Crowdsourcing Hypotheses Tests: Methods and Results of the Preregistered Bayesian Analyses

Quentin F. Gronau
University of Amsterdam

Alexander Ly
University of Amsterdam
Centrum Wiskunde & Informatica

Don van den Bergh
University of Amsterdam

Maarten Marsman
University of Amsterdam

Koen Derks
Nyenrode Business University

Eric-Jan Wagenmakers
University of Amsterdam

Abstract

This is the methods and results section for the Bayesian analysis of the "Crowdsourcing hypotheses tests" data set. The methods section is a copy from the preregistration document that can also be found at `https://osf.io/9jzy4/`.

Keywords: Bayes factor, model averaging, preregistration

Methods

The "Crowdsourcing hypotheses tests" project studied five empirical phenomena (i.e., $q = 1, 2, \ldots, 5$), each of which was subject to replication attempts from the same set of $l = 1, 2, \ldots, 13$ 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 $\delta_{\text{pop},q}$ 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 $\delta_{l,q} - \frac{1}{2}\delta_{\text{pop},q}$ and the PureProfile effect size is given by $\delta_{l,q} + \frac{1}{2}\delta_{\text{pop},q}$. Thus, $\delta_{\text{pop},q}$ 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 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 draws 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:

$$\delta_{l,q} \sim \text{Normal}(\mu_q, \tau_q^2) \tag{1}$$

$$d_{1,l,q} \sim \text{Normal}(\delta_{l,q} - \tfrac{1}{2}\,\delta_{\text{pop},q}, \text{SE}_{1,l,q}^2) \tag{2}$$

$$d_{2,l,q} \sim \text{Normal}(\delta_{l,q} + \tfrac{1}{2}\,\delta_{\text{pop},q}, \text{SE}_{2,l,q}^2), \tag{3}$$

where $d_{p,l,q}$ denotes the observed effect size of the lth team, the pth population, and the qth question, and $\text{SE}_{p,l,q}$ 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 $\delta_{\text{pop},q}$ quantifies the impact of "population", that is, the difference in effect size between the MTurk population and the PureProfile population. If $\delta_{\text{pop},q} = 0$, 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 $\delta_{\text{pop},q}$ are assigned smooth prior distributions and no prior plausibility is assigned to the special cases where $\mu_q = 0$, $\tau_q = 0$, or $\delta_{\text{pop},q} = 0$. For this estimation approach we use the following priors: $\mu_q \sim \text{Cauchy}(0, 1/\sqrt{2})$, $\tau_q \sim \text{InvGamma}(1, 0.15)$ (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 $\delta_{\text{pop},q} \sim \text{Normal}(0, 0.5^2)$. 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 $\delta_{\text{pop},q} = 0$. Specifically, for each question q, we will assess the predictive adequacy of the following eight models:

$$
\begin{aligned}
&\mathcal{H}_1 : \mu_q = 0, \tau_q = 0, \delta_{\text{pop},q} = 0, \\
&\mathcal{H}_2 : \mu_q = 0, \tau_q = 0, \delta_{\text{pop},q} \sim \text{Normal}(0, 0.15^2), \\
&\mathcal{H}_3 : \mu_q = 0, \tau_q \sim \text{InvGamma}(1, 0.15), \delta_{\text{pop},q} = 0, \\
&\mathcal{H}_4 : \mu_q = 0, \tau_q \sim \text{InvGamma}(1, 0.15), \delta_{\text{pop},q} \sim \text{Normal}(0, 0.15^2), \\
&\mathcal{H}_5 : \mu_q \sim t(0.35, 0.102, 3)\text{I}(0, \infty), \tau_q = 0, \delta_{\text{pop},q} = 0, \\
&\mathcal{H}_6 : \mu_q \sim t(0.35, 0.102, 3)\text{I}(0, \infty), \tau_q = 0, \delta_{\text{pop},q} \sim \text{Normal}(0, 0.15^2), \\
&\mathcal{H}_7 : \mu_q \sim t(0.35, 0.102, 3)\text{I}(0, \infty), \tau_q \sim \text{InvGamma}(1, 0.15), \delta_{\text{pop},q} = 0, \\
&\mathcal{H}_8 : \mu_q \sim t(0.35, 0.102, 3)\text{I}(0, \infty), \tau_q \sim \text{InvGamma}(1, 0.15), \delta_{\text{pop},q} \sim \text{Normal}(0, 0.15^2).
\end{aligned}
\tag{4}
$$

