more lucrative corporate sponsorships, which can mean millions of dollars for top athletes. Unfortunately, now more than ever, athletes are looking for any advantage they can get, including performance enhancing drugs. The USARA is currently investigating the case of Scott Evans, a young racer who recently burst onto the scene. Multiple samples taken at races in March and May of 2016 indicate that Evans has been taking a compound called adrexiphol, which is known to boost lung capacity and blood oxygen levels.

Adrexiphol is a synthetic compound that was created by researchers who were experimenting with ways to improve lung function among elderly patients with chronic obstructive pulmonary diseases (COPD), such as the type of emphysema caused by longtime tobacco use. Because adrexiphol is synthetic, there are no natural causes of having adrexiphol in a person's bloodstream; someone must ingest it directly.

Because adrexiphol is a new compound and quite different from any other family of substances, it does not fall under any classification set by the US Food and Drug Administration (FDA). Therefore, it is legal to possess and consume in the United States, even without doctor

supervision. However, the USARA banned adrexiphol in February, 2015 because it was concerned it could be used as a performance enhancing drug.

Athletes who violate the USARA policy on performance enhancing drugs are disqualified from races for a period of time that depends on the severity of the offense. For example, use of human growth hormone or anabolic steroids typically yields a 12-24 month ban, whereas improper use of other banned substances may yield a 1-12 month ban (e.g., Albuterol, which is used to treat asthma, or Adderall, which is used to treat ADHD - attention deficit hyperactivity disorder). In other cases, athletes may be formally reprimanded and put on a special watch list but not banned from competition for any period of time.

2_4_Ban_DV1 Should Scott Evans be banned from competition for taking adrexiphol? (Enter the number of months of the ban from 0 to 24)

2_4_Ban_DV2 How severe is Scott Evans's offense?
- ❍ 1 Not at all severe
- ❍ 2
- ❍ 3
- ❍ 4
- ❍ 5
- ❍ 6
- ❍ 7 Extremely severe

2_4_Ban_DV3 How acceptable or unacceptable is it for racers to use adrexiphol?
- ❍ very acceptable
- ❍ moderately acceptable
- ❍ slightly acceptable
- ❍ neither acceptable nor unacceptable
- ❍ slightly unacceptable
- ❍ moderately unacceptable
- ❍ very unacceptable

2_4_Ban_MC1 According to the article, is adrexiphol legal or illegal according to the US Food and Drug Administration?
- ❍ legal
- ❍ illegal

❍ I don't know

2_4_Ban_MC2 According to the article, is adrexiphol permitted or not permitted under the rules of the US Adventure Racing Association?

❍ permissible

❍ impermissible

❍ I don't know

Illegal But Not Banned Condition

2_4_Ill_Intro The sport of Adventure Racing is sweeping the world, with popularity growing by leaps and bounds in recent years. Each race is a unique expedition on a course designed to test athletes for up to 10 days of non-stop racing in the disciplines of trekking, mountain biking, kayaking, navigation and more. Adventure racers may find themselves ripping down rapids in a canoe, rappelling off a 100 foot rock face, and then shredding through tight single track on a mountain bike, all in one day.

The United States Adventure Racing Association (USARA) is the official governing body of adventure racing in the USA. It originally formed as a means to bring interested athletes together and popularize the sport. As interest in Adventure Racing grew around the world, the USARA partnered with similar organizations around the world and created the Adventure Racing (AR) World Series. The AR World Series leads up to the AR World Championship, which has crowned the world's top adventure racers since 2001.

As popularity has grown, so too has the prize money and even more lucrative corporate sponsorships, which can mean millions of dollars for top athletes. Unfortunately, now more than ever, athletes are looking for any advantage they can get, including performance enhancing drugs. The USARA is currently investigating the case of Scott Evans, a young racer who recently burst onto the scene. Multiple samples taken at races in March and May of 2016 indicate that Evans has been taking a compound called adrexiphol, which is known to boost lung capacity and blood oxygen levels.

Adrexiphol is a synthetic compound that was created by researchers who were experimenting with ways to improve lung function among elderly patients with chronic obstructive pulmonary diseases (COPD), such as the type of emphysema caused by longtime tobacco use. Because adrexiphol is synthetic, there are no natural causes of having adrexiphol in a person's bloodstream; someone must ingest it directly.

Because adrexiphol is a new compound and quite different from any other family of substances, it does not fall under any classification for performance enhancing drugs set by the USARA and is not on the list of banned substances. Therefore, it is permissible to use according to USARA rules. In February, 2015, however, the US Food and Drug Administration (FDA) made adrexiphol illegal to possess and consume in the United States, unless under a doctor's supervision for treatment of a chronic lung disease.

Athletes who violate the USARA policy on performance enhancing drugs are disqualified from races for a period of time that depends on the severity of the offense. For example, use of human growth hormone or anabolic steroids typically yields a 12-24 month ban, whereas improper use of other banned substances may yield a 1-12 month ban (e.g., Albuterol, which is used to treat asthma, or Adderall, which is used to treat ADHD - attention deficit hyperactivity disorder). In other cases, athletes may be formally reprimanded and put on a special watch list but not banned from competition for any period of time.

2_4_Ill_DV1 Should Scott Evans be banned from competition for taking adrexiphol? (Enter the number of months of the ban from 0 to 24)

2_4_Ill_DV2 How severe is Scott Evans's offense?

- ❍ 1 Not at all severe
- ❍ 2
- ❍ 3
- ❍ 4
- ❍ 5
- ❍ 6
- ❍ 7 Extremely severe

2_4_Ill_DV3 How acceptable or unacceptable is it for racers to use adrexiphol?

- ❍ very acceptable
- ❍ moderately acceptable
- ❍ slightly acceptable
- ❍ neither acceptable nor unacceptable
- ❍ slightly unacceptable
- ❍ moderately unacceptable
- ❍ very unacceptable

2_4_Ill_MC1 According to the article, is adrexiphol legal or illegal according to the US Food and Drug Admininstration?

- ❍ legal
- ❍ illegal
- ❍ I don't know

2_4_Ill_MC2 According to the article, is adrexiphol permitted or not permitted under the rules of the US Adventure Racing Association?

- ❍ permissible
- ❍ impermissible
- ❍ I don't know

Team 2 Analysis Plan

Responses to all three *DV* questions (i.e., *2_4_Ban_DV1*, *2_4_Ban_DV2*, and *2_4_Ban_DV3*, or *2_4_Ill_DV1*, *2_4_Ill_DV2*, and *2_4_Ill_DV3*, depending on condition) will be z-scored. The DV will be the mean of these three z-scores. We will compare the legal but banned condition and the illegal but not banned condition using an independent-samples t-test. The effect size will be an independent-groups Cohen's *d*. If this composite DV shows poor internal reliability ($\alpha <$ .70), we will use responses to *2_4_Ban_DV1* and *2_4_Ill_DV1*.

Team 3 Materials

Prohibited Condition

3_4_Pro_Intro Tom is an ambitious athlete who participates in triathlons all over the US. Triathlon is an athletic competition that involves swimming, cycling, and running. To perform well in an upcoming triathlon, Tom exercises two times per day. To further enhance his performance, Tom takes a performance enhancing supplement called Prolexa. Prolexa increases overall endurance, reduces recovery time, and supports the immune system. The product is legal according to the US federal law, but according to the International Triathlon Union, it is prohibited to take the product before or during competition.

How do you feel about Tom taking Prolexa? Please rate the following statements:

3_4_Pro_DV1 What do you think about Tom's behavior?

I think Tom's behavior is…

- ❍ 1 …not at all acceptable
- ❍ 2
- ❍ 3
- ❍ 4
- ❍ 5

❍ 6

❍ 7 …definitely acceptable

3_4_Pro_DV2 I think Tom's behavior is…
❍ 1 …extremely morally reprehensible

❍ 2

❍ 3

❍ 4

❍ 5

❍ 6

❍ 7 …morally justifiable

3_4_Pro_DV3 I think Tom's behavior is…
❍ 1 …definitely wrong

❍ 2

❍ 3

❍ 4

❍ 5

❍ 6

❍ 7 …definitely ok

Illegal Condition
3_4_Ill_Intro Tom is an ambitious athlete who participates in triathlons all over the US. Triathlon is an athletic competition that involves swimming, cycling, and running. To perform well in an upcoming triathlon, Tom exercises two times per day. To further enhance his performance, Tom takes a performance enhancing supplement called Prolexa. Prolexa increases overall endurance, reduces recovery time, and supports the immune system. The product is illegal according to US federal law, but it is available on the black market. How do you feel about Tom taking Prolexa? Please answer the following statements:

3_4_Ill_DV1 What do you think about Tom's behavior?

I think Tom's behavior is…
❍ 1 …not at all acceptable

❍ 2

❍ 3

❍ 4

❍ 5

❍ 6

❍ 7 …definitely acceptable

3_4_Ill_DV2 I think Tom's behavior is…
❍ 1 …extremely morally reprehensible

❍ 2

❍ 3

❍ 4

❍ 5

❍ 6

❍ 7 …morally justifiable

3_4_Ill_DV3 I think Tom's behavior is…
❍ 1 …definitely wrong

❍ 2

❍ 3

❍ 4

❍ 5

❍ 6

❍ 7 …definitely ok

Team 3 Analysis Plan
To maintain consistency with other calculated effect sizes for Research Question 4, responses to all *DV* questions be reverse-scored, such that higher numbers indicate greater opposition. The DV will be the mean of responses to *3_4_Pro_DV1*, *3_4_Pro_DV2*, and *3_4_Pro_DV3*, or *3_4_Ill_DV1*, *3_4_Ill_DV2*, and *3_4_Ill_DV3*, depending on condition. We will compare the prohibited and illegal conditions using an independent-samples t-test. The effect size will be an independent-groups Cohen's *d*. If this composite DV shows poor internal reliability ($\alpha < .70$), we will use responses to *3_4_Ban_DV1* and *3_4_Ill_DV1*.

Team 4 Materials
Not Against the Law + Against Local Norm Condition

4_4_Norm_Intro This questionnaire is interested in your judgements regarding competitive sports activities. Please read the fictional scenario carefully and answer the questions given below.

James is a basketballer for his high school team who is training for a tournament game in a month's time. He needs to improve his performance and win the tournament game in order to gain a place in his preferred college squad. James wants to do well but even though he has been training, he still isn't at the level he wants. James is approached by his close friend who takes a performance enhancing drug (PED) called Teason and offers them to him. However, Teason is banned under the state high school basketball league, but is legal under Section 2.1 of the Federal Anti-Doping Act. James accepts, takes the steroids and wins the tournament game over other clean competitors, getting him selected for his preferred college squad.

4_4_Norm_DV I believe James' behavior is….

	1 strongly disagree	2	3	4	5	6	7 strongly agree
Unethical	❍	❍	❍	❍	❍	❍	❍
Bad	❍	❍	❍	❍	❍	❍	❍
Immoral	❍	❍	❍	❍	❍	❍	❍

Against the Law + Not Against Local Norm Condition
4_4_Law_Intro This questionnaire is interested in your judgements regarding competitive sports activities. Please read the fictional scenario carefully and answer the questions given below.

James is a basketballer for his high school team who is training for a tournament game in a month's time. He needs to improve his performance and win the tournament game in order to gain a place in his preferred college squad. James wants to do well but even though he has been training, he still isn't at the level he wants. James is approached by his close friend who takes a performance enhancing drug (PED) called Teason and offers them to him. Teason is not banned under the state high school basketball league but it is illegal as it contravenes Section 2.1 of the Federal Anti-Doping Act. James accepts, takes the steroids and wins the tournament game over other clean competitors, getting him selected for his preferred college squad.

4_4_Law_DV I believe James' behavior is….

	1 strongly disagree	2	3	4	5	6	7 strongly agree

Unethical	❍	❍	❍	❍	❍	❍	❍
Bad	❍	❍	❍	❍	❍	❍	❍
Immoral	❍	❍	❍	❍	❍	❍	❍

Team 4 Analysis Plan

The DV will be the mean of responses to the three items in *4_4_Norm_DV* or *4_4_Law_DV*, depending on condition. We will compare the "not against the law + against local norm" condition and the "against the law + not against local norm" condition using an independent-samples t-test. The effect size will be an independent-groups Cohen's *d*. If this composite DV shows poor internal reliability ($\alpha < .70$), we will use responses to item 3 (immoral)

Team 5 Materials

Against the League Condition

5_4_Lea_Intro John is a track and field athlete. Recently, he was found to use performance enhancing drugs in a competition. However, the use of performance enhancing drugs in sports is prohibited by the rules set by the Association of Athletics Federation.

5_4_Lea_DV1 How strongly are you opposed to John's use of performance enhancing drugs in sports?

❍ 1 Not at all opposed

❍ 2

❍ 3

❍ 4

❍ 5

❍ 6

❍ 7 Very opposed

5_4_Lea_DV2 How wrong do you think John's action is to use of performance enhancing drugs in sports?

❍ 1 Not at all wrong

❍ 2

❍ 3

❍ 4

❍ 5

❍ 6

❍ 7 Very wrong

Against the Law Condition
5_4_Law_Intro John is a track and field athlete. Recently, he was found to use performance enhancing drugs in a competition. However, the use of performance enhancing drugs in sports is prohibited by law.

5_4_Law_DV1 How strongly are you opposed to John's use of performance enhancing drugs in sports?
❍ 1 Not at all opposed
❍ 2
❍ 3
❍ 4
❍ 5
❍ 6
❍ 7 Very opposed

5_4_Law_DV2 How wrong do you think John's action is to use of performance enhancing drugs in sports?
❍ 1 Not at all wrong
❍ 2
❍ 3
❍ 4
❍ 5
❍ 6
❍ 7 Very wrong

Team 5 Analysis Plan
The DV will be the mean of responses to *5_4_Lea_DV1* and *5_4_Lea_DV2*, or *5_4_Law_DV1* and *5_4_Law_DV2*, depending on condition. We will compare the "against the league" condition and the "against the law" condition using an independent-samples t-test. The effect size will be an independent-groups Cohen's *d*. If this composite DV shows poor internal reliability ($\alpha < .70$), we will use responses to *5_4_Lea_DV1* and *5_4_Law_DV1*.

Team 6 Materials
Against the League Condition
6_4_Lea_Intro1 In this study, we are interested in your opinions of different things. On the next page you will be provided with a piece of news on performance enhancing drugs (PEDs) in sports. Please read the news story carefully and then answer questions.

P.S. we refer to the athlete as John Doe for confidentiality.

6_4_Lea_Intro2 After a routine drug test at the 2012 Australian Open tennis tournament, John Doe tested positive for a banned substance. Upon being notified of the result, he called a press conference, accepting personal responsibility for an inadvertent infringement of the Tennis Anti-Doping Programme (TADP). The TADP is a unified set of rules that is administered and enforced by the International Tennis Federation (ITF) on behalf of the governing bodies of professional tennis (i.e. the ITF, ATP, WTA and the four Grand Slams).

6_4_Lea_DV1 How serious do you think John Doe's misconduct is?
- ❍ 1 not at all
- ❍ 2
- ❍ 3
- ❍ 4
- ❍ 5
- ❍ 6
- ❍ 7 very much

6_4_Lea_DV2 How many years do you think John Doe should be suspended?
- ❍ one year
- ❍ two years
- ❍ three years
- ❍ four years
- ❍ five years
- ❍ six years
- ❍ seven years

Against the Law Condition
6_4_Law_Intro1 In this study, we are interested in your opinions of different things. On the next page you will be provided with a piece of news on performance enhancing drugs (PEDs) in sports. Please read the news story carefully and then answer questions. P.S. we refer to the athlete as John Doe for confidentiality.

6_4_Law_Intro2 After a routine drug test at the 2012 Australian Open tennis tournament, John Doe tested positive for a banned substance. Upon being notified of the result, he called a press conference, accepting personal responsibility for an inadvertent infringement of the World Anti-Doping Agency (WADA) Code. WADA code is the document regulating anti-doping policies in all sports and all countries. It is administered by WADA, an international independent agency composed and funded equally by the athletic leagues and governments of the world.

6_4_Law_DV1 How serious do you think John Doe's misconduct is?

- ❍ 1 not at all
- ❍ 2
- ❍ 3
- ❍ 4
- ❍ 5
- ❍ 6
- ❍ 7 very much

6_4_Law_DV2 How many years do you think John Doe should be suspended?

- ❍ one year
- ❍ two years
- ❍ three years
- ❍ four years
- ❍ five years
- ❍ six years
- ❍ seven years

Team 6 Analysis Plan

The DV will be the mean of responses to *6_4_Lea_DV1* and *6_4_Lea_DV2*, or *6_4_Law_DV1* and *6_4_Law_DV2*, depending on condition. We will compare the "against the league" condition and the "against the law" condition using an independent-samples t-test. The effect size will be an independent-groups Cohen's *d*. If this composite DV shows poor internal reliability ($\alpha < .70$), we will use responses to *6_4_Lea_DV2* and *6_4_Law_DV2*.

Team 7 Materials

Violate Rules of the Sport Condition

7_4_Rule_Intro Daniel is a French professional cyclist who has been racing professionally for the past 6 years. Within the professional ranks, he usually places somewhat highly in races (often in the top 20 out of 100 or more cyclists), but he's looking to get to the next level and start getting some top-10 results.

Daniel has heard from other cyclists that meldonium, a drug that boosts blood and oxygen flow in the body, can give you a slight edge in races, especially the ones that are very long and fatiguing. Daniel starts taking meldonium, but he doesn't tell anyone and is careful to burn the packaging that the drug comes in.

Although using meldonium is not illegal, it is a banned substance within the sport of cycling.

7_4_Rule_DV How morally acceptable are Daniel's actions?

- ❍ 1 totally unacceptable
- ❍ 2
- ❍ 3
- ❍ 4
- ❍ 5
- ❍ 6
- ❍ 7 totally acceptable

Violate the Law Condition

7_4_Law_Intro Daniel is a French professional cyclist who has been racing professionally for the past 6 years. Within the professional ranks, he usually places somewhat highly in races (often in the top 20 out of 100 or more cyclists), but he's looking to get to the next level and start getting some top-10 results. Daniel has heard from other cyclists that meldonium, a drug that boosts blood and oxygen flow in the body, can give you a slight edge in races, especially the ones that are very long and fatiguing. Daniel starts taking meldonium, but he doesn't tell anyone and is careful to burn the packaging that the drug comes in. Meldonium is not only a banned substance within the sport of cycling, but taking performance-enhancing drugs is also against the law in France.

7_4_Law_DV How morally acceptable are Daniel's actions?

- ❍ 1 totally unacceptable
- ❍ 2
- ❍ 3
- ❍ 4
- ❍ 5
- ❍ 6
- ❍ 7 totally acceptable

Team 7 Analysis Plan

Responses to *DV* questions will be reverse-scored, such that higher scores indicate greater opposition, to be consistent with other effect sizes calculated for Research Question 4. The DV will be the reverse-scored responses to *7_4_Rule_DV* or *7_4_Law_DV*, depending on condition. We will compare the "violate rules of the sport" condition and the "violate the law" condition using an independent-samples t-test. The effect size will be an independent-groups Cohen's *d*.

Team 8 Materials

Against the League Condition

8_4_Lea_Intro We are interested in gathering information about people's opinions about the use of performance-enhancing drugs in sports. Below, we'd like you to read about a debate about performance-enhancing drugs in sports and provide your opinion about this specific case.

Should Oxopedrone Be Allowed in American Soccer?

June 29, 2016 - Washington DC - Yahoo Sports

Use of Oxopedrone stimulant injections is a controversial topic in American soccer. While the US Government has no regulations prohibiting the use of the drug, the sport's official governing body currently forbids its use in official competitions.

Now, some players and fans want to repeal the United States Soccer Federation's ban on the use of Oxopedrone. Preliminary medical studies have not demonstrated any negative side effects to Oxopedrone use; however, several athletes known to be using the stimulant have recently exhibited tremendous upswings in their performance....

8_4_Lea_DV Based on this excerpt of the article, please give your opinion on whether Oxopedrone injections should be permitted in American soccer. Please rate your agreement with the following statements using the following scale.

	Strongly Disagree	Disagree	Slightly Disagree	Neither Agree Nor Disagree	Slightly Agree	Agree	Strongly Agree
Oxopedrone injections should be permitted in American Soccer	❍	❍	❍	❍	❍	❍	❍
It's wrong for athletes	❍	❍	❍	❍	❍	❍	❍

to use Oxopedrone to enhance their performance I support the movement to repeal the ban on Oxopedrone use	❍	❍	❍	❍	❍	❍	❍

8_4_Lea_MC How familiar would you say you are with soccer?

❍ Extremely Unfamiliar

❍ Rather Unfamiliar

❍ Slightly Unfamiliar

❍ Neither familiar nor unfamiliar

❍ Slightly Familiar

❍ Rather Familiar

❍ Extremely Familiar

Against the Law Condition

8_4_Law_Intro We are interested in gathering information about people's opinions about the use of performance-enhancing drugs in sports. Below, we'd like you to read about a debate about performance-enhancing drugs in sports and provide your opinion about this specific case.

Should Oxopedrone Be Allowed in American Soccer?

June 29, 2016 - Washington DC - Yahoo Sports

Use of Oxopedrone stimulant injections is a controversial topic in American soccer. While the sport's official governing body has no regulations prohibiting the use of the drug in official competitions, the US Government currently forbids its use.

Now, some players and fans want to repeal the national ban on the use of Oxopedrone. Preliminary medical studies have not demonstrated any negative side effects to Oxopedrone use; however, several athletes known to be using the stimulant have recently exhibited tremendous upswings in their performance....

8_4_Law_DV Based on this excerpt of the article, please give your opinion on whether Oxopedrone injections should be permitted in American soccer. Please rate your agreement with the following statements using the following scale.

	Strongly Disagree	Disagree	Slightly Disagree	Neither Agree Nor Disagree	Slightly Agree	Agree	Strongly Agree
Oxopedrone injections should be permitted in American Soccer	❍	❍	❍	❍	❍	❍	❍
It's wrong for athletes to use Oxopedrone to enhance their performance	❍	❍	❍	❍	❍	❍	❍
I support the movement to repeal the ban on Oxopedrone use	❍	❍	❍	❍	❍	❍	❍

8_4_Law_MC How familiar would you say you are with soccer?
- ❍ Extremely Unfamiliar
- ❍ Rather Unfamiliar
- ❍ Slightly Unfamiliar
- ❍ Neither familiar nor unfamiliar
- ❍ Slightly Familiar
- ❍ Rather Familiar
- ❍ Extremely Familiar

Team 8 Analysis Plan

Participants who say that they are "Extremely familiar" with soccer in *8_4_Lea_MC* or *8_4_Law_MC* will be excluded from analysis. The DV will be the mean of responses to the three items in *8_4_Lea_DV* or *8_4_Law_DV*, depending on condition (items 1 and 3 will be reverse-scored, so that higher scores indicate greater opposition to performance enhancers). We will compare the "against the league" condition and the "against the law" condition using an independent-samples t-test. The effect size will be an independent-groups Cohen's *d*. If this composite DV shows poor internal reliability ($\alpha < .70$), we will use the reverse-scored responses to item 1 ("Oxopedrone injections should be permitted in American soccer").

Team 9 Materials

Against Proximal Authority Condition

9_4_Aut_Intro In the following task you will read a vignette and will be asked about your position.

Roy is an American professional runner that has won several marathons and has been a recognized athlete in the American Running Association during the past years. He uses performance enhancing drugs to stimulate his body and perform at optimal levels by increasing focus and energy.

What Roy does is against the rules of the American Running Association, while it is not against the laws of his country.

9_4_Aut_DV Under these conditions, to what extent are you opposed to the use of performance enhancing drugs in sports because it is against the rules? (where 1 means "none at all" and 7 "absolutely")
- ❍ 1
- ❍ 2
- ❍ 3
- ❍ 4

❍ 5

❍ 6

❍ 7

Against Law Condition
9_4_Law_Intro In the following task you will read a vignette and will be asked about your position.

Roy is an American professional runner that has won several marathons and has been a recognized athlete in the American Running Association during the past years. He uses performance enhancing drugs to stimulate his body and perform at optimal levels by increasing focus and energy.

What Roy does is against the laws of his country, while it is not against the rules of the American Running Association.

9_4_Law_DV Under these conditions, to what extent are you opposed to the use of performance enhancing drugs in sports because it is against the rules? (where 1 means "none at all" and 7 "absolutely")
❍ 1

❍ 2

❍ 3

❍ 4

❍ 5

❍ 6

❍ 7

Team 9 Analysis Plan
The DV will be responses to *9_4_Aut_DV* or *9_4_Law_DV*, depending on condition. We will compare the "against proximal authority" condition and the "against the law" condition using an independent-samples t-test. The effect size will be an independent-groups Cohen's *d*.

Team 10 did not develop materials for this research question.

Team 11 Materials

Banned Condition
11_4_Ban_Intro On the next page, you will read a scenario about a person doing something, which may or may not be considered wrong. We would like you to tell us how wrong you consider the person's action in the scenario to be. Some of the details of the scenario may seem

implausible, but **please assume that all of the provided information is true, and that the person described in the scenario is fully aware of this.**

There are no right or wrong answers in this survey, we are simply interested in your opinions. All of your responses are anonymous, so please respond as honestly as possible.

11_4_Ban_DV Joe is a professional competitive weightlifter. Joe does not use performance-enhancing substances like steroids, but he is considering starting to use a performance enhancing substance to improve his performance. Joe is fully aware that he is the only competitor in his competitive weightlifting circuit who is not currently using steroids. *All of the people Joe competes against* are currently using steroids that are legal, but are banned under their competitive circuit's rules. Knowing this, Joe decides to start using anabolic steroids to improve his performance, since it is not against the law, even though it does violate the rules of his circuit.

How wrong is it for Joe to use steroids to improve his performance?

- ❍ Not at all wrong
- ❍
- ❍
- ❍
- ❍ Moderately wrong
- ❍
- ❍
- ❍
- ❍ Extremely wrong

11_4_Ban_MC1 Did Joe break any law by taking anabolic steroids?

- ❍ Yes
- ❍ No
- ❍ Not Sure

11_4_Ban_MC2 Did Joe break any rules of his competitive circuit by taking anabolic steroids?

- ❍ Yes
- ❍ No

❍ Not Sure

Illegal Condition
11_4_Illl_Intro On the next page, you will read a scenario about a person doing something, which may or may not be considered wrong. We would like you to tell us how wrong you consider the person's action in the scenario to be. Some of the details of the scenario may seem implausible, but **please assume that all of the provided information is true, and that the person described in the scenario is fully aware of this.**

There are no right or wrong answers in this survey, we are simply interested in your opinions. All of your responses are anonymous, so please respond as honestly as possible.

11_4_Ill_DV Joe is a professional competitive weightlifter. Joe does not use performance-enhancing substances like steroids, but he is considering starting to use a performance enhancing substance to improve his performance. Joe is fully aware that he is the only competitor in his competitive weightlifting circuit who is not currently using steroids. *All of the people Joe competes against* are currently using steroids that are illegal, but are permitted under their competitive circuit's rules. Knowing this, Joe decides to start using anabolic steroids to improve his performance, since it does not violate the rules of his circuit, even though it is against the law.

How wrong is it for Joe to use steroids to improve his performance?
❍ Not at all wrong
❍
❍
❍
❍ Moderately wrong
❍
❍
❍
❍ Extremely wrong

11_4_Ill_MC1 Did Joe break any law by taking anabolic steroids?
❍ Yes
❍ No
❍ Not Sure

11_4_Ill_MC2 Did Joe break any rules of his competitive circuit by taking anabolic steroids?
- ❍ Yes
- ❍ No
- ❍ Not Sure

Team 11 Analysis Plan
The DV will be responses to *11_4_Ban_DV* or *11_4_Ill_DV*, depending on condition. We will compare the "banned" condition and the "illegal" condition using an independent-samples t-test. The effect size will be an independent-groups Cohen's *d*.

Team 12 Materials
Proximal Authority Condition
12_4_PA_Intro On April 1, 2016, US sprinter Jimmy Chambers became the first person to test positive for the steroid THG (tetrahydrogestrinone) in a drug test during the Olympic trials.

The use of the steroid THG is against the rules set by the International Olympic Committee (IOC).

12_4_PA_DV1 On a scale of 1 to 7, how wrong was it for Chambers to take the steroid THG?
- ❍ 1 Not Wrong At All
- ❍ 2
- ❍ 3
- ❍ 4
- ❍ 5
- ❍ 6
- ❍ 7 Extremely Wrong

12_4_PA_DV2 On a scale of 1 to 7, how severely should Chambers be punished in your opinion?
- ❍ 1 No Punishment
- ❍ 2
- ❍ 3
- ❍ 4
- ❍ 5
- ❍ 6
- ❍ 7 Severe Punishment

Legal Authority Condition
12_4_LA_Intro On April 1, 2016, US sprinter Jimmy Chambers became the first person to test positive for the steroid THG (tetrahydrogestrinone) in a drug test during the Olympic trials. The use of the steroid THG is against US law.

12_4_LA_DV1 On a scale of 1 to 7, how wrong was it for Chambers to take the steroid THG?
❍ 1 Not Wrong At All

❍ 2

❍ 3

❍ 4

❍ 5

❍ 6

❍ 7 Extremely Wrong

12_4_LA_DV2 On a scale of 1 to 7, how severely should Chambers be punished in your opinion?
❍ 1 No Punishment

❍ 2

❍ 3

❍ 4

❍ 5

❍ 6

❍ 7 Severe Punishment

Team 12 Analysis Plan
The DV will be the mean of responses to *12_4_PA_DV1* and *12_4_PA_DV2*, or *12_4_LA_DV1* and *12_4_LA_DV2*, depending on condition. We will compare the "proximal authority" condition and the "legal authority" condition using an independent-samples t-test. The effect size will be an independent-groups Cohen's *d*. If this composite DV shows poor internal reliability ($\alpha < .70$), we will use responses to *12_4_PA_DV1* and *12_4_LA_DV1*.

Team 13 Materials
General Note: These materials consist of a "banned" version and an "illegal" version of six different scenarios. Participants will be randomly assigned to see one version of each scenario, for a total of six scenarios, three of which will be "banned" and three of which will be "illegal". The order of presentation of the six scenarios will be randomized.

13_4 _Intro There's been a lot of talk about performance enhancing drugs in sports. These drugs increase performance but have health implications such as increased rates of cancers, heart attacks, strokes, and organ failures. Some of these drugs are illegal in various countries and many are banned by the international organizations that govern individual sports. We are interested in your opinions of the following athletes who have tested positive for performance enhancing drugs.

13_4_Ban_DV1 A cyclist recently tested positive for taking a newly developed drug that increases red blood cell production and helps athletes train harder. This drug is legal in this cyclist's home country and is banned by the governing body of the sport (the Union Cycliste Internationale). Thus, this athlete took a drug that was legal and banned.

How wrong was it for this athlete to take this drug?

- ❍ Not at all morally wrong
- ❍ very slightly morally wrong
- ❍ slightly morally wrong
- ❍ somewhat morally wrong
- ❍ mostly morally wrong
- ❍ almost completely morally wrong
- ❍ Completely morally wrong

13_4_Ill_DV1 A cyclist recently tested positive for taking a newly developed drug that increases red blood cell production and helps athletes train harder. This drug is illegal in this cyclist's home country and is not banned by the governing body of the sport (the Union Cycliste Internationale). Thus, this athlete took a drug that was illegal and not banned.

How wrong was it for this athlete to take this drug?

- ❍ Not at all morally wrong
- ❍ very slightly morally wrong
- ❍ slightly morally wrong
- ❍ somewhat morally wrong
- ❍ mostly morally wrong
- ❍ almost completely morally wrong
- ❍ Completely morally wrong

13_4_Ban_DV2 A tennis player recently tested positive for taking a drug that increases blood flow to muscles to aid in recovery. This drug is legal in the player's home country and is banned

by the governing body of the sport (the International Tennis Federation). Thus, this athlete took a drug that was legal and banned.

How wrong was it for this athlete to take this drug?

- ❍ Not at all morally wrong
- ❍ very slightly morally wrong
- ❍ slightly morally wrong
- ❍ somewhat morally wrong
- ❍ mostly morally wrong
- ❍ almost completely morally wrong
- ❍ Completely morally wrong

13_4_Ill_DV2 A tennis player recently tested positive for taking a drug that increases blood flow to muscles to aid in recovery. This drug is illegal in the player's home country and is not banned by the governing body of the sport (the International Tennis Federation). Thus, this athlete took a drug that was illegal and not banned.

How wrong was it for this athlete to take this drug?

- ❍ Not at all morally wrong
- ❍ very slightly morally wrong
- ❍ slightly morally wrong
- ❍ somewhat morally wrong
- ❍ mostly morally wrong
- ❍ almost completely morally wrong
- ❍ Completely morally wrong

13_4_Ban_DV3 An Olympian recently tested positive for taking a stimulant drug that reduces fatigue and increases concentration. Taking the drug is legal in the Olympian's home country and is banned by the governing body (the International Olympic Committee). Thus, this athlete took a drug that was legal and banned.

How wrong was it for this athlete to take this drug?

- ❍ Not at all morally wrong
- ❍ very slightly morally wrong
- ❍ slightly morally wrong
- ❍ somewhat morally wrong

- ❍ mostly morally wrong
- ❍ almost completely morally wrong
- ❍ Completely morally wrong

13_4_Ill_DV3 An Olympian recently tested positive for taking a stimulant drug that reduces fatigue and increases concentration. Taking the drug is illegal in the Olympian's home country and is not banned by the governing body (the International Olympic Committee). Thus, this athlete took a drug that was illegal and not banned.

How wrong was it for this athlete to take this drug?

- ❍ Not at all morally wrong
- ❍ very slightly morally wrong
- ❍ slightly morally wrong
- ❍ somewhat morally wrong
- ❍ mostly morally wrong
- ❍ almost completely morally wrong
- ❍ Completely morally wrong

13_4_Ban_DV4 A middle distance runner recently tested positive for taking a drug that relaxes bronchial tubes to increase airflow. Taking the drug is legal in the runner's home country and is banned by the governing body of the sport (the International Associations of Athletics Federation). Thus, this athlete took a drug that was legal and banned.

How wrong was it for this athlete to take this drug?

- ❍ Not at all morally wrong
- ❍ very slightly morally wrong
- ❍ slightly morally wrong
- ❍ somewhat morally wrong
- ❍ mostly morally wrong
- ❍ almost completely morally wrong
- ❍ Completely morally wrong

13_4_Ill_DV4 A middle distance runner recently tested positive for taking a drug that relaxes bronchial tubes to increase airflow. Taking the drug is illegal in the runner's home country and is not banned by the governing body of the sport (the International Associations of Athletics Federation). Thus, this athlete took a drug that was illegal and not banned.

How wrong was it for this athlete to take this drug?

- ❍ Not at all morally wrong
- ❍ very slightly morally wrong
- ❍ slightly morally wrong
- ❍ somewhat morally wrong
- ❍ mostly morally wrong
- ❍ almost completely morally wrong
- ❍ Completely morally wrong

13_4_Ban_DV5 A soccer (football) player recently tested positive for taking a new synthetic growth hormone, a drug that builds muscle. Taking the drug is legal in the player's home country and is banned by the governing body of the sport (the Federation Interntionale de Football Association or FIFA). Thus, this athlete took a drug that was legal and banned.

How wrong was it for this athlete to take this drug?

- ❍ Not at all morally wrong
- ❍ very slightly morally wrong
- ❍ slightly morally wrong
- ❍ somewhat morally wrong
- ❍ mostly morally wrong
- ❍ almost completely morally wrong
- ❍ Completely morally wrong

13_4_Ill_DV5 A soccer (football) player recently tested positive for taking a new synthetic growth hormone, a drug that builds muscle. Taking the drug is illegal in the player's home country and is not banned by the governing body of the sport (the Federation Interntionale de Football Association or FIFA). Thus, this athlete took a drug that was illegal and not banned.

How wrong was it for this athlete to take this drug?

- ❍ Not at all morally wrong
- ❍ very slightly morally wrong
- ❍ slightly morally wrong
- ❍ somewhat morally wrong
- ❍ mostly morally wrong
- ❍ almost completely morally wrong

❍ Completely morally wrong

13_4_Ban_DV6 An international baseball player recently tested positive for taking a new synthetic steroid. Taking the drug is legal in the player's home country and is banned by the governing body of the sport (Major League Baseball). Thus, this athlete took a drug that was legal and banned.

How wrong was it for this athlete to take this drug?
❍ Not at all morally wrong

❍ very slightly morally wrong

❍ slightly morally wrong

❍ somewhat morally wrong

❍ mostly morally wrong

❍ almost completely morally wrong

❍ Completely morally wrong

13_4_Ill_DV6 An international baseball player recently tested positive for taking a new synthetic steroid. Taking the drug is illegal in the player's home country and is not banned by the governing body of the sport (Major League Baseball). Thus, this athlete took a drug that was illegal and not banned.

How wrong was it for this athlete to take this drug?
❍ Not at all morally wrong

❍ very slightly morally wrong

❍ slightly morally wrong

❍ somewhat morally wrong

❍ mostly morally wrong

❍ almost completely morally wrong

❍ Completely morally wrong

Team 13 Analysis Plan
Participants will respond to three *DV* questions in the "banned" condition, and three *DV* questions in the "illegal" condition. The DV will be the mean of responses to each of these questions. We will compare the banned and illegal conditions using a paired-samples t-test. The effect size will be a repeated-measures Cohen's *d,* which will be converted to an independent-groups *d* for comparison to other effect sizes.

Research Question 5: Is a utilitarian vs. deontological moral orientation related to personal happiness?

Team 1 Materials

Moral Dilemmas

1_5_Moral_DV1 You are driving through a busy city street when all of a sudden a young mother carrying a child trips and falls into the path of your vehicle. You are going too fast to brake in time; your only hope is to swerve out of the way. Unfortunately, the only place you can swerve is currently occupied by a little old lady. If you swerve to avoid the young mother and baby, you will seriously injure or kill the old lady.

Is it appropriate to swerve and hit the old lady in order to avoid the young mother and child?

- Inappropriate 1
- 2
- 3
- 4
- 5
- 6
- Appropriate 7

1_5_Moral_DV2 You are the head of a poor household in a developing country. Your crops have failed for the second year in a row, and it appears that you have no way to feed your family. Your sons, ages eight and ten, are too young to go off to the city where there are jobs, but your daughter could fare better. You know a man from your village who lives in the city and who makes sexually explicit films featuring girls such as your daughter. In front of your daughter, he tells you that in one year of working in his studio your daughter could earn enough money to keep your family fed for several growing seasons.

Is it appropriate for you to employ your daughter in the pornography industry in order to feed your family?

- Inappropriate 1
- 2
- 3
- 4
- 5
- 6
- Appropriate 7

1_5_Moral_DV3 You are a police officer, and have recently caught a criminal you have been hunting for some time. He is allegedly responsible for rigging a series of explosive devices: some that have already gone off and some that have yet to detonate. He places explosives outside city cafes and sets them to go off at a time when people are drinking coffee on the patios. In this manner, he has injured many people and might injure many more. Now that the criminal is in custody, you want to know where the unexploded bombs are so you can defuse them. He refuses to talk, so you decide to use "aggressive interrogation techniques" like holding his head under water and beating him.

Is it appropriate for you to use "aggressive interrogation techniques" in order to find and defuse the unexploded bombs?

- ❍ Inappropriate 1
- ❍ 2
- ❍ 3
- ❍ 4
- ❍ 5
- ❍ 6
- ❍ Appropriate 7

Happiness

1_5_Happy_DV Please imagine a ladder with steps numbered from zero at the bottom to ten at the top. Suppose we say that the top of the ladder represents the best possible life for you and the bottom of the ladder represents the worst possible life for you. If the top step is 10 and the bottom step is 0, on which step of the ladder do you feel you personally stand at the present time?

- ❍ Worst possible life 0
- ❍ 1
- ❍ 2
- ❍ 3
- ❍ 4
- ❍ 5
- ❍ 6
- ❍ 7
- ❍ 8
- ❍ 9
- ❍ Best possible life 10

Team 1 Analysis Plan

The measure of moral orientation will be the mean of responses to *1_5_Moral_DV1*, *1_5_Moral_DV2*, and *1_5_Moral_DV3*. If this composite measure shows poor internal reliability ($\alpha < .70$), we will use responses to *1_5_Moral_DV1*. The measure of happiness will be responses to *1_5_Happy_DV*. The effect size will be the Pearson correlation (*r*) between these two measures.

Team 2 Materials

Ethical Viewpoints

2_5_Moral_DV How important to you are the following character traits? (1 = not at all important; 7 = extremely important)

	1 not at all important	2	3	4	5	6	7 extremely important
innovative	❍	❍	❍	❍	❍	❍	❍
principled	❍	❍	❍	❍	❍	❍	❍
benevolent	❍	❍	❍	❍	❍	❍	❍
dependable	❍	❍	❍	❍	❍	❍	❍
resourceful	❍	❍	❍	❍	❍	❍	❍
trustworthy	❍	❍	❍	❍	❍	❍	❍
effective	❍	❍	❍	❍	❍	❍	❍
honest	❍	❍	❍	❍	❍	❍	❍
influential	❍	❍	❍	❍	❍	❍	❍
dutiful	❍	❍	❍	❍	❍	❍	❍
independent	❍	❍	❍	❍	❍	❍	❍
dedicated to cause	❍	❍	❍	❍	❍	❍	❍
results-oriented	❍	❍	❍	❍	❍	❍	❍
good intentioned	❍	❍	❍	❍	❍	❍	❍

productive	❍	❍	❍	❍	❍	❍	❍
noted for integrity	❍	❍	❍	❍	❍	❍	❍
compassionate	❍	❍	❍	❍	❍	❍	❍
financially secure	❍	❍	❍	❍	❍	❍	❍
law-abiding	❍	❍	❍	❍	❍	❍	❍
a winner	❍	❍	❍	❍	❍	❍	❍

Happiness
2_5_Happy_Intro For each of the following statements and/or questions, please select the point on the scale that you feel is most appropriate in describing you.

2_5_Happy_DV1 In general, I consider myself:
❍ 1 Not a very happy person

❍ 2

❍ 3

❍ 4

❍ 5

❍ 6

❍ 7 A very happy person

2_5_Happy_DV2 Compared with most of my peers, I consider myself:
❍ 1 Less happy

❍ 2

❍ 3

❍ 4

❍ 5

❍ 6

❍ 7 More happy

2_5_Happy_DV3 Some people are generally very happy. They enjoy life regardless of what is going on, getting the most out of everything. To what extent does this characterization describe you?

❍ 1 Not at all

❍ 2

❍ 3

❍ 4

❍ 5

❍ 6

❍ 7 A great deal

2_5_Happy_DV4 Some people are generally not very happy. Although they are not depressed, they never seem as happy as they might be. To what extent does this characterization describe you?

❍ 1 Not at all

❍ 2

❍ 3

❍ 4

❍ 5

❍ 6

❍ 7 A great deal

Team 2 Analysis Plan

The measure of moral orientation will be the mean of responses to items 2, 4, 6, 8, 16, and 19 of *2_5_Moral_DV* (i.e., the "deontology index: *principled, dependable, trustworthy, honest, noted for integrity, law-abiding*), reverse-scored, such that higher scores indicate a less deontological moral orientation. If this composite measure shows poor internal reliability ($\alpha < .70$), we will use responses to item 6 (trustworthy). The measure of happiness will be the mean of responses to *2_5_Happy_DV1, 2_5_Happy_DV2, 2_5_Happy_DV3,* and *2_5_Happy_DV4.* If this composite measure shows poor internal reliability ($\alpha < .70$), we will use responses to *2_5_Happy_DV1*. The effect size will be the Pearson correlation (*r*) between these two measures.

Team 3 Materials

Moral Dilemmas

3_5_Moral_DV It is war time. Enemy soldiers have taken over your village. They have orders to kill all remaining civilians. You and some of your townspeople have sought refuge in the cellar of a large house. Outside you hear the voices of soldiers who have come to search the house for

valuables. A baby with no parents begins to cry loudly. You cover her mouth to block the sound. If you remove your hand from the baby's mouth her crying will summon the attention of the soldiers who will kill you and the others hiding out in the cellar. To save yourself and the others you must smother the child to death.
Is it acceptable for you to smother the child in order to save yourself and the other townspeople from being killed?

❍ 1…not at all acceptable

❍ 2

❍ 3

❍ 4

❍ 5

❍ 6

❍ 7…definitely acceptable

Happiness
3_5_Happy_DV Below are five statements that you may agree or disagree with. Using the 1 - 7 scale below, indicate your agreement with each item. Please be open and honest in your responding.

	1 - Strongly disagree	2 - Disagree	3 - Slightly disagree	4 - Neither agree nor disagree	5 - Slightly agree	6 - Agree	7 - Strongly agree
In most ways my life is close to my ideal.	❍	❍	❍	❍	❍	❍	❍
The conditions of my life are excellent.	❍	❍	❍	❍	❍	❍	❍
I am satisfied with my life.	❍	❍	❍	❍	❍	❍	❍

So far I have gotten the important things I want in life.	❍	❍	❍	❍	❍	❍	❍
If I could live my life over, I would change almost nothing.	❍	❍	❍	❍	❍	❍	❍

Team 3 Analysis Plan

The measure of moral orientation will be responses to *3_5_Moral_DV*. The measure of happiness will be the mean of responses to the five items in *3_5_Happy_DV*. If this composite measure shows poor internal reliability ($\alpha < .70$), we will use responses to item 1 ("In most ways my life is close to ideal"). The effect size will be the Pearson correlation (*r*) between these two measures.

Team 4 Materials

Moral Orientation

4_5_Moral_Intro Below are some scenarios that assess your preferences in certain ambiguous situations. When you are completing this measure, please remember that there are no right or wrong answers to these questions. Indicate the way which would best fit your way of thinking about the situation.

4_5_Moral_DV1 In front of the cafeteria on a major university campus is a busy two-lane road with a cross walk and a traffic light. There is no intersection, but the light can be controlled by a pedestrian button on each sidewalk. When there is a little traffic, a person would either press the button and wait for the light or just walk across without the light.

❍ No harm is done just to go ahead; it's inconvenient to wait when there is little or no traffic

❍ One should obey all traffic laws

4_5_Moral_DV2 You are the instructor of an evening class which meets every Wednesday night. One part of the course is a library tour, in which you acquaint the students with various materials and sources for study. Unfortunately, you have just received a memo from the library director which notifies you that the tour must be conducted on the following Thursday night. When you take the proposed change to the students, all are still very interested in going on the new date except for two students who are unable to attend. Both have previous commitments, but it would be valuable for all to attend. The tour has always been part of the course.

- The class would be better off if a majority went on the tour than if none did
- Not even one student should be treated unfairly

4_5_Moral_DV3 Many people think that abortions should be allowed; others think they should largely be prohibited.

- Thousands of children are born into homes where they are unwanted and where they add to existing financial and emotional problems
- Aborting a fetus is equivalent to (or very close to) the taking of a human life

4_5_Moral_DV4 You are a sales representative for an electronics manufacturing firm. You have scheduled dinner with an important client for tomorrow and would very much like to impress him. A good friend of yours is a member of an exclusive country club near town. You could really impress your client if you took him to dinner at the club. You consider asking your friend to loan you his membership card.

- The product you are selling is good, and everyone would win if the deal goes through
- People should never ask their friends to be dishonest

4_5_Moral_DV5 One of your employees has accidentally come across a copy of your chief competitor's product price changes for next month. The booklet is on your desk in a manila envelope.

- The price guide will give you a temporary advantage over your competitor
- Using the information would be basically unfair and dishonest

4_5_Moral_DV6 You are middle aged and have been out of work for nearly two months. You need a job to support your family, and you have just been notified that you have a promising interview in three days with a company for which you would very much like to work. Unfortunately, you are well aware that youth is favoured in today's job market and you are afraid that your age might work against you. So, you are thinking of dying your hair to get rid of some of the grey and temporarily reporting your age as several years younger than your true age. After all, you are vigorous, healthy, and highly competent, and you have often been told you look young for your age.

- You need the job to support your family, and you would be good for the company
- One should always be honest

4_5_Moral_DV7 You work for a state auditor's office which has a policy against accepting gifts from anyone with whom the state may have business. Your birthday is in one week, and a very good friend of your father's has just dropped by with a pair of fine leather gloves and a birthday card. This person also works for a construction firm which has built city facilities in the past.

- Both the person and your father might be upset if you do not accept the gift
- Employees have an obligation to follow state policy

4_5_Moral_DV8 Some people believe in capital punishment; others do not.

- ❍ There is always the possibility that a mistake was made in convicting him/her
- ❍ Justice requires the death of the murderer; anything less is unfair to the victim and the victim's family

Happiness

4_5_Happy_Intro For each of the following statements and/or questions, please circle the point on the scale that you feel is most appropriate in describing you.

4_5_Happy_DV1 In general, I consider myself:

- ❍ 1 Not a very happy person
- ❍ 2
- ❍ 3
- ❍ 4
- ❍ 5
- ❍ 6
- ❍ 7 A very happy person

4_5_Happy_DV2 Compared to most of my peers, I consider myself:

- ❍ 1 Less happy
- ❍ 2
- ❍ 3
- ❍ 4
- ❍ 5
- ❍ 6
- ❍ 7 More happy

4_5_Happy_DV3 Some people are generally very happy. They enjoy life regardless of what is going on, getting the most out of everything. To what extent does this characterization describe you?

- ❍ 1 Not at all
- ❍ 2
- ❍ 3

❍ 4

❍ 5

❍ 6

❍ 7 A great deal

4_5_Happy_DV4 Some people are generally not very happy. Although they are not depressed, they never seem as happy as they might be. To what extend does this characterization describe you?

❍ 1 Not at all

❍ 2

❍ 3

❍ 4

❍ 5

❍ 6

❍ 7 A great deal

Team 4 Analysis Plan

The measure of moral orientation will be the mean of responses to *4_5_Moral_DV1*, *4_5_Moral_DV2, 4_5_Moral_DV3, 4_5_Moral_DV4, 4_5_Moral_DV5, 4_5_Moral_DV6, 4_5_Moral_DV7*, and *4_5_Moral_DV8* (the first [utilitarian] option will be scored as 1, while the second [deontological] option will be scored as 0 for each item). If this composite measure shows poor internal reliability ($\alpha < .70$), we will use responses to *4_5_Moral_DV1*. The measure of happiness will be the mean of responses to *4_5_Happy_DV1, 4_5_Happy_DV2, 4_5_Happy_DV3,* and *4_5_Happy_DV4* (reverse-scored). If this composite measure shows poor internal reliability ($\alpha < .70$), we will use responses to *4_5_Happy_DV1*. The effect size will be the Pearson correlation (*r*) between these two measures.

Team 5 Materials

Moral Dilemmas

5_5_Moral_DV You find a time machine and travel back to the year 1920. While checking into a hotel, you meet a young Austrian artist and veteran of the First World War. You realize this is Adolf Hitler before his rise to power in Nazi Germany. He is staying in the hotel room next to yours and the doors are not locked. It would be easy to simply smother him with a pillow in his sleep and disappear, stopping the Second World War and the Nazi party before they even start. However, he has not committed any crimes yet and it seems wrong to hurt an innocent person. How appropriate is it for you to kill an innocent young Hitler in order to prevent the Second World War?

❍ 1 Not at all appropriate

❍ 2

❍ 3

❍ 4

❍ 5

❍ 6

❍ 7 Very appropriate

Happiness
5_5_Happy_Intro Below are five statements with which you may agree or disagree. Using the 1 – 7 scale below, indicate your agreement with each item. Please be open and honest in your response.

5_5_Happy_DV1 In most ways my life is close to my ideal.
❍ 1 Strongly disagree

❍ 2

❍ 3

❍ 4

❍ 5

❍ 6

❍ 7 Strongly agree

5_5_Happy_DV2 The conditions of my life are excellent.
❍ 1 Strongly disagree

❍ 2

❍ 3

❍ 4

❍ 5

❍ 6

❍ 7 Strongly agree

5_5_Happy_DV3 I am satisfied with my life.
❍ 1 Strongly disagree

❍ 2

❍ 3

❍ 4

❍ 5

❍ 6

❍ 7 Strongly agree

5_5_Happy_DV4 So far I have gotten the important things I want in life.
❍ 1 Strongly disagree

❍ 2

❍ 3

❍ 4

❍ 5

❍ 6

❍ 7 Strongly agree

5_5_Happy_DV5 If I could live my life over, I would change almost nothing.
❍ 1 Strongly disagree

❍ 2

❍ 3

❍ 4

❍ 5

❍ 6

❍ 7 Strongly agree

Team 5 Analysis Plan
The measure of moral orientation will be responses to *5_5_Moral_DV*. The measure of happiness will be the mean of responses to *5_5_Happy_DV1, 5_5_Happy_DV2, 5_5_Happy_DV3, 5_5_Happy_DV4,* and *5_5_Happy_DV5*. If this composite measure shows poor internal reliability ($\alpha < .70$), we will use responses to *5_5_Happy_DV1*. The effect size will be the Pearson correlation (*r*) between these two measures.

Team 6 Materials
Moral Dilemmas
6_5_Moral_Intro In this study, we are interested in people's response to different scenarios. On the next page you will be provided with a particular scenario. Please read the information carefully and then answer questions.

6_5_Moral_DV1 A teacher maintains the following rule: If a child in class behaves well, the teacher rewards him or her. Alex has done his homework and therefore, following the rules, he needs to be rewarded. The teacher, however, suspects that Alex has cheated in doing his homework and therefore does not deserve a reward. Yet, the teacher is not sure and worries that breaking the rule might lead to disorder.
Imagine that you are the teacher, what will you do?

❍ 1 definitely reward Alex

❍ 2

❍ 3

❍ 4

❍ 5

❍ 6

❍ 7 definitely not reward Alex

6_5_Happy_DV For each of the following statements and/or questions, please circle the point on the scale that you feel is most appropriate in describing you.

	1 strongly disagree	2	3	4	5	6	7 strongly agree
In most ways my life is close to my ideal.	❍	❍	❍	❍	❍	❍	❍
The conditions of my life are excellent.	❍	❍	❍	❍	❍	❍	❍
I am satisfied with my life.	❍	❍	❍	❍	❍	❍	❍
So far I have gotten the important things I	❍	❍	❍	❍	❍	❍	❍

want in life. If I could live my life over, I would change almost nothing.	❍	❍	❍	❍	❍	❍	❍

Team 6 Analysis Plan

The measure of moral orientation will be responses to *6_5_Moral_DV* (reverse-scored, such that higher scores indicate more utilitarian judgments). The measure of happiness will be the mean of responses to the five items in *6_5_Happy_DV*. If this composite measure shows poor internal reliability ($\alpha < .70$), we will use responses to item 3 ("I am satisfied with my life"). The effect size will be the Pearson correlation (r) between these two measures.

Team 7 Materials

Moral Dilemmas

7_5_Moral_Intro Please read the following four scenarios, and answer the single-item question that follows each scenario:

7_5_Moral_DV1 You are at the wheel of a runaway trolley quickly approaching a fork in the tracks. On the tracks extending to the left is a group of five railway workmen. On the tracks extending to the right is a single railway workman.

If you do nothing the trolley will proceed to the left, causing the deaths of the five workmen. The only way to avoid the deaths of these workmen is to hit a switch on your dashboard that will cause the trolley to proceed to the right, causing the death of the single workman.

Is it appropriate for you to hit the switch in order to avoid the deaths of the five workmen?

❍ 1 not appropriate at all

❍ 2

❍ 3

❍ 4

❍ 5

❍ 6 absolutely appropriate

7_5_Moral_DV2 A runaway trolley is heading down the tracks toward five workmen who will be killed if the trolley proceeds on its present course. You are on a footbridge over the tracks, in between the approaching trolley and the five workmen. Next to you on this footbridge is a stranger who happens to be very large.

The only way to save the lives of the five workmen is to push this stranger off the bridge and onto the tracks below where his large body will stop the trolley. The stranger will die if you do this, but the five workmen will be saved.
Is it appropriate for you to push the stranger onto the tracks in order to save the five workmen?
❍ 1 not appropriate at all

❍ 2

❍ 3

❍ 4

❍ 5

❍ 6 absolutely appropriate

7_5_Moral_DV3 You are driving along a country road when you hear a plea for help coming from some roadside bushes. You pull over and encounter a man whose legs are covered with blood. The man explains that he has had an accident while hiking and asks you to take him to a nearby hospital.
Your initial inclination is to help this man, who will probably lose his leg if he does not get to the hospital soon. However, if you give this man a lift, his blood will ruin the leather upholstery of your car.
Is it appropriate for you to leave this man by the side of the road in order to preserve your leather upholstery?
❍ 1 not appropriate at all

❍ 2

❍ 3

❍ 4

❍ 5

❍ 6 absolutely appropriate

7_5_Moral_DV4 While on vacation on a remote island, you are fishing from a seaside dock. You observe a group of tourists board a small boat and set sail for a nearby island. Soon after their departure, you hear over the radio that there is a violent storm brewing, a storm that is sure to intercept them.
The only way that you can ensure their safety is to warn them by borrowing a nearby speedboat. The speedboat belongs to a miserly tycoon who would not take kindly to your borrowing his property.
Is it appropriate for you to borrow the speedboat in order to warn the tourists about the storm?
❍ 1 not appropriate at all

❍ 2

❍ 3

❍ 4

❍ 5

❍ 6 absolutely appropriate

7_5_Happy_DV Please indicate your agreement with the following statements:

	1 strongly disagree	2 disagree	3 slightly disagree	4 neither agree nor disagree	5 slightly agree	6 agree	7 strongly agree
In most ways my life is close to my ideal.	❍	❍	❍	❍	❍	❍	❍
The conditions of my life are excellent.	❍	❍	❍	❍	❍	❍	❍
I am satisfied with my life.	❍	❍	❍	❍	❍	❍	❍
So far I have gotten the important things I want in life.	❍	❍	❍	❍	❍	❍	❍
If I could live my life over, I would change almost nothing.	❍	❍	❍	❍	❍	❍	❍

Team 7 Analysis Plan

The measure of moral orientation will be the mean of responses to *7_5_Moral_DV1, 7_5_Moral_DV2, 7_5_Moral_DV3,* and *7_5_Moral_DV4.* If this composite measure shows poor internal reliability ($\alpha < .70$), we will use responses to *7_5_Moral_DV1.* The measure of happiness will be the mean of responses to the five items in *7_5_Happy_DV.* If this composite measure shows poor internal reliability ($\alpha < .70$), we will use responses to item 3, ("I am satisfied with my life"). The effect size will be the Pearson correlation (*r*) between these two measures.

Team 8 Materials

Moral Dilemmas

8_5_Moral_DV Please answer the following items using the 1 - 7 scale below. While we understand each often depends on the specific situation, we ask you to indicate your overall reaction here.

	1 - Strongly Disagree	2 - Disagree	3 - Slightly Disagree	4 - Neither Agree Nor Disagree	5 - Slightly Agree	6 - Agree	7 - Strongly Agree
It is sometimes OK to lie if no one gets hurt as a result.	❍	❍	❍	❍	❍	❍	❍
There is something wrong about the fact that the money I sometimes spend on a single fancy dinner would keep a whole family alive for months in other parts of the world.	❍	❍	❍	❍	❍	❍	❍
It is OK to build a fuel-burning power plant	❍	❍	❍	❍	❍	❍	❍

that will create respiratory problem for a small group of citizen who live nearby if it provides cheap energy that will improve the lives of a many more people.							
It is wrong to make animals suffer in medical research even if it leads to findings that could save human lives.	❍	❍	❍	❍	❍	❍	❍
The money spent in the US on keeping severely disabled people alive would be better spent helping many more healthy children survive in poor countries.	❍	❍	❍	❍	❍	❍	❍

It is never OK for a US company to pay bribes and collaborate with corrupt governments, even if it means that the citizens of that country will therefore be deprived of the company's jobs and products that would have substantially improved their quality of life.	❍	❍	❍	❍	❍	❍	❍
Extreme interrogation measures are never justified even when the lives of thousands of citizens might be at stake.	❍	❍	❍	❍	❍	❍	❍
If someone is caught for a crime they committed years ago, they should be sent to jail even if they have since	❍	❍	❍	❍	❍	❍	❍

become a productive member of society. In ethical decisions, it's less important to respect abstract principles than to make sure everyone is better off. It is important to act ethically even if no one seems better off as a result.	❍ ❍	❍ ❍	❍ ❍	❍ ❍	❍ ❍	❍ ❍	❍ ❍

Happiness

8_5_Happy_DV In this study you will be asked a series of questions about your life and your beliefs. Below are five statements that you may agree or disagree with. Using the 1 - 7 scale below, indicate your agreement with each item. Please be open and honest in your responding.

	1 - Strongly Disagree	2 - Disagree	3 - Slightly Disagree	4 - Neither Agree Nor Disagree	5 - Slightly Agree	6 - Agree	7 - Strongly Agree
In most ways my life is close to my ideal. The conditions of my life	❍ ❍	❍ ❍	❍ ❍	❍ ❍	❍ ❍	❍ ❍	❍ ❍

are excellent.							
I am satisfied with my life.	❍	❍	❍	❍	❍	❍	❍
So far I have gotten the important things I want in life.	❍	❍	❍	❍	❍	❍	❍
If I could live my life over, I would change almost nothing.	❍	❍	❍	❍	❍	❍	❍

Team 8 Analysis Plan

The measure of moral orientation will be the mean of responses to the ten items in *8_5_Moral_DV* (items 4, 6, 7, 8, and 10 will be reverse-scored, such that higher scores indicate more utilitarian judgments). If this composite measure shows poor internal reliability ($\alpha < .70$), we will use responses to item 10 ("It is important to act ethically even if no one seems better off as a result" [reverse-scored]). The measure of happiness will be the mean of responses to the five items in *7_5_Happy_DV*. If this composite measure shows poor internal reliability ($\alpha < .70$), we will use responses to item 3, ("I am satisfied with my life"). The effect size will be the Pearson correlation (*r*) between these two measures.

Team 9 Materials

Footbridge Dilemma

9_5_Moral_DV A runaway trolley is heading down the tracks toward five workmen who will be killed if the trolley proceeds on its present course. You are on a footbridge over the tracks, in between the approaching trolley and the five workmen. Next to you on this footbridge is a stranger who happens to be very large.

The only way to save the lives of the five workmen is to push this stranger off the bridge and onto the tracks below where his large body will stop the trolley. The stranger will die if you do this, but the five workmen will be saved. Is it appropriate for you to push the stranger on to the tracks in order to save the five workmen?

❍ Yes

❍ No

Happiness

9_5_Happy_Intro For each of the following statements and/or questions, please select the point on the scale that you feel is most appropriate in describing you.

9_5_Happy_DV1 In general, I consider myself:
❍ 1 Not a very happy person

❍ 2

❍ 3

❍ 4

❍ 5

❍ 6

❍ 7 A very happy person

9_5_Happy_DV2 Compared with most of my peers, I consider myself:
❍ 1 Less happy

❍ 2

❍ 3

❍ 4

❍ 5

❍ 6

❍ 7 More happy

9_5_Happy_DV3 Some people are generally very happy. They enjoy life regardless of what is going on, getting the most out of everything. To what extent does this characterization describe you?
❍ 1 Not at all

❍ 2

❍ 3

❍ 4

❍ 5

❍ 6

❍ 7 A great deal

9_5_Happy_DV4 Some people are generally not very happy. Although they are not depressed, they never seem as happy as they might be. To what extent does this characterization describe you?

❍ 1 Not at all

❍ 2

❍ 3

❍ 4

❍ 5

❍ 6

❍ 7 A great deal

Team 9 Analysis Plan

The measure of moral orientation will be responses to *9_5_Moral_DV* ("Yes" will be coded as 1, and "No" will be coded as 0). The measure of happiness will be the mean of responses to *9_5_Happy_DV1, 9_5_Happy_DV2, 9_5_Happy_DV3,* and *9_5_Happy_DV4.* If this composite measure shows poor internal reliability ($\alpha < .70$), we will use responses to *9_5_Happy_DV2*. The effect size will be the Pearson correlation (*r*) between these two measures.

Team 10 Materials

Moral Dilemmas

10_5_Moral_DV1 Suppose you are a doctor in a health clinic overrun by patients with a serious disease. You just received a shipment of drugs that can cure the disease but the drugs have their own severe side effects.

If you administer the drugs to your patients, a small number will die from the side effects but most will live. If you do not, most will die from the disease. Is it appropriate for you to administer the drug to your patients?

❍ 1 Definitely inappropriate

❍ 2

❍ 3

❍ 4

❍ 5

❍ 6

❍ 7 Definitely appropriate

10_5_Moral_DV2 Suppose you are a soldier guarding a border checkpoint between your nation and one troubled by insurgent violence. You notice a young man in a cheap car approaching the

checkpoint with a determined look on his face. You suspect he means to bomb the checkpoint, killing all the soldiers inside. He is rapidly approaching your station. Is it appropriate for you to shoot and kill the approaching man?

❍ 1 Definitely inappropriate

❍ 2

❍ 3

❍ 4

❍ 5

❍ 6

❍ 7 Definitely appropriate

10_5_Moral_DV3 It is war time. Enemy soldiers have taken over your village. They have orders to kill all remaining civilians. You and some of your townspeople have sought refuge in the cellar of a large house. Outside you hear the voices of soldiers who have come to search the house for valuables. A baby with no parents begins to cry loudly. You cover her mouth to block the sound. If you remove your hand from the baby's mouth her crying will summon the attention of the soldiers who will kill you and the others hiding out in the cellar. To save yourself and the others you must smother the child to death. Should you smother the child in order to save yourself and the other townspeople from being killed?

❍ 1 Definitely inappropriate

❍ 2

❍ 3

❍ 4

❍ 5

❍ 6

❍ 7 Definitely appropriate

Happiness

10_5_Happy_Intro Please answer the following questions.

10_5_Happy_DV1 In general, I consider myself:

❍ 1 Not a very happy person

❍ 2

❍ 3

❍ 4

❍ 5

❍ 6

❍ 7 A very happy person

10_5_Happy_DV2 Please rate how happy you are right now.
❍ 1 Not very happy

❍ 2

❍ 3

❍ 4

❍ 5

❍ 6

❍ 7 Very happy

Team 10 Analysis Plan
The measure of moral orientation will be the mean of responses to *10_5_Moral_DV1*, *10_5_Moral_DV2*, and *10_5_Moral_DV3*. If this composite measure shows poor internal reliability ($\alpha < .70$), we will use responses to *10_5_Moral_DV1*. The measure of happiness will be the mean of responses to *10_5_Happy_DV1* and *10_5_Happy_DV2*. If this composite measure shows poor internal reliability ($\alpha < .70$), we will use responses to *10_5_Happy_DV1*. The effect size will be the Pearson correlation (*r*) between these two measures.

Team 11 Materials
Morality
11_5_Moral_DV1 You are the coach of a children's soccer team. The morning of the big game, you realize that you forgot to reserve a field, which is your responsibility. You drive out to the local sports complex and see that none of the fields are being used. The security guard says that you cannot use a field unless you have reserved one ahead of time. You could tell the guard the truth, in which case the children will not get to play their game and will be very disappointed, or you could lie and tell the guard that you reserved the field a week ago, in which case the children will get to play their game. In this situation, should you tell the truth, or lie about having reserved the field?
❍ Definitely should tell the truth

❍

❍

❍

❍ I'm completely divided about what to do

❍

❍

❍

❍ Definitely should lie

11_5_Moral_DV2 Your professor at college has paired you with another classmate to work with on an assignment. The pairs were formed randomly, and although the professor did not record who is working together, she states emphatically that she does not want anyone swapping partners. However, after class a classmate asks if you would be willing to switch partners. The classmate was assigned to work with a good friend of yours, and your assigned partner is a good friend of the classmate. If you disobey your professor and switch partners, everyone will have a much more enjoyable time working on the project. Plus, everyone will be working with someone with whom they collaborate well, so everyone's grades would probably end up being higher as well. The professor would never find out that you switched partners, but you would be directly defying the professor's orders. In this situation, should you obey the professor and refuse to switch partners, or disobey the professor and switch partners?

❍ Definitely should refuse to switch partners

❍

❍

❍

❍ I'm completely divided about what to do

❍

❍

❍

❍ Definitely should switch partners

Happiness

11_5_Happy_DV Please think about what you have been doing and experiencing during the past four weeks. Then report how much you experienced each of the following feelings, using the scale below.

	Very Rarely or Never	Rarely	Sometimes	Often	Very Often or Always
Positive	❍	❍	❍	❍	❍
Good	❍	❍	❍	❍	❍
Pleasant	❍	❍	❍	❍	❍
Happy	❍	❍	❍	❍	❍

Joyful	❍	❍	❍	❍	❍
Contented	❍	❍	❍	❍	❍

Team 11 Analysis Plan

The measure of moral orientation will be the mean of responses to *11_5_Moral_DV1* and *11_5_Moral_DV2*. If this composite measure shows poor internal reliability ($\alpha < .70$), we will use responses to *11_5_Moral_DV1*. The measure of happiness will be the mean of responses to the six items in *11_5_Happy_DV*. If this composite measure shows poor internal reliability ($\alpha < .70$), we will use responses to item 4 ("Happy"). The effect size will be the Pearson correlation (*r*) between these two measures.

Team 12 did not develop materials for this research question.

Team 13 Materials

Moral Dilemmas

13_5_Moral_DV1 A runaway trolley is heading down the tracks toward five workmen who will be killed if the trolley keeps going. A bystander was on a footbridge over the tracks in between the approaching trolley and the five workmen. Next to the bystander was a very large stranger. The only way to save the lives of the five workmen was to push the stranger off the bridge and onto the tracks below where his large body stopped the trolley. The stranger died but the five workmen were saved.

How morally wrong was it for the bystander to push the man off the footbridge?

❍ Not at all morally wrong

❍ very slightly morally wrong

❍ slightly morally wrong

❍ somewhat morally wrong

❍ mostly morally wrong

❍ almost completely morally wrong

❍ Completely morally wrong

13_5_Moral_DV2 There is an accident and deadly fumes in the ventilation system are traveling to hospital rooms. In one room there are three patients. In another room there is a single man. Without intervention, the fumes would have gone into the room with the three patients and killed them. The hospital's night watchman avoided this by hitting a switch to reroute the fumes into room with the single man. The single man was killed but the other three patients were saved.

How morally wrong was it for the night watchman to reroute the fumes?

❍ Not at all morally wrong

❍ very slightly morally wrong

- ❍ slightly morally wrong
- ❍ somewhat morally wrong
- ❍ mostly morally wrong
- ❍ almost completely morally wrong
- ❍ Completely morally wrong

13_5_Moral_DV3 Enemy soldiers have taken over a village. They have orders to kill everyone. A large group hides in the basement of a house to avoid being slaughtered. The group included a mother with a fussy newborn baby. The mother had to cover the baby's mouth to stop the baby from crying out and thus alerting the soldiers to her group. Eventually the mother had to suffocate the baby to prevent the soldiers from discovering and killing the large group.

How morally wrong was it for the mother to suffocate the baby?

- ❍ Not at all morally wrong
- ❍ very slightly morally wrong
- ❍ slightly morally wrong
- ❍ somewhat morally wrong
- ❍ mostly morally wrong
- ❍ almost completely morally wrong
- ❍ Completely morally wrong

13_5_Moral_DV4 A doctor has five patients, each of whom is about to die due to a failing organ of some kind. The doctor had another healthy young patient. The only way to save the lives of the first five patients was to transplant the organs from the healthy patient (against his will) into their bodies. The families and staff begged the doctor to conduct the operations but the doctor refused.

How morally wrong was it for the doctor to refuse to conduct the operations?

- ❍ Not at all morally wrong
- ❍ very slightly morally wrong
- ❍ slightly morally wrong
- ❍ somewhat morally wrong
- ❍ mostly morally wrong
- ❍ almost completely morally wrong
- ❍ Completely morally wrong

13_5_Moral_DV5 An ecology professor leads a research team to a remote jungle. The team is large and includes spouses and children of the researchers. A terrorist group takes the group hostage. Over the course of their detainment, the second-in-command of the terrorist group develops a relationship with the professor. He offers him a deal for their escape. If the professor will torture and kill one of the other members picked at random as show of respect, he will let the rest of the group go. The professor refuses the deal and the entire group is massacred.

How morally wrong was it for the professor to refuse this deal?

- ❍ Not at all morally wrong
- ❍ very slightly morally wrong
- ❍ slightly morally wrong
- ❍ somewhat morally wrong
- ❍ mostly morally wrong
- ❍ almost completely morally wrong
- ❍ Completely morally wrong

Happiness

13_5_Happy_DV1 Below are five statements that you may agree or disagree with. Using the 1 - 7 scale below, indicate your agreement with each item. Please be open and honest in your responding.

	1 - Strongly Disagree	2 - Disagree	3 - Slightly Disagree	4 - Neither Agree Nor Disagree	5 - Slightly Agree	6 - Agree	7 - Strongly Agree
In most ways my life is close to my ideal.	❍	❍	❍	❍	❍	❍	❍
The conditions of my life are excellent.	❍	❍	❍	❍	❍	❍	❍
I am satisfied	❍	❍	❍	❍	❍	❍	❍

with my life. So far I have gotten the important things I want in life. If I could live my life over, I would change almost nothing.	❍ ❍	❍ ❍	❍ ❍	❍ ❍	❍ ❍	❍ ❍	❍ ❍

13_5_Happy_DV2 All things considered, would you say you are…

❍ 1 Very happy

❍ 2

❍ 3

❍ 4

❍ 5 Not happy at all

Team 13 Analysis Plan

The measure of moral orientation will be the mean of responses to *13_5_Moral_DV1* (reverse scored), *13_5_Moral_DV2* (reverse scored), *13_5_Moral_DV3* (reverse-scored), *13_4_Moral_DV4,* and *10_5_Moral_DV5.* If this composite measure shows poor internal reliability ($\alpha < .70$), we will use responses to *13_5_Moral_DV3* (reverse scored). The measure of happiness will be the mean of responses to the five items in *13_5_Happy_DV1* (*13_5_Happy_DV2* will be ignored for this project, but full data will be publicly posted). If this composite measure shows poor internal reliability ($\alpha < .70$), we will use responses to item 3 ("I am satisfied with my life"). The effect size will be the Pearson correlation (*r*) between these two measures.

Team 7 Materials (Long Version/Original Materials)

Moral Dilemmas

7_5L_Moral_Intro Please read the following scenarios, and answer the single-item question that follows each scenario:

7_5L_Moral_DV1 You are at the wheel of a runaway trolley quickly approaching a fork in the tracks. On the tracks extending to the left is a group of five railway workmen. On the tracks extending to the right is a single railway workman. If you do nothing the trolley will proceed to the left, causing the deaths of the five workmen. The only way to avoid the deaths of these workmen is to hit a switch on your dashboard that will cause the trolley to proceed to the right, causing the death of the single workman. Is it appropriate for you to hit the switch in order to avoid the deaths of the five workmen?

❍ Yes

❍ No

7_5L_Moral_DV2 While on vacation on a remote island, you are fishing from a seaside dock. You observe a group of tourists board a small boat and set sail for a nearby island. Soon after their departure you hear over the radio that there is a violent storm brewing, a storm that is sure to intercept them. The only way that you can ensure their safety is to warn them by borrowing a nearby speedboat. The speedboat belongs to a miserly tycoon who has hired a fiercely loyal guard to make sure that no one uses his boat without permission. To get to the speedboat you will have to lie to the guard. Is it appropriate for you to lie to the guard in order to borrow the speedboat and warn the tourists about the storm?

❍ Yes

❍ No

7_5L_Moral_DV3 You are a member of a government legislature. The legislature is deciding between two different policies concerning environmental hazards. Policy A has a 90% chance of causing no deaths at all and has a 10% chance of causing 1000 deaths. Policy B has an 88% chance of causing no deaths and a 12% chance of causing 10 deaths. Is it appropriate for you to vote for Policy B over Policy A?

❍ Yes

❍ No

7_5L_LMoral_DV4 You are a member of a government legislature. The legislature is deciding between two different policies concerning environmental hazards. Policy A has a 90% chance of causing no deaths at all and has a 10% chance of causing 1000 deaths. Policy B has an 88% chance of causing no deaths and a 12% chance of causing 10 deaths. Is it appropriate for you to vote for Policy A over Policy B?

❍ Yes

❍ No

7_5L_Moral_DV5 You are a member of a government legislature. The legislature is deciding between two different policies concerning environmental hazards. Policy A has a 90% chance of causing no deaths at all and has a 10% chance of causing 1000 deaths. Policy B has a 92% chance of causing no deaths and an 8% chance of causing 10,000 deaths. Is it appropriate for you to vote for Policy A over Policy B?
❍ Yes

❍ No

7_5L_Moral_DV6 You are the late-night watchman in a hospital. Due to an accident in the building next door, there are deadly fumes rising up through the hospital's ventilation system. In a certain room of the hospital are three patients. In another room there is a single patient. If you do nothing the fumes will rise up into the room containing the three patients and cause their deaths. The only way to avoid the deaths of these patients is to hit a certain switch, which will cause the fumes to bypass the room containing the three patients. As a result of doing this the fumes will enter the room containing the single patient, causing his death. Is it appropriate for you to hit the switch in order to avoid the deaths of the three patients?
❍ Yes

❍ No

7_5L_Moral_DV7 You are a member of a government legislature. The legislature is deciding between two different policies concerning environmental hazards. Policy A has a 90% chance of causing no deaths at all and has a 10% chance of causing 1000 deaths. Policy B has a 92% chance of causing no deaths and an 8% chance of causing 10,000 deaths. Is it appropriate for you to vote for Policy B over Policy A?
❍ Yes

❍ No

7_5L_Moral_DV8 You work for the Bureau of Health, a government agency. You are deciding whether or not your agency should encourage the use of a certain recently developed vaccine. The vast majority of people who take the vaccine develop an immunity to a certain deadly disease, but a very small number of people who take the vaccine will actually get the disease that the vaccine is designed to prevent. All the available evidence, which is very strong, suggests that the chances of getting the disease due to lack of vaccination are much higher than the chances of getting the disease by taking the vaccine. Is it appropriate for you to direct your agency to encourage the use of this vaccine in order to promote national health?
❍ Yes

❍ No

7_5L_Moral_DV9 You are on a cruise ship when there is a fire on board, and the ship has to be abandoned. The lifeboats are carrying many more people than they were designed to carry. The lifeboat you're in is sitting dangerously low in the water—a few inches lower and it will sink. The seas start to get rough, and the boat begins to fill with water. If nothing is done it will sink before the rescue boats arrive and everyone on board will die. However, there is an injured person who will not survive in any case. If you throw that person overboard the boat will stay afloat and the remaining passengers will be saved. Is it appropriate for you to throw this person overboard in order to save the lives of the remaining passengers?

❍ Yes

❍ No

7_5L_Moral_DV10 A viral epidemic has spread across the globe killing millions of people. You have developed two substances in your home laboratory. You know that one of them is a vaccine, but you don't know which one. You also know that the other one is deadly. Once you figure out which substance is the vaccine you can use it to save millions of lives. You have with you two people who are under your care, and the only way to identify the vaccine is to inject each of these people with one of the two substances. One person will live, the other will die, and you will be able to start saving lives with your vaccine. Is it appropriate for you to kill one of these people with a deadly injection in order to identify a vaccine that will save millions of lives?

❍ Yes

❍ No

7_5L_Moral_DV11 You are the leader of a small group of soldiers. You are on your way back from a completed mission deep in enemy territory when one of your men has stepped in trap that has been set by the enemy and is badly injured. The trap is connected to a radio device that by now has alerted the enemy to your presence. They will soon be on their way. If the enemy finds your injured man they will torture him and kill him. He begs you not to leave him behind, but if you try to take him with you your entire group will be captured. The only way to prevent this injured soldier from being tortured is to shoot him yourself. Is it appropriate for you to shoot this soldier in order to prevent him from being tortured by the enemy?

❍ Yes

❍ No

7_5L_Moral_DV12 You are the leader of a small army that consists of warriors from two tribes, the hill tribe and the river tribe. You belong to neither tribe. During the night a hill tribesman got into an argument with a river tribesman and murdered him. The river tribe will attack the hill tribe unless the murderer is put to death, but the hill tribe refuses to kill one of its own warriors. The only way for you to avoid a war between the two tribes that will costs hundreds of lives is to publicly execute the murderer by cutting off is head with your sword. Is it appropriate for you to cut off this man's head in order to prevent the two tribes from fighting a war that will cost hundreds of lives?

❍ Yes

❍ No

7_5L_Moral_DV13 You are part of a group of ecologists who live in a remote stretch of jungle. The entire group, which includes eight children, has been taken hostage by a group of paramilitary terrorists. One of the terrorists takes a liking to you. He informs you that his leader intends to kill you and the rest of the hostages the following morning. He is willing to help you and the children escape, but as an act of good faith he wants you to kill one of your fellow hostages whom he does not like. If you refuse his offer all the hostages including the children and yourself will die. If you accept his offer then the others will die in the morning but you and the eight children will escape. Is it appropriate for you to kill one of your fellow hostages in order to escape from the terrorists and save the lives of the eight children?

❍ Yes

❍ No

7_5L_Moral_DV14 You are the leader of a mountaineering expedition that is stranded in the wilderness. Your expedition includes a family of six that has a genetically caused vitamin deficiency. A few people's kidneys contain large amounts of this vitamin. There is one such person in your party. The only way to save the lives of the six members of this family is to remove one of this man's kidneys so that the necessary vitamins may be extracted from it. The man will not die if you do this, but his health will be compromised. The man is opposed to this plan, but you have the power to do as you see fit. Is it appropriate for you to forcibly remove this man's kidney in order to save the lives of the six vitamin-deficient people?

❍ Yes

❍ No

7_5L_Moral_DV15 It is wartime and you and your two children, ages eight and five, are living in a territory that has been occupied by the enemy. At the enemy's headquarters is a doctor who performs painful experiments on humans that inevitably lead to death. He intends to perform experiments on one of your children, but he will allow you to choose which of your children will be experimented upon. You have twenty-four hours to bring one of your children to his laboratory. If you refuse to bring one of your children to his laboratory he will find them both and experiment on both of them. Is it appropriate for you to bring one of your children to the laboratory in order to avoid having them both die?

❍ Yes

❍ No

7_5L_Moral_DV16 You, your husband, and your four children are crossing a mountain range on your return journey to your homeland. You have inadvertently set up camp on a local clan's sacred burial ground. The leader of the clan says that according to the local laws, you and your family must be put to death. However, he will let yourself, your husband, and your three other children live if you yourself will kill your oldest son. Is it appropriate for you to kill your oldest son in order to save your husband and your other three children?

❍ Yes

❍ No

7_5L_Moral_DV17 A runaway trolley is heading down the tracks toward five workmen who will be killed if the trolley proceeds on its present course. You are on a footbridge over the tracks, in between the approaching trolley and the five workmen. Next to you on this footbridge is a stranger who happens to be very large. The only way to save the lives of the five workmen is to push this stranger off the bridge and onto the tracks below where his large body will stop the trolley. The stranger will die if you do this, but the five workmen will be saved. Is it appropriate for you to push the stranger on to the tracks in order to save the five workmen?
❍ Yes

❍ No

7_5L_Moral_DV18 Enemy soldiers have taken over your village. They have orders to kill all remaining civilians. You and some of your townspeople have sought refuge in the cellar of a large house. Outside you hear the voices of soldiers who have come to search the house for valuables. Your baby begins to cry loudly. You cover his mouth to block the sound. If you remove your hand from his mouth his crying will summon the attention of the soldiers who will kill you, your child, and the others hiding out in the cellar. To save yourself and the others you must smother your child to death. Is it appropriate for you to smother your child in order to save yourself and the other townspeople?
❍ Yes

❍ No

7_5L_Moral_DV19 You are a doctor. You have five patients, each of whom is about to die due to a failing organ of some kind. You have another patient who is healthy. The only way that you can save the lives of the first five patients is to transplant five of this young man's organs (against his will) into the bodies of the other five patients. If you do this, the young man will die, but the other five patients will live. Is it appropriate for you to perform this transplant in order to save five of your patients?
❍ Yes

❍ No

7_5L_Moral_DV20 Your plane has crashed in the Himalayas. The only survivors are yourself, another man, and a young boy. The three of you travel for days, battling extreme cold and wind. Your only chance at survival is to find your way to small a village on the other side of the mountain, several days away. The boy has a broken leg and cannot move very quickly. His chances of surviving the journey are essentially zero. Without food, you and the other man will probably die as well. The other man suggests that you sacrifice the boy and eat his remains over the next few days. Is it appropriate to kill this boy so that you and the other man may survive your journey to safety?
❍ Yes

❍ No

7_5L_Moral_DV21 You are a young architect visiting one of your construction sites with your boss. Your boss is a despicable individual who makes everyone around him miserable including you. It occurs to you that if you were to push him off of the building you are inspecting he would fall to his death and everyone would think it was an accident. Is it appropriate for you to push your boss off of the building in order to get him out of your life?
❍ Yes

❍ No

7_5L_Moral_DV22 You are in hospital lounge waiting to visit a sick friend. A young man sitting next to you explains that his father is very ill. The doctors believe that he has a week to live at most. He explains further that his father has a substantial life insurance policy that expires at midnight. If his father dies before midnight, this young man will receive a very large sum of money. He says that the money would mean a great deal to him and that no good will come from his father's living a few more days. He offers you half a million dollars to go up to his father's room and smother his father with a pillow. Is it appropriate for you to kill this man's father in order to get money for yourself and this young man?
❍ Yes

❍ No

7_5L_Moral_DV23 You are a fifteen-year-old girl who has become pregnant. By wearing loose clothing and deliberately putting on weight you have managed to keep your pregnancy a secret. One day, while at school, your water breaks. You run to the girls' locker room and hide for several hours while you deliver the baby. You are sure that you are not prepared to care for this baby. You think to yourself that it would be such a relief to simply clean up the mess you've made in the locker room, wrap the baby in some towels, throw the baby in the dumpster behind the school, and act as if nothing had ever happened. Is it appropriate for you to throw your baby in the dumpster in order to move on with your life?
❍ Yes

❍ No

7_5L_Moral_DV24 You are driving along a country road when you hear a plea for help coming from some roadside bushes. You pull over and encounter a man whose legs are covered with blood. The man explains that he has had an accident while hiking and asks you to take him to a nearby hospital. Your initial inclination is to help this man, who will probably lose his leg if he does not get to the hospital soon. However, if you give this man a lift, his blood will ruin the leather upholstery of your car. Is it appropriate for you to leave this man by the side of the road in order to preserve your leather upholstery?
❍ Yes

❍ No

7_5L_Hap_DV1 Please indicate your agreement with the following statements:

	1 strongly disagree	2 disagree	3 slightly disagree	4 neither agree nor disagree	5 slightly agree	6 agree	7 strongly agree
In most ways my life is close to my ideal.	❍	❍	❍	❍	❍	❍	❍
The conditions of my life are excellent.	❍	❍	❍	❍	❍	❍	❍
I am satisfied with my life.	❍	❍	❍	❍	❍	❍	❍
So far I have gotten the important things I want in life.	❍	❍	❍	❍	❍	❍	❍
If I could live my life over, I would change almost nothing.	❍	❍	❍	❍	❍	❍	❍

7_5L_Hap_DV2 This scale consists of a number of words that describe different feelings and emotions. Read each item and then select the number from the scale below. Indicate the extent you generally feel this way, that is, how you feel on the average.

	1 very slightly or not at all	2 a little	3 moderately	4 quite a bit	5 extremely
Interested	❍	❍	❍	❍	❍
Distressed	❍	❍	❍	❍	❍
Excited	❍	❍	❍	❍	❍
Upset	❍	❍	❍	❍	❍
Strong	❍	❍	❍	❍	❍
Guilty	❍	❍	❍	❍	❍
Scared	❍	❍	❍	❍	❍
Hostile	❍	❍	❍	❍	❍
Enthusiastic	❍	❍	❍	❍	❍
Proud	❍	❍	❍	❍	❍
Irritable	❍	❍	❍	❍	❍
Alert	❍	❍	❍	❍	❍
Ashamed	❍	❍	❍	❍	❍
Inspired	❍	❍	❍	❍	❍
Nervous	❍	❍	❍	❍	❍
Determined	❍	❍	❍	❍	❍
Attentive	❍	❍	❍	❍	❍
Jittery	❍	❍	❍	❍	❍
Active	❍	❍	❍	❍	❍
Afraid	❍	❍	❍	❍	❍

7_5L_Hap_DV3 This questionnaire contains a series of statements that refer to how you may feel things have been going in your life. Read each statement and decide the extent to which you agree or disagree with it. Try to respond to each statement according to your own feelings about how things are actually going, rather than how you might wish them to be.

Please use the following scale when responding to each statement.

	0 Strongly Disagree	1	2	3	4 Strongly Agree
I find I get intensely involved in many of the things I do each day.	❍	❍	❍	❍	❍
I believe I have discovered who I really am.	❍	❍	❍	❍	❍
I think it would be ideal if things came easily to me in my life.	❍	❍	❍	❍	❍
My life is centered around a set of core beliefs that give meaning to my life.	❍	❍	❍	❍	❍
It is more important that I really enjoy what I do than that other people are impressed by it.	❍	❍	❍	❍	❍

I believe I know what my best potentials are and I try to develop them whenever possible.	❍	❍	❍	❍	❍
Other people usually know better what would be good for me to do than I know myself.	❍	❍	❍	❍	❍
I feel best when I'm doing something worth investing a great deal of effort in.	❍	❍	❍	❍	❍
I can say that I have found my purpose in life.	❍	❍	❍	❍	❍
If I did not find what I was doing rewarding for me, I do not think I could continue doing it.	❍	❍	❍	❍	❍
As yet, I've not figured out what to do with my life.	❍	❍	❍	❍	❍

I can't understand why some people want to work so hard on the things that they do.	❍	❍	❍	❍	❍
I believe it is important to know how what I'm doing fits with purposes worth pursuing.	❍	❍	❍	❍	❍
I usually know what I should do because some actions just feel right to me.	❍	❍	❍	❍	❍
When I engage in activities that involve my best potentials, I have this sense of really being alive.	❍	❍	❍	❍	❍
I am confused about what my talents really are.	❍	❍	❍	❍	❍
I find a lot of the things I do are personally	❍	❍	❍	❍	❍

expressive for me.					
It is important to me that I feel fulfilled by the activities that I engage in.	❍	❍	❍	❍	❍
If something is really difficult, it probably isn't worth doing.	❍	❍	❍	❍	❍
I find it hard to get really invested in the things that I do.	❍	❍	❍	❍	❍
I believe I know what I was meant to do in life.	❍	❍	❍	❍	❍

Team 7 Original Materials Analysis Plan

The measure of moral orientation will be the total number of "Yes" responses to *7_5L_Moral_DV1* through *7_5L_Moral_DV24*. The measure of happiness will be the mean of a "hedonic happiness" index and a "eudaimonic happiness" index. The hedonic happiness index will be the mean of two z-scores: first, the mean responses to the five items in *7_5L_Hap_DV1* will be z-scored. Second, we will take the mean of the positive affect items in *7_5L_Hap_DV2* (Interested, Excited, Strong, Enthusiastic, Proud, Alert, Inspired, Determined, Attentive, Active) and subtract from it the mean of the negative affect items (Distressed, Upset, Guilty, Scared, Hostile, Irritable, Ashamed, Nervous, Jittery, Afraid) (see Watson, Clark, & Tellegen, 1988), then z-score the resulting "positive affect score". These two z-scores will then be averaged to compute the hedonic happiness index. The eudaimonic happiness index will be the z-score of the mean of responses to the 21 items in *7_5L_Hap_DV3* (items 3, 7, 11, 12, 16, 19, and 20 will be reverse-scored, see Waterman et al., 2010). The composite happiness measure will be the mean of the hedonic happiness index and the eudaimonic happiness index. The effect size will be the Pearson correlation (*r*) between the measure of moral orientation and the composite happiness measure.

Demographics

Demo_Intro We would now like to collect some general information about you.

Sex What is your biological sex?
- ❍ Male
- ❍ Female

Age What is your age?

Height What is your height in feet and inches (e.g.,5 ft 3 in)?
Feet:
Inches:

Weight What is your weight in pounds (e.g., 170 lbs)?