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., $\text{I}(0, \infty)$; 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 $\text{InvGamma}(1, 0.15)$ distribution (Gronau, van Erp, et al., 2017; van Erp et al.,

2017). Finally, parameter $\delta_{\text{pop},q}$ 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 $P(\mathcal{H}_j) = 1/8 = 0.125$, $j = 1, 2, \ldots, 8$. 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 $\mathcal{H}_5, \mathcal{H}_6, \mathcal{H}_7$, and $\mathcal{H}_8$ (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 $\mathcal{H}_1, \mathcal{H}_2, \mathcal{H}_3$, and $\mathcal{H}_4$ (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 $\mathcal{H}_3, \mathcal{H}_4, \mathcal{H}_7$, and $\mathcal{H}_8$, models for which $\tau_q \neq 0$) versus the evidence for homogeneity (i.e., the summed posterior model probabilities for $\mathcal{H}_1, \mathcal{H}_2, \mathcal{H}_5$, and $\mathcal{H}_6$, 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 $\mathcal{H}_2, \mathcal{H}_4, \mathcal{H}_6$, and $\mathcal{H}_8$ (i.e., models for which $\delta_{\text{pop},q} \neq 0$) versus the summed posterior

model probabilities for $\mathcal{H}_1, \mathcal{H}_3, \mathcal{H}_5$, and $\mathcal{H}_7$ (i.e., models for which $\delta_{\text{pop},q} = 0$). 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 $\delta_{\text{pop},q}$ across all eight models (including a spike at zero, the height of which equals the summed posterior model probability across the four models where $\delta_{\text{pop},q} = 0$).

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 $\text{SE}_{p,l,q}$ are viewed as repeated measurements of the labs across the two populations. Hence, laboratory membership $l = 1, 2, \ldots, 13$ 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 $q = 1, 2, \ldots, 5$ 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 $\mathcal{M}_0$, while the alternative model $\mathcal{M}_1$ 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, 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,

$$d_{p,l,q} \sim \text{Normal}(\delta_{p,l,q}, \text{SE}^2_{p,l,q}), \tag{5}$$

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

$$t_{p,l,q} = \frac{d_{p,l,q} - \bar{d}_{\text{overall}}}{\text{SE}_{p,l,q}} \tag{6}$$

where $\bar{d}_{\text{overall}}$ 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 $\bar{d}_{\text{overall}}$ 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; 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 $\delta_{\text{pop},1}$ between the MTurk and the PureProfile populations.

First we present the results of the *unfiltered* data. Figure 1 shows the prior and posterior distributions from the model with all three parameters free to vary. The top panel of Figure 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 2 shows the prior and posterior distributions from the model with all three parameters free to vary. The top panel of Figure 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.