BirthCntry What country were you born in?
- ❍ Afghanistan
- ❍ Albania
- ❍ Algeria
- ❍ Andorra
- ❍ Angola
- ❍ Antigua and Barbuda
- ❍ Argentina
- ❍ Armenia
- ❍ Australia
- ❍ Austria
- ❍ Azerbaijan
- ❍ Bahamas
- ❍ Bahrain
- ❍ Bangladesh
- ❍ Barbados
- ❍ Belarus
- ❍ Belgium
- ❍ Belize

- Benin
- Bhutan
- Bolivia
- Bosnia and Herzegovina
- Botswana
- Brazil
- Brunei Darussalam
- Bulgaria
- Burkina Faso
- Burundi
- Cambodia
- Cameroon
- Canada
- Cape Verde
- Central African Republic
- Chad
- Chile
- China
- Colombia
- Comoros
- Congo, Republic of the...
- Costa Rica
- Côte d'Ivoire
- Croatia
- Cuba
- Cyprus
- Czech Republic
- Democratic People's Republic of Korea
- Democratic Republic of the Congo
- Denmark

- Djibouti
- Dominica
- Dominican Republic
- Ecuador
- Egypt
- El Salvador
- Equatorial Guinea
- Eritrea
- Estonia
- Ethiopia
- Fiji
- Finland
- France
- Gabon
- Gambia
- Georgia
- Germany
- Ghana
- Greece
- Grenada
- Guatemala
- Guinea
- Guinea-Bissau
- Guyana
- Haiti
- Honduras
- Hong Kong (S.A.R.)
- Hungary
- Iceland
- India

- Indonesia
- Iran, Islamic Republic of...
- Iraq
- Ireland
- Israel
- Italy
- Jamaica
- Japan
- Jordan
- Kazakhstan
- Kenya
- Kiribati
- Kuwait
- Kyrgyzstan
- Lao People's Democratic Republic
- Latvia
- Lebanon
- Lesotho
- Liberia
- Libyan Arab Jamahiriya
- Liechtenstein
- Lithuania
- Luxembourg
- Madagascar
- Malawi
- Malaysia
- Maldives
- Mali
- Malta
- Marshall Islands

- Mauritania
- Mauritius
- Mexico
- Micronesia, Federated States of...
- Monaco
- Mongolia
- Montenegro
- Morocco
- Mozambique
- Myanmar
- Namibia
- Nauru
- Nepal
- Netherlands
- New Zealand
- Nicaragua
- Niger
- Nigeria
- North Korea
- Norway
- Oman
- Pakistan
- Palau
- Panama
- Papua New Guinea
- Paraguay
- Peru
- Philippines
- Poland
- Portugal

- Qatar
- Republic of Korea
- Republic of Moldova
- Romania
- Russian Federation
- Rwanda
- Saint Kitts and Nevis
- Saint Lucia
- Saint Vincent and the Grenadines
- Samoa
- San Marino
- Sao Tome and Principe
- Saudi Arabia
- Senegal
- Serbia
- Seychelles
- Sierra Leone
- Singapore
- Slovakia
- Slovenia
- Solomon Islands
- Somalia
- South Africa
- South Korea
- Spain
- Sri Lanka
- Sudan
- Suriname
- Swaziland
- Sweden

- ❍ Switzerland
- ❍ Syrian Arab Republic
- ❍ Tajikistan
- ❍ Thailand
- ❍ The former Yugoslav Republic of Macedonia
- ❍ Timor-Leste
- ❍ Togo
- ❍ Tonga
- ❍ Trinidad and Tobago
- ❍ Tunisia
- ❍ Turkey
- ❍ Turkmenistan
- ❍ Tuvalu
- ❍ Uganda
- ❍ Ukraine
- ❍ United Arab Emirates
- ❍ United Kingdom of Great Britain and Northern Ireland
- ❍ United Republic of Tanzania
- ❍ United States of America
- ❍ Uruguay
- ❍ Uzbekistan
- ❍ Vanuatu
- ❍ Venezuela, Bolivarian Republic of...
- ❍ Viet Nam
- ❍ Yemen
- ❍ Zambia
- ❍ Zimbabwe

YearsEngl How many years of experience with English do you have?

Ethnicity What is your ethnicity?

- ❑ White
- ❑ Black or African American
- ❑ Hispanic or Latino/a
- ❑ American Indian or Alaska Native
- ❑ Asian
- ❑ Native Hawaiian or Pacific Islander
- ❑ Other _____

Educ What is the highest level of school you have completed or the highest degree you have received?

- ❍ Less than high school degree
- ❍ High school graduate (high school diploma or equivalent including GED)
- ❍ Some college but no degree
- ❍ Associate degree in college (2-year)
- ❍ Bachelor's degree in college (4-year)
- ❍ Master's degree
- ❍ Professional degree (JD, MD)
- ❍ Doctoral degree

PolIdeol Please indicate your own political ideology:

- ❍ 1 very liberal
- ❍ 2
- ❍ 3
- ❍ 4 moderate
- ❍ 5
- ❍ 6
- ❍ 7 very conservative

PolParty What is your political party affiliation?

- ❍ Strongly support Democrats
- ❍ Moderately support Democrats
- ❍ Slightly support Democrats

- ❍ Slightly support Republicans
- ❍ Moderately support Republicans
- ❍ Strongly support Republicans
- ❍ Independent
- ❍ Libertarian
- ❍ None
- ❍ Other political party (please indicate) _____

Religion What is your religion?

- ❍ Christian (Including Church of England, Catholic, Protestant and all other Christian denominations)
- ❍ Buddhist
- ❍ Hindu
- ❍ Jewish
- ❍ Muslim
- ❍ Sikh
- ❍ No religion
- ❍ Others (please indicate) _____

Income Please indicate your entire household income last year before taxes.

- ❍ Less than $10,000
- ❍ $10,000 to $19,999
- ❍ $20,000 to $29,999
- ❍ $30,000 to $39,999
- ❍ $40,000 to $49,999
- ❍ $50,000 to $59,999
- ❍ $60,000 to $69,999
- ❍ $70,000 to $79,999
- ❍ $80,000 to $89,999
- ❍ $90,000 to $99,999
- ❍ $100,000 to $149,999

❍ $150,000 or more

Employ Which statement best describes your current employment status?
❍ Working (paid employee)

❍ Working (self-employed)

❍ Not working (temporary layoff from a job)

❍ Not working (looking for work)

❍ Not working (retired)

❍ Not working (disabled)

❍ Not working (student)

❍ Not working (other) _____

❍ Prefer not to answer

SexOrientation What is your sexual orientation?
❍ Heterosexual (straight)

❍ Homosexual (gay or lesbian)

❍ Bisexual

❍ Prefer not to answer

❍ Other (Specify, if desired) _____

SUPPLEMENT 2 - Deviations from pre-analysis plans

Below we outline instances in which the reported analyses departed from our initial pre-registration plans.

1. Discontinuation of plan to run a second forecasting survey

We originally planned to ask all individuals who completed the prediction study to participate in a second forecasting survey, in which we planned to provide feedback on the distribution of answers of everyone that took part in the first survey. Other than this feedback, the second forecasting survey was meant to be an exact replication of the first one. Our goal would have been to test whether feedback had a positive effect on the accuracy of predictions. However, given the notable workload of the first forecasting survey in terms of effort and time for respondents, we realized that a low retention rate among forecasters would pose a threat for the power of the secondary tests. Therefore, we decided not to run the second round of the survey, and as a consequence we could not test the set of the secondary hypotheses as written in the pre-analysis plan.

2. Cutoff used to determine a high-quality study

We originally pre-registered that we would use a 1-10 (10-point) scale and cutoff of 6 or above to identify studies rated as higher in quality by independent raters. However, after already collecting the independent ratings we realized we had actually employed a 0-10 (11-point) scale.

Therefore we used a cutoff of 5 or above when selecting higher quality studies. This revised cutoff was not pre-registered for our analysis of the Main Studies, but was pre-registered for the analysis of the Replication Studies.

3. Number of categories under "job rank"

The category "Tenure-track Assistant Professor" was accidentally omitted from the pre-analysis plan for the forecasting survey for the Main Studies, but included in the pre-analysis plan for the Replication Studies where we again examined forecasting accuracy.

4. Tau-squared statistics in the Main Studies and Replication Studies Statistics

We did not pre-register that we would compute and report the tau-squared statistic in our examination of heterogeneity in the Main Studies and Replication Studies. The usefulness of this statistic was pointed out to us by an anonymous reviewer, so we report it in the main text.

5. Meta-regression analysis of the Main Studies

The reported meta-regression predicting Main Studies' effect sizes from team and hypothesis was not pre-registered. We had pre-registered the intraclass correlation coefficient analyses, but the usefulness of examining the predictive effects of team and hypothesis, controlling for the other predictor, did not occur to us until we were already analyzing the data. This analysis was, however, pre-registered for the Replication Studies.

6. Forecasts and meta-analyzed effects

We pre-registered and carried out analyses relating scientists' forecasts to the estimated effect sizes for each of the 64 study designs for the Main Studies and Replication Studies. It later occurred to us to repeat the forecasting analyses after meta-analyzing across the Main Studies and Replication Studies for each design to have the most accurate effect size estimates. Note that although sample sizes are approximately doubled for the meta-analyzed effect sizes, the forecasts are most relevant to the Main Studies, since forecasters were provided details on this participant population (MTurk workers) and sample sizes.

7. Probit models for testing forecasters' sensitivity to design choices

We pre-registered and carried out linear models to test whether forecasters are sensitive to different design choices (equations 1a and 2a in the manuscript, regression Table S5.4 in Supplement 5). As additional robustness checks, we estimated the same regressions using a probit model (refer to Table S5.5 in Supplement 5). The coefficients and the standard errors estimated via probit model are in line with those estimated via ordinary least squares estimation.

8. Multivariate regression for the effect of monetary incentives on prediction accuracy

In addition to the models specified in the pre-analysis plan, for which the results estimated independently are reported in Table S5.6 of Supplement 5, we also report in Table S5.18 the

coefficients of (3) and (4) estimated jointly through the multivariate regression technique to take into account potential correlations between the forecasts regarding statistical significance and regarding effect sizes. As expected, the coefficients estimated jointly are consistent with those estimated independently, but they are more precisely estimated.

9. One-stage multivariate meta-analysis of Main Studies and Replication Studies

The one-stage multivariate meta-analytic model of the Main Studies' and Replication Studies' data presented in Supplement 8 was not pre-registered. It was conducted based on feedback from an anonymous reviewer.

SUPPLEMENT 3 - Materials for forecasting survey

Below are the instructions and dependent measures for the forecasting survey, including the information on monetary incentives for accuracy, which were presented only in the incentivized condition. Note that we provide example materials for only one study design (Hypothesis 1 materials from Team 1); each forecaster provided her predictions about the research results as well as expert assessments of quality for all 64 sets of study materials.

Can you tell if a study will produce a significant effect from just looking at the materials?

BACKGROUND

Can you predict whether a hypothesis will be supported from just looking at the study materials? In this project, we are "crowdsourcing" a hypothesis test by having different research teams independently create their own versions of the materials to test 5 different hypotheses. On average, each hypothesis is tested by 13 teams. We will then run a large online data collection on Mechanical Turk randomly assigning thousands of MTurk workers to participate as research subjects in one of the different versions of each of the 5 studies. There will be around 550 Mechanical Turk subjects per study version to provide adequate power. For more on Mechanical Turk samples, please see Paolacci et al., (2010). The goal of the project is to examine the extent to which differences in how independent research teams choose to operationalize the same hypothesis influence the final effect size estimates.

In this prediction survey, we are asking independent scientists (in this case you!) to look at each set of materials for each hypothesis and try to predict the results that will be obtained in the online data collection. You will be asked to make predictions about 1) the probability that each hypothesis was supported with a significant effect, and 2) the effect size in terms of Cohen's d or in terms of Pearson r.

Quoting Wikipedia on effect sizes: an effect size is a quantitative measure of the strength of a phenomenon. Examples of effect sizes are the correlation between two variables, the regression coefficient in a regression, the mean difference, or even the risk with which something happens, such as how many people survive after a heart attack for every one person that does not survive. For each type of effect size, a larger absolute value always indicates a stronger effect. In the social sciences, a Cohen's *d* of .20 is considered to be a small effect, .50 is considered to be a medium effect, and .80 is considered to be a large effect. Similarly, a Pearson correlation *r* of .10 is considered to be a small effect, .30 is considered to be a medium effect, and .50 is considered to be a large effect.

COMPENSATION

In compensation for your time completing this survey, you will be listed as a co-author on the final report on the Crowdsourcing a Hypothesis Test project that we will submit for publication. You will have one month to finish the survey.

YOUR TASK

For each set of study materials (which includes the description of the Dependent Variable – DV henceforth – and the planned analysis) and for each hypothesis, we will ask you five questions:

A) To what extent does this set of materials provide a scientifically informative and valid test of the research hypothesis (0= not at all informative, 10= extremely informative)?

B) Do you predict this hypothesis will find statistically significant support ($p < .05$) when tested using this set of materials? Here we ask you for the probability that you assign the binary outcome: whether the effect for this set of study materials will be in the same direction as the hypothesis, and will be statistically significant with a *p*-value smaller than 0.05.

C) How confident are you in your prediction in question B), on a scale from 0 to 10, where 0 is "not confident at all" and 10 is "very confident"?

D) What do you predict will be the effect size for this hypothesis when tested with this set of materials? Here we will ask about the effect size either in terms of Cohen's *d* or in terms of Pearson *r*, which we will specify for each hypothesis. Please put a negative sign (-) in front of the effect size or correlation if you think it will be the opposite direction from the original hypothesis.

E) How confident are you in your prediction in question D), on a scale from 0 to 10, where 0 is "not confident at all" and 10 is "very confident"?

Moreover, for each of the 5 hypotheses, we will also ask you the following question: How familiar are you with this area of research, where 0 is "not at all" and 10 is "extremely familiar"?

Please note that:

- the material used to test the hypotheses is always expressed in *Italic font*;
- the hypothesis tested is reported in blue on each page;
- you are required to input your answers in the "Your Predictions" section, always highlighted in red;
- your answers are saved in real time, hence if needed you can complete the survey in more than one session. To come back simply click on the survey link: the survey will automatically continue where you stopped at the end of your previous session;
- for each hypothesis tested, the "back button" on the bottom left allows you to go back and update the answers that you submitted previously;
- in the footer of each page you can find a link leading to the instructions.

Please click the button on the bottom right to continue.

Incentives scheme

This prediction survey includes a direct monetary incentive for making accurate predictions. Once the full results of the “crowdsourcing a hypothesis test” project are known, we will randomly select one specific version (out of 13) of one specific hypothesis (out of 5) to be the basis for the computation of payoffs for all the participants. There will be only one randomly selected version-hypothesis pair that will apply to all participants. Each participant will be rewarded with a monetary payoff based on the accuracy of her answers about statistically significant support or not and about the predicted effect size of the chosen version-hypothesis pair. In other words, only the predictions referring to one specific version (out of 13) of one specific hypothesis (out of 5) will be the basis for the computation of the monetary amount that will be transferred to each participant. This specific version-hypothesis pair will be randomly selected from the pool of all possible version-hypothesis pairs.

The timing of the randomization will work as follows:

- One specific version-hypothesis pair is randomly selected to be the basis for payoff calculations for all participants;
- For each participant, the payoff for statistically significant support and the payoff for effect size are calculated according to the proposed rules;
- Final payoff, determined as the sum of the payoff for statistically significant support and the payoff for effect size, will be transferred to the participants.

You can find detailed descriptions of the mechanisms determining payoff in the following sections **Payoffs for Predicting Statistically Significant Support (or Not) Accurately** and **Payoffs for Predicting Effect Size Accurately**. In sum, you are incentivized to provide us with your true beliefs: the expected earnings are the highest when you report what you really believe.

Payoffs for Predicting Statistically Significant Support (or Not) Accurately

Monetary amounts are calculated based on a quadratic score rule, starting from the answer that the participant gave to the question about whether significant support for the hypothesis would be obtained using that set of materials. The paid amount (in USD) is determined by the following formula:

$$\mathbf{A=10(1-(p^*-p_i)^2)}$$

where p* is a binary variable being 1 if the study obtained a significant result and 0 if the study obtained a nonsignificant result and p_i is the probability that individual i assigned to the event "the study version x to test hypothesis y produced a significant result".

To give you an example, let's assume that the version-hypothesis randomly selected to determine the payoffs is composed by study version x to test hypothesis y. The following table summarizes the amounts paid (per capita), as a function of your reported probability in two possible cases: x-y produces a significant result or x-y does not produce a significant result.

	Payoff in USD	
p_i	Study version "x" to test hypothesis "y" produced a significant result	Study version "x" to test hypothesis "y" did not produce a significant result
0	0	10
0.1	1.9	9.9
0.2	3.6	9.6
0.3	5.1	9.1
0.4	6.4	8.4
0.5	7.5	7.5
0.6	8.4	6.4
0.7	9.1	5.1
0.8	9.6	3.6
0.9	9.9	1.9
1	10	0

You should read the table above as follows: if the study produces a significant result and your predicted probability is p_i=0.6, then you earn 8.4 USD; if it does not produce a significant result, you earn only 6.4 USD. Note that the more accurate your estimate, the higher your payoff. Moreover, due to the way the formula is constructed, the expected earnings are highest when you report what you really believe.

Payoffs for Predicting Effect Size Accurately

Individual payoffs are determined applying the step-function represented in following tables (the first one refers to effects measured in terms of Cohen's d, the second one to effects measured in terms of Pearson r: tables are not identical due to the non-linearity of the relation between the two measures of effect size). In both tables, the left column shows the intervals around the effect size of the hypothesis tested; the right column shows the payoffs associated with each interval. If your

prediction falls between the effect size and the boundaries specified in the left column, then you are entitled to receive the payoff specified in the right column. Only the smallest interval containing both the observed effect size and your prediction contributes to the determination of your payoff.

Cohen's d		Pearson r	
Distance of prediction from effect size	Payoff in USD	Distance of prediction from effect size	Payoff in USD
± 0.05	50	± 0.02	50
± 0.10	25	± 0.04	25
± 0.15	15	± 0.06	15
± 0.20	10	± 0.08	10
± 0.25	5	± 0.10	5
± 0.50	2	± 0.20	2
± 1.00	1	± 0.40	1
> 1	0	> 0.40	0

You should read the tables in the following way: if the absolute distance between the effect size measured as Cohen's *d* and your prediction is lower than 0.05, then you earn 50 USD; if it is between 0.15 and 0.20, then you earn 10 USD; if it is higher than 1, then you are not entitled to any payment. Similarly, if the effect is measured in terms of Pearson *r*. Note that your payoff is independent from the predictions of other participants. As before, the more accurate your estimate, the higher your payoff.

Prediction survey: Implicit and automatic cognition

How familiar are you with research on implicit and automatic cognition, where 0 is "not at all" and 10 is "extremely familiar"?

- ○ 0 (not at all)
- ○ 1

- ...
- 9
- 10 (extremely familiar)

If you wish to see the information about monetary incentives again, please click the Scheme of Incentives, otherwise please click the button on the bottom right to continue with the prediction survey. You will able to access this information throughout the survey.

Team 1 hypothesis 1

Below, you are going to see an ORIGINAL research hypothesis and one of a dozen different versions of materials designed by research teams to test that hypothesis. For each set of materials, please predict whether you think the study will result in support for the hypothesis.

Hypothesis 1: People explicitly self-report an awareness of harboring negative automatic associations with members of negatively stereotyped social groups.

Material and Planned Analyses

521 American Mechanical Turk subjects took this study. Please find below how the material was presented to the MTurkers when testing the hypothesis, followed by the planned analyses.

Material: Regardless of my explicit (i.e. conscious) beliefs about social equality, I possess automatic (i.e. unconscious) negative associations towards members of stigmatized social groups such as African Americans, members of the LBGT community, and the poor.

Possible Choices: 5-point Likert scale ranging from "Strongly disagree" to "Strongly agree"

Planned Analyses: For Hypothesis 1 materials, the DV is the response to the statement "Regardless of my explicit (i.e. conscious) beliefs about social equality, I possess automatic (i.e. unconscious) negative associations towards members of stigmatized social groups such as African Americans, members of the LBGT community, and the poor."

Team 1 compares the DV to a null hypothesis of $\mu = 3$ with a one-sample t-test. The effect size estimate is a single-sample Cohen's d (the difference between the sample mean and the null hypothesis, divided by the sample standard deviation).

Based on the material and the planned analyses presented above, please answer the following questions.

Your Predictions

A) To what extent does this set of materials provide an informative test of research hypothesis 1?

- 0 (not at all informative)
- 1
- ...
- 9
- 10 (extremely informative)

B) Do you predict hypothesis 1 will find statistically significant support ($p < .05$) when tested using this set of materials? Here we ask you for the probability that you assign the binary outcome: whether the effect for this set of study materials will be in the same direction as the hypothesis, and will be statistically significant with a p-value smaller than 0.05.

Please insert in the box below the probability that the study will produce a significant effect in the direction of the hypothesis. Note that probabilities should range from 0.0 to 1.0

C) How confident are you in your prediction in question B), on a scale from 0 to 10, where 0 is "not confident at all" and 10 is "very confident"?

- o 0 (not confident at all)
- o 1
- o ...
- o 9
- o 10 (very confident)

D) What do you predict will be the effect size for this hypothesis when tested with this set of materials? Here we ask about the effect size in terms of Cohen's d. Please put a negative sign (-) in front of the effect size or correlation if you think it will be the opposite direction from the original hypothesis.

Demographics

Completing assessments of all of the over 60 sets of study materials qualifies you for authorship on this project. Would you like to be listed as a co-author on the crowdsourcing a hypothesis test project when it is submitted?

- o Yes, I would like to be listed as co-author
- o No, I would not like to be listed as co-author

Please fill out these demographic measures.

- First name as you would like it to appear on the final project report

- Last name as you would like it to appear on the final project report
- Middle initial as you would like it to appear on the final project report
- What is your age?
- What is your gender?
- What is your ethnicity?
- What country were you born in?
- What country do you currently reside in?
- How many years of experience with English do you have?
- What is your institution affiliation?
- What department are you in at your institution (e.g., psychology, organizational behavior,
- statistics)?
- What is your job rank? (please select one) [Undergraduate research assistant, Research assistant, Lab manager, Masters student, Doctoral student, Postdoctoral researcher, Non tenure-track Lecturer, Tenure-track Assistant Professor, Untenured Associate Professor, Tenured Associate Professor, Tenured Full Professor, Dean, Other (please indicate)]
- If relevant, what year did you receive, or do you expect to receive, your doctoral degree? Please leave this item blank if you do not have or do not intend to pursue a doctoral degree
- How many total peer-reviewed academic articles have you published?
- How many peer-reviewed academic articles have you published specifically on the topic of moral judgments?
- How many peer-reviewed academic articles have you published specifically on the topic of research methods or statistics?
- How many peer-reviewed academic articles have you published specifically on the topic of implicit or automatic attitudes?
- How many peer-reviewed academic articles have you published specifically on the topic of human happiness?
- How many times have you taught a graduate level statistics or methods course?
- Rate your proficiency in statistics, on a scale from 1 (extremely low) to 10 (extremely high)
- Please paste the link to your Google Scholar profile here if you have one
- Please paste the link to your professional website here if you have one
- Please upload your most current CV
- What is your email address?
- What is your work address?

SUPPLEMENT 4 - Online advertisements for the project

FACEBOOK VERSION OF ADVERTISEMENT:

Forecasting survey: Can you predict a study's results just from looking at the materials?

Can you predict a priori whether a hypothesis will be empirically supported by simply examining the study materials (scenarios, manipulation, and dependent variables)? Please join us on this crowdsourced project in which you and others will attempt to do just that. In exchange for your time, you will have the option to join us as a collaborator and co-author on the paper reporting the project.

In this forecasting survey, we are asking colleagues across the world to assess approximately 65 sets of materials designed by up to 15 different research teams to test 5 experimental hypotheses from the field of social psychology. Respondents are asked to predict a priori the results— effect sizes and significance levels— that will be obtained in the online data collection using each of the 65 sets of materials. In return for making these predictions, and for evaluating the quality of each set of materials and filling out a set of demographic measures, you will be credited as an author on the final report of this large scale crowdsourced project. You will have up to one month from when you begin the survey to examine the materials and make your predictions, and can begin anytime up to January 1st, 2018. You can complete the survey in separate sittings if you wish.

Please complete the forecasting survey here

https://hhs.qualtrics.com/jfe/form/SV_429KTVl1tvp7VGZ

EMAIL TO COLLEAGUES WHO TEACH GRADUATE METHODS COURSES

Forecasting survey: Crowdsourcing a hypothesis test project

Hi [ADD NAME]

I am wondering if you or someone you know is teaching a methods class and may have students interested in taking part in the forecasting survey for our "Crowdsourcing a hypothesis test" project. Or perhaps you know some individual graduate students or colleagues who might want to participate? If so please forward them our advertisement below with the survey link.

Forecasting survey: Can you predict a study's results just from looking at the materials?

Can you predict a priori whether a hypothesis will be empirically supported by simply examining the study materials (scenarios, manipulation, and dependent variables)? Please join us on this crowdsourced project in which you and others will attempt to do just that. In exchange for your time, you will have the option to join us as a collaborator and co-author on the paper reporting the project.

In this forecasting survey, we are asking colleagues across the world to assess approximately 65

sets of materials designed by up to 15 different research teams to test 5 experimental hypotheses from the field of social psychology. Respondents are asked to predict a priori the results— effect sizes and significance levels— that will be obtained in the online data collection using each of the 65 sets of materials. In return for making these predictions, and for evaluating the quality of each set of materials and filling out a set of demographic measures, you will be credited as an author on the final report of this large scale crowdsourced project. You will have up to one month from when you begin the survey to examine the materials and make your predictions, and can begin anytime up to January 1st, 2018. You can complete the survey in separate sittings if you wish.

Please complete the forecasting survey here

https://hhs.qualtrics.com/jfe/form/SV_429KTVl1tvp7VGZ

EMAIL TO RESEARCH TEAMS WHO DESIGNED MATERIALS

Dear [Researcher's Name],

Thank you again for being a part of the Crowdsourcing a Hypothesis Test project. We (the project team) contacted you a couple of months ago to let you know that we had finished collecting all of the data from the large, primary study, in which thousands of participants on Mechanical Turk were randomly assigned to different sets of study materials designed by researchers from all around the world. We also launched the prediction arm of the project, in which we recruited researchers to predict whether each set of materials will support its

hypothesis. We were hoping to recruit at least 100 scholars to participate in this study, but we only managed to collect a sample of about half that size. So, we are launching another wave of data collection, and we are hoping that you can help us one more time. Below, you will find a description of the prediction study. If you have any friends, collaborators, or students that you think would be interested in this project, we would really appreciate it if you could forward this description to them. The only requirement for participation are that they be involved in academic behavioral science/psychology or a related field in some capacity (undergraduate research assistants, tenured professors, and everyone in between are welcome).

Making predictions about the results of what amounts to 60 different studies is no small feat, so we are happy to offer participants in the forecasting study chance to join us as authors on the eventual write-up of this project, if they would like to.

Thank you again for being a part of this exciting project! We are nearly ready to start writing up the results and sharing them with everyone involved!

EMAIL TO POTENTIAL TWEETERS

Re: Twitter ad: Predicting study results from materials

Hi [add name]

Hope things are good!

My collaborators and I have launched a forecasting survey to supplement our "Crowdsourcing a hypothesis test" project.

In the main project, we have "crowdsourced" a hypothesis test by having up to 15 different research teams create their own versions of the materials to test 5 different hypotheses. We have completed running a large online data collection randomly assigning thousands of participants to complete one of the up to 15 different versions of each of the 5 studies, with over 300 Mechanical Turk subjects per study version to provide adequate power. This will allow us to examine empirically the extent to which differences in how independent research teams choose to operationalize the same hypothesis influence the final effect size estimates.

In a forecasting survey, we are testing whether colleagues around the world can guess whether a hypothesis will be empirically supported by simply examining the study materials (scenarios, manipulation, and dependent variables) created by the different research teams. Respondents to the forecasting survey are asked to assess the approximately 65 sets of materials, and to predict the results— effect sizes and significance levels— that will be obtained in the online data collection using each of the sets of materials. In return for making these predictions, and for evaluating the quality of each set of materials and filling out a set of demographic measures, they are credited as an author on the final report of this large scale crowdsourced project. Respondents have until January 1, 2018 to complete the survey.

Would you be willing to tweet the link to our advertisement to help us recruit forecasters? This

would be a huge help for this research.

The tweet should look something like this:

Collaborative project: Can you predict a study's results just from looking at the materials?

https://hhs.qualtrics.com/jfe/form/SV_429KTVl1tvp7VGZ

SUPPLEMENT 5 – More detailed methods and results from the forecasting study

Methodological details

Materials. We recruited respondents for the forecasting survey, almost all academics, and provided them with the materials and the research designs created by the 15 materials-maker teams from the Main Studies and Replication Studies for each of the five original hypotheses. For each of the 64 study versions (up to 13 for each of the five hypotheses), we asked for their predictions about the effect size as well as significance level (whether $p < .05$ or not). All the relevant study materials were fully disclosed to the forecasters, including detailed information about the sample sizes employed in each version, the exact framing of the questions, the research designs, vignettes, and the directional versions of the original hypotheses.

In addition to their forecasts, respondents were also asked for their confidence in each of the predictions they made on a scale ranging from 0 (not at all confident) to 10 (very confident). They were further asked to self-report their degree of familiarity with research on moral judgments, negotiation, and implicit cognition (0 = not at all, 10 = extremely familiar), and the extent to which each set of materials provides an informative test of the research hypothesis (0 = not at all informative, 10 = extremely informative). This last question had the purpose of capturing independent ratings of the quality of each set of materials. Forecasters also completed a battery of demographic measures, including their area of specialization and job rank (e.g., research assistant, graduate student, postdoctoral researcher, assistant professor, associate professor, full professor), and degree of familiarity with statistics (0 = extremely low; 10 = extremely high). (See Supplement 3 for the forecasting survey materials).

Each participant randomized in the monetary incentives treatment received a financial bonus based on the accuracy of her predictions. The incentives were computed on the predictions

of both the effect size and statistical significance (in terms of $p < .05$ or not) of a randomly selected set of materials. Realized payoff ranged between $1 and $59.90; the average payment was $12.50.

The survey consisted of reviewing 64 sets of study materials and providing over 300 quality assessments and predictions about the empirical results that would be obtained. Forecasters therefore had one month to submit their answers, and they could complete the survey in more than one session. On average respondents took more than eight hours, often divided across multiple sessions, to complete the survey. We randomized the order in which predictions were made for each study version.

Recruiting forecasters. We posted the link to the forecasting survey on various academic websites, platforms, and Facebook pages oriented towards psychology, judgment and decision making, and research methodology (SPSP, JDM Society, SJDM mailing list, Psych Science Accelerator network, Psych MAP, ISCON, Psychological Methods Discussion Group, Many Labs network). We also emailed the materials-makers teams in the project, colleagues at our own institutions (e.g., INSEAD and the Stockholm School of Economics) and colleagues who teach methods classes, asking them to forward the forecasting survey to students and anyone else who might be interested. Finally, we asked colleagues with a large number of followers on Twitter to post on their accounts the link to the survey. Supplement 4 presents the online advertisements and emails used to recruit forecasters.

One hundred forty-one individuals (78% male, 22% female) completed the forecasting survey. The median age of respondents was 32 years. One third of the forecasters (41 out of 141) were born in the United States; two thirds were born in other countries, including Germany (25), Canada (8), Slovakia (8), Italy (7), France (5) and the United Kingdom (5), with an additional 42

participants originally from 23 further countries. The forecasters currently reside in 24 countries. Unsurprisingly, given the length of the survey and time commitment involved in making the predictions and expert assessments, the completion rate was slightly less than 20% (141 completed surveys out of 712 individuals who initially clicked on the survey link). Among those who did not complete the survey, 269 potential respondents spent a negligible amount of time on it (less than five minutes), suggesting that many initial clicks were due to curiosity rather than to a deep interest in taking the survey. Out of 141 individuals that completed the survey, more than 97% worked in academia at the time of the data collection, with job ranks ranging from undergraduate research assistants to tenured full professors (as summarized in Table S5.1).

Table S5.1. Forecasters by job rank.

Job Rank	Forecasters	Percentage
Undergraduate research assistant	1	0.70
Research assistant	2	1.40
Lab manager	1	0.70
Masters student	5	3.50
Doctoral student	40	28.40
Postdoctoral researcher	20	14.20
Non tenure-track lecturer	8	5.70
Tenure-track assistant professor	27	19.10
Untenured associate professor	3	2.10
Tenured associate professor	19	13.50
Tenured full professor	6	4.30
Other job rank	9	6.40

Only five respondents categorized themselves as working outside academia (one consultant, one data scientist, one director, and two research scientists). Removing these individuals from the sample had no substantial effect on the reported results, which include both academics and non-academics unless stated otherwise. The median number of publications of the forecasters was 7 (*IQR* = [2.5, 15]), and the mean self-reported proficiency in statistics 5.49 (*SD* = 1.31; 0 = extremely low; 10 = extremely high). Forecasters reported moderate familiarity (0 = not at all, 10 = extremely familiar) with the research topics of "implicit and automatic cognition" (*M* = 4.87, *SD* = 2.42), "negotiations" (*M* = 3.19, *SD* = 2.29), and "moral judgments" (*M* = 5.04, *SD* = 2.47).

Table S5.2. Descriptive statistics for the forecasting survey

Variable	Mean	SD
Confidence: Significance	5.49	0.43
Confidence: Effect size	5.02	0.31
Familiarity: Moral Judgment	5.04	2.47
Familiarity: Negotiation	3.19	2.29
Familiarity: Implicit Cognition	4.87	2.42
Familiarity: Statistics	5.49	1.31
Informativeness of the Materials	5.22	0.74

Notes. Means and standard deviations (*SD*) reported in the table are computed based on the study-level averages of each variable.

Table S5.3: Forecasters by country of birth and residence

	Country of Birth	Country of Residence		Country of Birth	Country of Residence
Australia	-	2	Netherlands	3	4
Austria	4	2	Norway	-	2
Belarus	1	-	Poland	4	1
Belgium	-	3	Portugal	1	3
Canada	8	6	Russia	1	-
China	1	1	Serbia	3	3
Colombia	2	1	Singapore	1	2
France	5	4	Slovakia	8	7
Germany	25	21	Spain	-	1
Indonesia	1	-	Sweden	2	2
Ireland	2	-	Switzerland	1	3
Israel	1	-	Taiwan	2	-
Italy	7	6	Thailand	-	1
Japan	1	-	Ukraine	1	-
Kazakhstan	1	-	UK	5	11
Lithuania	1	1	USA	47	53
Malaysia	1	-	Uruguay	1	1

Additional Analyses

Robustness checks. The results reported in the main text are robust to several alternative approaches to analyzing the data pre-registered in our analysis plan. They are likewise robust to using the Replication Studies' and meta-analyzed effect sizes as the objective outcomes predicted by the forecasters. The complete datasets from the Main Studies, Replication Studies, and Forecasting Survey are publicly posted online (https://osf.io/9jzy4/) to facilitate re-analysis. In the spirit of crowdsourcing, we welcome alternative perspectives on all of our results.

Individual regression analyses and non-parametric Spearman correlation tests confirm that there is a positive and significant correlation between scientists' forecasts and realized outcomes of the studies, both for significance levels and effect sizes (refer to tables S5.4 and S5.5 for individual regression analysis).

The Spearman test on the correlation between the aggregated predictions of directional significance and the vector collecting realized statistical significance for each of the 64 versions of the materials is positive and significant: $rho(62) = 0.599$, $p = 1.748 \times 10^{-7}$. The Spearman correlation between scientists' beliefs and realized effect sizes is likewise statistically significant: $r(62) = 0.699$, $p = 2.2 \times 10^{-16}$. Results are confirmed when restricting the sample to incentivized forecasts (significance: $rho(62) = 0.665$, $p = 1.998 \times 10^{-9}$, effect size: $rho(62) = 0.747$, $p = 2.2 \times 10^{-16}$), and to non-incentivized forecasts (significance: $rho(62) = 0.516$, $p = 1.305 \times 10^{-5}$, effect size: $rho(62) = 0.584$, $p = 4.16 \times 10^{-7}$).

Table S5.4a: Forecasting results controlling for different sets of fixed effects (linear model)

	Dependent Variable: Realized Statistical Significance			
	(1)	(2)	(3)	(4)
Predicted Sign.	0.309***	0.148***	0.255***	0.089**
	(0.054)	(0.036)	(0.058)	(0.032)
Constant	0.430***	0.463***	0.483*	0.509*
	(0.072)	(0.140)	(0.220)	(0.203)
Team FE	No	No	Yes	Yes
Hypothesis FE	No	Yes	No	Yes
Observations	9,024	9,024	9,024	9,024
R^2	0.032	0.258	0.229	0.473
F Statistic	295.764***	627.956***	167.145***	403.601***

Notes. *$p < 0.05$; **$p < 0.01$; ***$p < 0.001$. Standard errors clustered at individual and team x hypothesis level.

Table S5.4b: Forecasting results controlling for different sets of fixed effects (probit model)

	Dependent Variable: Realized Statistical Significance			
	(1)	(2)	(3)	(4)
Predicted Sign.	0.805***	0.482***	0.818***	0.464***
	(0.140)	(0.098)	(0.151)	(0.088)
Constant	-0.182***	-0.149	-0.117	-0.095
	(0.177)	(0.354)	(0.576)	(0.688)
Team FE	No	No	Yes	Yes
Hypothesis FE	No	Yes	No	Yes
Observations	9,024	9,024	9,024	9,024
Log Likelihood	-5,999.396	-4,850.174	-4,950.452	-3,382.261
Akaike Inf. Crit.	12,002.790	9,712.348	9,934.903	6,806.521

Notes. *$p < 0.05$; **$p < 0.01$; ***$p < 0.001$. Standard errors clustered at individual and team x hypothesis level.

Table S5.5: Forecasting results controlling for different sets of fixed effects (linear model)

	Dependent Variable: Realized Effect Size			
	(1)	(2)	(3)	(4)
Predicted Eff. Size	0.241*	0.097*	0.228*	0.091**
	(0.101)	(0.045)	(0.085)	(0.033)
Constant	0.252***	0.052	0.155	-0.062
	(0.065)	(0.152)	(0.234)	(0.255)
Team FE	No	No	Yes	Yes
Hypothesis FE	No	Yes	No	Yes
Observations	9,024	9,024	9,024	9,024
R^2	0.030	0.418	0.177	0.518
F Statistic	283.557***	1,295.698***	121.313***	483.321***

Notes. *$p < 0.05$; **$p < 0.01$; ***$p < 0.001$. Standard errors clustered at individual and team x hypothesis level.

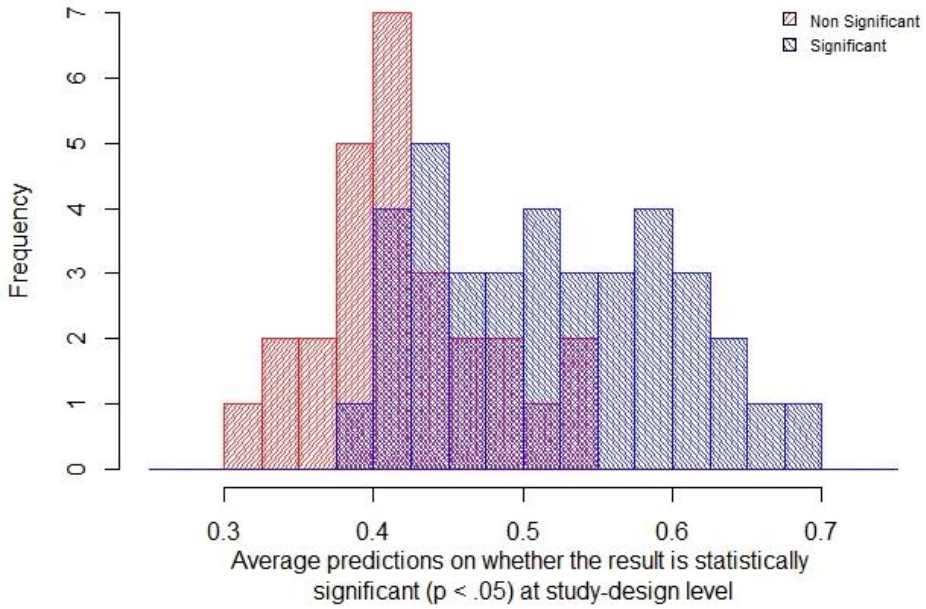


Figure S5.1. Histograms of average predictions regarding whether the result is statistically significant ($p < .05$) or not at study-design level. Study designs that yielded statistically non-significant results (red with horizontal texture), received less support from forecasters than the study designs that yielded statistically significant results (blue with vertical texture; the overlap of the two distributions is shown with both textures and colors). This is in terms of the average study-design level probability assigned to the binary outcome: "whether the effect for this set of study materials will be in the same direction as the hypothesis, and will be statistically significant with a *p*-value smaller than 0.05."

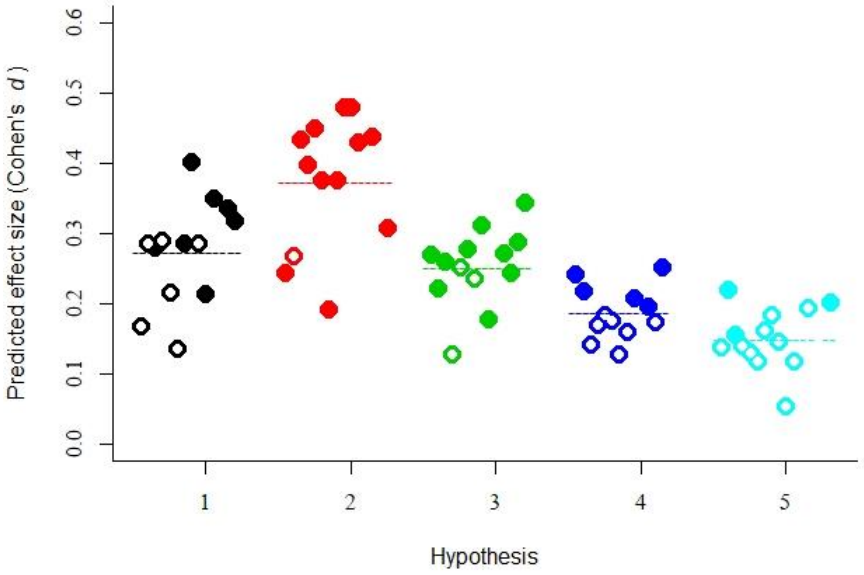


Figure S5.2. Dispersion of average predicted effect sizes (Cohen's *d*) by hypothesis. Hollow dots refer to study versions that yielded realized results inconsistent with the original hypothesis in the sense of not being statistically significant in the expected direction. Filled dots indicate study designs whose obtained results were consistent with the original hypothesis (i.e., statistically significant in the expected direction). H1: Awareness of automatic prejudice, H2: Extreme offers reduce trust, H3: Moral praise for needless work, H4: Proximal authorities drive legitimacy of performance enhancers, H5: Deontological judgments predict happiness.

Incentives and forecasting accuracy. In individual level regressions in which the dependent variable is a measure of the accuracy of prediction (absolute prediction error for the main pre-specified hypothesis, squared prediction error for the robustness checks), we find that monetary incentives do not statistically significantly improve the accuracy of forecasts. This holds true for predictions of directional statistical significance as well as for predictions of effect size.

(S5.1) $$y_{ith} = \beta_0 + \beta_1 T_i + Team_t + Hyp_h + \varepsilon_{ith}$$

(S5.2) $$\hat{y}_{ith} = \beta_0 + \beta_1 T_i + Team_t + Hyp_h + \varepsilon_{ith}$$

In the most parsimonious models (S5.1) and (S5.2) in which the absolute prediction errors (y_{ith} and $\hat{y}_{ith}$ for whether the results are statistically significant ($p < .05$) and effect size predictions, respectively) are regressed on the treatment incentives dummy T_i (positive values identify the participants that were assigned to the monetary incentives group) and on the hypothesis and teams fixed effects, the coefficient of the monetary incentives treatment is $\beta_1 = -.011$, $t(9003) = -0.89$, $p = .233$ for the prediction of statistical significance in terms of $p < 0.05$, and $\beta_1 = -.009$, $t(9003) = -0.89$, $p = .372$ for the prediction of effect sizes. See Table S5.6 for the full regression tables.

The correlations between scientists' forecasts and realized statistical significance were similar when calculated separately for the incentivized condition, $r(62) = .62$, 95% *CI* [.45, .75], $p < .001$, and the non-incentivized condition, $r(62) = .53$, 95% *CI* [.32 .68], $p < .001$. These correlations were computed at team-hypothesis level, for a total of 64 observations. Likewise, correlations between scientists' forecasts and observed effect sizes estimates were comparable in the incentivized, $r(62) = .72$, 95% *CI* [.57 , .82], $p < .001$, and non-incentivized conditions, $r(62) = .62$, 95% *CI* [.44, .75], $p < .001$.

Incentivized and non-incentivized scientists also exhibited similar means and standard deviations for their forecasted outcomes, which in turn showed similar correspondence with the observed overall outcomes of the studies. In the incentivized condition, means and standard deviations for forecasts about statistical significance ($M = 0.47$, $SD = 0.09$; paired t to compare with realized outcomes $t(63) = -1.93$, 95% *CI* [-0.22, 0.00], $p = .059$), are comparable to the non-incentivized condition ($M = 0.48$, $SD = 0.09$; paired $t(63) = -1.64$, 95% *CI* [-0.21, 0.02], $p = .106$). In the incentivized condition, forecasts for effect sizes ($M = 0.25$, $SD = 0.12$; paired $t(63) = -1.05$, 95% *CI* [-0.18, 0.06], $p = .298$) are likewise very similar to the non-incentivized

condition (M = 0.25, SD = 0.91; paired $t(63)$ = -0.99, 95% CI [-0.19, 0.06], p = .326). We observed no substantive differences in forecasts between scientists who received monetary incentives for accuracy and those who did not (see Table S5.6).

Table S5.6: Monetary incentives and forecasting accuracy

	Dependent Variable: Absolute Predicted Error	
	Significance	Effect Size
	(1)	(2)
Treatment	-0.011	-0.009
	(0.009)	(0.010)
Team FE	Yes	Yes
Hypothesis FE	Yes	Yes
Observations	9,024	9,024
R^2	0.030	0.221
F Statistic	13.772***	127.853***

Notes. *p < 0.05; **p < 0.01; ***p < 0.001. Standard errors clustered at individual and team x hypothesis level.

As a robustness check for the (null) effect of monetary incentives on the accuracy of predictions, we measured the quality of predictions using as the dependent variable squared prediction errors rather than absolute prediction errors. Next, we regressed individual squared prediction errors on the treatment dummy and on teams and hypotheses fixed effects (Table S5.7). The estimated coefficients for the treatment variable were not statistically significant (β = -.009, $t(9003)$ = -0.688, p = .491 for the predictions of significance levels, and β = -.080, $t(9003)$ = -0.697, p = .486 for the predictions of effect sizes), in line with the results presented earlier.

Table S5.7: Monetary incentives and forecasting accuracy

	Dependent Variable: Squared Prediction Error	
	Significance	Effect Size
	(1)	(2)
Treatment	-0.009	-0.080
	(0.013)	(0.115)
Team FE	Yes	Yes
Hypothesis FE	Yes	Yes
Observations	9,024	9,024
R^2	0.026	0.008
F Statistic	11.921***	3.465***

Notes. *$p < 0.05$; **$p < 0.01$; ***$p < 0.001$. Standard errors clustered at individual and team x hypothesis level.

Which hypothesis elicits the most accuracy? The correlations presented in Figure 3b in the main text suggest that scientists' predictions were more accurate for Hypothesis 5 relative to Hypotheses 1 to 4. We can only speculate that this finding is a result of the lower variability that characterizes the observed effect sizes estimated for Hypothesis 5 if compared to the observed effect sizes estimated for the other Hypotheses. However, with only a small number of observations (12 or 13) for each hypothesis, we consider this finding tentative.

Who is most accurate? We further pre-registered some additional regressions with the purpose of identifying if any individual characteristics are associated with more accurate forecasts. We did not make strong directional predictions about these potential demographic and individual-level moderators of forecasting accuracy, which included gender, age, job rank, number of publications, confidence in the prediction, familiarity with the academic area of study, and familiarity with statistics.

We estimated the linear regression model (S5.3) and (S5.4) where, in line with previous models, y_{ith} and $\hat{y_{ith}}$ are the absolute prediction error for the forecasts about statistical significance

and about effect size, T_i is the treatment incentives dummy, $Moderator_k$ refers to the aforementioned variables, and $Team_t$ and $Hypothesis_h$ refer to the team- and hypothesis-level fixed effects:

$$y_{ith} = \beta_0 + \beta_1 T_i + \beta_2 Moderator_k + Team_t + Hyp_h + \varepsilon_{ith} \text{ for } k = 1, \ldots, 7 \quad \text{(S5.3)}$$

$$\hat{y}_{ith} = \beta_0 + \beta_1 T_i + \beta_2 Moderator_k + Team_t + Hyp_h + \varepsilon_{ith} \text{ for } k = 1, \ldots, 7 \quad \text{(S5.4)}$$

Tables S5.8 and S5.9 show the estimated coefficients obtained from equations (S5.3) and (S5.4). First, we included one moderator at a time (columns 1 to 7) to account for potential multicollinearity among the regressors, then further ran a comprehensive model (column 8). The two-way clustered standard errors, at individual and at team-hypothesis level, are reported in parenthesis.

As seen in Table S5.8, the variables "Job rank" and "Confidence about the expressed forecast" are the only variables that are correlated with more accurate predictions about the statistical significance levels of the studies. Higher confidence in the forecast was associated with lower absolute prediction errors (β = -0.012, $t(9002)$ = -2.51, p = .012). We controlled for job ranks by including a dummy for each of the 13 possible job rank categories. Higher job ranks tended to be associated with more accurate forecasts when compared to the reference category "Undergraduate Research Assistant" for the predictions about a statistically significant result or not. Notably, however, three further variables that aimed to capture the seniority and eminence of the researcher (total number of articles published in peer-reviewed journals, reported familiarity with the area of study, and reported proficiency in statistics) were not statistically significantly correlated with more accurate predictions.

Table S5.8: Forecaster characteristics and accuracy of predictions about statistical significance

	Dependent Variable: Absolute Prediction Error (Statistical Significance)							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Treatment	-0.012	-0.012	-0.013	-0.011	-0.016	-0.012	-0.013	-0.017
	(0.009)	(0.009)	(0.010)	(0.010)	(0.010)	(0.009)	(0.009)	(0.012)
Gender	-0.017							-0.014
	(0.013)							(0.013)
Age		0.0004						0.001
		(0.001)						(0.001)
Research Assistant			-0.054					-0.064
			(0.041)					(0.034)
Lab Manager			-0.058***					-0.051*
			(0.007)					(0.025)
Master St.			-0.110***					-0.129***
			(0.009)					(0.018)
Doctoral St.			-0.093***					-0.117***
			(0.016)					(0.021)
Post Doc			-0.088***					-0.112***
			(0.017)					(0.022)
Lecturer NTT			-0.085***					-0.117***
			(0.024)					(0.033)
Assistant Prof TT			-0.097***					-0.131***
			(0.016)					(0.023)
Associate Prof Unt			-0.095**					-0.120**
			(0.035)					(0.045)
Associate Prof Ten			-0.081***					-0.121***
			(0.018)					(0.026)
Full Prof			-0.106**					-0.144**
			(0.037)					(0.054)
Other			-0.094***					-0.144***
			(0.010)					(0.028)
Publications				-0.0001				-0.0002
				(0.0002)				(0.0005)
Confidence					-0.012*			-0.014*
					(0.005)			(0.006)
Familiarity						-0.002		0.001
						(0.002)		(0.002)
Proficiency stat							-0.003	0.003
							(0.004)	(0.004)
Obs.	9,024	8,896	9,024	8,896	9,024	9,024	8,960	8,768
R^2	0.030	0.029	0.031	0.031	0.037	0.030	0.030	0.041
F Statistic	13.391***	12.552***	9.396***	13.298***	16.430***	13.212***	13.378***	10.006***

Notes. *$p < 0.05$; **$p < 0.01$; ***$p < 0.001$. Standard errors clustered at individual and team x hypothesis levels. All the regressions control for Team and Hypothesis fixed effects. Omitted category for job rank: 'undergraduate research assistant.'

Parallel analyses were conducted for effect size predictions (see Table S5.9). No clear pattern emerges in the correlation between job rank and accuracy; moreover, statistical significance tends to disappear once controlling for the full set of individual moderators. At the same time, confidence in effect size predictions is associated with *less* accurate forecasts ($\beta = 0.014$, $t(9002) = 3.87$, $p < .001$); yet note this is the opposite pattern to that observed for confidence about predictions regarding statistical significance levels, rendering the results conflicting and inconclusive. Familiarity with the topic of the study is not associated with more accurate predictions, nor are gender, age, total number of publications in peer-reviewed journals, and reported proficiency in statistics. In sum, we generally failed to identify individual differences consistently associated with making more accurate forecasts about research findings, with the partial exception of academic seniority which was associated with more accurate forecasts about statistical significance levels but not effect sizes.

Table S5.9: Forecaster characteristics and accuracy of predictions about effect sizes

	Dependent Variable: Absolute Prediction Error - Effect Size							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Treatment	-0.009	-0.008	-0.005	-0.009	-0.006	-0.009	-0.009	-0.005
	(0.010)	(0.011)	(0.011)	(0.011)	(0.010)	(0.010)	(0.011)	(0.012)
Gender	0.0003							0.001
	(0.008)							(0.010)
Age		-0.001						0.001
		(0.001)						(0.001)
Research Assistant			-0.031**					-0.047*
			(0.010)					(0.019)
Lab Manager			0.002					0.024
			(0.022)					(0.021)
Master St.			0.041					0.076
			(0.046)					(0.047)
Doctoral St.			-0.014					0.022
			(0.014)					(0.018)
Post Doc			-0.005					0.023
			(0.015)					(0.017)
Lecturer NTT			-0.030**					0.007
			(0.010)					(0.014)
Assistant Prof TT			-0.031**					0.001
			(0.011)					(0.013)
Associate Prof Unt			0.037**					0.065***
			(0.011)					(0.013)
Associate Prof Ten			-0.027**					0.002
			(0.010)					(0.014)
Full Prof			-0.044***					-0.019
			(0.011)					(0.020)
Other			-0.017					0.015
			(0.010)					(0.013)
Publications				-0.0003				-0.0001
				(0.0002)				(0.0004)
Confidence					0.014***			0.016**
					(0.004)			(0.005)
Familiarity						-0.001		-0.005
						(0.003)		(0.003)
Proficiency stat							0.001	-0.005
							(0.003)	(0.004)
Obs.	9,024	8,896	9,024	8,896	9,024	9,024	8,960	8,768
R2	0.221	0.219	0.222	0.220	0.224	0.221	0.220	0.222
F Statistic	121.751***	118.406***	82.970***	118.922***	124.019***	121.763***	120.209***	67.398***

Notes. *$p < 0.05$; **$p < 0.01$; ***$p < 0.001$. Standard errors clustered at individual and team x hypothesis levels. All the regressions control for Team and Hypothesis fixed effects. Omitted category for job rank: 'undergraduate research assistant.'

Forecasts and the Replication Studies' effect sizes. Although predictions in the forecasting survey were specifically aimed at the Mechanical Turk participants in the Main Studies, forecasted effect sizes and statistical significance levels also predicted outcomes in the Replication Studies using the PureProfile participant pool. The correlation between scientists' forecasts and the results being statistically significant in the predicted direction is positive and itself statistically significant: $r(62) = 0.42$, 95% *CI* [0.19, 0.60], $p < .001$; the same holds for the correlation between scientists' forecasts and observed effect sizes: $r(62) = 0.59$, 95% *CI* [0.41 ,0.73], $p < .001$. Forecasters show sensitivity to how design choices affect research results, anticipating the results of different teams of materials-makers within each hypothesis, as well as different hypotheses within each team of materials designers (see Tables S5.10 and S5.11).

We also repeated our forecasting analyses using the effect sizes for each of the 64 study designs after meta-analytically combining the results of the Main Studies and Replication Studies. Scientists' forecasts were again related to both realized statistical significance levels, $r(62) = 0.51$, 95% *CI* [0.30, 0.67], $p < .001$, and observed effect sizes, $r(62) = 0.67$, 95% *CI* [0.51, 0.78], $p < .001$, and showed sensitivity to design choices within each hypothesis (Tables S5.14 and S5.15). Both when the Replication Studies' effect sizes and meta-analyzed effect sizes were used as the outcomes, monetary incentives and the assessed individual differences did not consistently moderate forecasting accuracy. The following tables repeat the forecasting analyses using as the targets of prediction the outcomes of the Replication Studies (tables S5.10, S5.11, S5.12, S5.13a and S5.13b), and the outcomes of the meta-analysis of the Main Studies and Replication Studies (tables S5.14, S5.15, S5.16, S5.17a and S5.17b), respectively. Forecasters'

predictions are again positively correlated with the realized significance and effect size of the studies; this finding is robust to the inclusion of different sets of fixed effects. Monetary incentives do not increase the accuracy of the predictions, and the assessed individual characteristics do not consistently moderate the accuracy of the forecasts.

Table S5.10: Monetary incentives and forecasting accuracy – Replication Studies

	Dependent Variable: Absolute Predicted Error	
	Significance	Effect Size
	(1)	(2)
Treatment	-0.008	-0.007
	(0.006)	(0.013)
Team FE	Yes	Yes
Hypothesis FE	Yes	Yes
Observations	9,024	9,024
R^2	0.023	0.121
F Statistic	10.431***	61.654***

Notes. *$p < 0.05$; **$p < 0.01$; ***$p < 0.001$. Standard errors clustered at individual and team x hypothesis level.

Table S5.11: Forecasting significance levels in the Replication Studies controlling for different sets of fixed effects

	Dependent Variable: Realized Significance			
	(1)	(2)	(3)	(4)
Predicted Sign.	0.222***	0.098*	0.187***	0.077*
	(0.062)	(0.040)	(0.054)	(0.032)
Constant	0.332***	0.334*	0.514*	0.506*
	(0.066)	(0.135)	(0.220)	(0.233)
Team FE	No	No	Yes	Yes
Hypothesis FE	No	Yes	No	Yes
Observations	9,024	9,024	9,024	9,024
R^2	0.016	0.137	0.271	0.372
F Statistic	148.120***	287.086***	208.850***	266.110***

Notes. *$p < 0.05$; **$p < 0.01$; ***$p < 0.001$. Standard errors clustered at individual and team x hypothesis level.

Table S5.12: Forecasting effect sizes in the Replication Studies controlling for different sets of fixed effects

	Dependent Variable: Realized Effect Size			
	(1)	(2)	(3)	(4)
Predicted Eff. Size	0.151*	0.065*	0.142**	0.058**
	(0.065)	(0.029)	(0.054)	(0.020)
Constant	0.131**	-0.087	0.043	-0.200
	(0.048)	(0.133)	(0.195)	(0.194)
Team FE	No	No	Yes	Yes
Hypothesis FE	No	Yes	No	Yes
Observations	9,024	9,024	9,024	9,024
R^2	0.022	0.344	0.155	0.473
F Statistic	198.294***	947.395***	103.420***	404.755***

Notes. *$p < 0.05$; **$p < 0.01$; ***$p < 0.001$. Standard errors clustered at individual and team x hypothesis level.

Tables S5.13a and S5.13b show the estimated coefficients obtained from equations (S5.3) and (S5.4), but now using the data about statistical significance and the effect size from each set of materials from the PureProfile participants (i.e., the Replication Studies). First we included one moderator at a time (columns 1 to 7; moderators specified in the first column), then we run a comprehensive model (column 8).

Table S5.13a: Forecaster characteristics and accuracy of predictions about statistical significance for the Replication Studies

	Dependent Variable: Absolute Prediction Error – Statistical Significance							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Treatment	-0.008	-0.007	-0.008	-0.009	-0.009	-0.008	-0.009	-0.010
	(0.006)	(0.006)	(0.007)	(0.006)	(0.006)	(0.006)	(0.006)	(0.007)
Gender	0.005							0.003
	(0.010)							(0.009)
Age		-0.001						-0.001
		(0.001)						(0.001)
Research Assistant			-0.103***					-0.096***
			(0.026)					(0.024)
Lab Manager			-0.065**					-0.075**
			(0.021)					(0.025)
Master St.			-0.079***					0.087***
			(0.019)					(0.022)
Doctoral St.			-0.091***					-0.097***
			(0.016)					(0.021)
Post Doc			-0.093***					-0.100***
			(0.015)					(0.021)
Lecturer NTT			-0.095***					-0.086**
			(0.023)					(0.032)
Assistant Prof TT			-0.092***					-0.098***
			(0.017)					(0.025)
Associate Prof Unt			-0.099**					-0.105***
			(0.029)					(0.032)
Associate Prof Ten			-0.091***					-0.096***
			(0.017)					(0.026)
Full Prof			-0.128**					-0.145**
			(0.032)					(0.039)
Other			-0.078***					-0.083***
			(0.014)					(0.030)
Publications				-0.00004				0.0004
				(0.0002)				(0.0004)
Confidence					-0.001			-0.002
					(0.005)			(0.007)
Familiarity						0.002		0.003
						(0.002)		(0.002)
Proficiency stat							0.0005	-0.0005
							(0.003)	(0.003)
Obs.	9,024	8,896	9,024	8,896	9,024	9,024	8,960	8,768
R^2	0.023	0.023	0.024	0.023	0.023	0.023	0.023	0.026
F Statistic	9.958***	9.851***	7.267***	10***	9.946***	10.13***	10***	6.227***

Notes. *$p < 0.05$; **$p < 0.01$; ***$p < 0.001$. Standard errors clustered at individual and team x hypothesis levels. All the regressions control for Team and Hypothesis fixed effects. Omitted category for job rank: 'undergraduate research assistant'

Table S5.13b: Forecaster characteristics and accuracy of predictions about effect sizes for the Replication Studies

	Dependent Variable: Absolute Prediction Error - Effect Size							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Treatment	-0.007	-0.005	0.002	-0.006	-0.003	-0.007	-0.007	-0.002
	(0.013)	(0.013)	(0.014)	(0.014)	(0.013)	(0.013)	(0.014)	(0.015)
Gender	0.004							0.006
	(0.013)							(0.014)
Age		-0.001						0.001
		(0.0007)						(0.002)
Research Assistant			-0.012					-0.026
			(0.009)					(0.022)
Lab Manager			0.018					0.039
			(0.019)					(0.020)
Master St.			0.108					0.148*
			(0.058)					(0.060)
Doctoral St.			0.026					0.068**
			(0.015)					(0.021)
Post Doc			0.036					0.068**
			(0.020)					(0.023)
Lecturer NTT			-0.006					0.042**
			(0.012)					(0.016)
Assistant Prof TT			-0.006					0.033*
			(0.011)					(0.015)
Associate Prof Unt			0.073***					0.107***
			(0.018)					(0.012)
Associate Prof Ten			0.009					0.046*
			(0.014)					(0.019)
Full Prof			-0.027*					0.003
			(0.011)					(0.025)
Other			0.016					0.059**
			(0.018)					(0.023)
Publications				-0.0004				-0.0001
				(0.0003)				(0.0005)
Confidence					0.016***			0.019***
					(0.003)			(0.005)
Familiarity						-0.00001		-0.004
						(0.003)		(0.003)
Proficiency stat							0.002	-0.006
							(0.005)	(0.006)
Obs.	9,024	8,896	9,024	8,896	9,024	9,024	8,960	8,768
R2	0.121	0.120	0.124	0.120	0.127	0.121	0.120	0.128
F Statistic	58.72***	57.39***	41.07***	57.45***	62.12***	58.71***	57.97***	34.74***

Notes. *$p < 0.05$; **$p < 0.01$; ***$p < 0.001$. Standard errors clustered at individual and team x hypothesis levels. All the regressions control for Team and Hypothesis fixed effects. Omitted category for job rank: 'Undergraduate Research Assistant.'

Forecasting Meta-analytic outcomes.

Table S5.14: Monetary incentives and forecasting accuracy - Meta-analytic outcomes

	Dependent Variable: Absolute Predicted Error	
	Significance	Effect Size
	(1)	(2)
Treatment	-0.008	-0.008
	(0.009)	(0.012)
Team FE	Yes	Yes
Hypothesis FE	Yes	Yes
Observations	9,024	9,024
R^2	0.023	0.150
F Statistic	10.811***	79.724***

Notes. $^*p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$. Standard errors clustered at individual and team x hypothesis level.

Table S5.15: Forecasting meta-analyzed results controlling for different sets of fixed effects

	Dependent Variable: Realized Significance			
	(1)	(2)	(3)	(4)
Predicted Sign.	0.268^{***}	0.143^{***}	0.212^{***}	0.086^{*}
	(0.057)	(0.037)	(0.057)	(0.033)
Constant	0.450^{***}	0.465^{*}	0.503^{*}	0.513^{*}
	(0.072)	(0.140)	(0.220)	(0.221)
Team FE	No	No	Yes	Yes
Hypothesis FE	No	Yes	No	Yes
Observations	9,024	9,024	9,024	9,024
R^2	0.024	0.175	0.222	0.374
F Statistic	220.8^{***}	383.212^{***}	161.054^{***}	269.361^{***}

Notes. $^*p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$. Standard errors clustered at individual and team x hypothesis level.

Table S5.16: Forecasting meta-analyzed results controlling for different sets of fixed effects

	Dependent Variable: Realized Effect Size			
	(1)	(2)	(3)	(4)
Predicted Eff. Size	0.193*	0.079*	0.181**	0.071**
	(0.081)	(0.036)	(0.068)	(0.025)
Constant	0.189***	-0.018	0.097	-0.129
	(0.055)	(0.133)	(0.213)	(0.221)
Team FE	No	No	Yes	Yes
Hypothesis FE	No	Yes	No	Yes
Observations	9,024	9,024	9,024	9,024
R^2	0.027	0.400	0.157	0.501
F Statistic	253.253***	1,200.482***	104.472***	451.128***

Notes. *$p < 0.05$; **$p < 0.01$; ***$p < 0.001$. Standard errors clustered at individual and team x hypothesis level.

Tables S5.17a and S5.17b show the estimated coefficients obtained from equations (S5.3) and (S5.4), but now using the data about statistical significance and the effect size from each set of materials obtained by using the dataset generated by meta-analyzing the Main Studies and the Replication Studies. In line with Tables S5.8 and S5.9 and Tables S5.13a and S5.13b, first we included one moderator at a time (columns 1 to 7; moderators specified in the first column), then we run a comprehensive model (column 8).

Table S5.17a: Forecaster characteristics and accuracy of predictions about statistical significance for the meta-analysis

	Dependent Variable: Absolute Prediction Error – Statistical Significance							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Treatment	-0.009	-0.009	-0.011	-0.008	-0.013	-0.008	-0.010	-0.015
	(0.009)	(0.009)	(0.010)	(0.009)	(0.009)	(0.009)	(0.009)	(0.012)
Gender	-0.011							-0.010
	(0.013)							(0.012)
Age		0.0004						0.001
		(0.001)						(0.001)
Research Assistant			-0.024					-0.034
			(0.037)					(0.033)
Lab Manager			-0.019*					-0.016
			(0.023)					(0.027)
Master St.			-0.096***					-0.114***
			(0.016)					(0.020)
Doctoral St.			-0.086***					-0.109***
			(0.018)					(0.023)
Post Doc			-0.086***					-0.110***
			(0.019)					(0.023)
Lecturer NTT			-0.073**					-0.104**
			(0.023)					(0.032)
Assistant Prof TT			-0.088***					-0.121***
			(0.019)					(0.025)
Associate Prof Unt			-0.084***					-0.109**
			(0.031)					(0.039)
Associate Prof Ten			-0.076***					-0.115***
			(0.020)					(0.027)
Full Prof			-0.099**					-0.140**
			(0.033)					(0.046)
Other			-0.085***					-0.130***
			(0.018)					(0.031)
Publications				-0.0001				-0.0001
				(0.0002)				(0.0004)
Confidence					-0.011*			-0.014*
					(0.005)			(0.006)
Familiarity						0.0003		0.003
						(0.002)		(0.002)
Proficiency stat							-0.003	0.003
							(0.004)	(0.004)
Obs.	9,024	8,896	9,024	8,896	9,024	9,024	8,960	8,768
R^2	0.024	0.023	0.026	0.024	0.030	0.023	0.024	0.035
F Statistic	10.42***	9.936***	7.596***	10.41***	13.27***	10.3***	10.47***	8.472***

Notes. *$p < 0.05$; **$p < 0.01$; ***$p < 0.001$. Standard errors clustered at individual and team x hypothesis levels. All the regressions control for Team and Hypothesis fixed effects. Omitted category for job rank: 'undergraduate research assistant'

Table S5.17b: Forecaster characteristics and accuracy of predictions about effect sizes for the meta-analysis

	Dependent Variable: Absolute Prediction Error - Effect Size							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Treatment	-0.008	-0.006	-0.001	-0.007	-0.004	-0.008	-0.007	-0.0004
	(0.012)	(0.012)	(0.013)	(0.012)	(0.012)	(0.012)	(0.013)	(0.013)
Gender	0.003							0.005
	(0.011)							(0.012)
Age		-0.001						0.001
		(0.001)						(0.001)
Research Assistant			-0.022*					-0.036
			(0.010)					(0.021)
Lab Manager			0.011					0.032
			(0.021)					(0.021)
Master St.			0.078					0.116*
			(0.054)					(0.056)
Doctoral St.			0.007					0.046*
			(0.015)					(0.020)
Post Doc			0.015					0.045*
			(0.018)					(0.021)
Lecturer NTT			-0.019					0.022
			(0.012)					(0.016)
Assistant Prof TT			-0.020**					0.015
			(0.012)					(0.015)
Associate Prof Unt			0.058***					0.089***
			(0.014)					(0.012)
Associate Prof Ten			-0.008					0.025
			(0.012)					(0.017)
Full Prof			-0.039**					-0.013
			(0.012)					(0.023)
Other			-0.0003					0.037
			(0.015)					(0.019)
Publications				-0.0004				-0.0001
				(0.0003)				(0.0005)
Confidence					0.015***			0.017***
					(0.003)			(0.005)
Familiarity						-0.0003		-0.004
						(0.003)		(0.003)
Proficiency stat							0.002	-0.005
							(0.004)	(0.005)
Obs.	9,024	8,896	9,024	8,896	9,024	9,024	8,960	8,768
R^2	0.151	0.149	0.153	0.150	0.155	0.151	0.150	0.155
F Statistic	75.92***	73.86***	52.39***	74.26***	78.64***	75.92***	74.98***	43.28***

Notes. *$p < 0.05$; **$p < 0.01$; ***$p < 0.001$. Standard errors clustered at individual and team x hypothesis levels. All the regressions control for Team and Hypothesis fixed effects. Omitted category for job rank: 'Undergraduate Research Assistant.'

Multivariate regression results for the Main Studies. In Table S5.18, we report the estimates of regressions (S5.1) and (S5.2) through a multivariate regression approach. This technique allows to jointly estimate the regressions with the same independent variables (monetary treatment dummy and fixed effects) but different dependent variables (absolute prediction error of the forecasts on significance and on effect size, respectively), and to take into account that the forecasts regarding significance levels and effect sizes might be correlated. As expected, the coefficients estimated jointly are consistent with those estimated independently (refer to Table S5.7), but the standard errors are lower.

Table S5.18: Monetary incentives and forecasting accuracy – Multivariate regressions

	Dependent Variable: Absolute Predicted Error	
	Significance	Effect Size
	(1)	(2)
Treatment	-0.011	-0.009
	(0.006)	(0.009)
Team FE	Yes	Yes
Hypothesis FE	Yes	Yes
Observations	9,024	9,024
R^2	0.030	0.221
F Statistic	13.772***	127.853***

Notes. *$p < 0.05$; **$p < 0.01$; ***$p < 0.001$. Standard errors clustered at individual and team x hypothesis level.

SUPPLEMENT 6 - Main Studies and Replication Studies analyses using high quality materials

As noted in the main text, we repeated all of our analyses, excluding 18 sets of materials that were rated as below 5 on a scale of 0 (not at all informative) to 10 (extremely informative) by independent raters in the Forecasting Study. The excluded materials sets consisted of the Hypothesis 1 materials designed by Teams 1, 4, 5, 6, 7, and 10, Hypothesis 2 materials designed by Team 7, Hypothesis 3 materials designed by Teams 4, 7, 9, and 12, Hypothesis 4 materials designed by Teams 6 and 7, and Hypothesis 5 materials designed by Teams 1, 2, 5, 6, and 11. We report the results of these more restrictive analyses below.

Null hypothesis significance tests. All five original materials sets were rated as "high-quality" (i.e., above the scale midpoint of 5), therefore removing the materials rated as lower in quality does not change the conclusions reported in the main text that the direct replications of the original findings were overall successful.

As seen in Table S6.1, the results of these analyses in the Main Studies are similar to those reported in the main text, with Hypotheses 2 and 3 receiving fairly consistent support, Hypothesis 4's results being more variable, and Hypothesis 5 receiving directional, but non-statistically-significant, support. Hypothesis 1 shows a somewhat different pattern, however, with most high-quality materials statistically significantly supporting the original finding, and only two showing the opposite result. Thus, the roughly even split between consistent and inconsistent results for Hypothesis 1 for the Main Studies when all study designs are included in the analyses may have been due, in part, to relatively lower quality materials producing results inconsistent with the original finding. However, results in the Replication Studies were very much in line with the results reported in the main text: Hypotheses 2 and 3 received consistent

support, Hypothesis 1 was split between consistent and inconsistent results, and Hypotheses 4 and 5 produced fairly variable results. Overall, excluding comparatively lower quality sets of materials does not much change the results of the null hypothesis significance tests, except perhaps for Hypothesis 1 and only in the Main Studies (not the Replication Studies).

Table S6.1. Summary of null hypothesis significance tests, high quality materials only.

Main Studies				
Hypothesis	**Consistent Results, $p < .05$**	**Consistent Results, $p > .05$**	**Inconsistent Results, $p > .05$**	**Inconsistent Results, $p < .05$**
1	71% (5)	0% (0)	0% (0)	29% (2)
2	92% (11)	8% (1)	0% (0)	0% (0)
3	89% (8)	0% (0)	0% (0)	11% (1)
4	50% (5)	20% (2)	20% (2)	10% (1)
5	25% (2)	50% (4)	25% (2)	0% (0)
Replication Studies				
Hypothesis	**Consistent Results, $p < .05$**	**Consistent Results, $p > .05$**	**Inconsistent Results, $p > .05$**	**Inconsistent Results, $p < .05$**
1	43% (3)	14% (1)	14% (1)	29% (2)
2	75% (9)	17% (2)	8% (1)	0% (0)
3	56% (5)	33% (3)	11% (1)	0% (0)
4	30% (3)	50% (5)	10% (1)	10% (1)

5	25% (2)	38% (3)	25% (2)	12% (1)

Meta-analytic statistics. We also repeated the meta-analytic analyses reported in the main text, excluding the materials sets rated as lower in quality. The results of these meta-analyses are generally consistent with those reported in the main text. In the Main Studies, they supported Hypotheses 2 and 3 (estimated mean effect sizes d = 1.11, 95% *CI* [0.66, 1.55], p < .001; d = 0.41, 95% *CI* [0.19, 0.63], p < .001), and did not support Hypotheses 1 and 4 (d = 0.14, 95% *CI* [-0.18, 0.46], p = .394; d = 0.10, 95% *CI* [-0.05, 0.24], p = .192). However, the estimated mean effect size for Hypothesis 5 was not statistically significant, r = .04, 95% *CI* [-.00, .08], p = .055, whereas, in the full analyses reported in the main text, it was statistically significant, though the estimate was small. In the Replication Studies, the analyses excluding the lower quality materials generally agreed with the full analyses, with Hypotheses 2 and 3 being supported (d = 0.64, 95% *CI* [0.33, 0.93], p < .001; d = 0.32, 95% *CI* [0.15, 0.49], p < .001), and Hypotheses 1, 4, and 5 not (d = -0.09, 95% *CI* [-0.44, 0.27], p = .634; d = 0.04, 95% *CI* [-0.07, 0.15], p = .452; r = .01, 95% *CI* [-.06, .07], ps > .801. Figures S6.1a-S6.1e present forest plots of these meta-analyses.

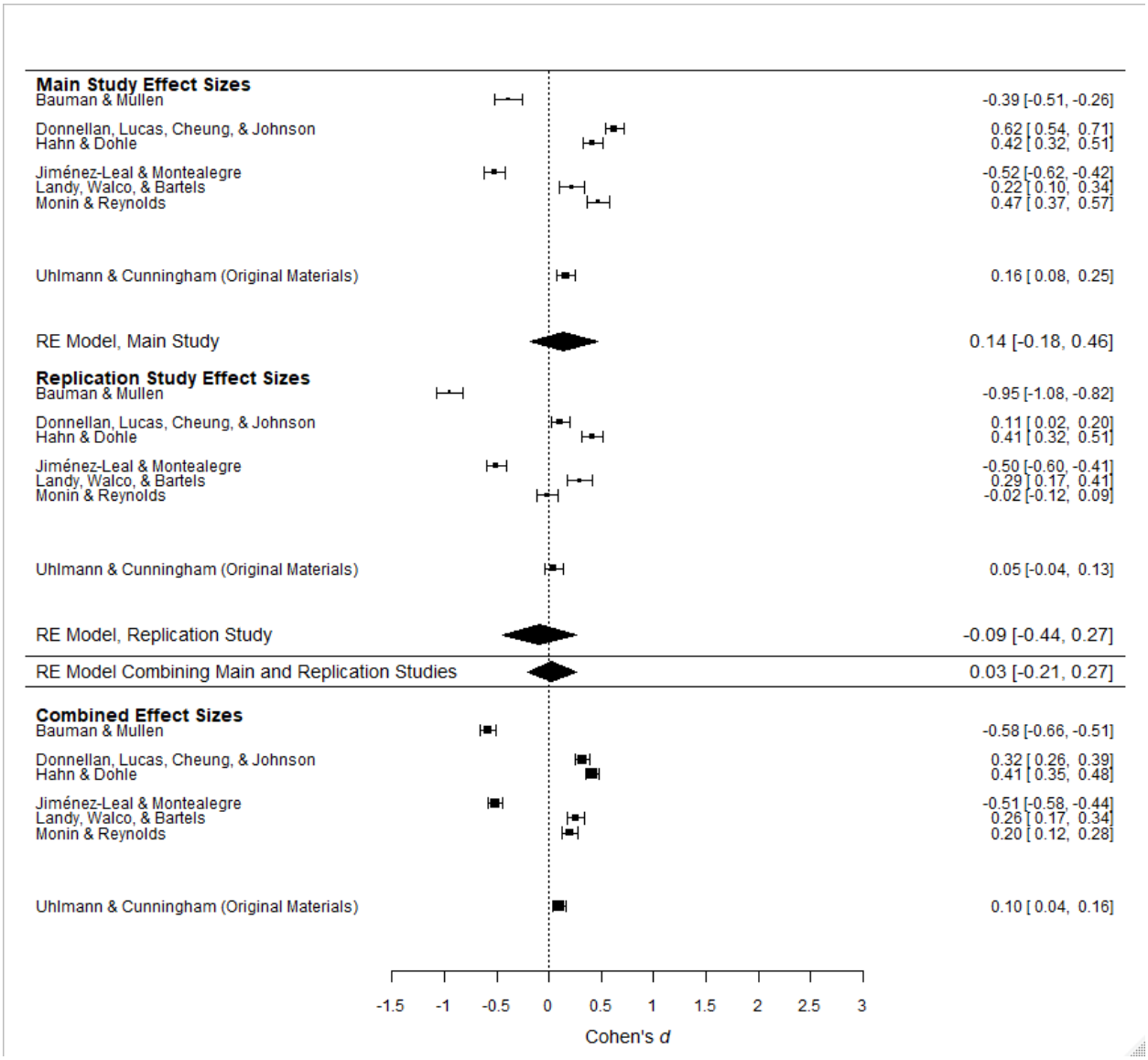


Figure S6.1a. Forest plot of observed effect sizes (independent-groups Cohen's *d*s) for Hypothesis 1, high quality materials only.

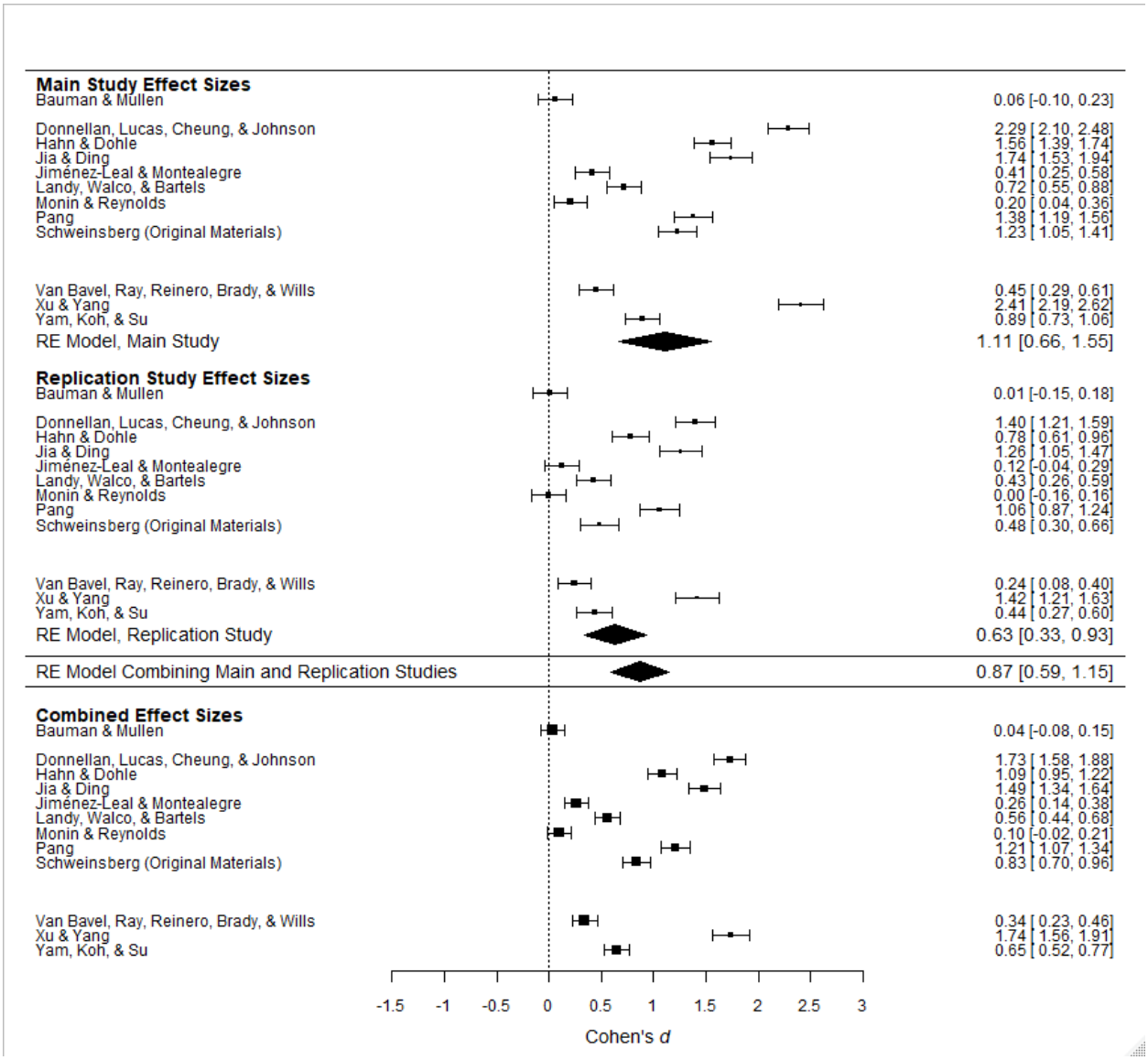


Figure S6.1b. Forest plot of observed effect sizes (independent-groups Cohen's *d*s) for Hypothesis 2, high quality materials only.

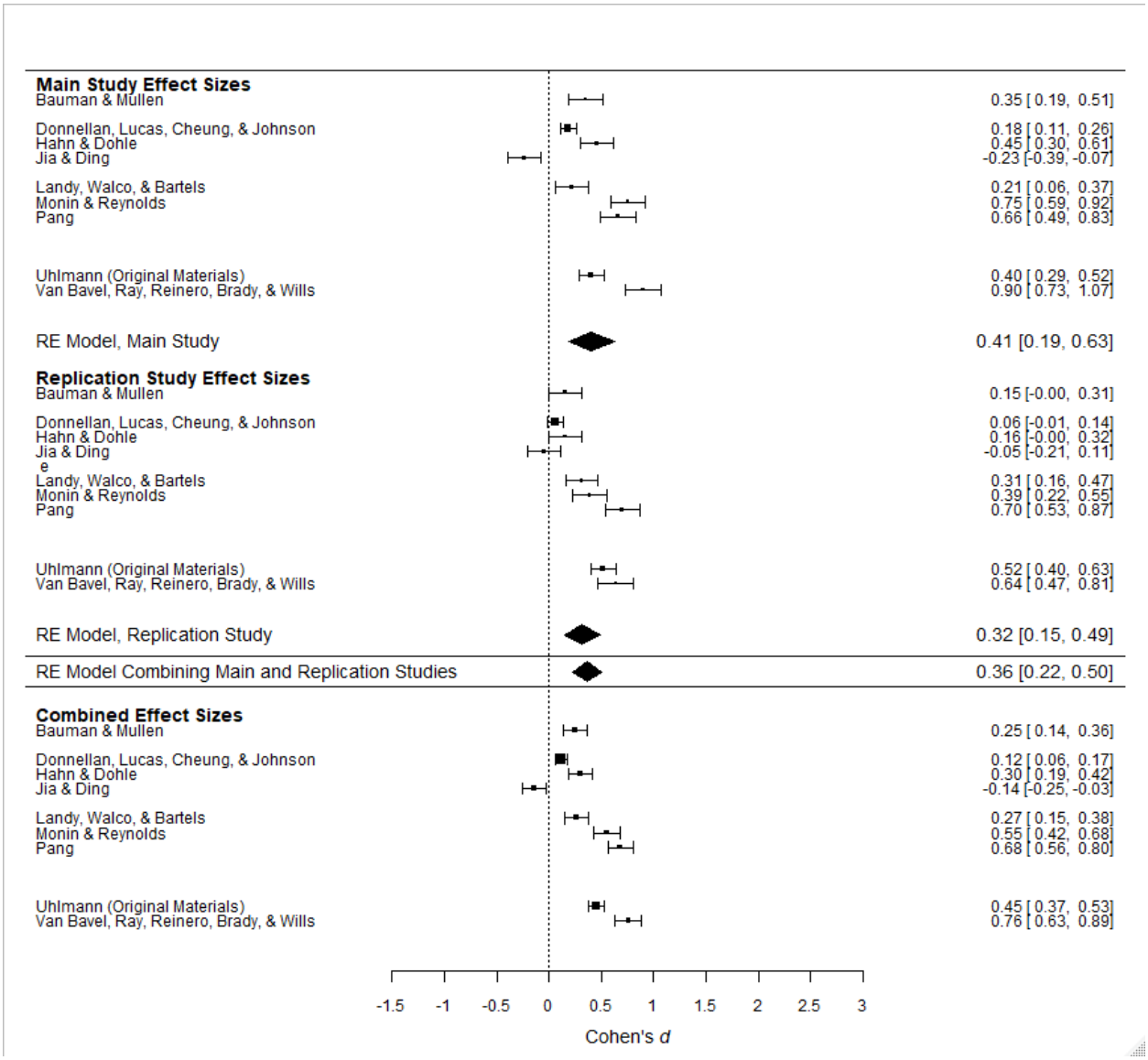


Figure S6.1c. Forest plot of observed effect sizes (independent-groups Cohen's *d*s) for Hypothesis 3, high quality materials only.

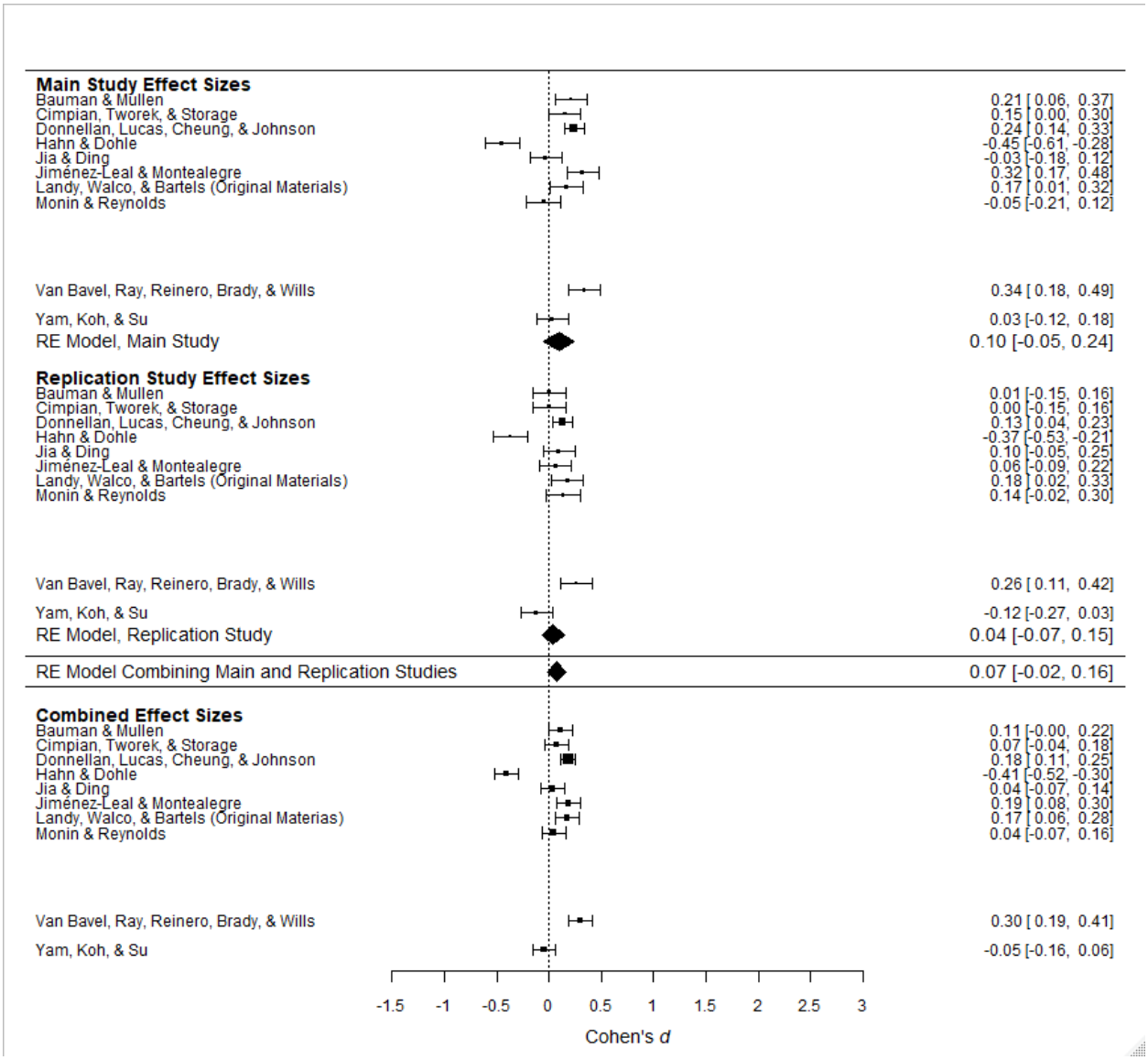


Figure S6.1d. Forest plot of observed effect sizes (independent-groups Cohen's *d*s) for Hypothesis 4, high quality materials only.

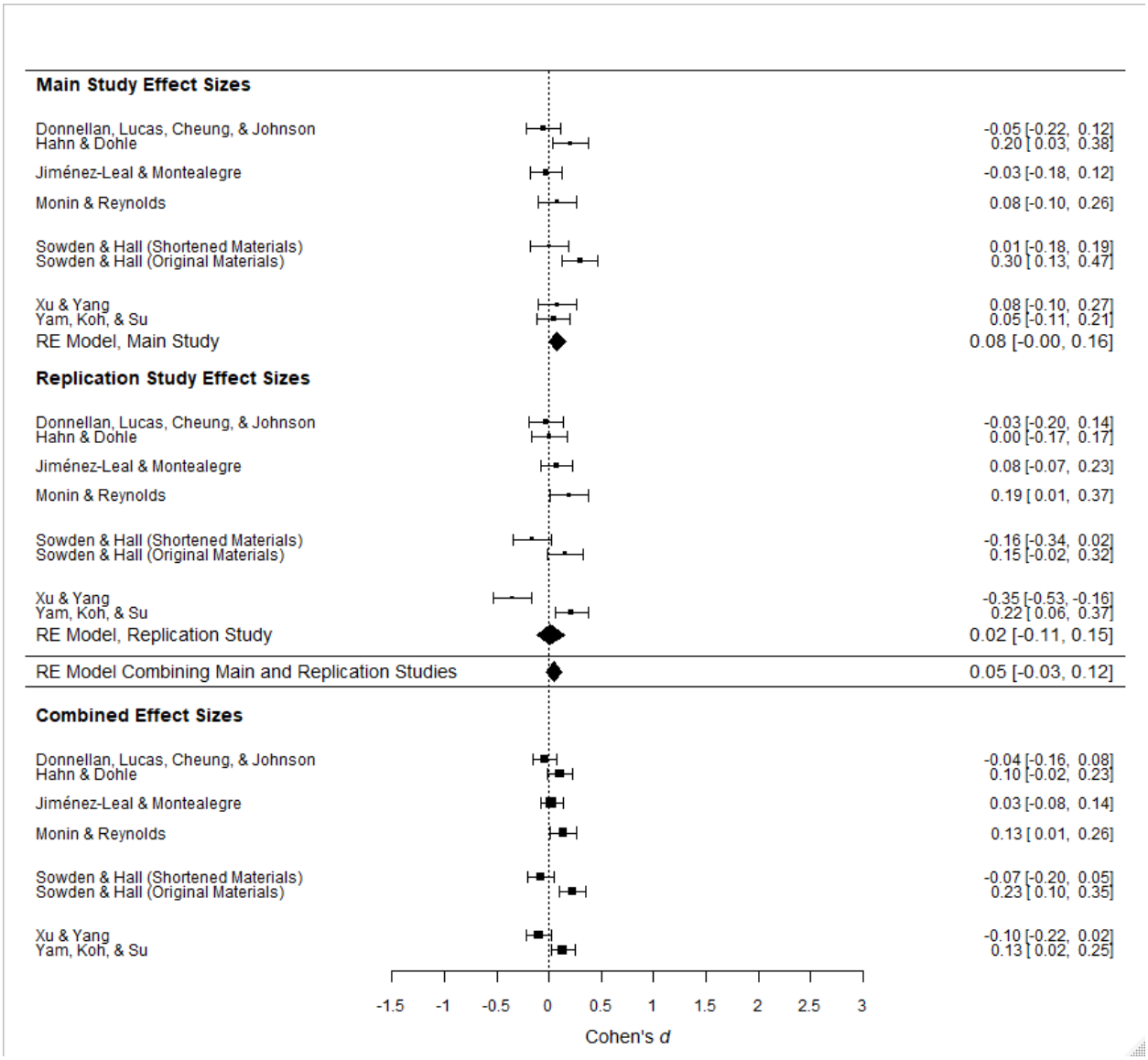


Figure S6.1e. Forest plot of observed effect sizes (converted to Cohen's *d*s) for Hypothesis 5, high quality materials only.