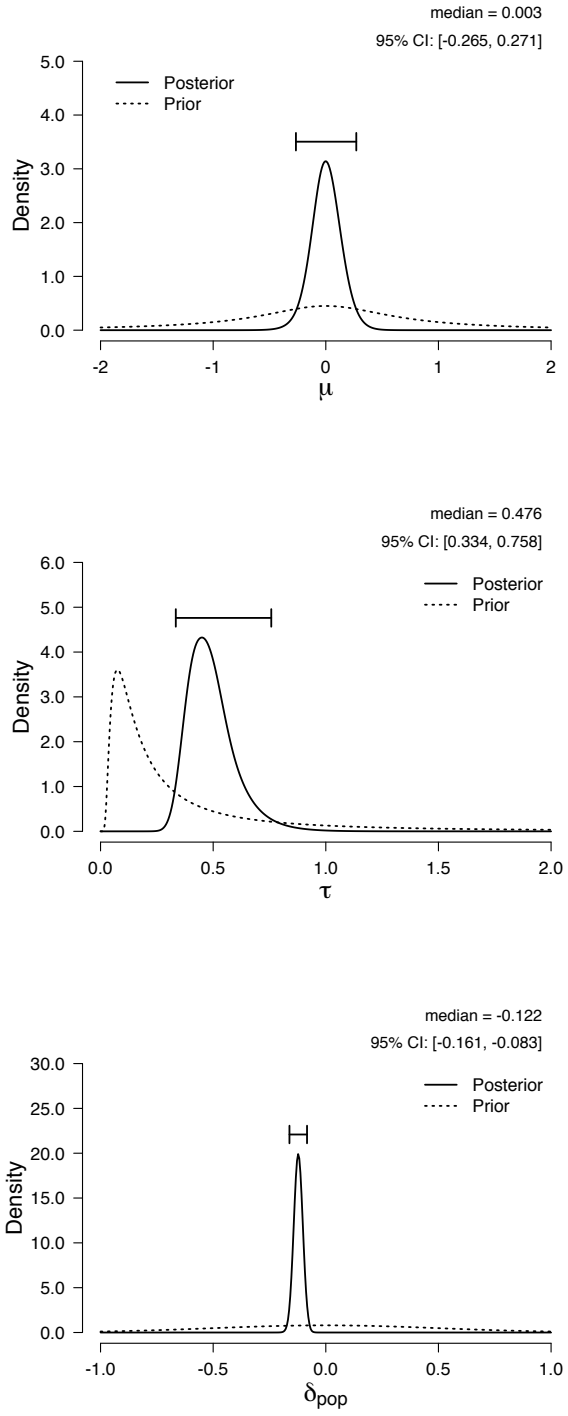

Figure 1. Estimation results for Q1 (*unfiltered* data). The upper panel displays the results for the group-level mean effect size μ_1, the middle panel displays the results for the across-team heterogeneity τ_1, and the lower panel displays the results for the difference $\delta_{\text{pop},1}$ between the MTurk and the PureProfile populations. Each panel shows the prior and posterior distribution, the posterior median, and a 95% posterior credible interval.

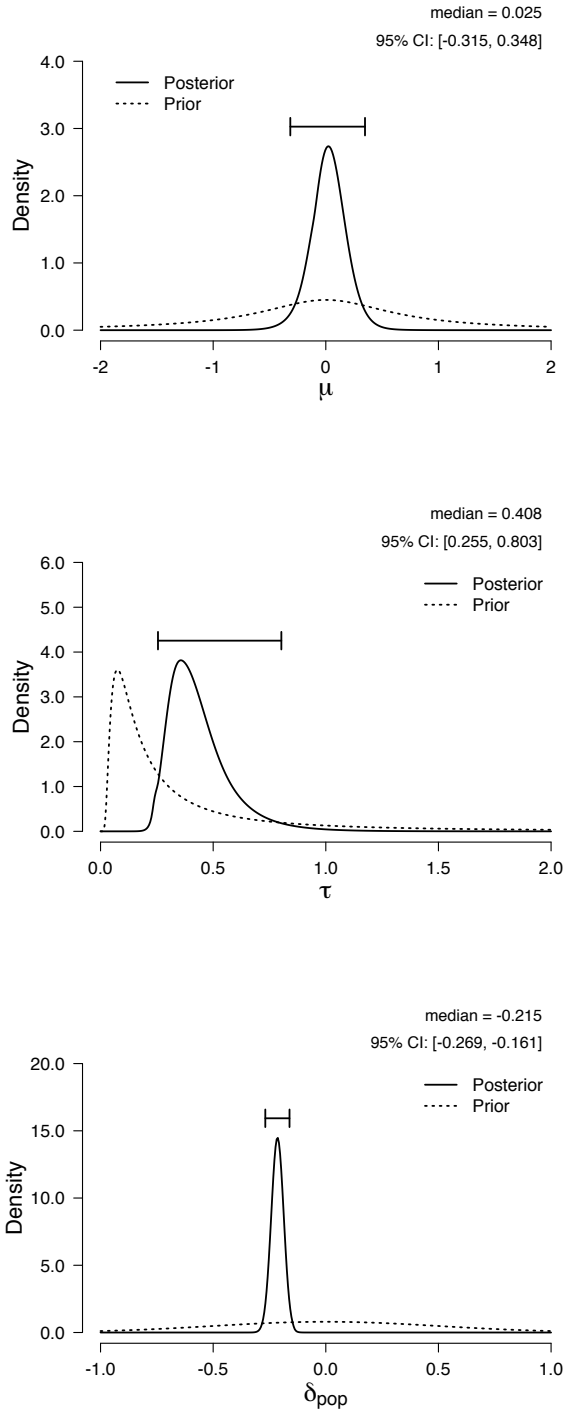

Figure 2. Estimation results for Q1 (*filtered* data). The upper panel displays the results for the group-level mean effect size μ_1, the middle panel displays the results for the across-team heterogeneity τ_1, and the lower panel displays the results for the difference $\delta_{\text{pop},1}$ between 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 $\delta_{\text{pop},1} = 0$. 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 1 presents the posterior model probabilities for Q1 based on the *unfiltered* data. The first column of Table 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 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 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 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 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.

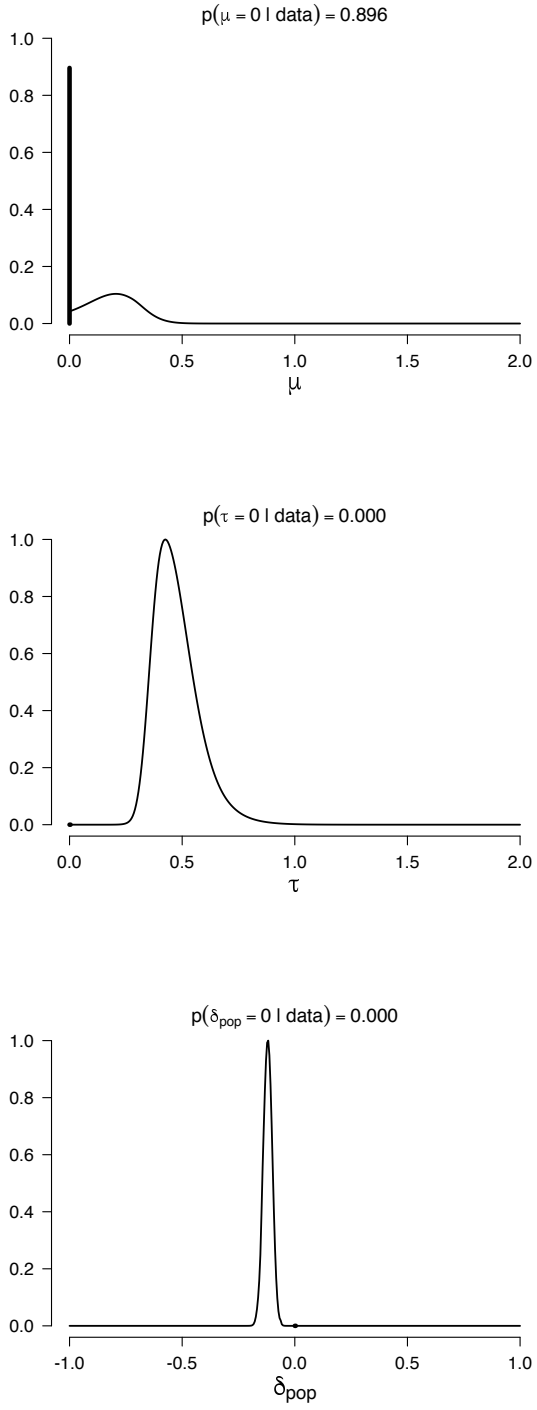

Figure 3. Model averaging results for Q1 (*unfiltered* data). The upper panel displays the results for the group-level mean effect size μ_1, the middle panel displays the results for the across-team heterogeneity τ_1, and the lower panel displays the results for the difference $\delta_{\text{pop},1}$ 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.