As in the full analyses, Hypotheses 1-4 showed significant and high levels of heterogeneity in the Main Studies, though Hypothesis 5 showed descriptively lower, and non-significant, levels of heterogeneity (see Table S6.2). In the Main Studies, only about 1%, 2%, 4%, 11%, and 51% of the variance across the effect sizes for Hypotheses 1, 2, 3, 4, and 5,

respectively, can be accounted for by chance, when lower quality materials are not included in the analyses. Similarly, in the Replication Studies, only about 1%, 3%, 7%, 18%, and 21% of the variance across the effect sizes for Hypotheses 1, 2, 3, 4, and 5, respectively, was likely due to chance variability, when lower quality materials are not included in the analyses. As in the full analyses, unexplained heterogeneity represented a vast majority of the observed variance across effect sizes.

Table S6.2. *Q* I^2, and τ^2 statistics from meta-analyses of high-quality materials, Main Studies and Replication Studies.

Main Studies						
Hypothesis	**Description**	***k***	**Effect Size [95% *CI*]**	***Q***	***I*^{2} [95% *CI*]**	**τ^2 [95% *CI*]**
1	Awareness of automatic prejudice	7	*d* = 0.14 [-0.18, 0.46]	*Q*(6) = 462.43***	89.06% [76.75, 96.76]	0.19 [0.08, 0.92]
2	Extreme offers reduce trust	12	*d* = 1.11 [0.66, 1.55]	*Q*(11) = 570.53***	98.42% [96.84, 99.46]	0.60 [0.30, 1.77]
3	Moral praise for needless work	9	*d* = 0.41 [0.19, 0.63]	*Q*(8) = 136.65***	95.51% [90.01, 98.80]	0.11 [0.05, 0.42]
4	Proximal authorities drive legitimacy of performance enhancers	10	*d* = 0.10 [-.05, 0.24]	*Q*(9) = 77.65***	89.06% [76.75, 96.76]	0.05 [0.02, 0.17]
5	Deontological judgments predict happiness	8	*r* = .04 [-0.00, 0.08]	*Q*(7) = 13.53 *ns*	49.21% [0.00, 87.82]	0.04 [0.00, 0.11]
Replication Studies						
Hypothesis	**Description**	***k***	**Effect Size [95% *CI*]**	***Q***	***I*^{2} [95% *CI*]**	**τ^2 [95% *CI*]**
1	Awareness of automatic prejudice	7	*d* = -0.09 [-0.44, 0.27]	*Q*(6) = 413.12***	98.85% [97.23, 99.77]	0.22 [0.09, 1.10]
2	Extreme offers reduce trust	12	*d* = 0.63 [0.33, 0.93]	*Q*(11) = 354.09***	97.13% [94.28, 99.01]	0.27 [0.13, 0.80]
3	Moral praise for needless work	9	*d* = 0.32 [0.15, 0.49]	*Q*(8) = 110.20***	92.61% [83.55, 98.01]	0.06 [0.03, 0.25]
4	Proximal authorities drive legitimacy of performance enhancers	10	*d* = 0.04 [-0.07, 0.15]	*Q*(9) = 46.55***	82.07% [61.58, 94.79]	0.03 [0.01, 0.10]
5	Deontological judgments predict happiness	8	*r* = 0.01, [-0.06, 0.07]	*Q*(7) = 32.81***	79.00% [51.61, 95.08]	0.01 [0.00, 0.04]

Note. ***$p < .001$.

Predicting effect sizes. We computed intraclass correlation coefficients predicting observed effect sizes from hypothesis and team, converting the Pearson *r*s from Hypothesis 5 to Cohen's *d*s, but excluding the lower quality materials. The results of these analyses were very consistent with the full analyses reported in the main text: hypothesis predicted a moderate amount of variance in both the Main Studies, *ICC* = .47, 95% CI [.17, .89] and the Replication Studies, *ICC* = .37, 95% *CI* [.10, .85], while team did not explain a significant amount of variance in either the Main Studies, *ICC* = -.12, 95% *CI* [-.33, .23], or the Replication Studies, *ICC* = -.06, 95% *CI* [-.29, .30]. Meta-regression also agreed with the full analyses: Hypothesis 2 produced larger effect sizes than the median hypothesis, Main Studies β = 1.114, 95% *CI* [0.584, 1.644], $p < .001$, Replication Studies β = 0.537, 95% *CI* [0.189, 0.884], $p = .003$, but no team produced significantly larger or smaller effect sizes than the median team in either study, *p*s > .222. Moreover, after accounting for both hypothesis and team, there was still significant residual heterogeneity, Main Studies $Q(26) = 870.51$, $p < .001$, $I^2 = 97.77\%$, 95% *CI* [96.40, 98.83], Replication Studies $Q(26) = 574.92$, $p < .001$, $I^2 = 96.10\%$, 95% *CI* [93.71, 97.97].

Aggregating results of the Main Studies and Replication Studies. Aggregating all of the effect sizes across the two studies, again excluding the lower quality materials, produced similar results to the meta-analyses above. Hypotheses 2 and 3 were strongly supported ($d = 0.87$, 95% *CI* [0.55, 1.08], $p < .001$; $d = 0.36$, 95% *CI* [0.23, 0.50], $p < .001$), Hypothesis 5 was not supported by a statistically significant directional effect ($r = .02$, 95% *CI* [-.01, .06], $p = .206$), unlike in the full analyses, and Hypotheses 1 and 4 were also not supported ($d = 0.03$, 95% *CI* [-0.21, 0.27], $p = .821$; $d = 0.07$, 95% *CI* [-0.02, 0.16], $p = .132$), consistent with the full analyses. As in the full analyses reported in the main text, the estimate for Hypothesis 5 was

close to zero, indicating a lack of overall empirical support for the predicted relationship between moral judgments and happiness.

Comparing the results of the Main Studies and Replication Studies. In 38 out of 46 cases, the Replication Studies' effect was directionally consistent with the effect size from the Main Studies. In 26 of those 38 cases, when new participants were run using the same study design, significant results were again significant again in the same direction, and non-significant effects were again non-significant. Breaking this down further, 11 of 35 significant findings from high quality materials in the Main Studies were not significant in the Replication Studies, and 3 of 11 non-significant findings from the Main Studies were statistically significant in the Replication Studies. Replication Studies' effect sizes were significantly smaller than the corresponding effect in the Main Studies, according to *z*-tests, in 18 out of 46 cases, and were not significantly greater in any cases, with no significant difference in 28 out of 46 cases. This generally agrees with the full analyses reported in the main text.

Lastly, the correlation between the Main Studies' and Replication Studies' effect sizes was again very substantial, $r(44) = .91$, 95% *CI* [.84, .95], $p < .001$. On the whole, the results of the supplemental analyses excluding the materials rated as lower in quality do not differ appreciably from the analyses including all of the materials.

SUPPLEMENT 7 - Bayesian analysis of project results

Abstract

This is the methods and results section for the Bayesian analysis of the "Crowdsourcing hypotheses tests" data set. The methods section follows on the preregistration document that can also be found at https://osf.io/9jzy4/.

Methods

The "Crowdsourcing hypotheses tests" project studied five empirical phenomena (i.e., $q = 1,2,\dots,5$), each of which was subject to replication attempts from the same set of $l = 1,2,\dots,13$ research teams. Each team, i.e., laboratory, l replicated each of the five phenomena twice: once in an MTurk population, and once in a PureProfile population. The following questions are of interest:

1. For each question q and across all of the replication attempts, what is the overall evidence for the presence of each of the five phenomena?
2. For each question q, what is the heterogeneity among the labs in the effect size estimates?
3. Over all questions q simultaneously, are some labs better than other labs in consistently producing large effect sizes?

Below we will deal with each of these questions in turn. In order to address the first two questions we apply a Bayesian model-averaging meta-analysis procedure (BAMAMA; e.g., Gronau, van Erp, et al., 2017; Scheibehenne, Gronau, Jamil, & Wagenmakers, 2017), separately for each of the five phenomena. In order to address the final question on "lab flair" we use an ANOVA model to take into account all phenomena simultaneously.

The Meta-Analytic Model

Below we outline the planned BAMAMA procedure for a specific phenomenon; the procedure will be carried out for each of the five phenomena separately. In our analysis for a specific phenomenon q, we assume that each team l has their own latent grand mean effect size, $\delta_{l,q}$. We also assume that there is a fixed effect $\delta_{pop,q}$ that quantifies the difference in effect size between the MTurk population and the PureProfile population. For a specific team l, the MTurk effect size is given by $\delta_{l,q} - \frac{1}{2}\delta_{pop,q}$ and the PureProfile effect size is given by $\delta_{l,q} + \frac{1}{2}\delta_{pop,q}$. Thus, $\delta_{pop,q}$ is the same for every team l.

Each team's latent grand mean effect size $\delta_{l,q}$ is assumed to be governed by a latent normal distribution with group mean μ_q μ_q and group heterogeneity (standard deviation) τ_q. The above parameters are not directly observed. We assume that the observed effect size $d_{1,l,q}$ (for the MTurk population) and $d_{2,l,q}$ (for the PureProfile population) are drawn from a normal distribution with mean equal to the latent true effect size and standard deviation equal to the standard error of the observed effect size. That is, the setup is as follows:

$$\delta_{l,q} \sim Normal(\mu_q, \tau_q^2) \quad (1)$$

$$d_{1,l,q} \sim Normal\left(\delta_{l,q} - \frac{1}{2}\delta_{pop,q}, SE_{1,l,q}^2\right) \quad (2)$$

$$d_{2,l,q} \sim Normal\left(\delta_{l,q} + \frac{1}{2}\delta_{pop,q}, SE_{2,l,q}^2\right) \quad (3)$$

where $d_{p,l,q}$ denotes the observed effect size of the lth team, the pth population, and the qth question, and $SE_{p,l,q}$ denotes the corresponding standard error; $p = 1$ corresponds to the MTurk population and $p = 2$ corresponds to the PureProfile population. For each question q, this leaves three main parameters:

1. Parameter μ_q quantifies the group-level mean effect size. If $\mu_q = 0$, the phenomenon at hand is absent on the group level, considered across all teams.
2. Parameter τ_q quantifies the heterogeneity across the teams. If $\tau_q = 0$, the teams have the same effect size.
3. Parameter $\delta_{pop,q}$quantifies the impact of "population", that is, the difference in effect size between the MTurk population and the PureProfile population. If $\delta_{pop,q} = 0$, the two populations have the same effect size.

Step 1: Estimation Using the Full Model

In a first step, we will explore the model parameters by estimating the full model, that is, a model in which the three key parameters μ_q, τ_q, and $\delta_{pop,q}$ are assigned smooth prior distributions and no prior plausibility is assigned to the special cases where $\mu_q = 0$, $\tau_q = 0$, or $\delta_{pop,q} = 0$. For this estimation approach we use the following priors: $\mu_q \sim Cauchy(0, \frac{1}{\sqrt{2}})$, $\tau_q \sim InvGamma(1, 0.15)$ (i.e., the primary prior for τ_q used in Gronau, van Erp, et al., 2017, based on empirical work reported in van Erp, Verhagen, Grasman, & Wagenmakers, 2017), and $\delta_{pop,q} \sim Normal(0, 0.5^2)$. The purpose of this first analysis is to get an indication of the size of the effects in case the effects are assumed to exist. The resulting posterior distributions will be plotted together with the priors, so that it is clear to what extent the data caused an update of the priors.

Step 2: Model Averaging

In BAMAMA we take seriously the hypothesis that either $\mu_q = 0$, $\tau_q = 0$, or $\delta_{pop,q} = 0$. Specifically, for each question q, we will assess the predictive adequacy of the following eight models:

$$H_1: \mu_q = 0, \tau_q = 0, \delta_{pop,q} = 0 \quad (4)$$

$$H_2: \mu_q = 0, \tau_q = 0, \delta_{pop,q} \sim Normal(0, 0.15^2),$$

$$H_3: \mu_q = 0, \tau_q \sim InvGamma(1, 0.15), \delta_{pop,q} = 0,$$

$$H_4: \mu_q = 0, \tau_q \sim InvGamma(1, 0.15), \delta_{pop,q} \sim Normal(0, 0.15^2),$$

$$H_5: \mu_q \sim t(0.35, 0.102, 3)I(0, \infty), \tau_q = 0, \delta_{pop,q} = 0,$$

$$H_6: \mu_q \sim t(0.35, 0.102, 3)I(0, \infty), \tau_q = 0, \delta_{pop,q} \sim Normal(0, 0.15^2),$$

$$H_7: \mu_q \sim t(0.35, 0.102, 3)I(0, \infty), \tau_q \sim InvGamma(1, 0.15), \delta_{pop,q} = 0,$$

$$H_8: \mu_q \sim t(0.35, 0.102, 3)I(0, \infty), \tau_q \sim InvGamma(1, 0.15), \delta_{pop,q} \sim Normal(0, 0.15^2).$$

In these models, μ_q is assigned the informative "Oosterwijk prior" (Gronau, Ly, & Wagenmakers, 2017), a shifted and scaled *t* distribution with location 0.35, scale 0.102, and three degrees of freedom, truncated to have mass only on positive effect sizes (i.e., $I(0, \infty)$; hence, this analysis assumes that the original experiments for the to-be-replicated effects reported a positive effect size). In our opinion, the Oosterwijk prior provides a reasonable specification for effects that are known to be of small-to medium size.

Parameter τ_q is assigned the same prior that was used for estimation, that is, an $InvGamma(1, 0.15)$ distribution (Gronau, van Erp, et al., 2017; van Erp et al., 2017). Finally, parameter $\delta_{pop,q}$ is assigned a normal prior with mean 0 and standard deviation 0.15, reflecting

the fact that we do not know the direction of the effect, but that the difference between the two populations, if present, is likely to be relatively small.

The eight models are assigned equal prior probability, such that $P(H_j) = 1/8 = 0.125$, $j = 1,2, \ldots ,8$. In this setup, it is a priori equally likely that each of the three parameters is present or absent.

Goal 1: Overall Evidence

For each question q separately, we will report the posterior model probability for all eight models. Of key interest with respect to the first goal is the summed posterior probability for models H_5, H_6, H_7, and H_8 (i.e., all models where $\mu_q \neq 0$); this posterior probability may be contrasted with its complement, that is, the summed posterior probability for models H_1, H_2, H_3, and H_4 (i.e., all models where $\mu_q = 0$). Dividing these two probabilities yields the posterior model odds; in this specific case, the prior odds is 1 (the summed prior probability for the models with $\mu_q \neq 0$ is 0.5), and therefore this posterior odds also equals the Bayes factor in favor of there being an effect $\mu_q \neq 0$ over there not being an effect $\mu_q = 0$, that is, the degree to which the data necessitate an update of our prior opinion.

Of secondary interest are the posterior distributions for μ_q, particularly for the models where $\mu_q \neq 0$. We will present model-averaged posterior distributions for μ_q across all eight models (including a spike at zero, the height of which equals the summed posterior model probability across the four models where $\mu_q = 0$).

Goal 2: Quantifying Heterogeneity

For each question q separately, we compare the fixed effects models against the random effects models. In order to quantify heterogeneity we proceed, first, to assess the evidence for heterogeneity (i.e., the summed posterior model probabilities for H_3, H_4, H_7, and H_8, models for which $\tau_q \neq 0$) versus the evidence for homogeneity (i.e., the summed posterior model probabilities for H_1, H_2, H_5, and H_6, models for which $\tau_q = 0$). The ratio of these probabilities gives the posterior odds, which in this case is the same as the Bayes factor in favor of there being a random effect over a fixed effect. Secondly, we provide the model-averaged posterior distributions for τ_q across all eight models (including a spike at zero, the height of which equals the summed posterior model probability across the four models where $\tau_q = 0$).

Extra Goal: Quantifying the Effect of Population

For each question q separately, we assess whether there is a population effect. Similar to the analyses above, we can quantify the evidence for a population effect (i.e., MTurk versus PureProfile) by contrasting the summed posterior model probabilities for H_2, H_4, H_6, and H_8 (i.e., models for which $\delta_{pop,q} \neq 0$) versus the summed posterior model probabilities for H_1, H_3, H_5, and H_7 (i.e., models for which $\delta_{pop,q} = 0$). The ratio of these probabilities gives the posterior odds, which in this case is the same as the Bayes factor in favor of there being an effect of the data being collected from MTurk or PureProfile. Secondly, we provide the model-averaged posterior distributions for $\delta_{pop,q}$ across all eight models (including a spike at zero, the height of which equals the summed posterior model probability across the four models where $\delta_{pop,q} = 0$).

BAMAMA Methodology

In order to execute the proposed analyses, we will rely on R (R Core Team, 2018) and implement all models using the rstan (Stan Development Team, 2018) package. To compute the posterior model probabilities, we will apply bridge sampling (Gronau, Sarafoglou, et al., 2017; Meng & Wong, 1996) as implemented in the bridgesampling package (Gronau, Singmann, & Wagenmakers, 2017).

Goal 3: Quantifying Effects of Lab Using ANOVA

To test the effect of laboratory we use a Bayesian ANOVA, where the observed effect sizes $d_{p,l,q}$, and the corresponding standard errors $SE_{p,l,q}$ are viewed as repeated measurements of the labs across the two populations. Hence, laboratory membership $l = 1, 2, \dots, 13$ is taken to be a random factor, the indicator that states from which population $p = 1, 2$ the measurements came from (i.e., MTurk or PureProfile) is viewed as a fixed factor, and the question indicator $q = 1, 2, \dots, 5$ is also a fixed factor. For added flexibility the interaction term between the populations and the questions is also included. As the goal is to infer whether the labs perform differently, the fixed factors population and questions, as well as the interaction, are entered in the null model M_0, while the alternative model M_1 is an extension of the null that also includes the random factor lab membership. The null model implies that the labs perform similarly, while the alternative model implies that their performances differ. The Bayes factor in favor of differential lab performance over the null is calculated using JASP (JASP Team, 2018; Wagenmakers, Love, et al., 2018; Wagenmakers, Marsman, et al., 2018), which makes use of the BayesFactor (Morey & Rouder, 2015) R package. In a secondary analysis, we provide plots of

the posterior distributions for each lab's latent effect size δ_l δ_l, that is, the latent average lab performance across the questions q and populations.

Adapting the Meta-Analytic Model for Use in a Repeated Measures ANOVA

The statistical difficulty stems from the fact that each observed effect size is normally distributed with a different standard error, that is,

$$d_{p,l,q} \sim Normal\left(\delta_{p,l,q}, SE^2_{p,l,q} \right) \quad (5)$$

while a core assumption of the ANOVA is that each observation is drawn from a normal population with the same (unknown) standard error. To account for standard errors that differ across populations, labs, and questions, we transform the observed effect sizes to

$$t_{p,l,q} = \frac{d_{p,l,q} - \bar{d}_{overall}}{SE_{p,l,q}} \quad (6)$$

where $\bar{d}_{overall}$ is the overall mean observed effect size averaged over the two populations p, the thirteen labs l, and the five questions q. The subtraction of $\bar{d}_{overall}$ is required to take out any "intercept" effects caused by possible effects of p and q, or a possible grand mean of lab performance, while the rescaling is needed to put all observations on the same scale. The simulation study shows that the Bayes factors behave as expected. Specifically, the Bayes factor indicates evidence for the null, when the data are generated under the null. Likewise, the Bayes factor indicates evidence for the alternative, when the data are generated under the alternative with a between labs variability that is large enough.

Results for BAMAMA Q1: Awareness of Automatic Negative Associations

Q1: "People explicitly self-report an awareness of harboring negative automatic associations with members of negatively stereotyped social groups." Below are the results from the preregistered BAMAMA analyses.

Full-Model Estimation for Q1

Three parameters are of interest: the group-level mean effect size μ_1, the across team heterogeneity τ_1, and the difference $\delta_{pop,1}$ between the MTurk and the PureProfile populations.

First, we present the results of the *unfiltered* data. Figure S7-1 shows the prior and posterior distributions from the model with all three parameters free to vary. The top panel of Figure S7-1 suggests that there is no effect on the group-level mean effect size; the middle panel suggests that there is considerable across-team heterogeneity; the bottom panel suggests that the MTurk population has a slightly higher effect size than the PureProfile population.

Next, we present the results of the *filtered* data. Figure S7-2 shows the prior and posterior distributions from the model with all three parameters free to vary. The top panel of Figure S7-2 suggests that there is no effect on the group-level mean effect size; the middle panel suggests that there is considerable across-team heterogeneity; the bottom panel suggests that the MTurk population has a higher effect size than the PureProfile population.

In order to quantify the degree of support that the data provide for and against the presence of each of these effects we now turn to a BAMAMA analysis.

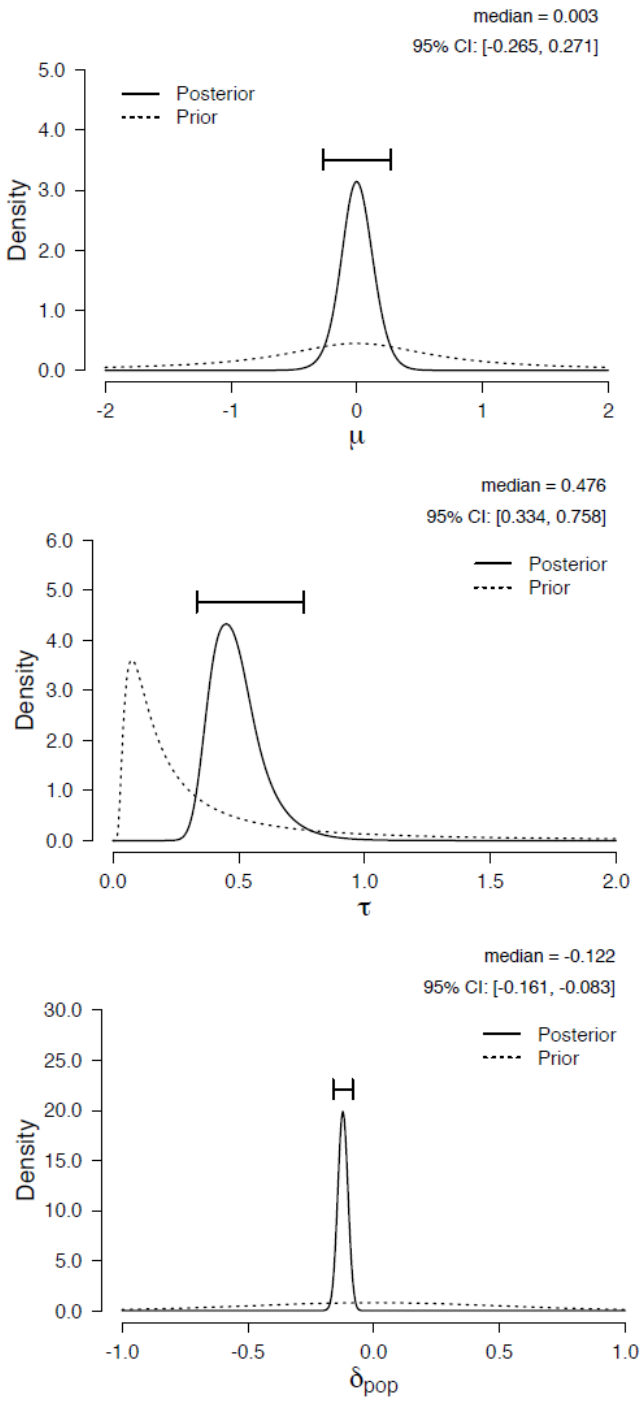


Figure S7-1. Estimation results for Q1 (*unfiltered* data). The upper panel displays the results for the group-level mean effect size μ_1, the middle panel displays the results for the across-team heterogeneity τ_q, and the lower panel displays the results for the difference $\delta_{pop,1}$ between the MTurk and the PureProfile populations. Each panel shows the prior and posterior distribution, the posterior median, and a 95% posterior credible interval.

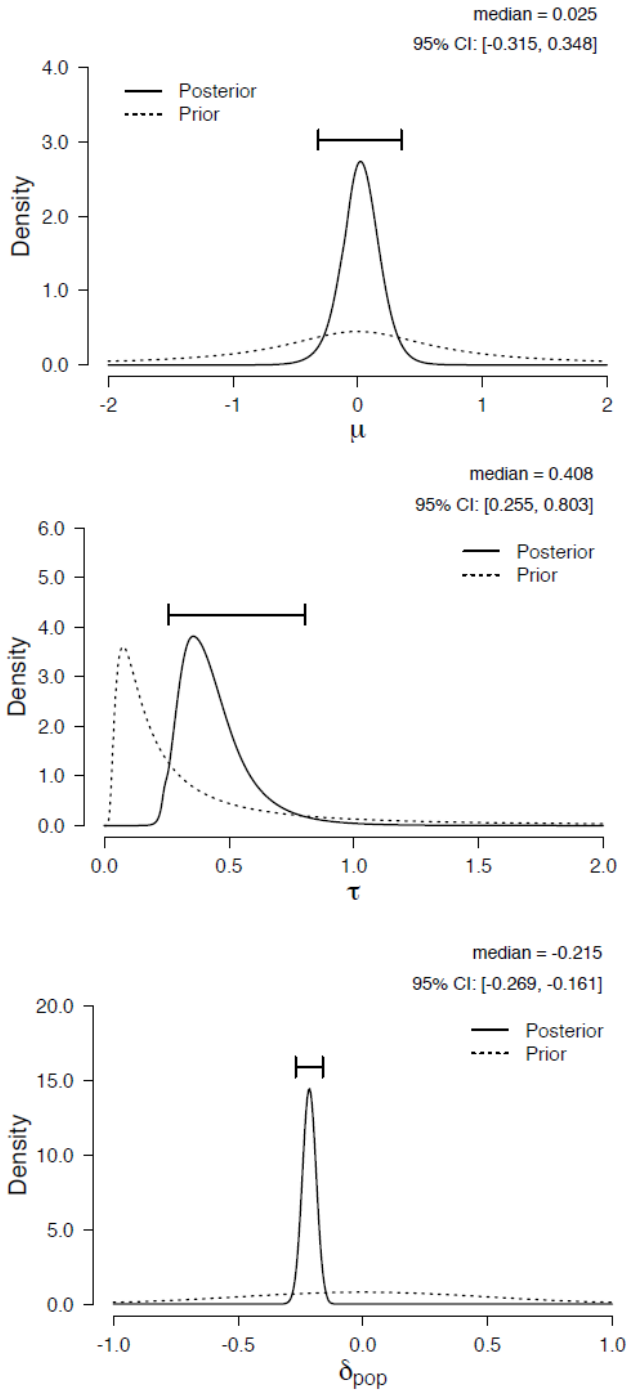


Figure S7-2. Estimation results for Q1 (*filtered* data). The upper panel displays the results for the group-level mean effect size μ_1, the middle panel displays the results for the across-team heterogeneity τ_1, and the lower panel displays the results for the difference $\delta_{pop,1}$ between the MTurk and the PureProfile populations. Each panel shows the prior and posterior distribution, the posterior median, and a 95% posterior credible interval.

Model Averaging for Q1

As outlined earlier, our model averaging approach considers eight models, constructed by the factorial combination of restrictions $\mu_1 = 0$, $\tau_1 = 0$, and $\delta_{pop,1} = 0$. Each model is assigned equal prior probability; hence, each restriction is a priori equally likely to hold. For each of the three restrictions, the inference is based on the evaluation of predictive performance for all eight models simultaneously. The first column of Table S7-1 presents the posterior model probabilities for Q1 based on the *unfiltered* data. The first column of Table S7-2 presents the posterior model probabilities for Q1 based on the *filtered* data.

Quantifying Overall Evidence for Q1. First we present the results of the *unfiltered* data. The Bayes factor and the posterior model odds both equal 8.615 in favor of the proposition that μ_1 equals 0. The summed posterior probability for the models in which $\mu_1 = 0$ equals 0.896. The top panel of Figure S7-3 shows the model averaged posterior distribution for μ_1 across all eight models, where the height of the spike at zero corresponds to the summed posterior probability that $\mu_1 = 0$. In sum, for Q1 the *unfiltered* data provide moderate evidence for the hypothesis that there is no effect on the group-level mean effect size.

Next we present the results for the *filtered* data. The Bayes factor and the posterior model odds both equal 4.916 in favor of the proposition that μ_1 equals 0. The summed posterior probability for the models in which $\mu_1 = 0$ equals 0.831. The top panel of Figure S7-4 shows the model-averaged posterior distribution for μ_1 across all eight models, where the height of the spike at zero corresponds to the summed posterior probability that $\mu_1 = 0$. In sum, for Q1 the *filtered* data provide moderate evidence for the hypothesis that there is no effect on the group-level mean effect size.

Quantifying Heterogeneity for Q1. First we present the results of the *unfiltered* data. The Bayes factor and the posterior model odds both equal 3.002×10^{14} in favor of the proposition that τ_1 does not equal 0. The summed posterior probability for the models in which $\tau_1 = 0$ equals 0.000. The middle panel of Figure S7-3 shows the model-averaged posterior distribution for τ_1 across all eight models, where the height of the spike at zero corresponds to the summed posterior probability that $\tau_1 = 0$. In sum, for Q1 the *unfiltered* data provide overwhelming evidence for the hypothesis that there is across-team heterogeneity.

Next, we present the results for the *filtered* data. The Bayes factor and the posterior model odds both equal ∞[1] in favor of the proposition that τ_1 does not equal 0. The summed posterior probability for the models in which $\tau_1 = 0$ equals 0.000. The middle panel of Figure S7-4 shows the model-averaged posterior distribution for τ_1 across all eight models, where the height of the spike at zero corresponds to the summed posterior probability that $\tau_1 = 0$. In sum, for Q1 the *filtered* data provide overwhelming evidence for the hypothesis that there is across-team heterogeneity.

[1] The true Bayes factor is so large that it exceeds the available numerical precision.

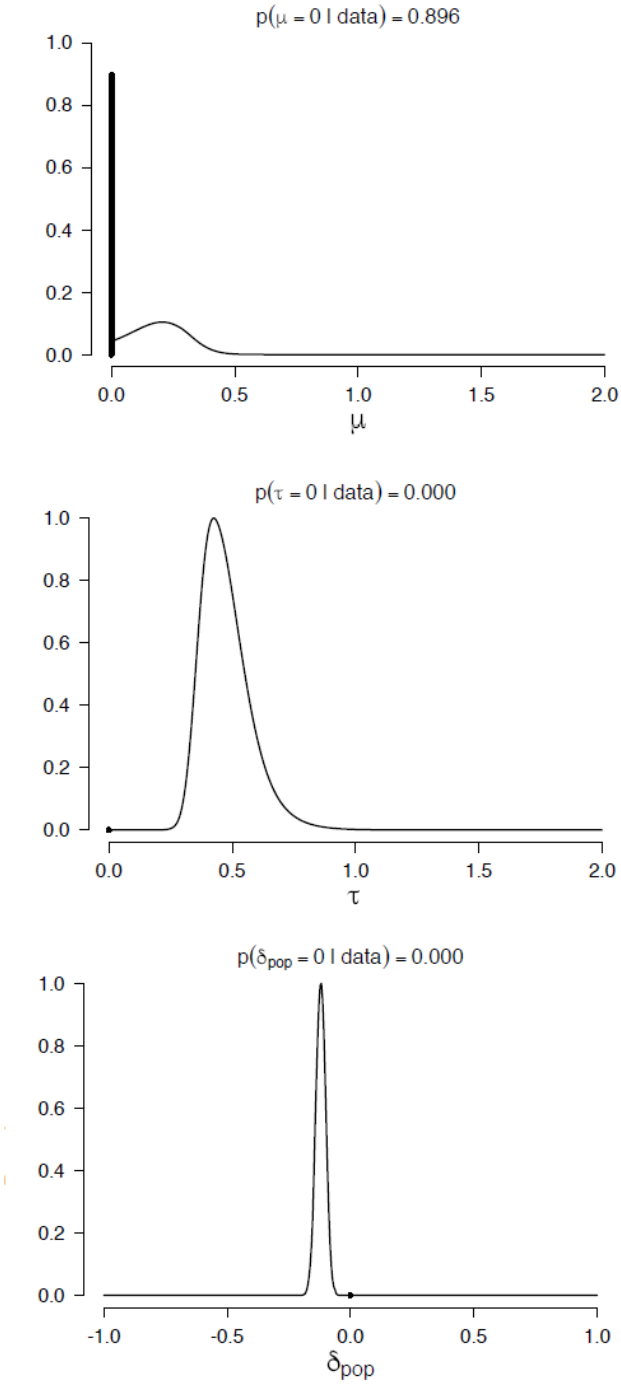


Figure S7-3. Model averaging results for Q1 (*unfiltered* data). The upper panel displays the results for the group-level mean effect size μ_1, the middle panel displays the results for the across-team heterogeneity τ_1, and the lower panel displays the results for the difference $\delta_{pop,1}$ between the MTurk and the PureProfile populations. Each panel shows the model-averaged posterior distribution for the parameter across all eight models, where the height of the spike at zero corresponds to the summed posterior probability that the parameter equals 0.

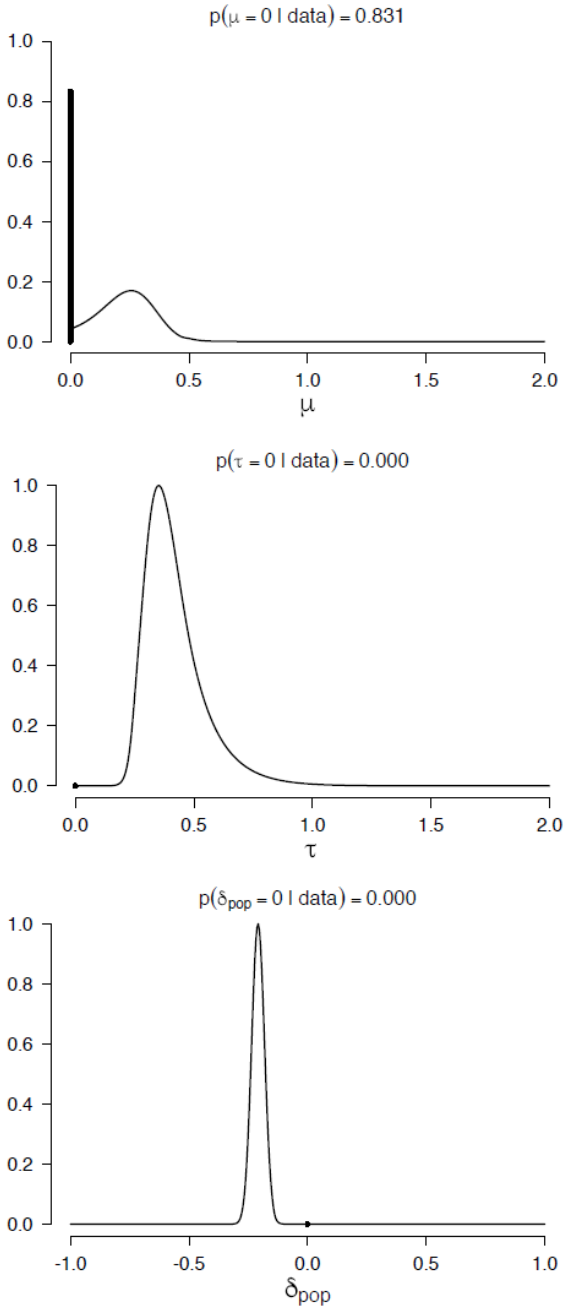


Figure S7-4. Model averaging results for Q1 (*filtered* data). The upper panel displays the results for the group-level mean effect size μ_1, the middle panel displays the results for the across-team heterogeneity τ_1, and the lower panel displays the results for the difference $\delta_{pop,1}$ between the MTurk and the PureProfile populations. Each panel shows the model-averaged posterior distribution for the parameter across all eight models, where the height of the spike at zero corresponds to the summed posterior probability that the parameter equals 0.

Table S7-1: Posterior Model Probabilities (*Unfiltered* Data)

Models	Question				
	1	2	3	4	5
H_1	0.000	0.000	0.000	0.000	0.000
H_2	0.000	0.000	0.000	0.000	0.000
H_3	0.000	0.000	**0.001**	**0.603**	**0.317**
H_4	**0.896**	**0.008**	**0.007**	**0.339**	**0.462**
H_5	0.000	0.000	0.000	0.000	0.000
H_6	0.000	0.000	0.000	0.000	0.000
H_7	0.000	0.000	**0.114**	**0.036**	**0.089**
H_8	**0.104**	**0.992**	**0.878**	**0.021**	**0.132**

Quantifying the Effect of Population for Q1. First we present the results of the *unfiltered* data. The Bayes factor and the posterior model odds both equal 1.040×10^7 in favor of the proposition that $\delta_{pop,1}$ does not equal 0. The summed posterior probability for the models in which $\delta_{pop,1} = 0$ equals 0.000. The lower panel of Figure S7-3 shows the model-averaged posterior distribution for $\delta_{pop,1}$ across all eight models, where the height of the spike at zero corresponds to the summed posterior probability that $\delta_{pop,1} = 0$. In sum, for Q1 the *unfiltered* data provide overwhelming evidence for the hypothesis that the MTurk population and the PureProfile population have different effect sizes.

Next we present the results for the *filtered* data. The Bayes factor and the posterior model odds both equal 1.406×10^{12} in favor of the proposition that $\delta_{pop,1}$ does not equal 0. The summed posterior probability for the models in which $\delta_{pop,1} = 0$ equals 0.000. The lower panel of Figure S7-4 shows the model-averaged posterior distribution for $\delta_{pop,1}$ across all eight models, where the height of the spike at zero corresponds to the summed posterior probability

that $\delta_{pop,1} = 0$. In sum, for Q1 the *filtered* data provide overwhelming evidence for the hypothesis that the MTurk population and the PureProfile population have different effect sizes.

Results for BAMAMA Q2: Lack of Trust Towards Negotiators Who Make Extreme First Offers

Q2: "Negotiators who make extreme first offers are trusted less, relative to negotiators who make moderate first offers." Below are the results from the preregistered BAMAMA analyses.

Full-Model Estimation for Q2

Three parameters are of interest: the group-level mean effect size μ_2, the across-team heterogeneity τ_2, and the difference $\delta_{pop,2}$ between the MTurk and the PurePro file populations.

Table S7-2: Posterior Model Probabilities (*Filtered* Data)

Models	Question				
	1	2	3	4	5
H_1	0.000	0.000	0.000	0.000	0.005
H_2	0.000	0.000	0.000	0.000	0.003
H_3	0.000	0.000	**0.003**	**0.419**	**0.582**
H_4	**0.831**	**0.010**	**0.008**	**0.483**	**0.349**
H_5	0.000	0.000	0.000	0.000	0.001
H_6	0.000	0.000	0.000	0.000	0.001
H_7	0.000	0.000	**0.290**	**0.045**	**0.037**
H_8	**0.169**	**0.990**	**0.700**	**0.053**	**0.022**

First we present the results of the *unfiltered* data. Figure S7-5 shows the prior and posterior distributions from the model with all three parameters free to vary. The top panel of Figure S7-5 suggests that there is an effect on the group-level mean effect size; the middle panel

suggests that there is considerable across-team heterogeneity; the bottom panel suggests that the MTurk population has a higher effect size than the PureProfile population.

Next we present the results of the *filtered* data. Figure S7-6 shows the prior and posterior distributions from the model with all three parameters free to vary. The top panel of Figure S7-6 suggests that there is an effect on the group-level mean effect size; the middle panel suggests that there is considerable across-team heterogeneity; the bottom panel suggests that the MTurk population has a higher effect size than the PureProfile population.

In order to quantify the degree of support that the data provide for and against the presence of each of these effects we now turn to a BAMAMA analysis.

Model Averaging for Q2

As outlined earlier, our model averaging approach considers eight models, constructed by the factorial combination of restrictions $\mu_2 = 0$, $\tau_2 = 0$, and $\delta_{pop,2} = 0$. Each model is assigned equal prior probability; hence, each restriction is a priori equally likely to hold. For each of the three restrictions, the inference is based on the evaluation of predictive performance for all eight models simultaneously. The second column of Table S7-1 presents the posterior model probabilities for Q2 based on the *unfiltered* data. The second column of Table S7-2 presents the posterior model probabilities for Q2 based on the *filtered* data.

Quantifying Overall Evidence for Q2. First we present the results of the *unfiltered* data. The Bayes factor and the posterior model odds both equal 125.851 in favor of the proposition that μ_2 does not equal 0. The summed posterior probability for the models in which $\mu_2 = 0$ equals 0.008. The top panel of Figure S7-7 shows the model-averaged posterior distribution for μ_2 across all eight models, where the height of the spike at zero corresponds to the summed posterior probability that $\mu_2 = 0$. In sum, for Q2 the *unfiltered* data provide compelling evidence for the hypothesis that there is an effect on the group-level mean effect size.

Next we present the results of the *filtered* data. The Bayes factor and the posterior model odds both equal 99.283 in favor of the proposition that μ_2 does not equal 0. The summed posterior probability for the models in which $\mu_2 = 0$ equals 0.010. The top panel of Figure S7-8 shows the model-averaged posterior distribution for μ_2 across all eight models, where the height of the spike at zero corresponds to the summed posterior probability that $\mu_2 = 0$. In sum, for Q2 the *filtered* data provide compelling evidence for the hypothesis that there is an effect on the group-level mean effect size.

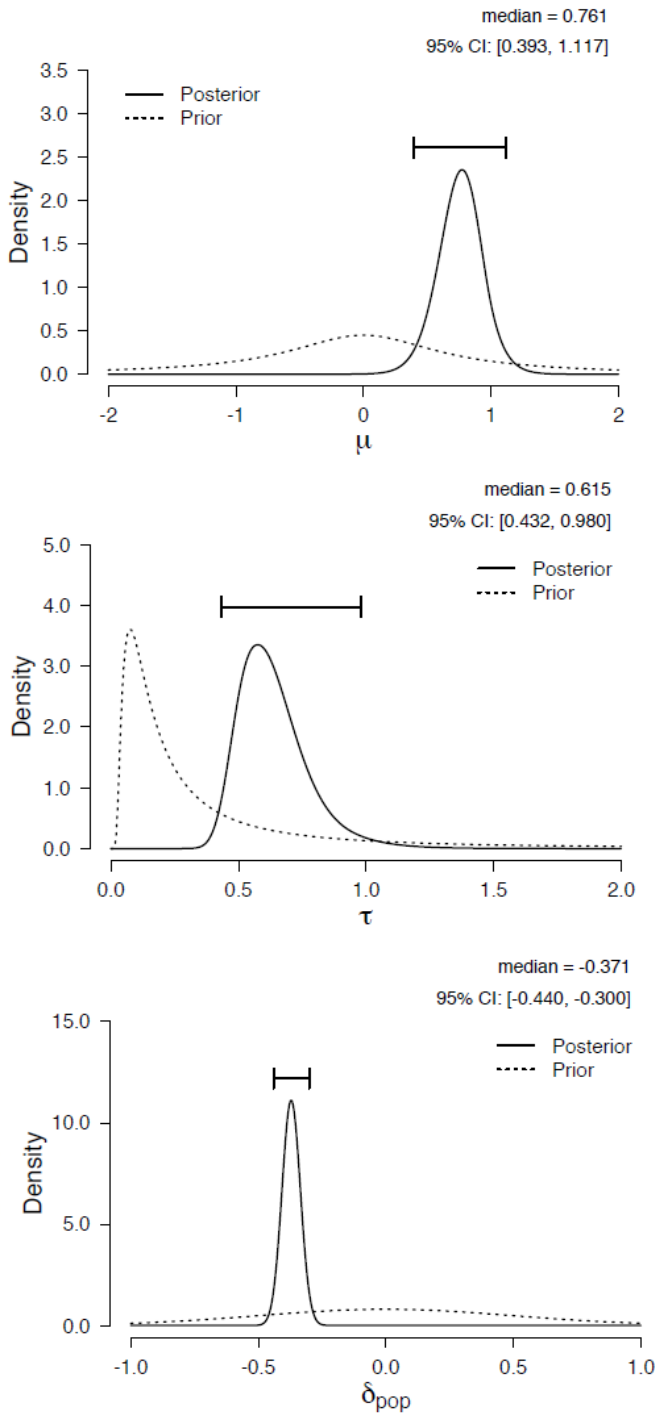


Figure S7-5. Estimation results for Q2 (*unfiltered* data). The upper panel displays the results for the group-level mean effect size μ_2, the middle panel displays the results for the across-team heterogeneity τ_2, and the lower panel displays the results for the difference $\delta_{pop,2}$ between the MTurk and the PureProfile populations. Each panel shows the prior and posterior distribution, the posterior median, and a 95% posterior credible interval.