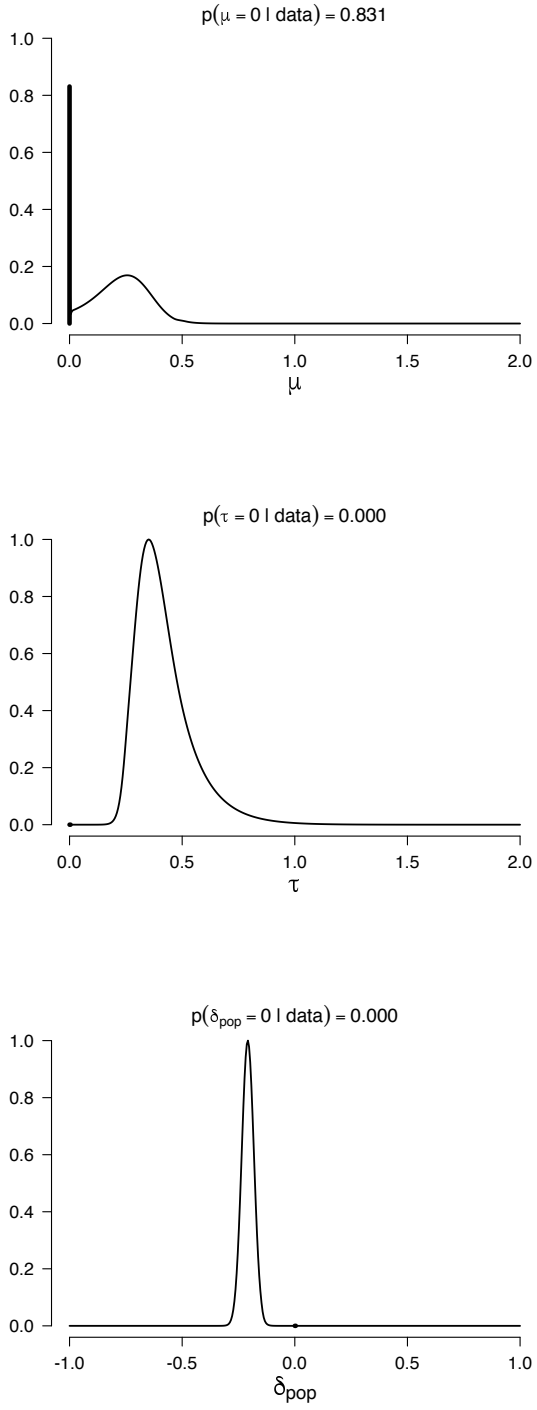

Figure 4. Model averaging results for Q1 (*filtered* data). The upper panel displays the results for the group-level mean effect size μ_1, the middle panel displays the results for the across-team heterogeneity τ_1, and the lower panel displays the results for the difference $\delta_{\text{pop},1}$ between 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 1
Posterior Model Probabilities (Unfiltered Data)

	Question				
Models	1	2	3	4	5
$\mathcal{H}_1$	0.000	0.000	0.000	0.000	0.000
$\mathcal{H}_2$	0.000	0.000	0.000	0.000	0.000
$\mathcal{H}_3$	0.000	0.000	**0.001**	**0.603**	**0.317**
$\mathcal{H}_4$	**0.896**	**0.008**	**0.007**	**0.339**	**0.462**
$\mathcal{H}_5$	0.000	0.000	0.000	0.000	0.000
$\mathcal{H}_6$	0.000	0.000	0.000	0.000	0.000
$\mathcal{H}_7$	0.000	0.000	**0.114**	**0.036**	**0.089**
$\mathcal{H}_8$	**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 $\delta_{\text{pop},1}$ does not equal 0. The summed posterior probability for the models in which $\delta_{\text{pop},1} = 0$ equals 0.000. The lower panel of Figure 3 shows the model-averaged posterior distribution for $\delta_{\text{pop},1}$ across all eight models, where the height of the spike at zero corresponds to the summed posterior probability that $\delta_{\text{pop},1} = 0$. 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 $\delta_{\text{pop},1}$ does not equal 0. The summed posterior probability for the models in which $\delta_{\text{pop},1} = 0$ equals 0.000. The lower panel of Figure 4 shows the model-averaged posterior distribution for $\delta_{\text{pop},1}$ across all eight models, where the height of the spike at zero corresponds to the summed posterior probability that $\delta_{\text{pop},1} = 0$. 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 $\delta_{\text{pop},2}$ between the MTurk and the PurePro-

Table 2
Posterior Model Probabilities (Filtered Data)

	Question				
Models	1	2	3	4	5
$\mathcal{H}_1$	0.000	0.000	0.000	0.000	**0.005**
$\mathcal{H}_2$	0.000	0.000	0.000	0.000	**0.003**
$\mathcal{H}_3$	0.000	0.000	**0.003**	**0.419**	**0.582**
$\mathcal{H}_4$	**0.831**	**0.010**	**0.008**	**0.483**	**0.349**
$\mathcal{H}_5$	0.000	0.000	0.000	0.000	**0.001**
$\mathcal{H}_6$	0.000	0.000	0.000	0.000	**0.001**
$\mathcal{H}_7$	0.000	0.000	**0.290**	**0.045**	**0.037**
$\mathcal{H}_8$	**0.169**	**0.990**	**0.700**	**0.053**	**0.022**

file populations.

First we present the results of the *unfiltered* data. Figure 5 shows the prior and posterior distributions from the model with all three parameters free to vary. The top panel of Figure 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 6 shows the prior and posterior distributions from the model with all three parameters free to vary. The top panel of Figure 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 $\delta_{\text{pop},2} = 0$. 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 1 presents the posterior model probabilities for Q2 based on the *unfiltered* data. The second column of Table 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

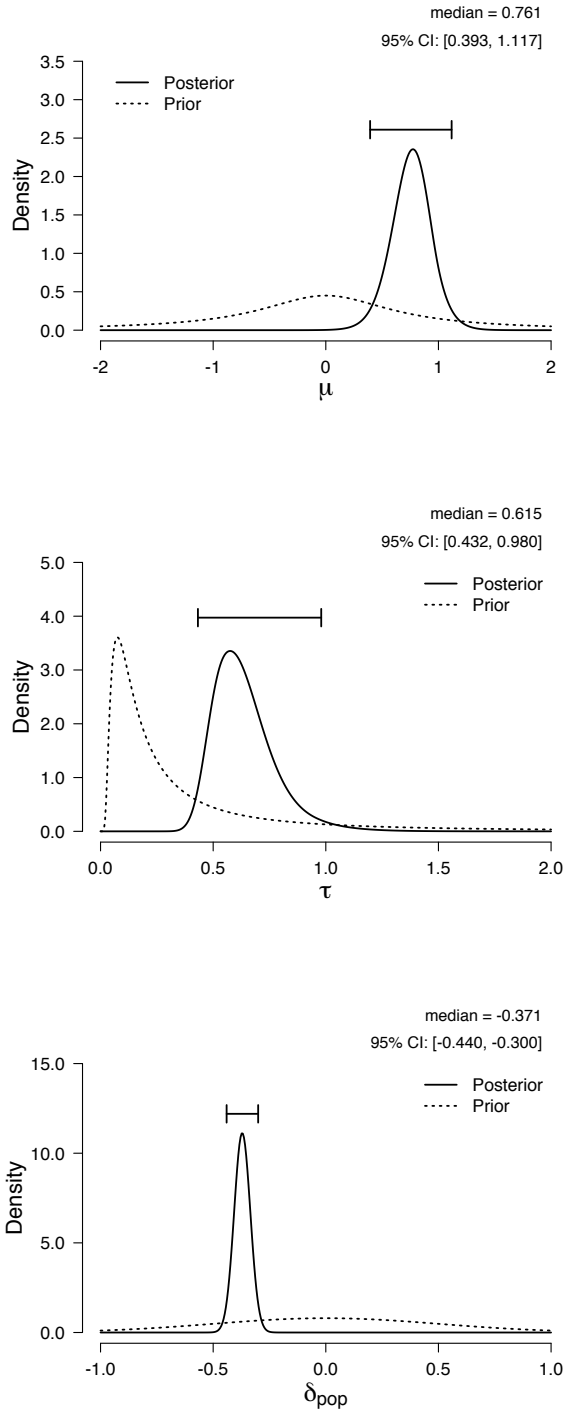

Figure 5. Estimation results for Q2 (*unfiltered* data). The upper panel displays the results for the group-level mean effect size μ_2, the middle panel displays the results for the across-team heterogeneity τ_2, and the lower panel displays the results for the difference $\delta_{\text{pop},2}$ between the MTurk and the PureProfile populations. Each panel shows the prior and posterior distribution, the posterior median, and a 95% posterior credible interval.