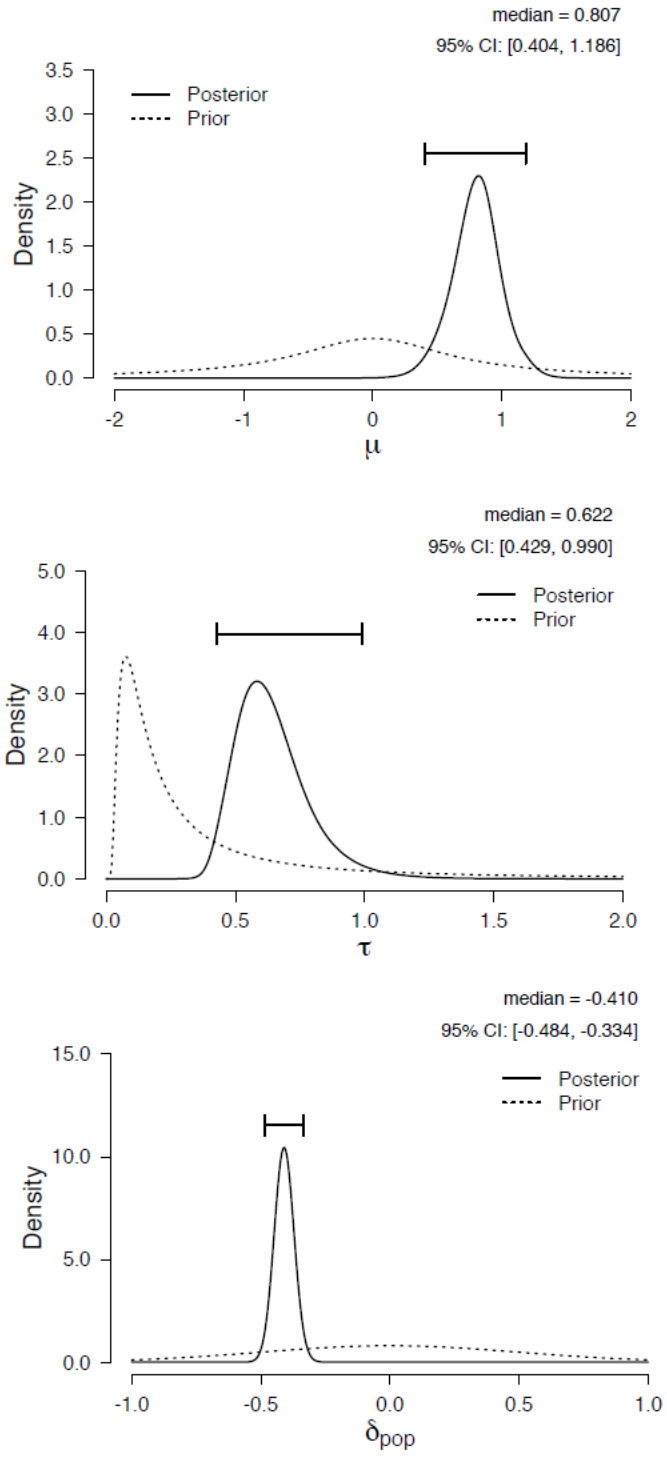


Figure S7-6. Estimation results for Q2 (*filtered* data). The upper panel displays the results for the group-level mean effect size μ_2, the middle panel displays the results for the across-team heterogeneity τ_2, and the lower panel displays the results for the difference $\delta_{pop,2}$ between the MTurk and the PureProfile populations. Each panel shows the prior and posterior distribution, the posterior median, and a 95% posterior credible interval.

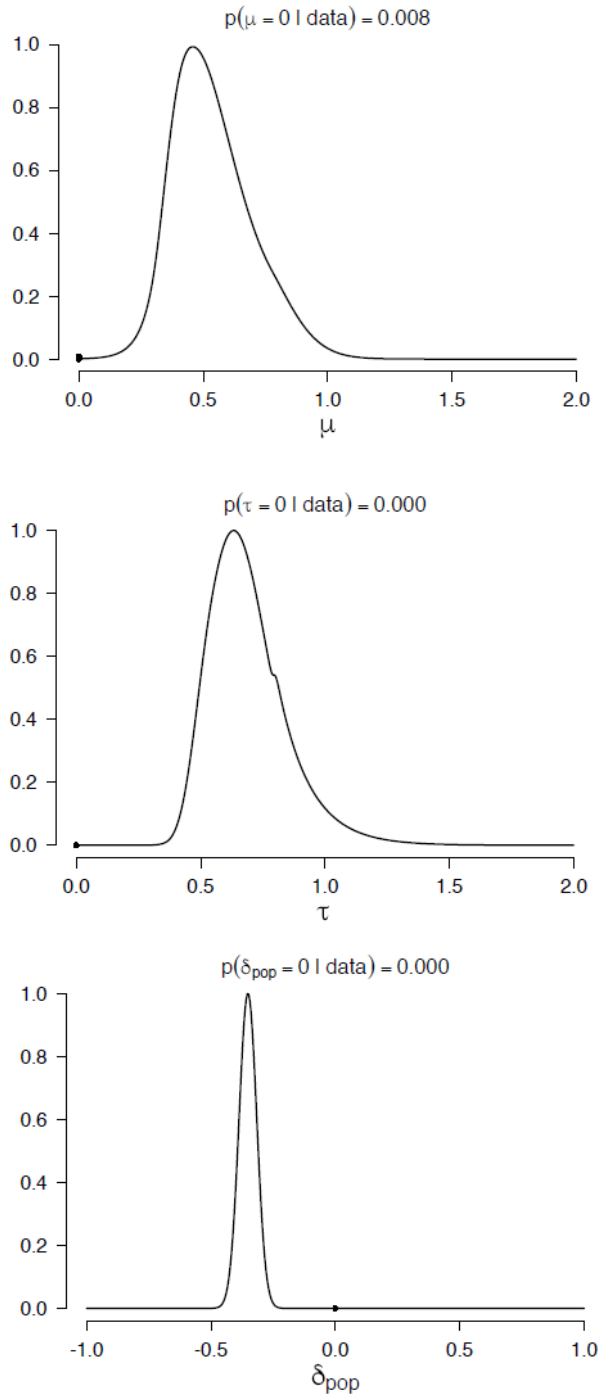


Figure S7-7. Model averaging results for Q2 (*unfiltered* data). The upper panel displays the results for the group-level mean effect size μ_2, the middle panel displays the results for the across-team heterogeneity τ_2, and the lower panel displays the results for the difference $\delta_{pop,2}$ between the MTurk and the PureProfile populations. Each panel shows the model-averaged posterior distribution for the parameter across all eight models, where the height of the spike at zero corresponds to the summed posterior probability that the parameter equals 0.

Quantifying Heterogeneity for Q2. First we present the results of the *unfiltered* data. The Bayes factor and the posterior model odds both equal ∞[2] in favor of the proposition that τ_2 does not equal 0. The summed posterior probability for the models in which $\tau_2 = 0$ equals 0.000. The middle panel of Figure S7-7 shows the model-averaged posterior distribution for τ_2 across all eight models, where the height of the spike at zero corresponds to the summed posterior probability that $\tau_2 = 0$. In sum, for Q2 the *unfiltered* data provide overwhelming evidence for the hypothesis that there is across-team heterogeneity.

Next we present the results of the *filtered* data. The Bayes factor and the posterior model odds both equal 9.007×10^{14} in favor of the proposition that τ_2 does not equal 0. The summed posterior probability for the models in which $\tau_2 = 0$ equals 0.000. The middle panel of Figure S7-8 shows the model-averaged posterior distribution for τ_2 across all eight models, where the height of the spike at zero corresponds to the summed posterior probability that $\tau_2 = 0$. In sum, for Q2 the *filtered* data provide overwhelming evidence for the hypothesis that there is across-team heterogeneity.

Quantifying the Effect of Population for Q2. First we present the results of the *unfiltered* data. The Bayes factor and the posterior model odds both equal ∞[3] in favor of the proposition that $\delta_{pop,2}$ does not equal 0. The summed posterior probability for the models in which $\delta_{pop,2} = 0$ equals 0.000. The lower panel of Figure S7-7 shows the model-averaged posterior distribution for $\delta_{pop,2}$ across all eight models, where the height of the spike at zero corresponds to the summed posterior probability that $\delta_{pop,2} = 0$. In sum, for Q2 the *unfiltered*

[2] The true Bayes factor is so large that it exceeds the available numerical precision.
[3] The true Bayes factor is so large that it exceeds the available numerical precision.

data provide overwhelming evidence for the hypothesis that the MTurk population and the PureProfile population have different effect sizes.

Next we present the results of the *filtered* data. The Bayes factor and the posterior model odds both equal 9.007×10^{14} in favor of the proposition that $\delta_{pop,2}$ does not equal 0. The summed posterior probability for the models in which $\delta_{pop,2} = 0$ equals 0.000. The lower panel of Figure S7-8 shows the model-averaged posterior distribution for $\delta_{pop,2}$ across all eight models, where the height of the spike at zero corresponds to the summed posterior probability that $\delta_{pop,2} = 0$. In sum, for Q2 the *filtered* data provide overwhelming evidence for the hypothesis that the MTurk population and the PureProfile population have different effect sizes.

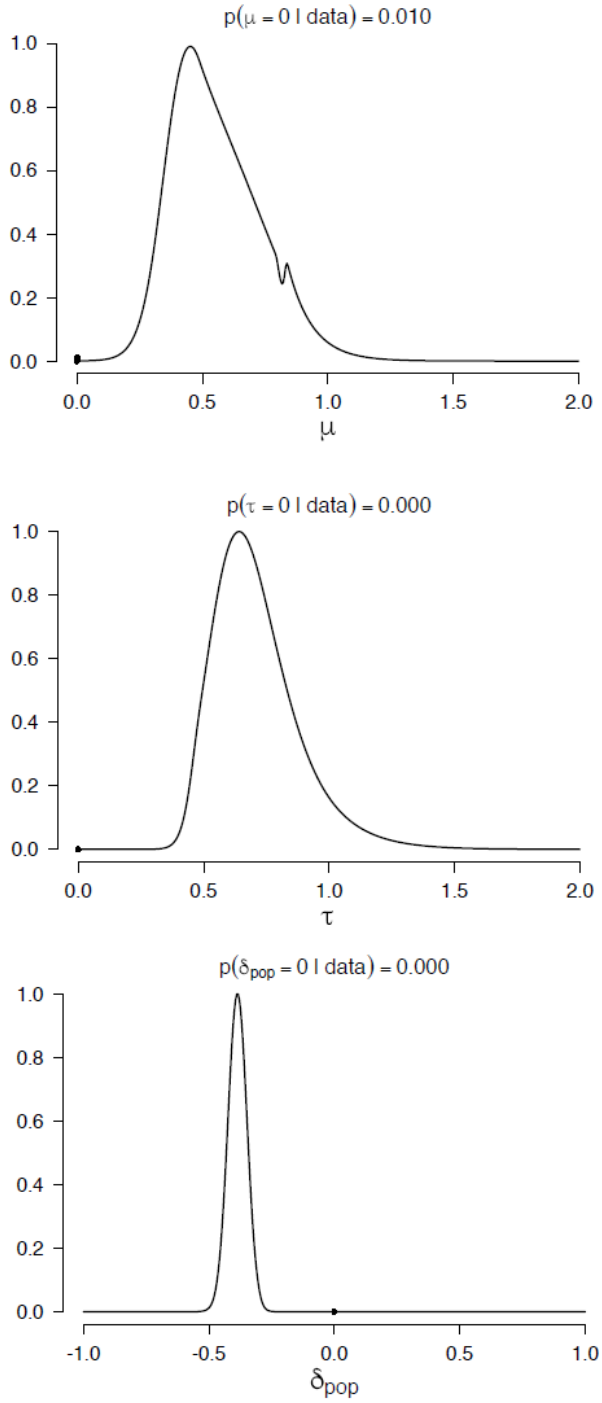


Figure S7-8. Model averaging results for Q2 (*filtered* data). The upper panel displays the results for the group-level mean effect size μ_2, the middle panel displays the results for the across-team heterogeneity τ_2, and the lower panel displays the results for the difference $\delta_{pop,2}$ between the MTurk and the PureProfile populations. Each panel shows the model-averaged posterior distribution for the parameter across all eight models, where the height of the spike at zero corresponds to the summed posterior probability that the parameter equals 0.

Results for BAMAMA Q3: Moral Judgments Towards Wealthy Workers

Q3: "A person continuing to work despite having no material/financial need to work has beneficial effects on moral judgments of that individual." Below are the results from the preregistered BAMAMA analyses.

Full-Model Estimation for Q3

Three parameters are of interest: the group-level mean effect size μ_3, the across-team heterogeneity τ_3, and the difference $\delta_{pop,3}$ between the MTurk and the PureProfile populations.

First we present the results of the *unfiltered* data. Figure S7-9 shows the prior and posterior distributions from the model with all three parameters free to vary. The top panel of Figure S7-9 suggests that there is a modest effect on the group-level mean effect size; the middle panel suggests that there is some across-team heterogeneity; the bottom panel suggests that the MTurk population has a slightly higher effect size than the PureProfile population.

Next we present the results of the *filtered* data. Figure S7-10 shows the prior and posterior distributions from the model with all three parameters free to vary. The top panel of Figure S7-10 suggests that there is a modest effect on the group-level mean effect size; the middle panel suggests that there is some across-team heterogeneity; the bottom panel suggests that the MTurk population has a slightly higher effect size than the PureProfile population.

In order to quantify the degree of support that the data provide for and against the presence of each of these effects we now turn to a BAMAMA analysis.

Model Averaging for Q3

As outlined earlier, our model averaging approach considers eight models, constructed by the factorial combination of restrictions $\mu_3 = 0$, $\tau_3 = 0$, and $\delta_{pop,3} = 0$. Each model is assigned equal prior probability; hence, each restriction is a priori equally likely to hold. For each of the three restrictions, the inference is based on the evaluation of predictive performance for all eight models simultaneously. The third column of Table S7-1 presents the posterior model probabilities for Q3 based on the *unfiltered* data. The third column of Table S7-2 presents the posterior model probabilities for Q3 based on the *filtered* data.

Quantifying Overall Evidence for Q3. First we present the results of the *unfiltered* data. The Bayes factor and the posterior model odds both equal 125.476 in favor of the proposition that μ_3 does not equal 0. The summed posterior probability for the models in which $\mu_3 = 0$ equals 0.008. The top panel of Figure S7-11 shows the model-averaged posterior distribution for μ_3 across all eight models, where the height of the spike at zero corresponds to the summed posterior probability that $\mu_3 = 0$. In sum, for Q3 the *unfiltered* data provide compelling evidence for the hypothesis that there is an effect on the group-level mean effect size.

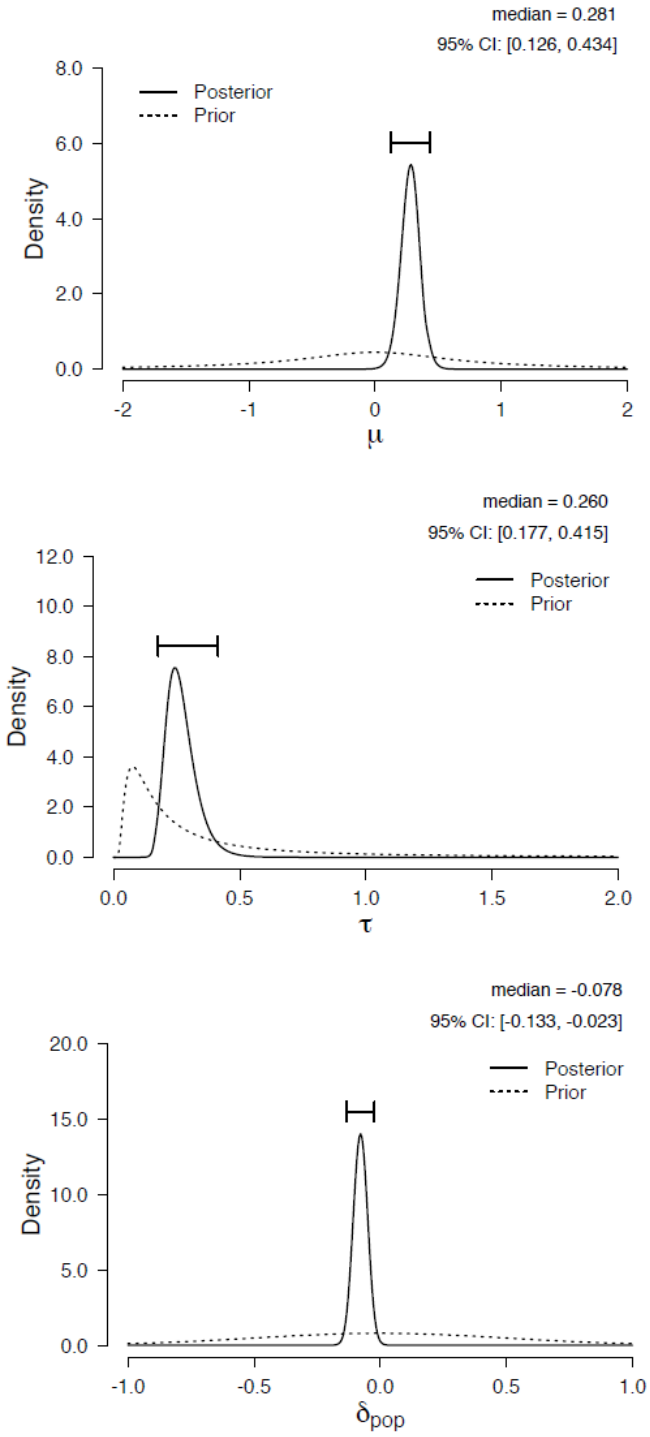


Figure S7-9. Estimation results for Q3 (*unfiltered* data). The upper panel displays the results for the group-level mean effect size μ_3 the middle panel displays the results for the across-team heterogeneity τ_3, and the lower panel displays the results for the difference $\delta_{pop,3}$ between the MTurk and the PureProfile populations. Each panel shows the prior and posterior distribution, the posterior median, and a 95% posterior credible interval.

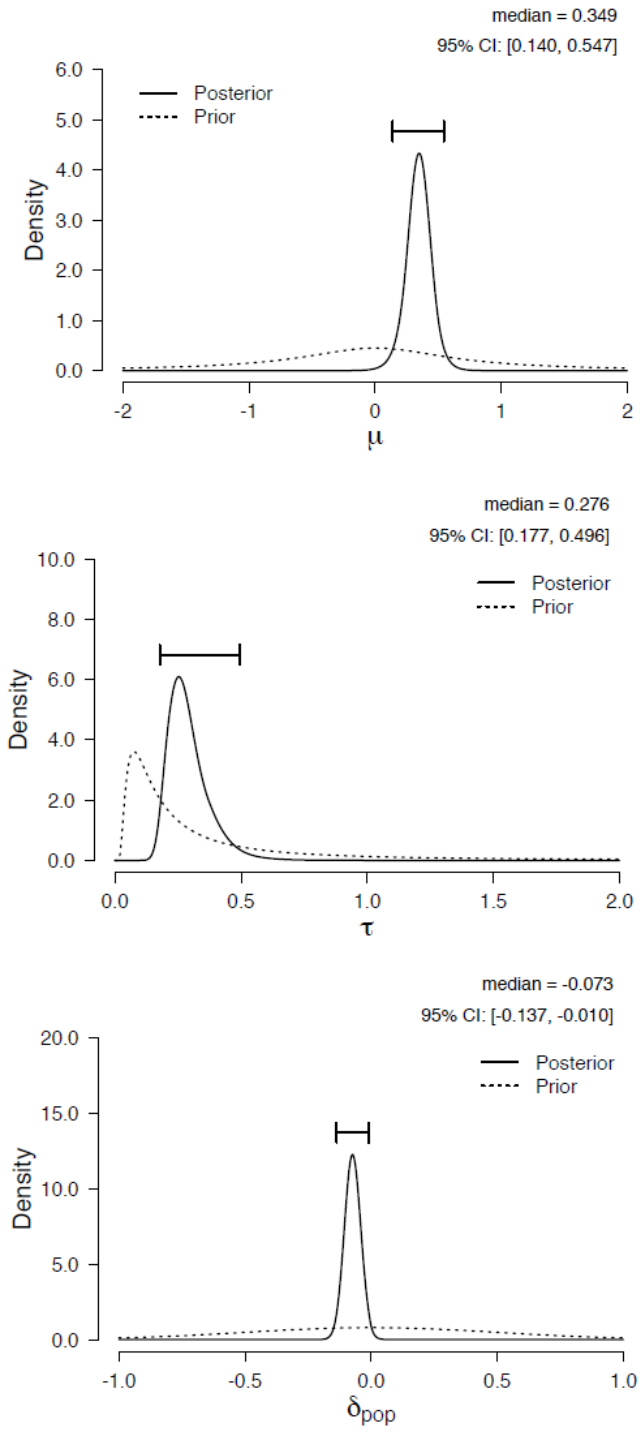


Figure S7-10. Estimation results for Q3 (*filtered* data). The upper panel displays the results for the group-level mean effect size μ_3, the middle panel displays the results for the across-team heterogeneity τ_3, and the lower panel displays the results for the difference $\delta_{pop,3}$ between the MTurk and the PureProfile populations. Each panel shows the prior and posterior distribution, the posterior median, and a 95% posterior credible interval.

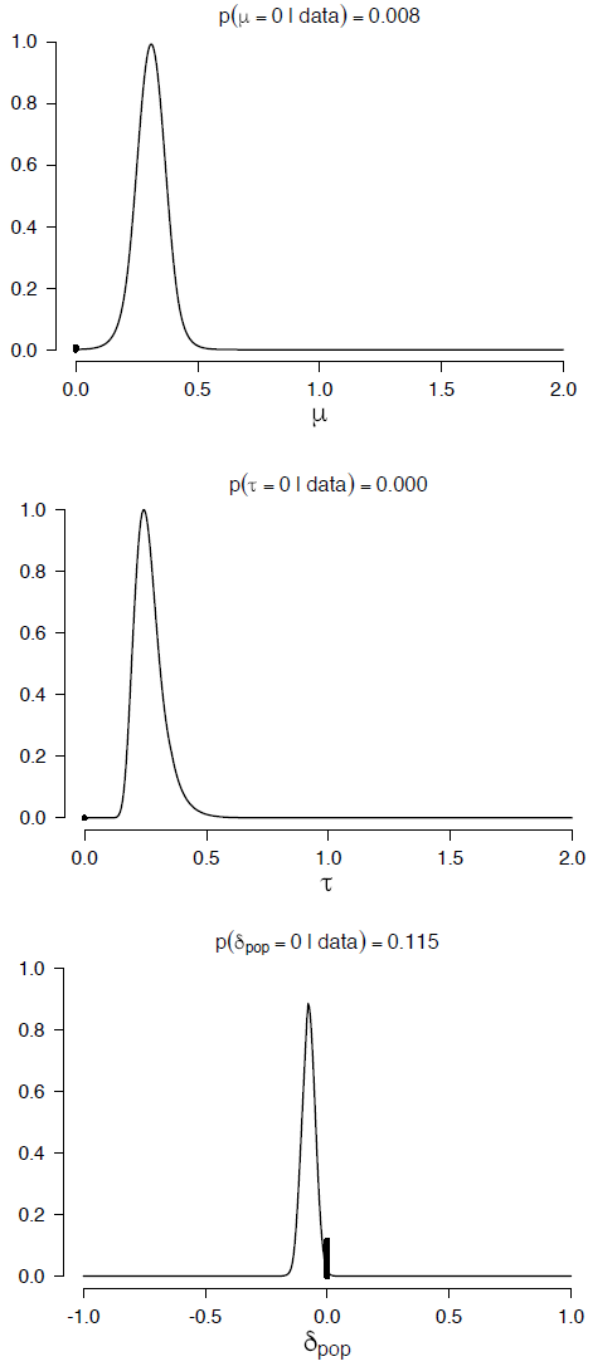


Figure S7-11. Model averaging results for Q3 (*unfiltered* data). The upper panel displays the results for the group-level mean effect size μ_3, the middle panel displays the results for the across-team heterogeneity τ_3, and the lower panel displays the results for the difference $\delta_{pop,3}$ between the MTurk and the PureProfile populations. Each panel shows the model-averaged posterior distribution for the parameter across all eight models, where the height of the spike at zero corresponds to the summed posterior probability that the parameter equals 0.

Next we present the results of the *filtered* data. The Bayes factor and the posterior model odds both equal 91.747 in favor of the proposition that μ_3 does not equal 0. The summed posterior probability for the models in which $\mu_3 = 0$ equals 0.011. The top panel of Figure S7-12 shows the model-averaged posterior distribution for μ_3 across all eight models, where the height of the spike at zero corresponds to the summed posterior probability that $\mu_3 = 0$. In sum, for Q3 the *filtered* data provide compelling evidence for the hypothesis that there is an effect on the group-level mean effect size.

Quantifying Heterogeneity for Q3. First we present the results of the *unfiltered* data. The Bayes factor and the posterior model odds both equal ∞[4] in favor of the proposition that τ_3 does not equal 0. The summed posterior probability for the models in which $\tau_3 = 0$ equals 0.000. The middle panel of Figure S7-11 shows the model-averaged posterior distribution for τ_3 across all eight models, where the height of the spike at zero corresponds to the summed posterior probability that $\tau_3 = 0$. In sum, for Q3 the *unfiltered* data provide overwhelming evidence for the hypothesis that there is across-team heterogeneity.

Next we present the results of the *filtered* data. The Bayes factor and the posterior model odds both equal ∞[5] in favor of the proposition that τ_3 does not equal 0. The summed posterior probability for the models in which $\tau_3 = 0$ equals 0.000. The middle panel of Figure S7-12 shows the model-averaged posterior distribution for τ_3 across all eight models, where the height of the spike at zero corresponds to the summed posterior probability that $\tau_3 = 0$. In sum, for Q3 the *filtered* data provide overwhelming evidence for the hypothesis that there is across-team heterogeneity.

[4] The true Bayes factor is so large that it exceeds the available numerical precision.
[5] The true Bayes factor is so large that it exceeds the available numerical precision.

Quantifying the Effect of Population for Q3. First we present the results of the *unfiltered* data. The Bayes factor and the posterior model odds both equal 7.694 in favor of the proposition that $\delta_{pop,3}$ does not equal 0. The summed posterior probability for the models in which $\delta_{pop,3} = 0$ equals 0.115. The lower panel of Figure S7-11 shows the model-averaged posterior distribution for $\delta_{pop,3}$ across all eight models, where the height of the spike at zero corresponds to the summed posterior probability that $\delta_{pop,3} = 0$. In sum, for Q3 the *unfiltered* data provide moderate evidence for the hypothesis that the MTurk population and the PureProfile population have different effect sizes. Next we present the results of the *filtered* data. The Bayes factor and the posterior model odds both equal 2.416 in favor of the proposition that $\delta_{pop,3}$ does not equal 0. The summed posterior probability for the models in which $\delta_{pop,3} = 0$ equals 0.293. The lower panel of Figure S7-12 shows the model-averaged posterior distribution for $\delta_{pop,3}$ across all eight models, where the height of the spike at zero corresponds to the summed posterior probability that $\delta_{pop,3} = 0$. In sum, for Q3 the *filtered* data provide weak evidence for the hypothesis that the MTurk population and the PureProfile population have different effect sizes.

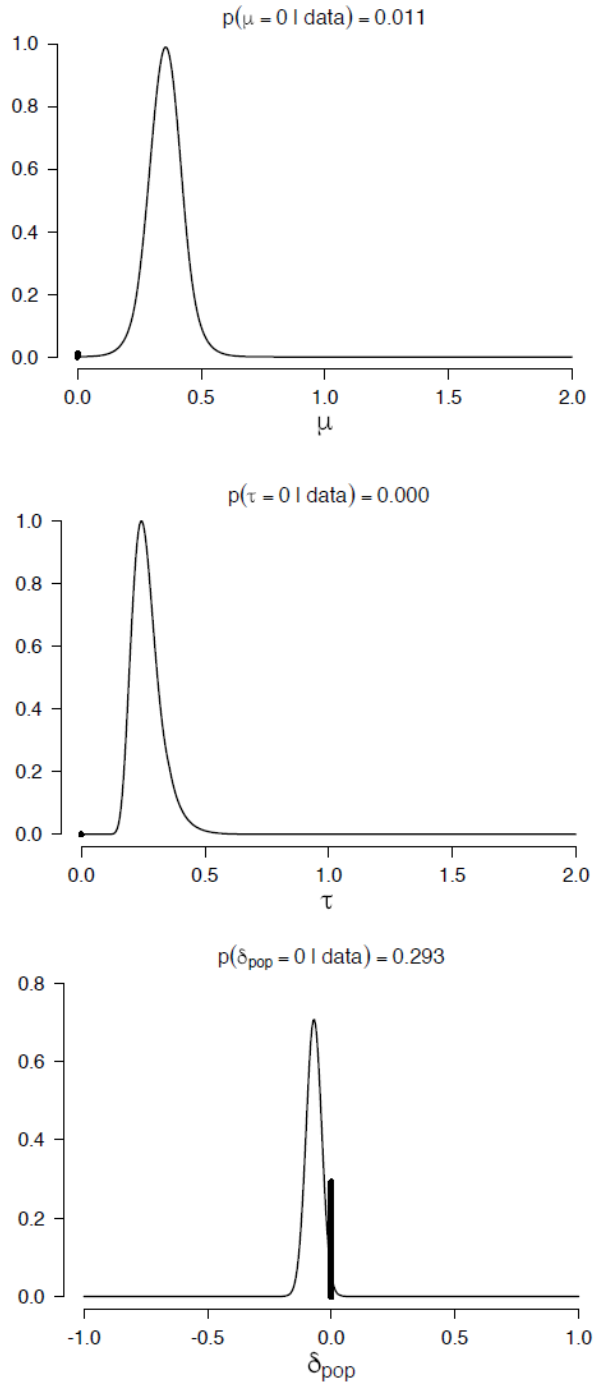


Figure S7-12. Model averaging results for Q3 (*filtered* data). The upper panel displays the results for the group-level mean effect size μ_3, the middle panel displays the results for the across-team heterogeneity τ_3, and the lower panel displays the results for the difference $\delta_{pop,3}$ between the MTurk and the PureProfile populations. Each panel shows the model-averaged posterior distribution for the parameter across all eight models, where the height of the spike at zero corresponds to the summed posterior probability that the parameter equals 0.

Results for BAMAMA Q4:

Opposition to Performance Enhancers Banned by Proximal Authority

Q4: "Part of why people are opposed to the use of performance enhancing drugs in sport is because they are 'against the rules.' But, whether the performance enhancer is against the rules established by a proximal authority (e.g., the league) contributes more to this judgment than whether it is against the law." Below are the results from the preregistered BAMAMA analyses.

Full-Model Estimation for Q4

Three parameters are of interest: the group-level mean effect size μ_4, the across-team heterogeneity τ_4, and the difference $\delta_{pop,4}$ between the MTurk and the PureProfile populations.

First we present the results of the *unfiltered* data. Figure S7-13 shows the prior and posterior distributions from the model with all three parameters free to vary. The top panel of Figure S7-13 suggests that if there exists an effect on the group-level mean effect size, it is likely to be very small; the middle panel suggests that there is some across-team heterogeneity; the bottom panel suggests that the MTurk population may have a slightly higher effect size than the PureProfile population, although the result does not appear conclusive.

Next we present the results of the *filtered* data. Figure S7-14 shows the prior and posterior distributions from the model with all three parameters free to vary. The top panel of Figure S7-14 suggests that if there exists an effect on the group-level mean effect size, it is likely to be very small; the middle panel suggests that there is some across-team heterogeneity; the bottom panel suggests that the MTurk population may have a slightly higher effect size than the PureProfile population, although the result does not appear conclusive.

In order to quantify the degree of support that the data provide for and against the presence of each of these effects we now turn to a BAMAMA analysis.

Model Averaging for Q4

As outlined earlier, our model averaging approach considers eight models, constructed by the factorial combination of restrictions $\mu_4 = 0$, $\tau_4 = 0$, and $\delta_{pop,4} = 0$. Each model is assigned equal prior probability; hence, each restriction is a priori equally likely to hold. For each of the three restrictions, the inference is based on the evaluation of predictive performance for all eight models simultaneously. The fourth column of Table S7-1 presents the posterior model probabilities for Q4 based on the *unfiltered* data. The fourth column of Table S7-2 presents the posterior model probabilities for Q4 based on the *filtered* data.

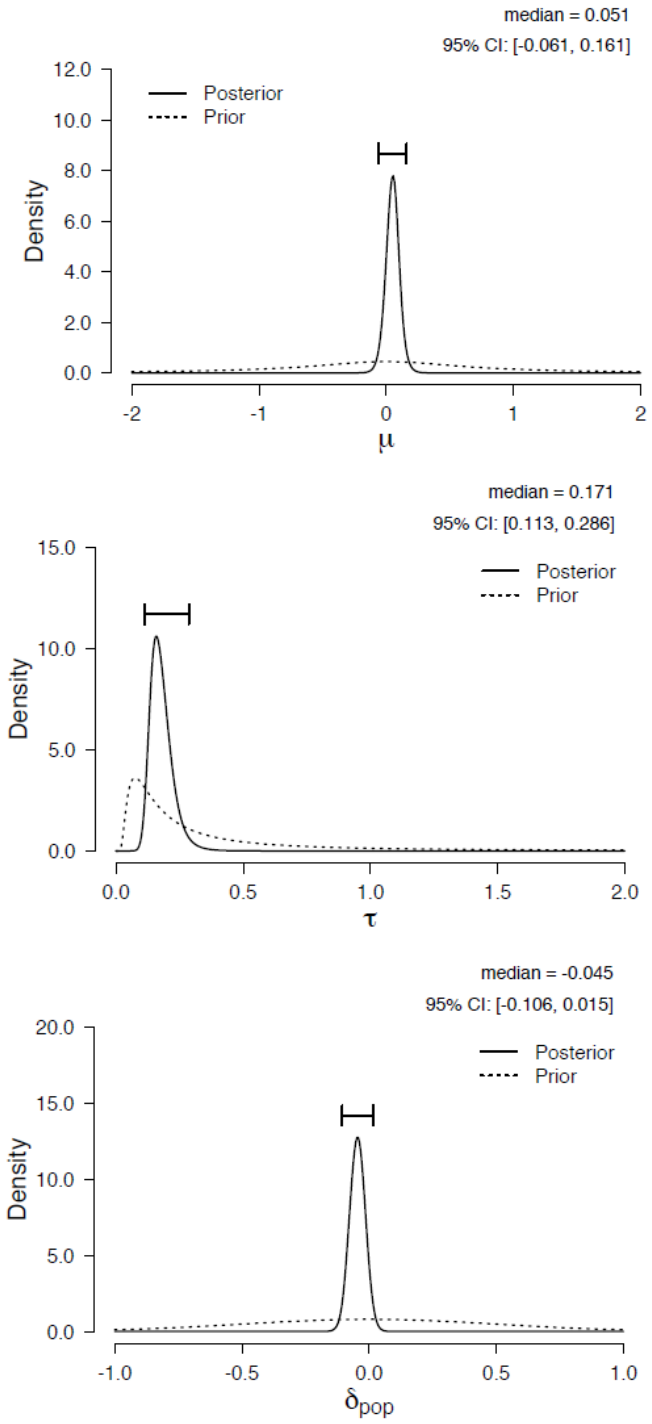


Figure S7-13. Estimation results for Q4 (*unfiltered* data). The upper panel displays the results for the group-level mean effect size μ_4, the middle panel displays the results for the across-team heterogeneity τ_4, and the lower panel displays the results for the difference $\delta_{pop,4}$ between the MTurk and the PureProfile populations. Each panel shows the prior and posterior distribution, the posterior median, and a 95% posterior credible interval.

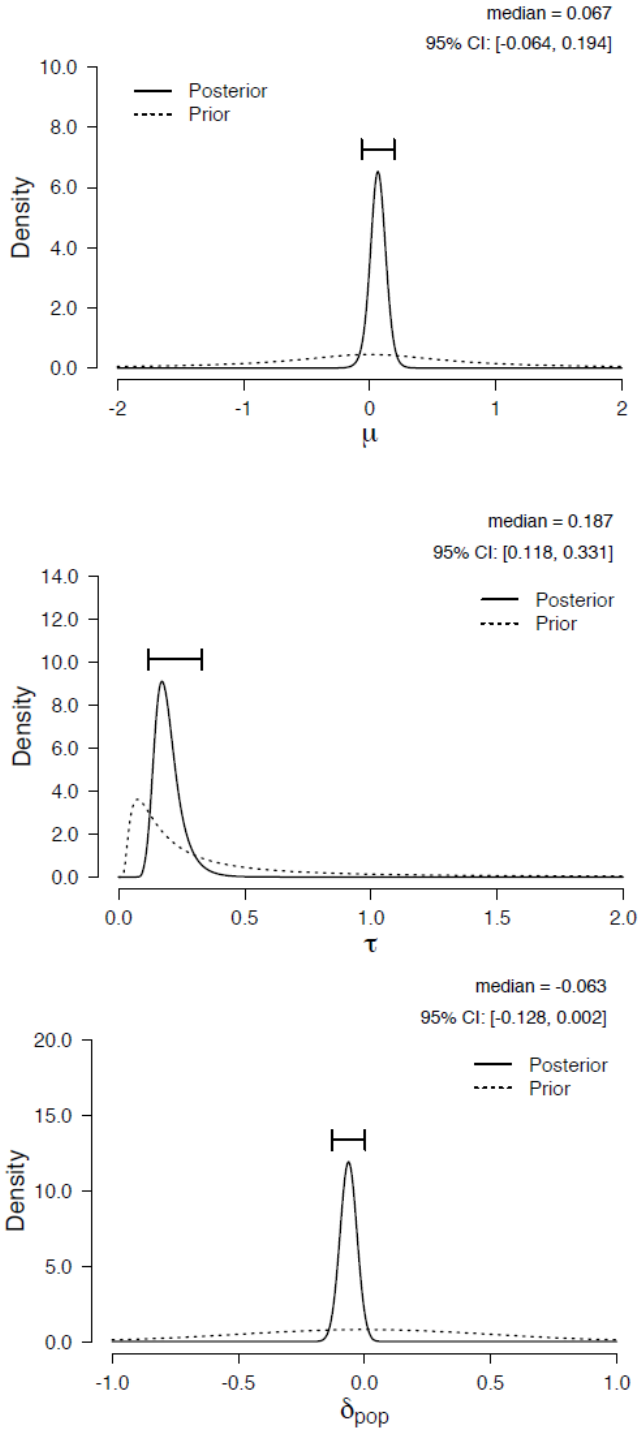


Figure S7-14. Estimation results for Q4 (*filtered* data). The upper panel displays the results for the group-level mean effect size μ_4, the middle panel displays the results for the across-team heterogeneity τ_4, and the lower panel displays the results for the difference $\delta_{pop,4}$ between the MTurk and the PureProfile populations. Each panel shows the prior and posterior distribution, the posterior median, and a 95% posterior credible interval.

Quantifying Overall Evidence for Q4. First we present the results of the *unfiltered* data. The Bayes factor and the posterior model odds both equal 16.421 in favor of the proposition that μ_4 equals 0. The summed posterior probability for the models in which $\mu_4 = 0$ equals 0.943. The top panel of Figure S7-15 shows the model-averaged posterior distribution for μ_4 across all eight models, where the height of the spike at zero corresponds to the summed posterior probability that $\mu_4 = 0$. In sum, for Q4 the *unfiltered* data provide strong evidence for the hypothesis that there is no effect on the group-level mean effect size.

Next we present the results of the *filtered* data. The Bayes factor and the posterior model odds both equal 9.263 in favor of the proposition that μ_4 equals 0. The summed posterior probability for the models in which $\mu_4 = 0$ equals 0.903. The top panel of Figure S7-16 shows the model-averaged posterior distribution for μ_4 across all eight models, where the height of the spike at zero corresponds to the summed posterior probability that $\mu_4 = 0$. In sum, for Q4 the *filtered* data provide moderate evidence for the hypothesis that there is no effect on the group-level mean effect size.

Quantifying Heterogeneity for Q4. First we present the results of the *unfiltered* data. The Bayes factor and the posterior model odds both equal ∞[6] in favor of the proposition that τ_4 does not equal 0. The summed posterior probability for the models in which $\tau_4 = 0$ equals 0.000. The middle panel of Figure S7-15 shows the model-averaged posterior distribution for τ_4 across all eight models, where the height of the spike at zero corresponds to the summed posterior probability that $\tau_4 = 0$. In sum, for Q4 the *unfiltered* data provide overwhelming evidence for the hypothesis that there is across-team heterogeneity.

[6] The true Bayes factor is so large that it exceeds the available numerical precision.

Next we present the results of the *filtered* data. The Bayes factor and the posterior model odds both equal 9.007×10^{15} in favor of the proposition that τ_4 does not equal 0. The summed posterior probability for the models in which $\tau_4 = 0$ equals 0.000. The middle panel of Figure S7-16 shows the model-averaged posterior distribution for τ_4 across all eight models, where the height of the spike at zero corresponds to the summed posterior probability that $\tau_4 = 0$. In sum, for Q4 the *filtered* data provide overwhelming evidence for the hypothesis that there is across-team heterogeneity.

Quantifying the Effect of Population for Q4. First we present the results of the *unfiltered* data. The Bayes factor and the posterior model odds both equal 1.771 in favor of the proposition that $\delta_{pop,4}$ equals 0. The summed posterior probability for the models in which $\delta_{pop,4} = 0$ equals 0.639. The lower panel of Figure S7-15 shows the model-averaged posterior distribution for $\delta_{pop,4}$ across all eight models, where the height of the spike at zero corresponds to the summed posterior probability that $\delta_{pop,4} = 0$. In sum, for Q4 the *unfiltered* data provide weak evidence for the hypothesis that the MTurk population and the PureProfile population have the same effect size.

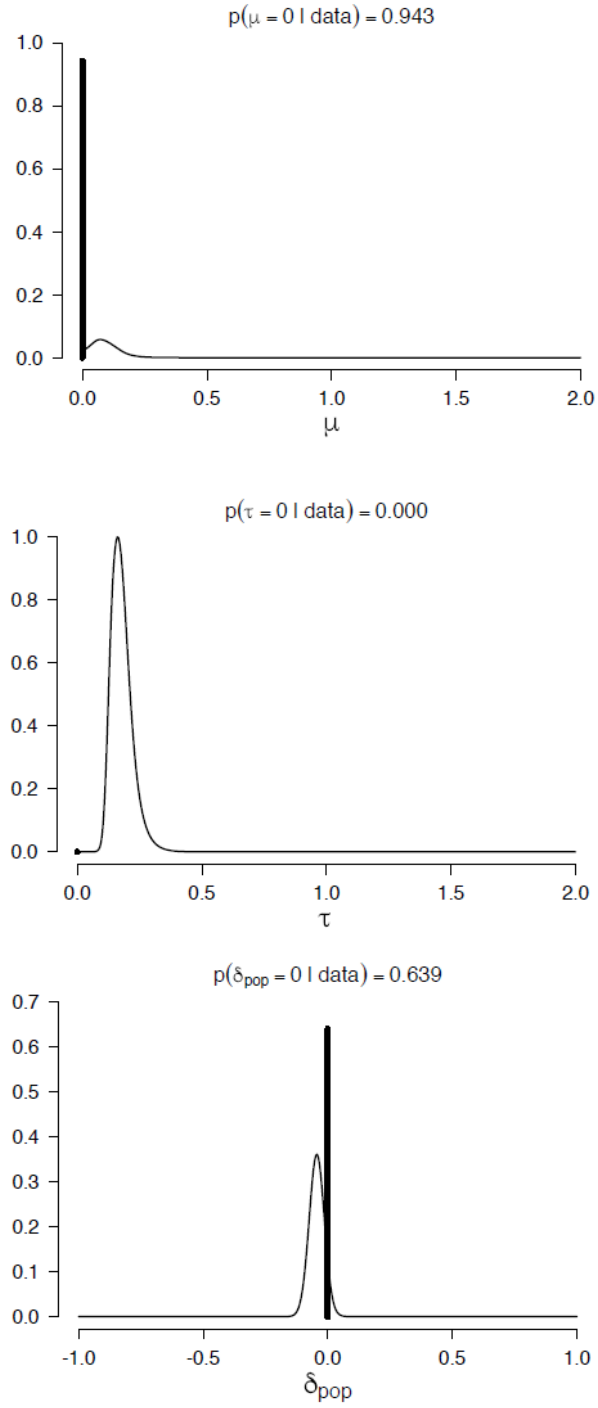


Figure S7-15. Model averaging results for Q4 (*unfiltered* data). The upper panel displays the results for the group-level mean effect size μ_4, the middle panel displays the results for the across-team heterogeneity τ_4, and the lower panel displays the results for the difference $\delta_{pop,4}$ between the MTurk and the PureProfile populations. Each panel shows the model-averaged posterior distribution for the parameter across all eight models, where the height of the spike at zero corresponds to the summed posterior probability that the parameter equals 0.

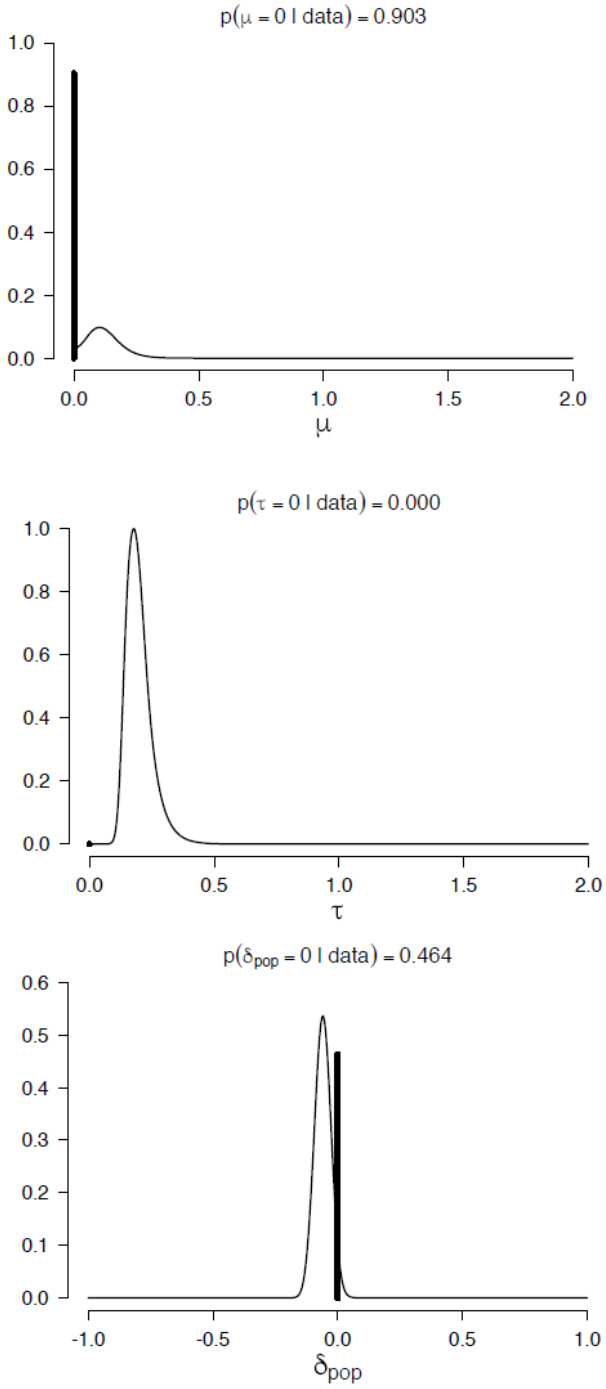


Figure S7-16. Model averaging results for Q4 (*filtered* data). The upper panel displays the results for the group-level mean effect size μ_4, the middle panel displays the results for the across-team heterogeneity τ_4, and the lower panel displays the results for the difference $\delta_{pop,4}$ between the MTurk and the PureProfile populations. Each panel shows the model-averaged posterior distribution for the parameter across all eight models, where the height of the spike at zero corresponds to the summed posterior probability that the parameter equals 0.