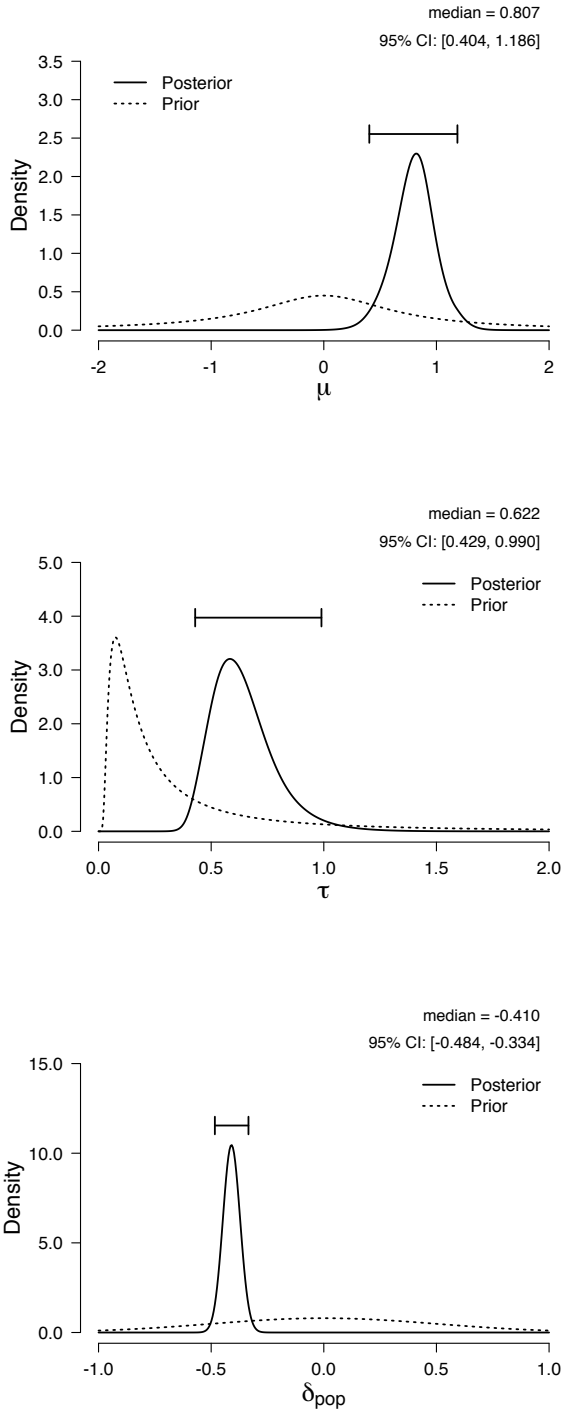

Figure 6. Estimation results for Q2 (*filtered* data). The upper panel displays the results for the group-level mean effect size μ_2, the middle panel displays the results for the across-team heterogeneity τ_2, and the lower panel displays the results for the difference $\delta_{\text{pop},2}$ between the MTurk and the PureProfile populations. Each panel shows the prior and posterior distribution, the posterior median, and a 95% posterior credible interval.

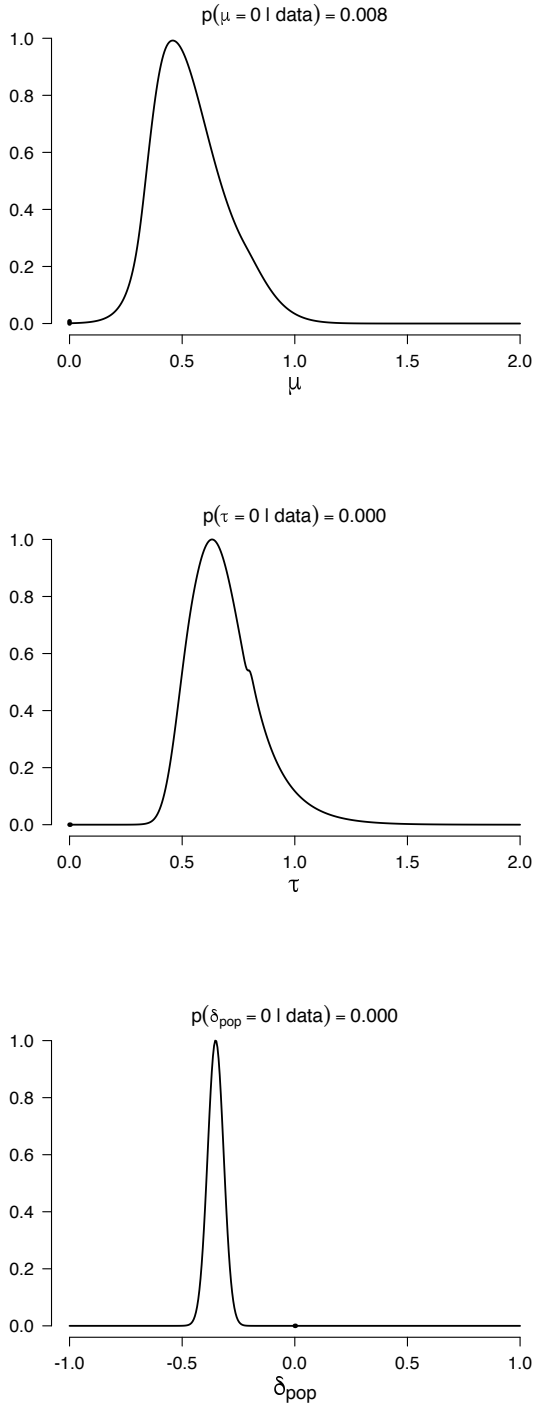

Figure 7. Model averaging results for Q2 (*unfiltered* data). The upper panel displays the results for the group-level mean effect size μ_2, the middle panel displays the results for the across-team heterogeneity τ_2, and the lower panel displays the results for the difference $\delta_{\text{pop},2}$ 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.