Next we present the results of the *filtered* data. The Bayes factor and the posterior model odds both equal 1.156 in favor of the proposition that $\delta_{pop,4}$ does not equal 0. The summed posterior probability for the models in which $\delta_{pop,4} = 0$ equals 0.464. The lower panel of Figure S7-16 shows the model-averaged posterior distribution for $\delta_{pop,4}$ across all eight models, where the height of the spike at zero corresponds to the summed posterior probability that $\delta_{pop,4} = 0$. In sum, for Q4 the *filtered* data provide weak evidence for the hypothesis that the MTurk population and the PureProfile population have different effect sizes.

Results for BAMAMA Q5: Deontological Moral Orientation and Happiness

Q5: "A deontological (as opposed to utilitarian) moral orientation is positively related to personal happiness." Below are the results from the preregistered BAMAMA analyses.

Full-Model Estimation for Q5

Three parameters are of interest: the group-level mean effect size μ_5, the across-team heterogeneity τ_5, and the difference $\delta_{pop,5}$ between the MTurk and the PureProfile populations.

First we present the results of the *unfiltered* data. Figure S7-17 shows the prior and posterior distributions from the model with all three parameters free to vary. The top panel of Figure S7-17 suggests that if there exists an effect on the group-level mean effect size, it is likely to be very small; the middle panel suggests that there is some across-team heterogeneity; the bottom panel suggests that the MTurk population has a slightly higher effect size than the PureProfile population.

Next we present the results of the *filtered* data. Figure S7-18 shows the prior and posterior distributions from the model with all three parameters free to vary. The top panel of

Figure S7-18 suggests that if there exists an effect on the group-level mean effect size, it is likely to be very small; the middle panel suggests that there is some across-team heterogeneity; the bottom panel suggests that the MTurk population may have a slightly higher effect size than the PureProfile population.

In order to quantify the degree of support that the data provide for and against the presence of each of these effects, we now turn to a BAMAMA analysis.

Model Averaging for Q5

As outlined earlier, our model averaging approach considers eight models, constructed by the factorial combination of restrictions $\mu_5 = 0$, $\tau_5 = 0$, and $\delta_{pop,5} = 0$. Each model is assigned equal prior probability; hence, each restriction is a priori equally likely to hold. For each of the three restrictions, the inference is based on the evaluation of predictive performance for all eight models simultaneously. The fifth column of Table S7-1 presents the posterior model probabilities for Q5 based on the *unfiltered* data. The fifth column of Table S7-2 presents the posterior model probabilities for Q5 based on the *filtered* data.

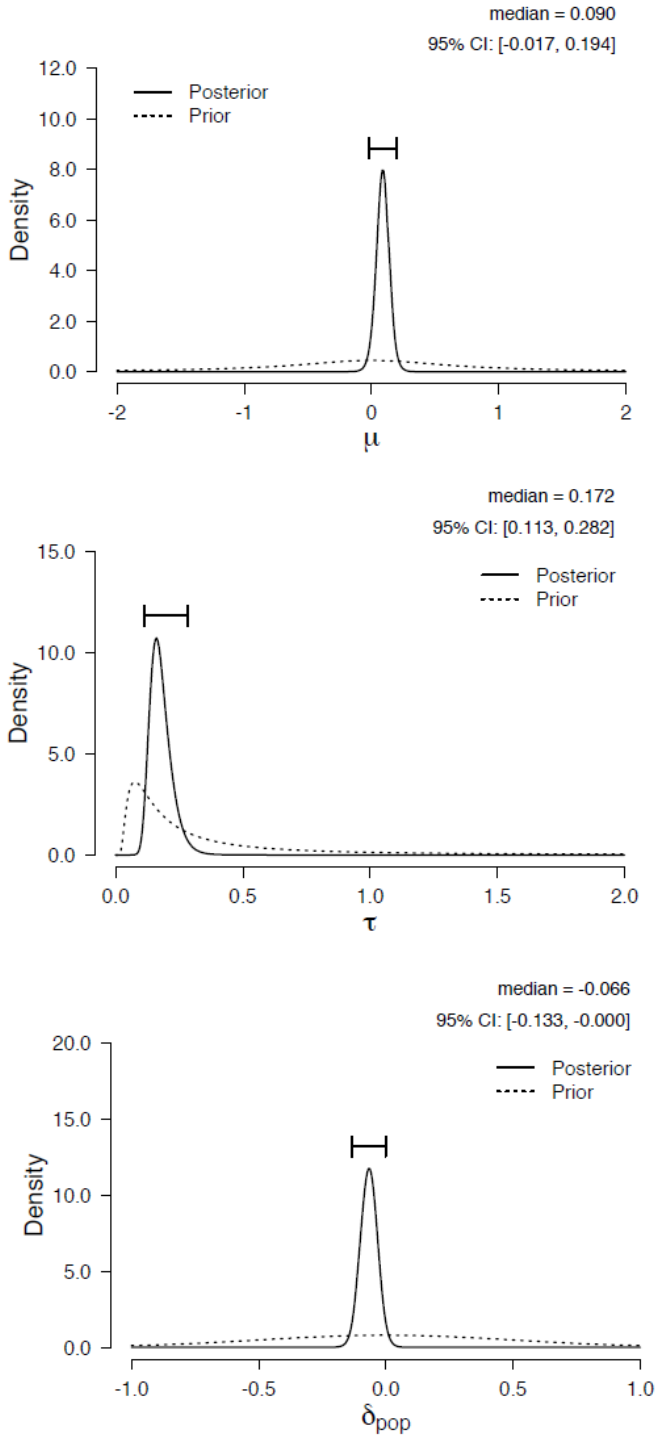


Figure S7-17. Estimation results for Q5 (*unfiltered* data). The upper panel displays the results for the group-level mean effect size μ_5, the middle panel displays the results for the across-team heterogeneity τ_5, and the lower panel displays the results for the difference $\delta_{pop,5}$ between the MTurk and the PureProfile populations. Each panel shows the prior and posterior distribution, the posterior median, and a 95% posterior credible interval.

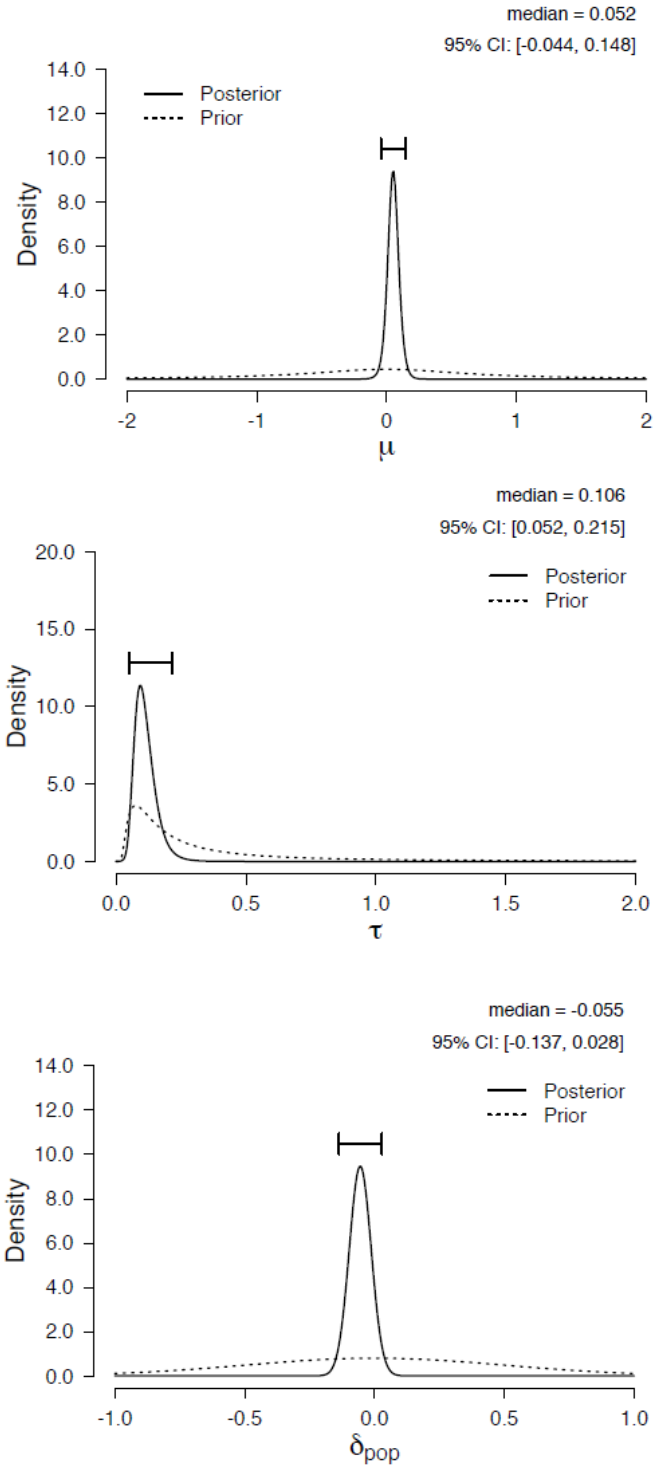


Figure S7-18. Estimation results for Q5 (*filtered* data). The upper panel displays the results for the group-level mean effect size μ_5, the middle panel displays the results for the across-team heterogeneity τ_5, and the lower panel displays the results for the difference $\delta_{pop,5}$ between the MTurk and the PureProfile populations. Each panel shows the prior and posterior distribution, the posterior median, and a 95% posterior credible interval.

Quantifying Overall Evidence for Q5. First we present the results of the *unfiltered* data. The Bayes factor and the posterior model odds both equal 3.519 in favor of the proposition that μ_5 equals 0. The summed posterior probability for the models in which $\mu_5 = 0$ equals 0.779. The top panel of Figure S7-19 shows the model-averaged posterior distribution for μ_5 across all eight models, where the height of the spike at zero corresponds to the summed posterior probability that $\mu_5 = 0$. In sum, for Q5 the *unfiltered* data provide moderate-to-weak evidence for the hypothesis that there is no effect on the group-level mean effect size.

Next we present the results of the *filtered* data. The Bayes factor and the posterior model odds both equal 15.239 in favor of the proposition that μ_5 equals 0. The summed posterior probability for the models in which $\mu_5 = 0$ equals 0.938. The top panel of Figure S7-20 shows the model-averaged posterior distribution for μ_5 across all eight models, where the height of the spike at zero corresponds to the summed posterior probability that $\mu_5 = 0$. In sum, for Q5 the *filtered* data provide strong evidence for the hypothesis that there is no effect on the group-level mean effect size.

Quantifying Heterogeneity for Q5. First we present the results of the *unfiltered* data. The Bayes factor and the posterior model odds both equal 6.929×10^{14} in favor of the proposition that τ_5 does not equal 0. The summed posterior probability for the models in which $\tau_5 = 0$ equals 0.000. The middle panel of Figure S7-19 shows the model-averaged posterior distribution for τ_5 across all eight models, where the height of the spike at zero corresponds to the summed posterior probability that $\tau_5 = 0$. In sum, for Q5 the *unfiltered* data provide overwhelming evidence for the hypothesis that there is across-team heterogeneity.

Next we present the results of the *filtered* data. The Bayes factor and the posterior model odds both equal 102.954 in favor of the proposition that τ_5 does not equal 0. The summed

posterior probability for the models in which $\tau_5 = 0$ equals 0.010. The middle panel of Figure S7-20 shows the model-averaged posterior distribution for τ_5 across all eight models, where the height of the spike at zero corresponds to the summed posterior probability that $\tau_5 = 0$. In sum, for Q5 the *filtered* data provide compelling evidence for the hypothesis that there is across-team heterogeneity.

Quantifying the Effect of Population for Q5. First we present the results of the *unfiltered* data. The Bayes factor and the posterior model odds both equal 1.461 in favor of the proposition that $\delta_{pop,5} = 0$ does not equal 0. The summed posterior probability for the models in which $\delta_{pop,5} = 0$ equals 0.406. The lower panel of Figure S7-19 shows the model-averaged posterior distribution for $\delta_{pop,5}$ across all eight models, where the height of the spike at zero corresponds to the summed posterior probability that $\delta_{pop,5} = 0$. In sum, for Q5 the *unfiltered* data provide weak evidence for the hypothesis that the MTurk population and the PureProfile population have different effect sizes.

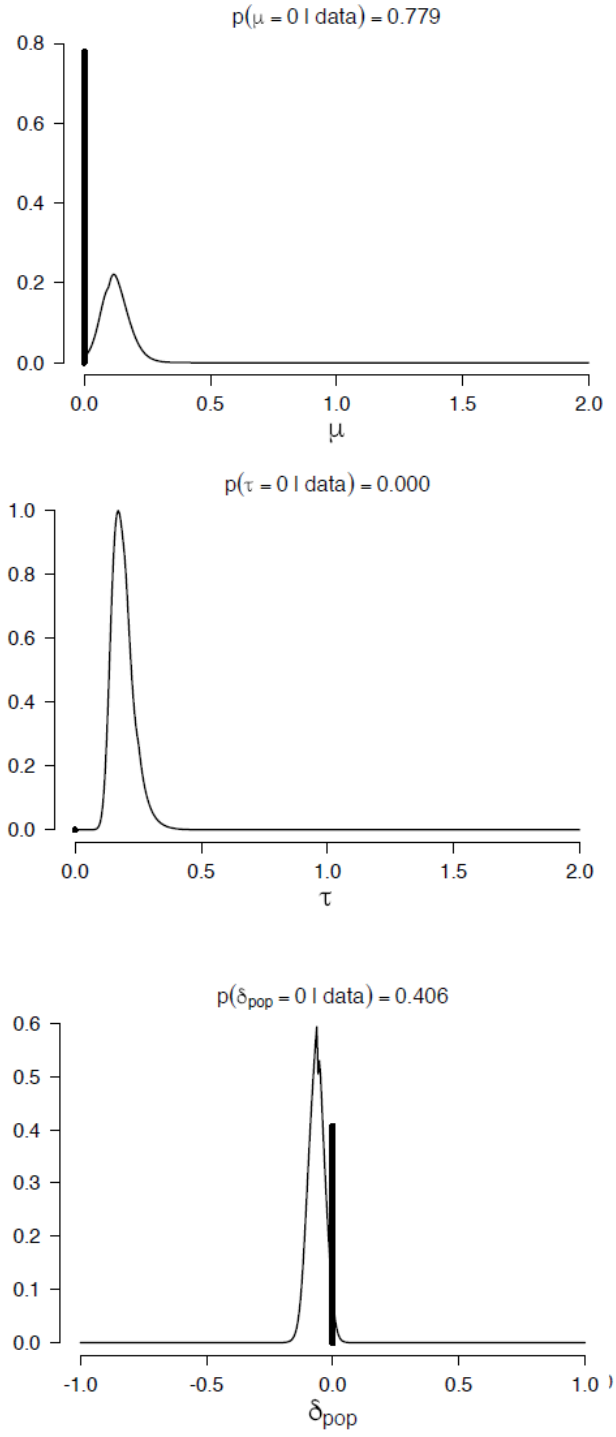


Figure S7-19. Model averaging results for Q5 (*unfiltered* data). The upper panel displays the results for the group-level mean effect size μ_5, the middle panel displays the results for the across-team heterogeneity τ_5, and the lower panel displays the results for the difference $\delta_{pop,5}$ between the MTurk and the PureProfile populations. Each panel shows the model-averaged posterior distribution for the parameter across all eight models, where the height of the spike at zero corresponds to the summed posterior probability that the parameter equals 0.

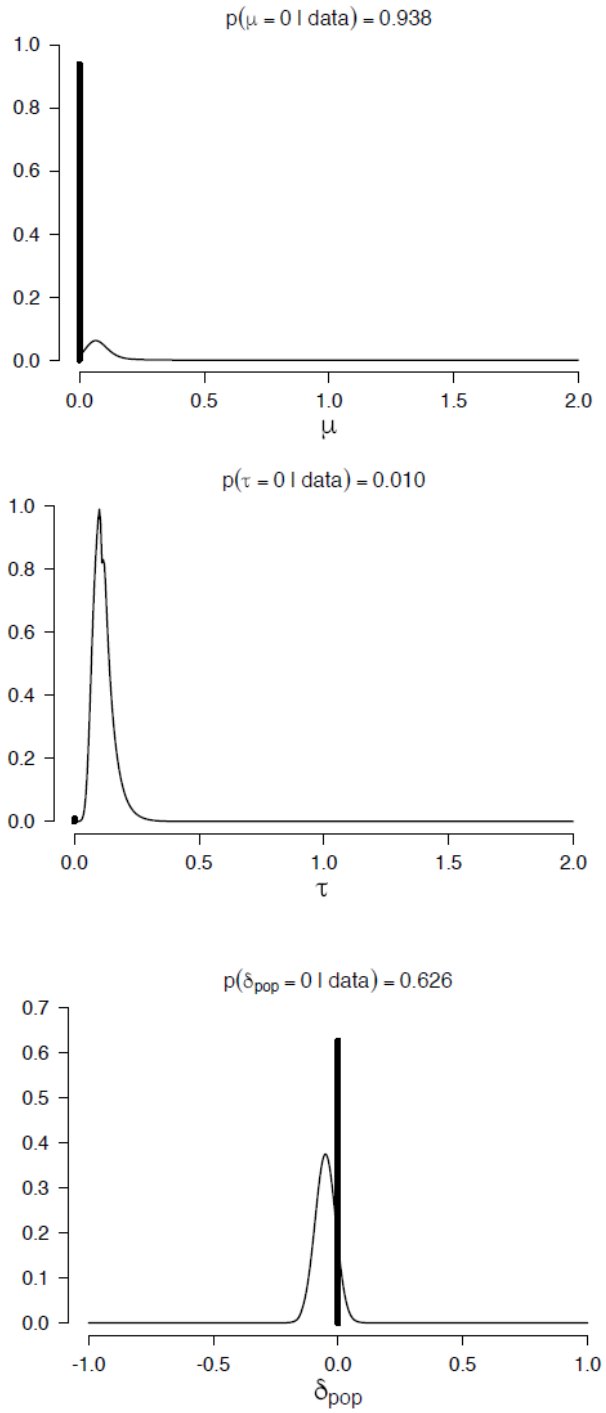


Figure S7-20. Model averaging results for Q5 (*filtered* data). The upper panel displays the results for the group-level mean effect size μ_5, the middle panel displays the results for the across-team heterogeneity τ_5, and the lower panel displays the results for the difference $\delta_{pop,5}$ between the MTurk and the PureProfile populations. Each panel shows the model-averaged posterior distribution for the parameter across all eight models, where the height of the spike at zero corresponds to the summed posterior probability that the parameter equals 0.

Next we present the results of the *filtered* data. The Bayes factor and the posterior model odds both equal 1.673 in favor of the proposition that $\delta_{pop,5}$ equals 0. The summed posterior probability for the models in which $\delta_{pop,5} = 0$ equals 0.626. The lower panel of Figure S7-20 shows the model-averaged posterior distribution for $\delta_{pop,5}$ across all eight models, where the height of the spike at zero corresponds to the summed posterior probability that $\delta_{pop,5} = 0$. In sum, for Q5 the *filtered* data provide weak evidence for the hypothesis that the MTurk population and the PureProfile population have the same effect size.

Results for Goal 3: Using the Bayesian ANOVA to Quantify the Evidence for or against a Lab Effect

To study whether or not there is a lab effect, we performed a Bayesian ANOVA in JASP (JASP Team, 2018) as described in the preregistration document (https://osf.io/9jzy4/).

For the unfiltered analyses, we first computed a new variable "standardisedAcrossAll" in JASP (version 0.9.1, or higher). This new variable centres the raw effect sizes at the overall mean (across questions q, labs l and populations p) and scaled with respect to the standard errors and sample sizes.[7]

The new variables was then specified as the dependent variable in a Bayesian ANOVA with random effect Lab, and fixed effects Question and Pop. The fixed effects Question and Pop, as well as its interaction were included in the null model, under the "Model" tab. The results of this analysis is summarized in Table S7-3. The Bayes factor of $BF_{01} = 12.03$ (2.69 % error) indicates evidence for absence over presence of a lab effect. A similar conclusion can be drawn

[7] For the reported two-sample tests the effective sample size were used.

from the descriptive plot of Figure S7-21, which plots the latent abilities of each lab separated by question with a 95% credible interval.

Table S7-3: Model Comparison – Standardized Across All

Models	P(M)	P(M|data)	BF_M	BF_{01}	Error %
Null model (incl. Pop, Question, Pop*Question)	0.50	0.92	12.03	1.00	
Lab	0.50	0.08	0.08	12.03	2.69

In addition to the analysis summarized in Table S7-3, we also performed the same analysis based on the unstandardized effect sizes. The results are provided by Table S7-4 and note that the evidence in favor of absence over presence of lab effect increases: $BF_{01} = 38.26$ (1.88 % error).

To explore whether any of the factors are relevant for the data at hand, we reran the Bayesian ANOVA (e.g., van den Bergh et al., 2019), but this time without adding terms to the null model, which includes only an intercept term. For this analysis, we considered ten models, which are listed in the left-most column of Table S7-5. Each of these ten models were given a prior probability of $P(M) = 0.10$, as shown in the second column. The third column shows the posterior model probability, that is, the probability for the model after data observation. For instance, the highest posterior probability $P(M|data) = 0.56$ is given to the model that, on top of the intercept term, also includes a main effect for the factor Question. The evidence of the "Question"-model relative to the null model can be found in the fourth column, whereas the evidence for the "Question"-model relative to all other models can be found in the third column. The fourth column shows that the model that includes a main effect for Question is $BF_{10} = 7.60e + 6$ times more likely than the intercept only model. In addition, the third column shows that the evidence for the "Question" model against all other models is increased by a factor of

$BF_M = 11.25$, that is, $P(M|data)/(1 - P(M|data) = 0.56/0.44$ divided by $P(M)/(1 - P(M) = 0.10/0.90$.

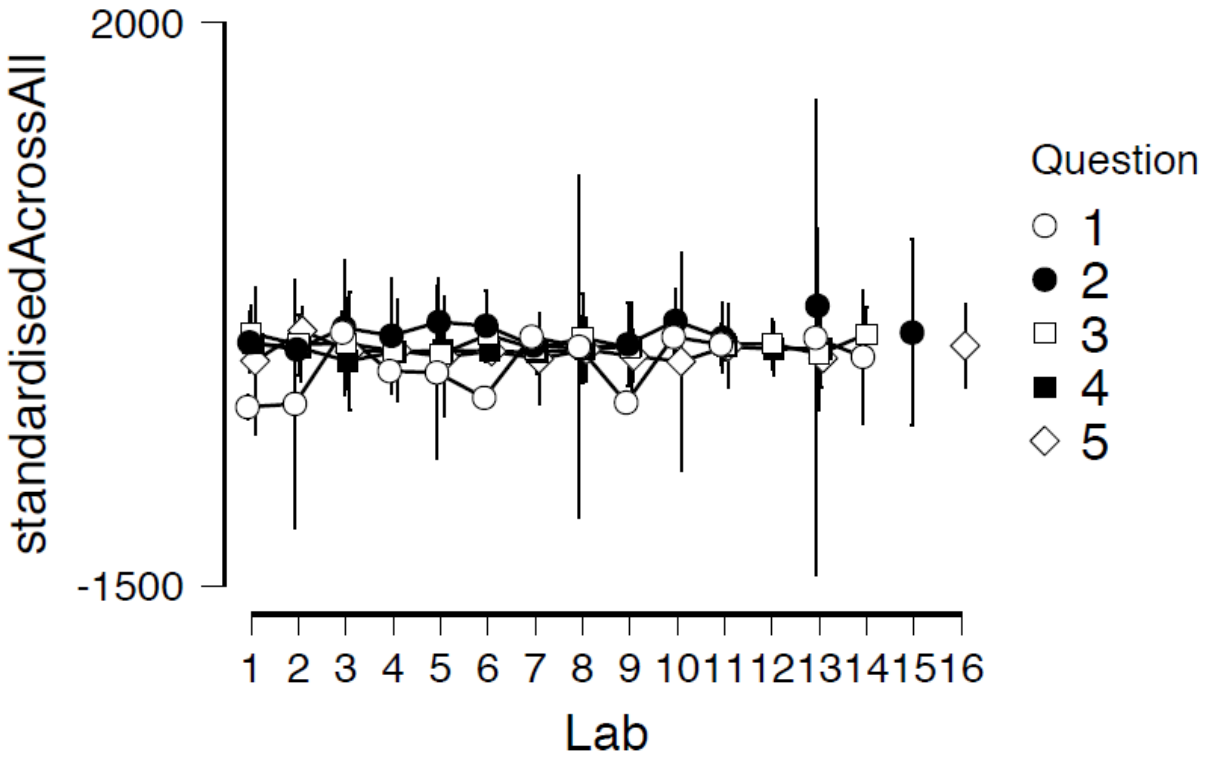


Figure S7-21. Descriptives plot with 95% credible interval, separated by questions and lab on the horizontal axis. Note that not all labs designed studies for all five questions.

Table S7-4: Model Comparison – Effect Size

Models	P(M)	P(M|data)	BF_M	BF_{01}	Error %
Null model (incl. Pop, Question, Pop*Question)	0.50	0.97	38.26	1.00	
Lab	0.50	0.03	0.03	38.26	1.88

Note that the factor Question appears in several models and one might be interested to study how effective it is to include this factor across all these models. Table S7-6 shows that the data indicate evidence in favor of including the factor Question to a model, i.e., $BF_{Inclusion} = 9.13e+6$, whereas the inclusion Bayes factor of $BF_{Inclusion} = 0.63$ and $BF_{Inclusion} = 0.07$ indicate evidence for excluding the factors Population and Lab, respectively, since they are both smaller than one. Hence, our exploratory analysis shows that most of the variability within the data can be explained by the factor Question alone.

Table S7-5: Model Comparison – Standardized Across All

Models	P(M)	P(M|data)	BF_M	BF_{10}	Error %
Null model	0.10	7.31e-8	6.58e-7	1.00	
Pop	0.10	3.20e-8	2.88e-7	0.44	7.63e-3
Question	0.10	0.56	11.25	7.60e+6	0.01
Pop + Question	0.10	0.35	4.83	4.78e+6	4.52
Pop + Question + Pop*Question	0.10	0.03	0.28	408953.57	1.84
Lab	0.10	6.08e-10	5.47e-9	8.31e-3	1.64e-4
Pop + Lab	0.10	2.78e-10	2.51e-9	3.81e-3	1.27
Question + Lab	0.10	0.04	0.35	507264.33	0.87
Pop + Question + Lab	0.10	0.03	0.24	357453.87	2.32
Pop + Question + Pop*Question + Lab	0.10	2.30e-3	0.02	31463.25	1.60

Table S7-6: Analysis of Effects – Standardized Across All

Models	P(M)	P(M|data)	BF_M
Pop	0.40	0.38	0.63
Question	0.40	0.97	9.13e+6
Lab	0.50	0.07	0.07
Pop*Question	0.20	0.03	0.09

Results for Goal 3: Filtered data

We reran the previously presented analyses with only study designs rated five or higher. This was done by activating a Filter in JASP and by computing a new variable "standardisedAcrossBetter." After filtering out the studies that were rated less than five, the evidence in favor of absence over presence of lab effect goes down from $BF_{01} \approx 12$ to $BF_{01} \approx 1$ $BF_{01} \approx 1$. The Bayes factor of $BF_{01} \approx 1.40$ with 2.67% error indicates neither evidence for or against a lab effect, see Table S7-7. Hence, restricting the confirmatory analysis to studies that were rated higher than five does not lead to evidence for a lab effect, see also Figure S7-22.

As before, to explore whether any of the factors are relevant for the filtered data, we reran the Bayesian ANOVA, but this time without adding additional terms into the null model. The results are summarized in Table S7-8, which shows that the "Question"-model is $BF_{10} = 238{,}598.75$ times more likely than the intercept only model. Similarly, Table S7-8 shows that after seeing the data, the evidence in favor of including the factor Question in the model went up, i.e., $BF_{Inclusion} = 349{,}323.04$, whereas the inclusion Bayes factors $BF_{Inclusion} = 0.91$ and $BF_{Inclusion} = 0.57$ indicate (little) evidence for excluding the factors Population and Lab, respectively, since they are both smaller than one. Hence, as before our exploratory analysis shows that most of the variability within the data can be explained by the factor Question alone.

Table S7-7: Model Comparison – Standardized Across Better

Models	P(M)	P(M|data)	BF_M	BF_{01}	Error %
Null model (incl. Pop, Question, Pop*Question)	0.50	0.58	1.40	1.00	
Lab	0.50	0.42	0.71	1.40	2.67

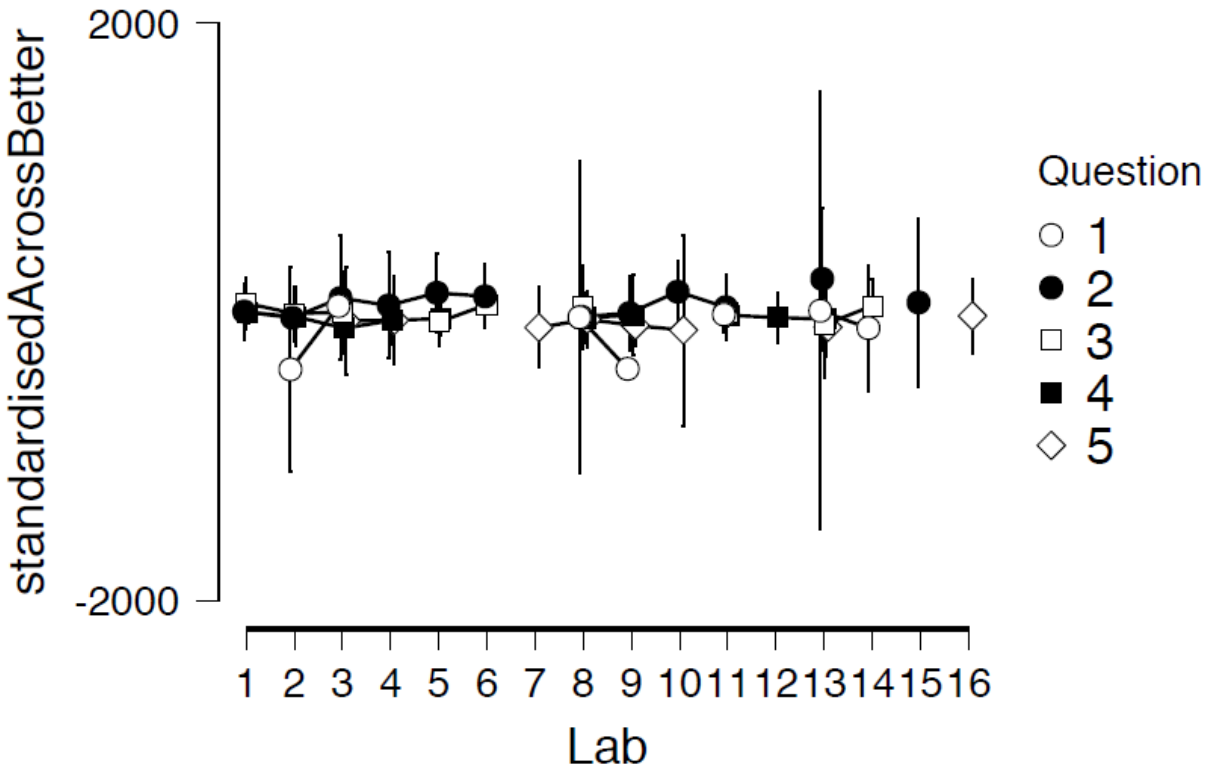


Figure S7-22. Descriptives plot with 95% credible interval, separated by questions and lab on the horizontal axis, based on studies that were rated five or higher.

Remaining Concerns

- Lab 16 is just Lab 7, but with the original materials for Question 5. This is unusual, especially when we want to test the effect of lab. Removing Lab 16 does not qualitatively change the results. Performing the analyses on only Labs 1 to 9, which designed materials for all five studies, also did not qualitatively change the results.

Table S7-8: Analysis of Effects – Standardized Across Better

Effects	P(incl)	P(incl|data)	$BF_{Inclusion}$
Lab	0.50	0.36	0.57
Pop	0.40	0.44	0.91
Question	0.40	0.93	349323.04
Pop*Question	0.20	0.07	0.15

- Q5: For the conversion from r to d, a point-biserial transformation is used, which assumes that one of the two continuous variables is dichotomised. This is unusual. The standard set-up is to Fisher *z*-transform the data. For the ANOVA test it would possibly be preferred to use the standard transformation from r to d instead.

- For the transformation used for the ANOVA we used the effect sample sizes instead of the sample sizes of the two groups.

References for Supplement 7

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

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

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

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

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

Meng, X.-L., & Wong, W. H. (1996). Simulating ratios of normalizing constants via a simple identity: A theoretical exploration. *Statistica Sinica, 6*, 831–860.

Morey, R. D., & Rouder, J. N. (2015). *BayesFactor 0.9.111.* Comprehensive R Archive Network. Retrieved from http://cran.r-project.org/web/packages/BayesFactor/index.html

R Core Team. (2018). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from https://www.R-project.org/

Scheibehenne, B., Gronau, Q. F., Jamil, T., & Wagenmakers, E.-J. (2017). Fixed or random? A resolution through model-averaging. Reply to Carlsson, Schimmack, Williams, and Burkner. *Psychological Science, 28*, 1698–1701.

Stan Development Team. (2018). *RStan: the R interface to Stan.* Retrieved from http://mc-stan.org/ (R package version 2.17.3)

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

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

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

Wagenmakers, E.-J., Marsman, M., Jamil, T., Ly, A., Verhagen, A. J., Love, J., ... Morey, R. D. (2018). Bayesian inference for psychology. Part I: Theoretical advantages and practical ramifications. *Psychonomic Bulletin & Review*, *25*(1), 35–57. doi: https://doi.org/10.3758/s13423-017-1343-3

SUPPLEMENT 8 - Multivariate meta-analysis of Main Study and Replication

In lieu of separately meta-analyzing each of the five hypotheses addressed in this paper, we can also consider each hypothesis as a potentially related outcome observed for a given individual. This results in a multivariate (multiple outcome) rather than a univariate (single outcome) meta-analysis. Multivariate meta-analysis can both improve the efficiency of estimation and allow for inference across outcomes (Riley et al., 2015). Because of the merit and applicability of multivariate meta-analysis to the several hypotheses of this paper, we also report results for such a model.

The nature of the experimental strategy employed in the Crowdsourcing Hypothesis Tests initiative has particular implications for the design of the multivariate meta-analysis. Broadly speaking, meta-analysis can be divided into two types of estimation approaches. The more common two-stage approach consists of a first-stage, in which effect sizes (and sampling variances) are calculated for each constituent study, and a second-stage, in which the effect size data from the first-stage is analyzed using random effects. Because the second-stage relies only on summary data, researchers can often perform meta-analysis on published studies wherein individual participant data is not available. When such data is available, as in the current analysis, one can instead perform a one-stage meta-analysis, in which a single mixed effects regression simultaneously estimates all of the study-specific effect sizes and produces the desired meta-analytic statistics.

Typical meta-analyses consider distinct “studies”, wherein research teams estimate effect sizes for the same research question using distinct samples of individuals. For multivariate analysis, outcomes for a given individual would all come from the same study, such that participants are nested in studies. Two-stage meta-analyses as a result conventionally estimate

effects assuming that there is no sample overlap across studies, and that all outcomes for a given study are observed for each individual present within the study. For the present paper, however, participants are re-randomized into a potentially different team's research design ("study" in the usual sense) for each hypothesis, hence neither of the usual two-stage assumptions about samples within and between studies are justified. As a consequence, we instead employ one-stage meta-analysis, allowing us to take into consideration correlations in individual outcomes that are non-nested within research designs.

In order to simultaneously estimate effect sizes in the one-stage model, we reparametrize effect sizes such that they are identified in a regression setting and are consistent across hypotheses. Cohen's *d* in particular does not directly emerge from regression analysis. A conceptually similar standardized effect size in regression settings with a single explanatory variable is the regression coefficient produced when first rescaling the dependent variable to have unit variance. For binary treatments, this regression coefficient (hereafter "standardized beta") estimates the mean difference in the outcome expressed in terms of the pooled standard deviation. We therefore perform this rescaling for all dependent variables in the model (separately for Main Studies and Replication Studies) in order to achieve the similar interpretation to Cohen's *d* of standardized beta.

Pearson's correlation, on the other hand, can be derived in univariate regression settings by first standardizing (demeaning and rescaling the variance) both the dependent and independent variables. For Hypothesis 5, which focuses on the correlation between personal happiness and moral judgments, we have hence centered both the dependent and independent variables and rescaled the independent variable measure, in addition to the rescaling of the dependent variable common to other effect sizes.

For many of the research designs, there is no experimental variation in treatment and effect sizes are instead defined as the within-individual difference in outcomes under two conditions. In a single study setting, regression analysis might estimate the effect size by regressing the difference in outcomes under the two conditions on a constant. Combining this design, wherein treatment is a constant, with the between-study design, where there is both an intercept and a multi-valued treatment, poses the typical econometric problem of collinearity. In order to address collinearity of trial specific intercepts with treatment terms that do not vary within trial (for many designs), intercepts are demeaned for the regression specification. Since between-subjects designs exclusively employ a binary treatment variable, the mean outcome for the control group is subtracted from the dependent variable. No adjustment for the treatment variable is needed since the treatment for the control group is valued as zero.

Having reparametrized the model, mixed effects regression then estimates the one-stage multivariate meta-analysis. Trials in the one-stage model (equivalent to "studies" in conventional meta-analyses) are defined as the factorial combination of research team and data collection effort (Main Studies or Replication Studies). For individual *i*, hypothesis *h*, and trial *j*, the model is specified as:

$$y_{ihj} = \beta_{hj}\left(\mathbb{1}\{\boldsymbol{h_j}\} \times T_{hj}\right) + e_{ihj},$$

where:

$$\beta_{hj} = \gamma_h + u_{hj},$$

$$u_{hj} \sim N(0, \Sigma_u) \text{ and } e_{ihj} \sim N(0, \Sigma_e)$$

The variance matrix of the trial-specific random effects specifies unstructured correlations between random effects for different hypotheses in a trial. The variance matrix of the individual residual specifies an unstructured individual correlation in outcomes across hypotheses.

Figures S8.1 through S8.5 report estimated standardized beta effect sizes by trial for each hypothesis in the one-stage multivariate analysis.[8] For Hypotheses 1-4, these standardized betas can be interpreted similarly to Cohen's *d*s, and for Hypothesis 5, it can be interpreted similarly to Pearson's *r*.

Table S8.1 reports for each hypothesize the overall estimated effect size, τ^2 (variance in effect sizes between different trials), and I^2 heterogeneity statistics. The qualitative findings for effect size direction and significance are similar to the two-stage, univariate models reported in the main text. Estimated between-study variances in effect sizes, τ^2, are also similar to those found in the univariate analysis and are again generally large relative to the effect size estimates. For all outcomes except for Hypothesis 5, the standard deviation of estimated effect size heterogeneity, τ, is larger in magnitude than the estimated effect size. Estimated I^2 statistics are likewise similar to the two-stage univariate models, with more than 80% of estimation variation in effect sizes arising due to between-study heterogeneity in effect sizes rather than sampling variance in the fixed effect size estimates.

In principal, a key advantage of multivariate meta-analysis would be to allow for an exploration of between-study heterogeneity jointly across all hypotheses. Due to unique features of this project, however, including the principal consideration that participants are non-nested in studies, the meta-analytic design does not readily lend itself to simple expressions of heterogeneity arising from either the variance partitioning approaches general to the mixed effects model or the direct multivariate extension of the I^2 formulated for non-nested aggregate data. Hence we restrict ourselves to the analysis of heterogeneity by hypothesis found above.

[8] Formally, τ^2 are the principal diagonals of Σ_u, the variance-covariance matrix for the random components, u_{hj}, of the treatment effects.

Given that results are very similar between the multivariate and univariate models and that the univariate model has a more conventional and readily interpretable structure, we moreover prefer to focus the primary presentation on the univariate meta-analyses present in the main text.

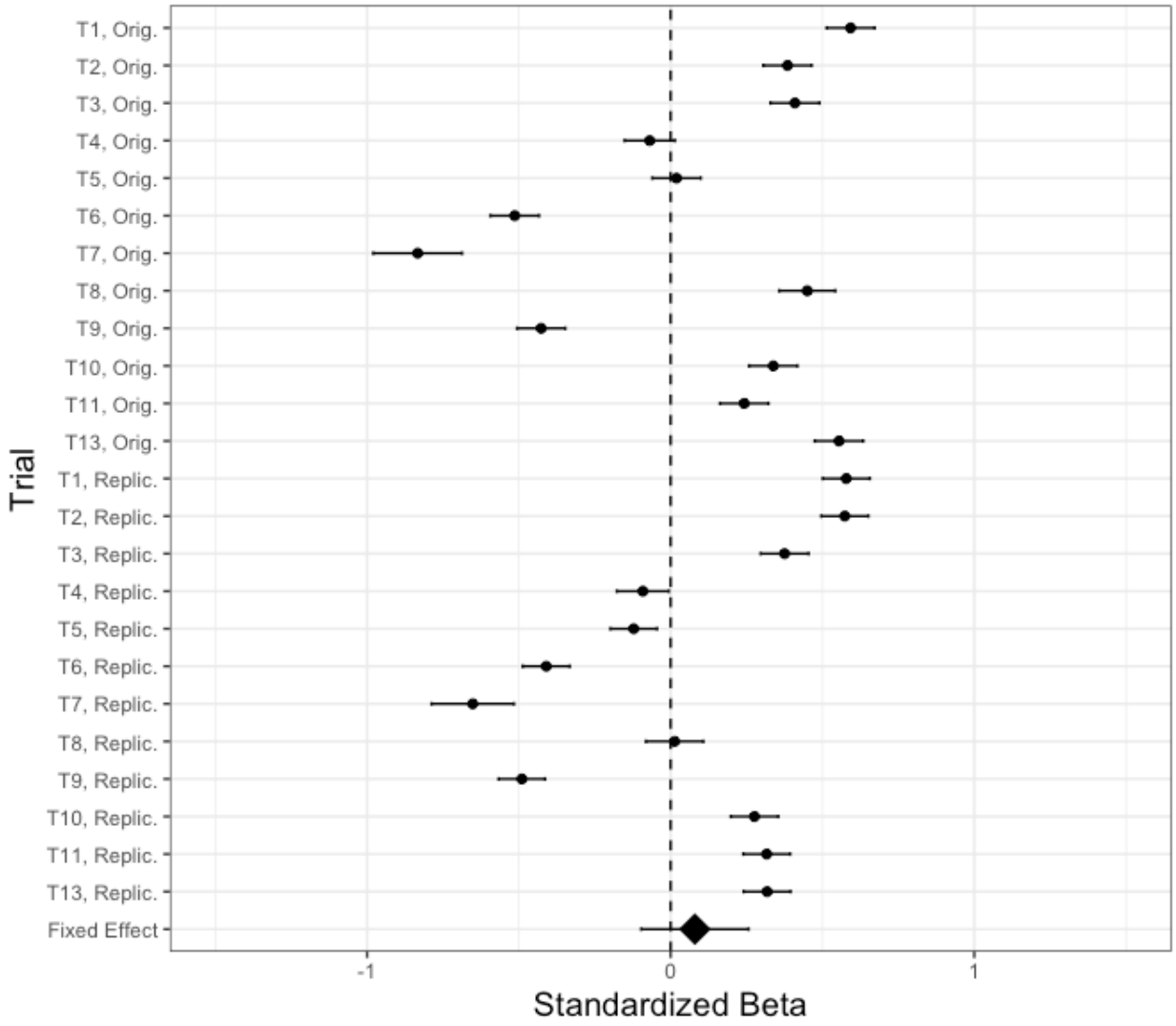


Figure S8.1. Estimated standardized beta effect sizes for each trial, Hypothesis 1. The research question was "When directly asked, do people explicitly self-report an awareness of harboring negative automatic associations with members of negatively stereotyped social groups?"

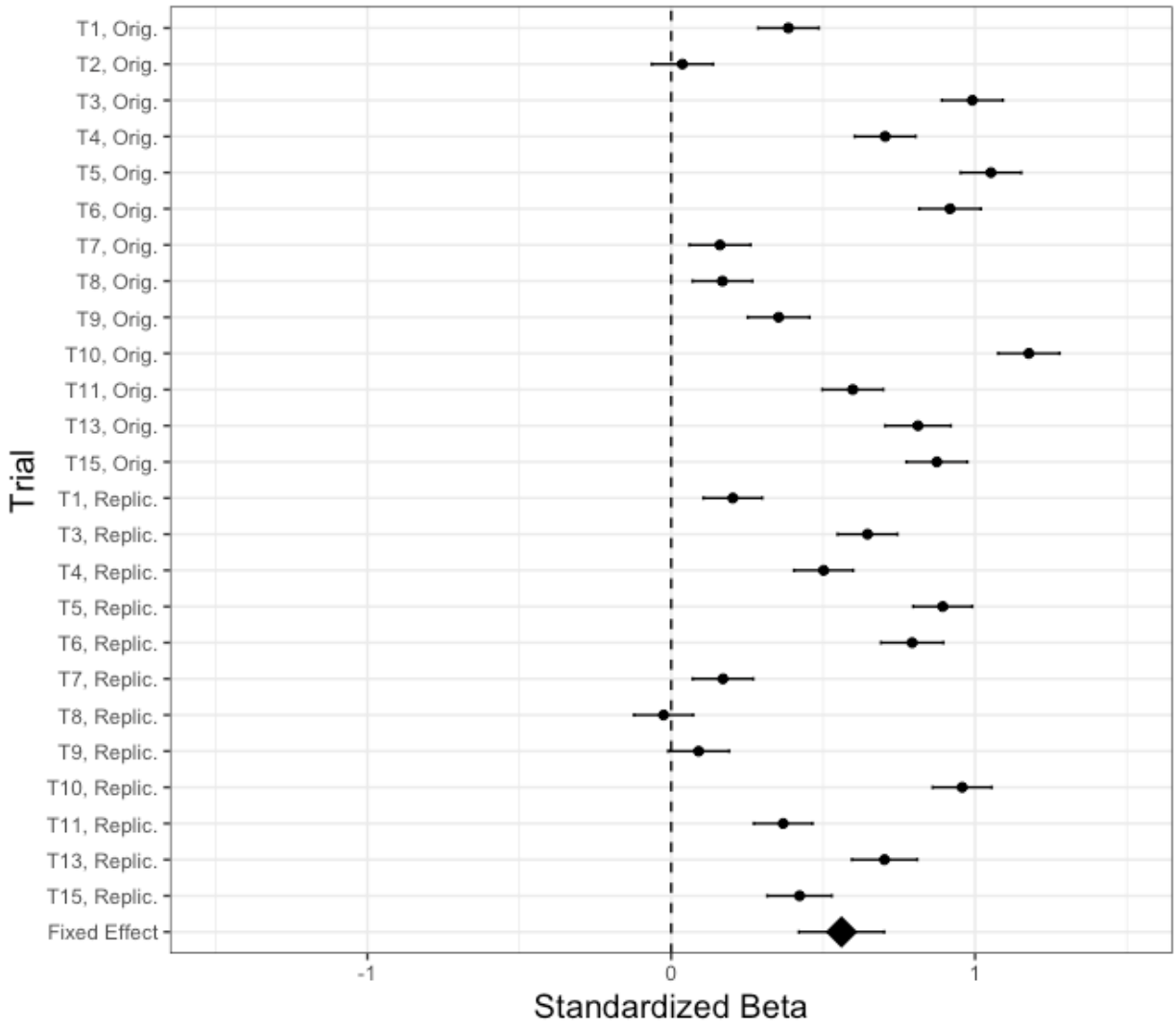


Figure S8.2. Estimated standardized beta effect sizes for each trial, Hypothesis 2. The research question was "Are negotiators who make extreme first offers trusted more, less, or the same relative to negotiators who make moderate first offers?"

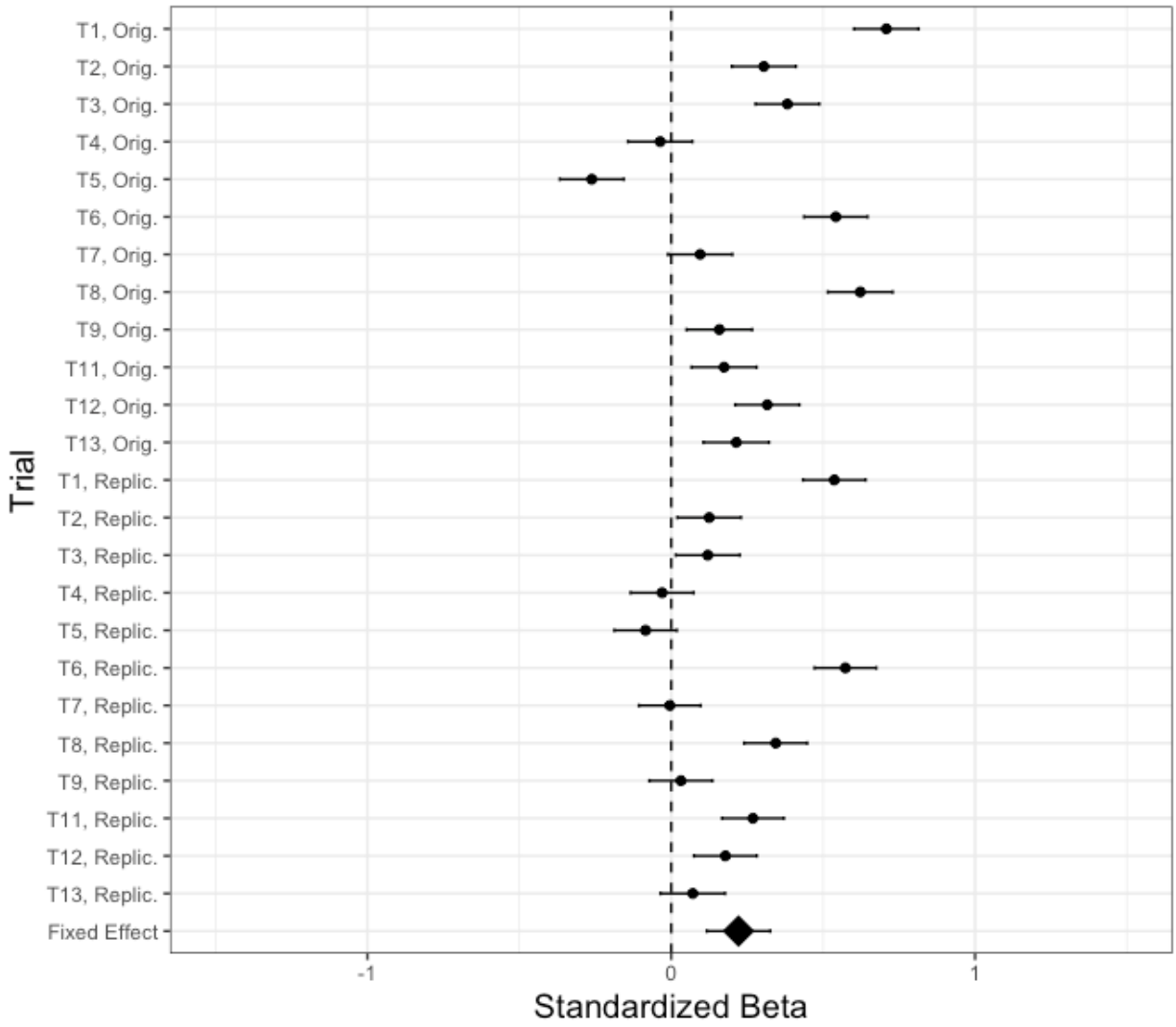


Figure S8.3. Estimated standardized beta effect sizes for each trial, Hypothesis 3. The research question was "What are the effects of continuing to work despite having no material/financial need to work on moral judgments of that individual - beneficial, detrimental, or no effect?"

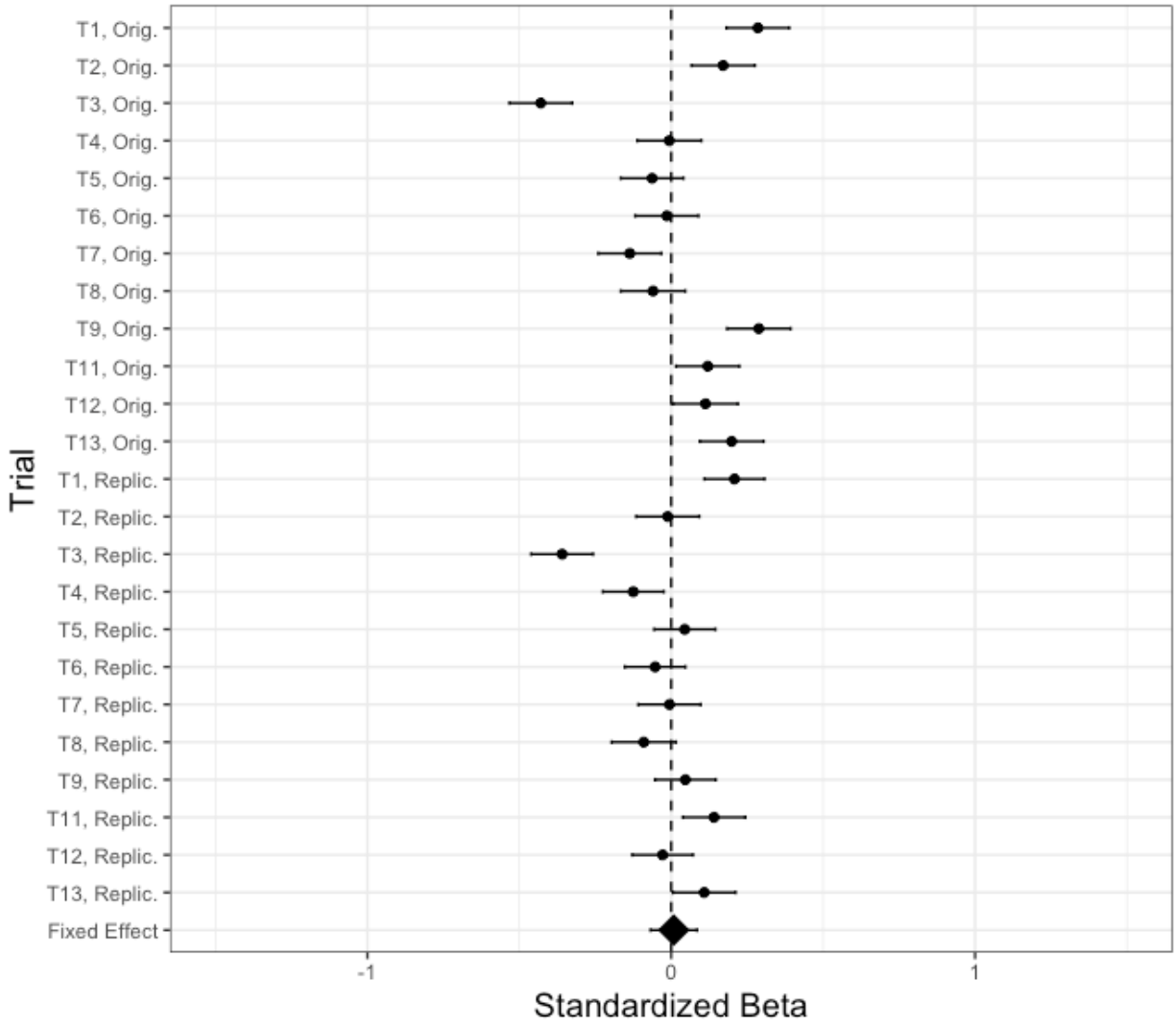


Figure S8.4. Estimated standardized beta effect sizes for each trial, Hypothesis 4. The research question was "Part of why people are opposed to the use of performance enhancing drugs in sports is because they are 'against the rules'. But which contributes more to this judgment - whether the performance enhancer is against the law, or whether it is against the rules established by a more proximal authority (e.g., the league)?"

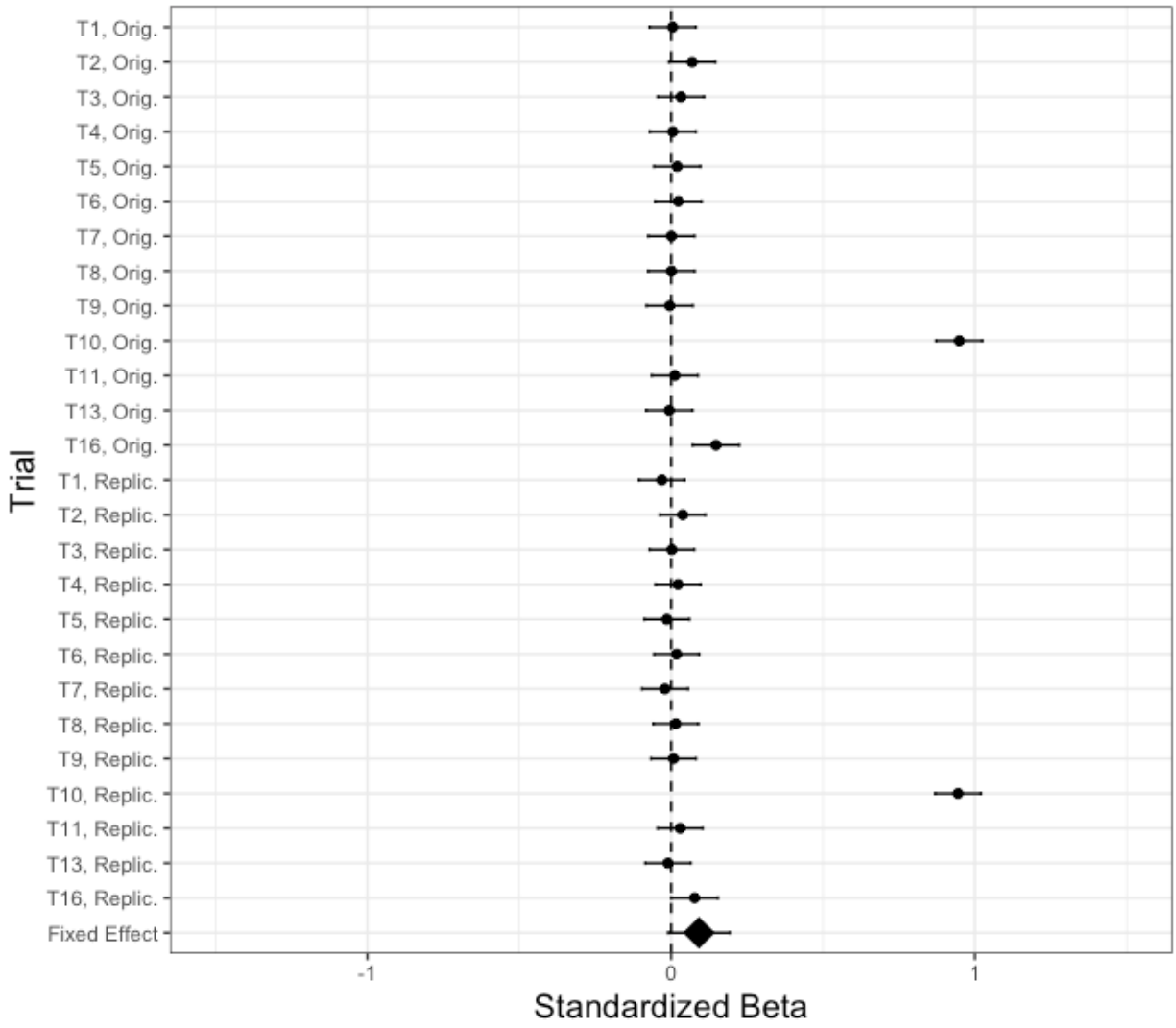


Figure S8.5. Estimated standardized beta effect sizes for each trial, Hypothesis 5. The research question was "Is a utilitarian vs. deontological moral orientation related to personal happiness?"

Table S8.1. Standardized effect sizes, I^2, and τ^2 statistics from multivariate meta-analyses (pooling Main Study and Replication).

Hypothesis	Description	Effect Size [95% *CI*]	I^2	τ^2
1	Awareness of automatic prejudice	β = 0.08 [-0.10, 0.26]	96.52%	0.20
2	Extreme offers reduce trust	β = 0.56 [0.42, 0.70]	95.33%	0.13
3	Moral praise for needless work	β = 0.22 [0.12, 0.33]	89.59%	0.07
4	Proximal authorities drive legitimacy of performance enhancers	β = 0.01 [-0.07, 0.08]	81.09%	0.03
5	Deontological judgments predict happiness	β = 0.09 [-0.01, 0.19]	90.44%	0.07

Reference for Supplement 8

Riley, R. D., Price, M. J., Jackson, D., Wardle, M., Gueyffier, F., Wang, J., … White, I. R. (2015). Multivariate meta-analysis using individual participant data. *Research Synthesis Methods, 6*(2), 157–174. doi:10.1002/jrsm.1129

SUPPLEMENT 9 – Additional analyses of Main Studies and Replication Studies

Descriptive Statistics

Table S9.1 presents descriptive statistics for each set of materials in the Main Studies and Replication Studies. It is important to note that the means cannot be directly compared across different materials sets, as different designs employed different dependent variables (see Supplement 1 for details).

Table S9.1. Descriptive statistics for all sets of materials, Main Studies and Replication Studies.

	Hypothesis 1					
	Main Studies			Replication Studies		
Research Team	Attitudes Toward Stigmatized Groups	Attitudes Toward Non-Stigmatized Groups	d_{IG}	Attitudes Toward Stigmatized Groups	Attitudes Toward Non-Stigmatized Groups	d_{IG}
Bauman & Mullen	-0.48 (1.24)	-	-0.39	-0.91 (0.95)	-	-0.95
Donnellan, Lucas, Cheung, & Johnson	-0.65 (1.05)	-	0.62	-0.09 (0.82)	-	0.11
Hahn & Dohle	0.61 (1.48)	-	0.42	0.56 (1.36)	-	0.41
Hall & Sowden	1.68 (0.89)	2.77 (0.98)	0.42	1.98 (1.16)	2.85 (1.06)	0.79
Jia & Ding	-.00 (0.50)	-	-0.01	-0.08 (0.49)	-	-0.15
Jiménez-Leal & Montealegre	-0.80 (1.54)	-	-0.52	-1.01 (2.00)	-	-0.50
Landy, Walco, & Bartels	2.95 (1.79)	2.57 (1.69)	0.22	2.99 (1.99)	2.43 (1.79)	0.29
Monin & Reynolds	0.58 (1.23)	-	0.47	-0.02 (1.10)	-	-0.02
Pang	-0.25 (0.38)	-	-0.66	-0.20 (0.40)	-	-0.50
Uhlmann & Cunningham (Original Materials)	0.26 (1.61)	-	0.16	0.08 (1.63)	-	0.05
Van Bavel, Ray, Reinero, Brady, & Wills	-0.85 (1.24)	-	-0.69	-0.84 (1.27)	-	-0.66
Xu & Yang	0.40 (1.24)	-	0.32	0.66 (1.92)	-	0.35
Yam, Koh, & Su	-0.19 (1.89)	-	-0.10	-0.21 (1.75)	-	-0.12

	Hypothesis 2					
	Main Studies			Replication Studies		
Research Team	Moderate Offer	Extreme Offer	d_{IG}	Moderate Offer	Extreme Offer	d_{IG}
Bauman & Mullen	74.74 (93.10)	68.89 (93.88)	0.06	89.66 (98.44)	88.43 (100.70)	0.01
Donnellan, Lucas, Cheung, & Johnson	4.72 (1.03)	2.34 (1.06)	2.29	4.54 (1.07)	2.87 (1.29)	1.40
Hahn & Dohle	3.71 (1.16)	1.99 (1.05)	1.56	3.80 (1.32)	2.64 (1.61)	0.78
Hall & Sowden	5.30 (1.06)	5.09 (1.08)	0.20	5.10 (1.31)	4.82 (1.35)	0.20
Jia & Ding	4.44 (1.15)	2.34 (1.26)	1.74	4.22 (1.33)	2.43 (1.50)	1.26
Jiménez-Leal & Montealegre	5.45 (3.33)	5.08 (3.22)	0.42	5.88 (2.97)	5.52 (2.91)	0.12
Landy, Walco, & Bartels	4.08 (1.72)	2.81 (1.81)	0.72	3.97 (2.04)	3.10 (2.05)	0.43
Monin & Reynolds	3.88 (0.88)	3.69 (1.00)	0.20	3.65 (1.04)	3.65 (1.02)	0.00
Pang	4.74 (0.96)	3.17 (1.30)	1.38	4.72 (1.22)	3.34 (1.40)	1.06
Schweinsberg (Original Materials)	4.48 (1.09)	2.97 (1.34)	1.23	4.06 (1.55)	3.26 (1.72)	0.48
Van Bavel, Ray, Reinero, Brady, & Wills	4.61 (1.35)	3.99 (1.41)	0.45	4.69 (1.61)	4.30 (1.60)	0.24
Xu & Yang	4.90 (1.24)	2.00 (1.18)	2.41	4.72 (1.38)	2.62 (1.59)	1.42
Yam, Koh, & Su	3.18 (1.08)	2.20 (1.11)	0.89	3.38 (1.53)	2.69 (1.74)	0.44

	Hypothesis 3					
	Main Studies			Replication Studies		
Research Team	Needless Work	No Work or Necessary Work	d_{IG}	Needless Work	No Work or Necessary Work	d_{IG}
Bauman & Mullen	3.95 (0.68)	3.71 (0.66)	0.35	3.71 (0.71)	3.60 (0.69)	0.15
Cimpian, Tworek, & Storage	6.46 (1.51)	5.90 (1.55)	0.37	6.56 (1.72)	6.20 (1.82)	0.21
Donnellan, Lucas, Cheung, & Johnson	5.45 (0.96)	5.27 (1.00)	0.18	5.35 (1.17)	5.28 (1.11)	0.06
Hahn & Dohle	5.91 (1.06)	5.36 (1.33)	0.45	5.77 (1.34)	5.56 (1.35)	0.16
Hall & Sowden	0.71 (0.85)	0.61 (0.76)	0.13	0.80 (0.97)	0.77 (0.90)	0.03
Jia & Ding	5.08 (1.34)	5.39 (1.32)	-0.23	5.31 (1.54)	5.38 (1.44)	-0.05
Jiménez-Leal & Montealegre	5.15 (1.01)	4.94 (1.20)	0.19	4.92 (1.46)	4.83 (1.38)	0.06
Landy, Walco, & Bartels	6.90 (1.48)	6.59 (1.51)	0.21	7.12 (1.71)	6.57 (1.77)	0.31
Monin & Reynolds	1.41 (1.05)	0.70 (0.82)	0.75	1.13 (1.33)	0.66 (1.07)	0.39
Pang	5.71 (1.00)	5.00 (1.16)	0.66	5.79 (1.20)	4.87 (1.41)	0.70
Uhlmann (Original Materials)	0.38 (0.93)	-	0.40	0.83 (1.60)	-	0.52
Van Bavel, Ray, Reinero, Brady, & Wills	5.24 (1.33)	4.14 (1.10)	0.90	5.03 (1.58)	4.14 (1.77)	0.64
Yam, Koh, & Su	4.83 (0.95)	4.84 (0.87)	-0.01	4.76 (1.06)	4.76 (1.19)	0.00

	Hypothesis 4					
	Main Studies			Replication Studies		
Research Team	Banned But Legal	Illegal But Not Banned	d_{IG}	Banned But Legal	Illegal But Not Banned	d_{IG}
Bauman & Mullen	-.09 (0.89)	0.09 (0.79)	0.21	0.00 (0.82)	-0.00 (0.83)	0.01
Cimpian, Tworek, & Storage	6.19 (1.16)	6.01 (1.24)	0.15	5.98 (1.62)	5.97 (1.46)	0.00
Donnellan, Lucas, Cheung, & Johnson	5.05 (0.39)	4.66 (1.48)	0.24	5.16 (1.55)	4.94 (1.67)	0.13
Hahn & Dohle	5.29 (1.46)	5.88 (1.17)	-0.45	5.20 (1.66)	5.80 (1.58)	-0.37
Hall & Sowden	6.23 (1.12)	6.37 (1.07)	-0.13	6.07 (1.69)	6.03 (1.71)	0.02
Jia & Ding	6.13 (1.19)	6.16 (1.17)	-0.03	6.25 (1.25)	6.12 (1.34)	0.10
Jiménez-Leal & Montealegre	6.18 (1.37)	5.69 (1.60)	0.32	5.80 (1.79)	5.68 (1.94)	0.06
Landy, Walco, & Bartels (Original Materials)	6.26 (2.23)	5.88 (2.33)	0.17	7.17 (2.02)	6.79 (2.25)	0.18
Monin & Reynolds	5.32 (1.31)	5.39 (1.37)	-0.05	2.58 (1.62)	2.36 (1.57)	0.14
Pang	2.27 (1.64)	2.25 (1.55)	0.01	2.76 (1.74)	2.82 (1.90)	-0.03
Van Bavel, Ray, Reinero, Brady, & Wills	5.98 (1.17)	5.52 (1.56)	0.34	6.18 (1.40)	5.79 (1.53)	0.26
Yam, Koh, & Su	5.50 (1.65)	5.45 (1.84)	0.03	5.40 (1.62)	5.60 (1.67)	-0.12

	Hypothesis 5					
	Main Studies			Replication Studies		
Research Team	Morality Measure	Happiness Measure	*r*	Morality Measure	Happiness Measure	*r*
Bauman & Mullen	2.02 (0.92)	4.59 (1.46)	.29	2.03 (0.94)	4.78 (1.21)	.24
Donnellan, Lucas, Cheung, & Johnson	3.10 (1.86)	4.41 (1.57)	-.03	2.56 (1.77)	4.12 (1.40)	-.01
Hahn & Dohle	3.78 (2.12)	4.43 (1.59)	.10	3.28 (2.00)	4.19 (1.43)	.00
Hall & Sowden (Shortened Materials)	3.93 (1.50)	4.52 (1.54)	.00	3.94 (1.58)	4.23 (1.41)	-.08
Hall & Sowden (Original Materials)	10.38 (4.59)	0.00 (0.86)	.15	8.59 (4.80)	-0.00 (0.84)	.08
Jia & Ding	3.94 (2.30)	4.31 (1.53)	.04	3.78 (2.36)	4.28 (1.49)	-.06
Jiménez-Leal & Montealegre	0.27 (0.44)	4.73 (1.34)	-.01	0.22 (0.41)	4.77 (1.32)	.04
Landy, Walco, & Bartels	4.36 (2.72)	3.58 (0.83)	.07	3.48 (2.63)	3.62 (0.79)	.14
Monin & Reynolds	2.62 (1.31)	4.47 (1.49)	.04	2.78 (1.50)	4.27 (1.43)	.10
Pang	5.03 (1.50)	4.28 (1.59)	.06	4.59 (1.69)	4.32 (1.50)	.05
Van Bavel, Ray, Reinero, Brady, & Wills	4.51 (1.57)	5.23 (1.95)	.01	4.04 (1.75)	6.05 (2.10)	-.12
Xu & Yang	4.45 (1.53)	5.06 (1.24)	.04	4.94 (1.65)	5.10 (1.38)	-.17
Yam, Koh, & Su	0.56 (0.50)	4.68 (1.39)	.02	0.37 (0.48)	4.80 (1.22)	.11

Null Hypothesis Significance Tests

As a further examination of the consistency of results across different teams' materials, we tested Hypotheses 1–4 using *t*-tests relating the manipulated independent variables to the dependent variables (either single-sample, independent-samples, or repeated-measures *t*-tests, as appropriate), and Hypothesis 5 using statistical significance tests for Pearson correlations between measures of moral orientation and happiness. Table S9.2 presents a summary of the results of these tests.

Several results are noteworthy. First, all five original hypotheses were supported in the Main Studies when the original materials were used; this was likewise true for three out of the five hypotheses in the Replication Studies (the results for Hypotheses 1 and 5 were directionally consistent with the original findings, but not statistically significant, in this sample). Meta-analytically combining the results from the Main Studies and Replication Studies, all five hypotheses were supported when the original materials were used. This suggests that the original findings were not merely false positives. Of course, that an effect directly replicates with the original materials does not necessarily mean that it will conceptually replicate using alternative study designs and materials — an effect may be an artifact of the original methodology, or perhaps very closely tied to specific operationalizations. Therefore, we examined variability in results across different sets of study materials designed to test the same hypothesis, categorizing each outcome as directionally consistent or inconsistent with the original effect, and as statistically significant ($p < .05$) or not.

As seen in Table S9.2, in the Main Studies, overall support for Hypotheses 2 and 3 across the array of conceptual replications was fairly consistent, though two sets of materials did show reverse effects for Hypothesis 3, one of them statistically significant. Materials testing

Hypothesis 1 split evenly between consistent (7, including the original materials) and inconsistent (6) with the original finding, and materials testing Hypothesis 4 were quite variable in their results, with the modal outcome being statistically significant support for the original finding. Lastly, materials testing Hypothesis 5 tended to be in the same direction as the original finding, but most effects were not statistically significant at the $\alpha = 0.05$ level. Results were similar in the Replication Studies, which had identical materials and procedures but a new sample of research participants. Support for Hypotheses 2 and 3 was quite consistent, with each effect only reversing, statistically non-significantly, in one instance. Hypothesis 1 was again split between consistent (6, including the original materials) and inconsistent (7) results, most of them statistically significant. Hypothesis 4 again produced quite variable results, though in the Replication Studies, the modal outcome was non-significant statistical support for the original finding. Lastly, the results for Hypothesis 5 were quite variable in the Replication Studies, with several results inconsistent with the original finding and statistically significant.

Table S9.2. Summary of null hypothesis significance tests.

Main Studies					
Hypothesis	**Description**	**Consistent Results, $p < .05$**	**Consistent Results, $p > .05$**	**Inconsistent Results, $p > .05$**	**Inconsistent Results, $p < .05$**
1	Awareness of automatic prejudice	54% (7)	0% (0)	8% (1)	38% (5)
2	Extreme offers reduce trust	92% (12)	8% (1)	0% (0)	0% (0)
3	Moral praise for needless work	77% (10)	8% (1)	8% (1)	8% (1)
4	Proximal authorities drive legitimacy of performance enhancers	42% (5)	25% (3)	25% (3)	8% (1)
5	Deontological judgments predict happiness	23% (3)	62% (8)	15% (2)	0% (0)
Replication Studies					
Hypothesis	**Description**	**Consistent Results, $p < .05$**	**Consistent Results, $p > .05$**	**Inconsistent Results, $p > .05$**	**Inconsistent Results, $p < .05$**
1	Awareness of automatic prejudice	38% (5)	8% (1)	8% (1)	46% (6)
2	Extreme offers reduce trust	77% (10)	15% (2)	8% (1)	0% (0)
3	Moral praise for needless work	46% (6)	46% (6)	8% (1)	0% (0)
4	Proximal authorities drive legitimacy of performance enhancers	25% (3)	50% (6)	17% (2)	8% (1)
5	Deontological judgments predict happiness	31% (4)	31% (4)	23% (3)	15% (2)

Explaining heterogeneity in the Replication Studies

When we repeated the analyses reported in the main text (see "Explaining heterogeneity in effect sizes") for the Replication Studies' data, hypothesis again explained a moderate amount of variance in observed effect size, *ICC* = .32, 95% *CI* [.09, .82], whereas team again did not explain statistically significant variance, *ICC* = -.18, 95% *CI* [-.26, .00]. Meta-regression agreed with these results: Hypothesis 2 produced larger effect sizes than the median hypothesis (Hypothesis 4, in the Replication Studies), $\beta = 0.55$, $p < .001$, 95% *CI* [0.24, 0.86], but no team produced reliably larger effect sizes than the median team (Hall & Sowden), *p*s > .307. Also, after accounting for hypothesis and team, there was again still substantial and statistically significant residual heterogeneity, $Q(44) = 1026.54$, $p < .001$, $I^2 = 96.34\%$, 95% *CI* [94.68, 97.74], $\tau^2 = 0.14$, 95% *CI* [0.09, 0.23].

"Flair" Analyses, Restricted to Hypotheses 3-5

In the main text, we conclude that our results fail to support the "flair" hypothesis (Baumeister, 2016) that some researchers are simply more talented at developing materials that produce large effect sizes (see "Explaining heterogeneity in effect sizes"). However, it might be argued that we should only expect researchers to demonstrate flair in a specific area of research. That is, there may not be any reason to expect that a research team's effect size in one area of research (e.g., prejudice research, as in Hypothesis 1) would be at all related to their effect size in another area (e.g., negotiations research, as in Hypothesis 2). Therefore, we re-ran the analyses reported in the main text, but restricted to Hypotheses 3, 4, and 5, which are all in the same area of research (moral judgment). If some research teams are particularly good at designing study materials for moral psychology research, and other research teams are not, these analyses are capable of demonstrating this.

Once again, however, we did not find support for the flair hypothesis. Observed effect sizes were significantly related to the hypothesis being tested in the Main Studies, *ICC* = .21, 95% *CI* [.00, .93], but not in the Replication Studies, *ICC* = .18, 95% *CI* [-.01, .92], though the magnitudes of the intraclass correlation coefficients were not substantially different. More importantly, observed effect sizes were not predicted by the team that designed the materials in the Main Studies, *ICC* = -.10, 95% *CI* [-.38, .32], nor the Replication Studies, *ICC* = .09, 95% *CI* [-.25, .50]. Meta-regression agreed with these results. In the Main Studies, Hypothesis 3 produced larger effect sizes than the median hypothesis (Hypothesis 5), β = .21, 95% *CI* [.01, .41], *p* = .044, but no team produced significantly different effect sizes from the median team (Cimpian, Tworek, & Storage), *p*s > .19. In the Replication Studies, neither Hypothesis 3 nor Hypothesis 4 produced different effect sizes from the median hypothesis (Hypothesis 5), *p*s > .09, and no team produced different effect sizes from the median team (Cimpian, Tworek, & Storage), *p*s > .13.

Publication Bias Analyses

Figure S9.1 presents funnel plots, sorted by hypothesis and sample (Main Studies, Replication Studies, and aggregating across both data collection efforts). Because we have reported the full results for each study, from every set of materials designed for each hypothesis, we would expect symmetric funnel plots, indicating an absence of publication bias (Egger, Smith, Schneider, & Minder, 1997). This is what we observed for Hypothesis 3 (Egger's tests: Main Studies *z* = 1.08, *p* = .282; Replication Studies *z* = 0.71, *p* = .477; Aggregated *z* = 1.46, *p* = .143) and Hypothesis 4 (Egger's tests: Main Studies *z* = - 0.84, *p* = .404; Replication Studies *z* = -0.73, *p* = .464; Aggregated *z* = -1.04, *p* = .301).

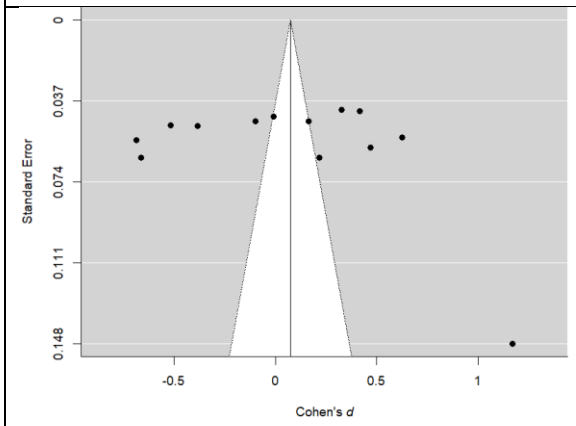
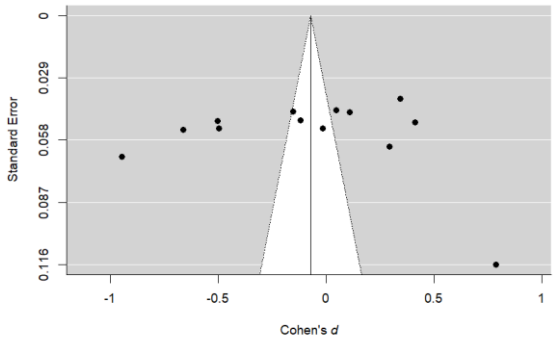
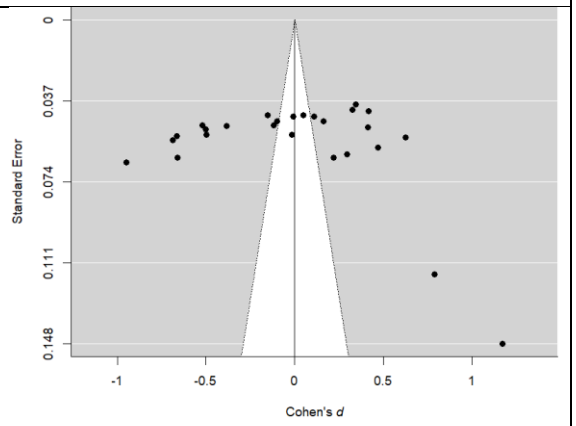
However, for Hypothesis 1, Egger's test suggests an asymmetric funnel plot for the Main Studies, $z = 2.03$, $p = .043$, no asymmetry in the Replication Studies, $z = 1.14$, $p = .256$, and significant asymmetry when aggregating across both data collection efforts, $z = 2.42$, $p = .016$. The observed asymmetries appear to be driven by the results from the team of Sowden and Hall, who operationalized "stigmatized groups" as political partisans with views opposing the participant's views. This design necessitated excluding participants with weak political affiliations, which resulted in a smaller sample size than other designs, but this design also resulted in the largest observed effect size for Hypothesis 1, producing the observed funnel plot asymmetry.

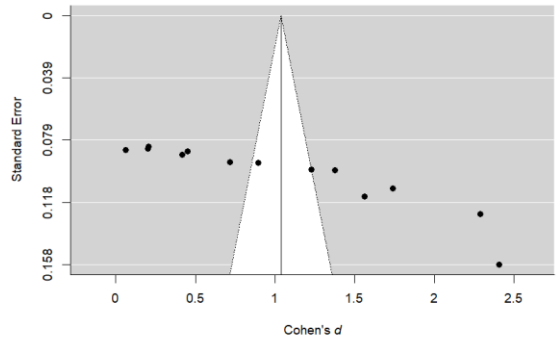
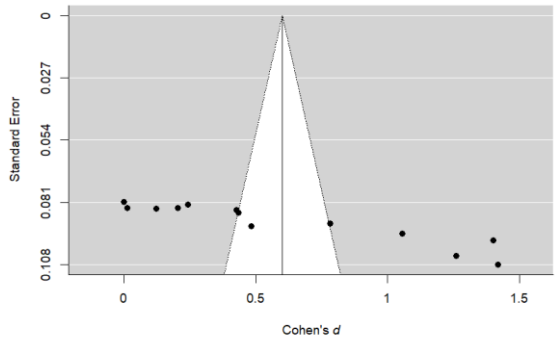
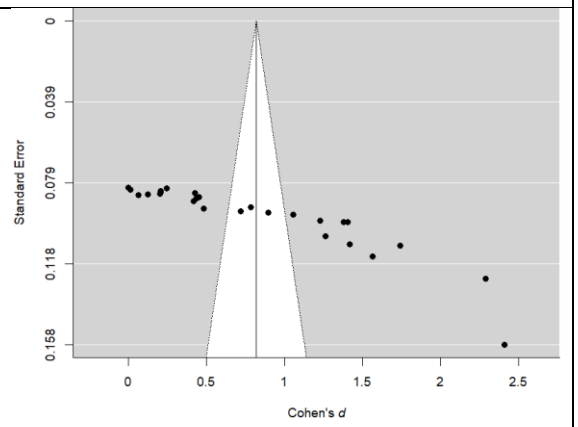
Moreover, for Hypothesis 2, Egger's test indicated funnel plot asymmetry in the Main Studies, $z = 7.15$, $p < .001$, the Replication Studies, $z = 8.73$, $p < .001$, and aggregating across both studies, $z = 11.10$, $p < .001$. For Hypothesis 5, there was no asymmetry in the Main Studies, $z = 1.41$, $p = .158$, significant asymmetry in the Replication Studies, $z = -2.65$, $p = .008$, and significant asymmetry when aggregating across both, $z = -.357$, $p < .001$. Note that for Hypothesis 5, Egger's test is *negative,* indicating that designs with better statistical power tend to show *larger* effects, rather than smaller effects, as would be expected in the presence of publication bias.

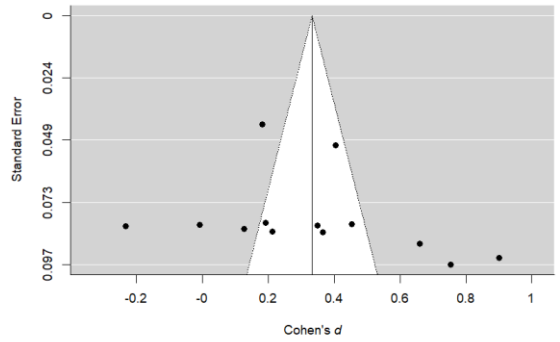
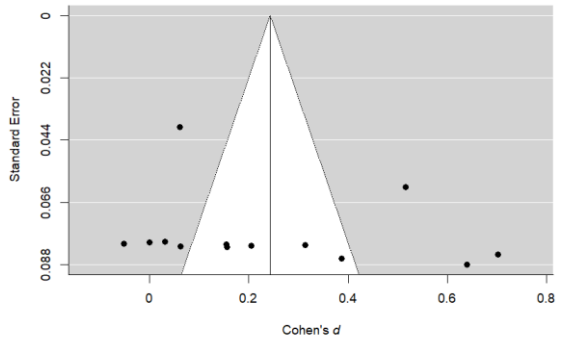
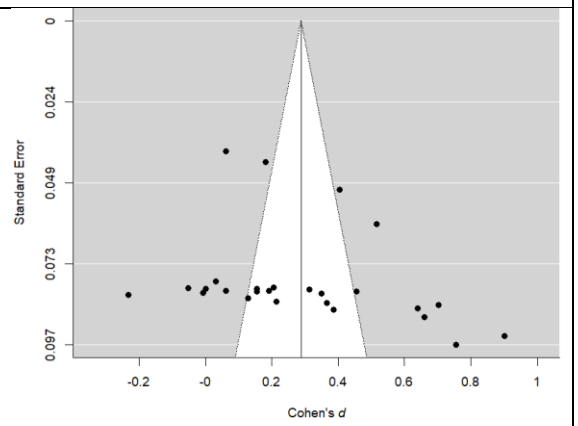
It is not entirely clear what underlies these observed funnel plot asymmetries. Given that the results from all of the crowdsourced research designs are presented, the asymmetries cannot be attributed to publication bias. They must therefore reflect some other "sample size effects" that are idiosyncratic to the designs tested in this project. This highlights one further advantage of crowdsourcing in comparison to the traditional practice of science: in a traditional meta-analysis of multiple studies conducted at different times, one cannot be certain whether funnel

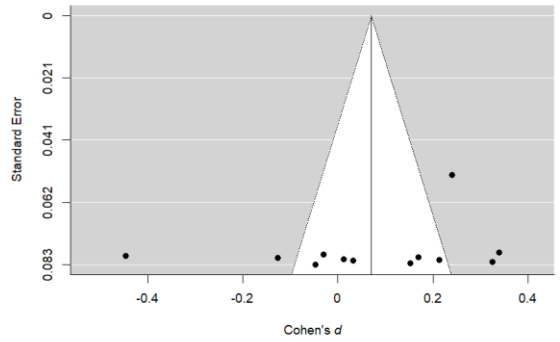
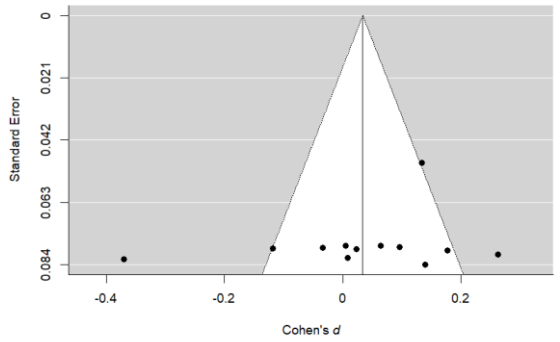
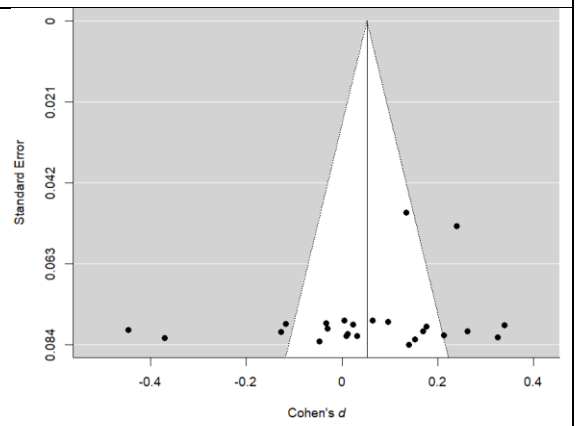
plot asymmetries reflect publication bias or some other sample size effect (see, e.g., Deeks, Macaskill, & Irwig, 2005), whereas in a crowdsourced project like this one, there is, by the very nature of the design, no publication bias.

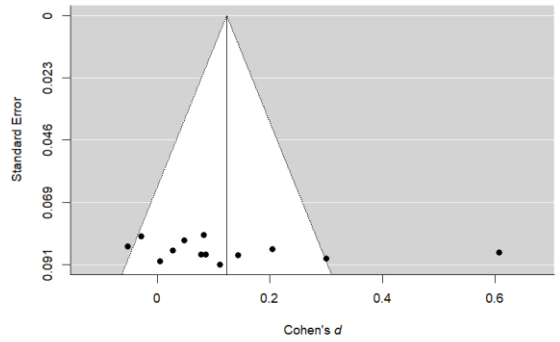
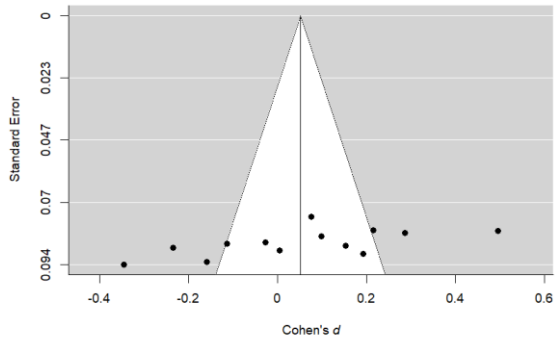
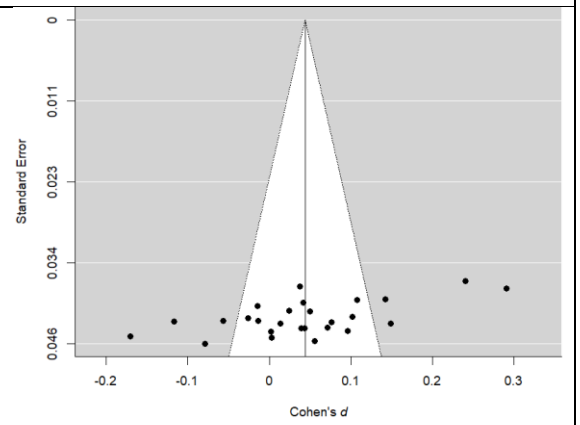
Figure S9.1. Funnel plots.

<table>
<tr><td colspan="3">Hypothesis 1 – Awareness of automatic prejudice</td></tr>
<tr><td>Main Studies</td><td>Replication Studies</td><td>Aggregated (Main Studies and Replication Studies)</td></tr>
<tr><td> 
</td><td> 
</td><td> 
</td></tr>
<tr><td></td><td></td><td></td></tr>
</table>

Hypothesis 2 – Extreme offers reduce trust		
Main Studies	Replication Studies	Aggregated (Main Studies and Replication Studies)
		

Hypothesis 3 – Moral praise for needless work		
Main Studies	Replication Studies	Aggregated (Main Studies and Replication Studies)
		

Hypothesis 4 – Proximal authorities drive legitimacy of performance enhancers		
Main Studies	Replication Studies	Aggregated (Main Studies and Replication Studies)
		

Hypothesis 5 – Deontological judgments predict happiness		
Main Studies	Replication Studies	Aggregated (Main Studies and Replication Studies)
		

Multivariate Meta-Analyses Nesting Study Within Hypothesis

In our main analyses combining the Main Studies and Replication Studies, we did not account for which data collection effort (Main Studies or Replication Studies) a given effect size came from. Therefore, we also conducted multivariate meta-analyses nesting study within each hypothesis, to account for the source of each effect size. The results were substantively similar to the simpler analyses reported in the main text. We again found support for Hypotheses 2 and 3, estimated mean effect sizes $d = 0.67$, $p < .001$, 95% CI [0.38, 0.97] and $d = 0.26$, $p < .001$, 95% CI [0.17, 0.34]. We found no support for Hypothesis 1, estimated mean effect size $d = -0.02$, $p = .643$, 95% CI [-0.12, 0.07]. The estimated mean effect sizes for Hypotheses 4 and 5 were statistically significant in these analyses, but remained small and similar to those reported in the main text, $d = 0.07$, $p = .001$, 95% CI [0.03, 0.11] and $r = .05$, $p < .001$, 95% CI [.02, .08], respectively. Therefore, including the study that an effect size came from in the analysis does not substantively alter the results.

References for Supplement 9

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

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

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