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 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 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.

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 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 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 $\delta_{\text{pop},2}$ does not equal 0. The summed posterior probability for the models in which $\delta_{\text{pop},2} = 0$ equals 0.000. The lower panel of Figure 7 shows the model-averaged posterior distribution for $\delta_{\text{pop},2}$ across all eight models, where the height of the spike at zero corresponds to the summed posterior probability that $\delta_{\text{pop},2} = 0$. In sum, for Q2 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 of the *filtered* data. The Bayes factor and the posterior model odds both equal 9.007×10^{14} in favor of the proposition that $\delta_{\text{pop},2}$

[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.

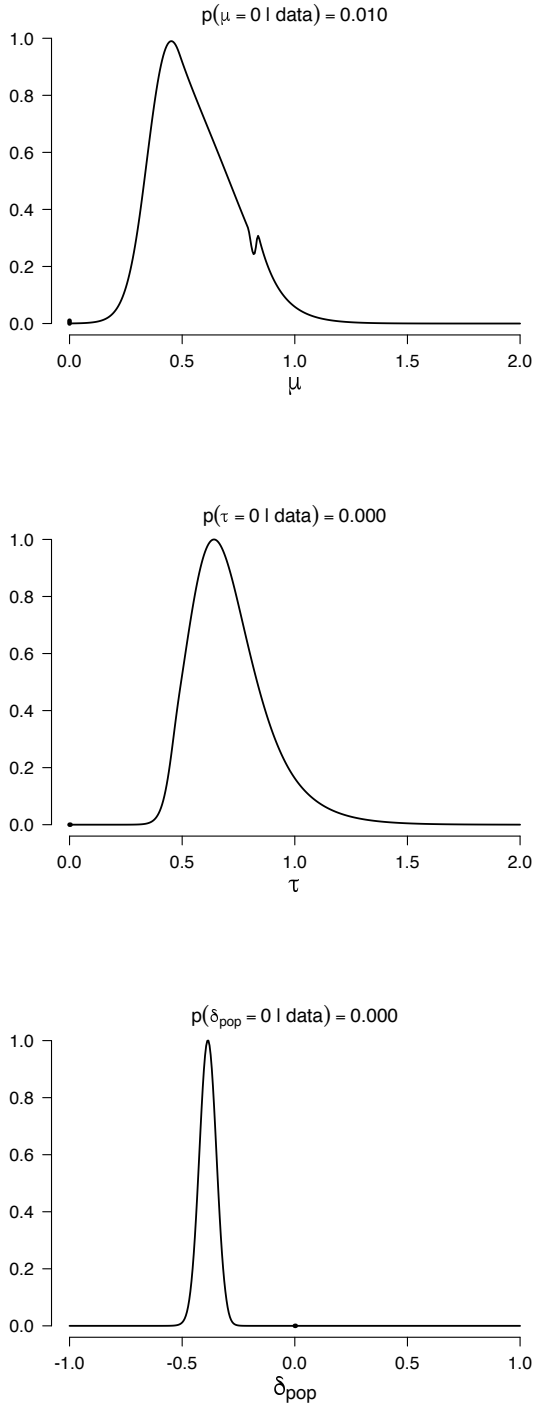

Figure 8. Model averaging results for Q2 (*filtered* data). The upper panel displays the results for the group-level mean effect size μ_2, the middle panel displays the results for the across-team heterogeneity τ_2, and the lower panel displays the results for the difference $\delta_{\text{pop},2}$ between 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.

does not equal 0. The summed posterior probability for the models in which $\delta_{\text{pop},2} = 0$ equals 0.000. The lower panel of Figure 8 shows the model-averaged posterior distribution for $\delta_{\text{pop},2}$ across all eight models, where the height of the spike at zero corresponds to the summed posterior probability that $\delta_{\text{pop},2} = 0$. 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.

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 $\delta_{\text{pop},3}$ between the MTurk and the PureProfile populations.

First we present the results of the *unfiltered* data. Figure 9 shows the prior and posterior distributions from the model with all three parameters free to vary. The top panel of Figure 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 10 shows the prior and posterior distributions from the model with all three parameters free to vary. The top panel of Figure 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 $\delta_{\text{pop},3} = 0$. 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 1 presents the posterior model probabilities for Q3 based on the *unfiltered* data. The third column of Table 2 presents the posterior model probabilities for Q3 based on the *filtered* data.

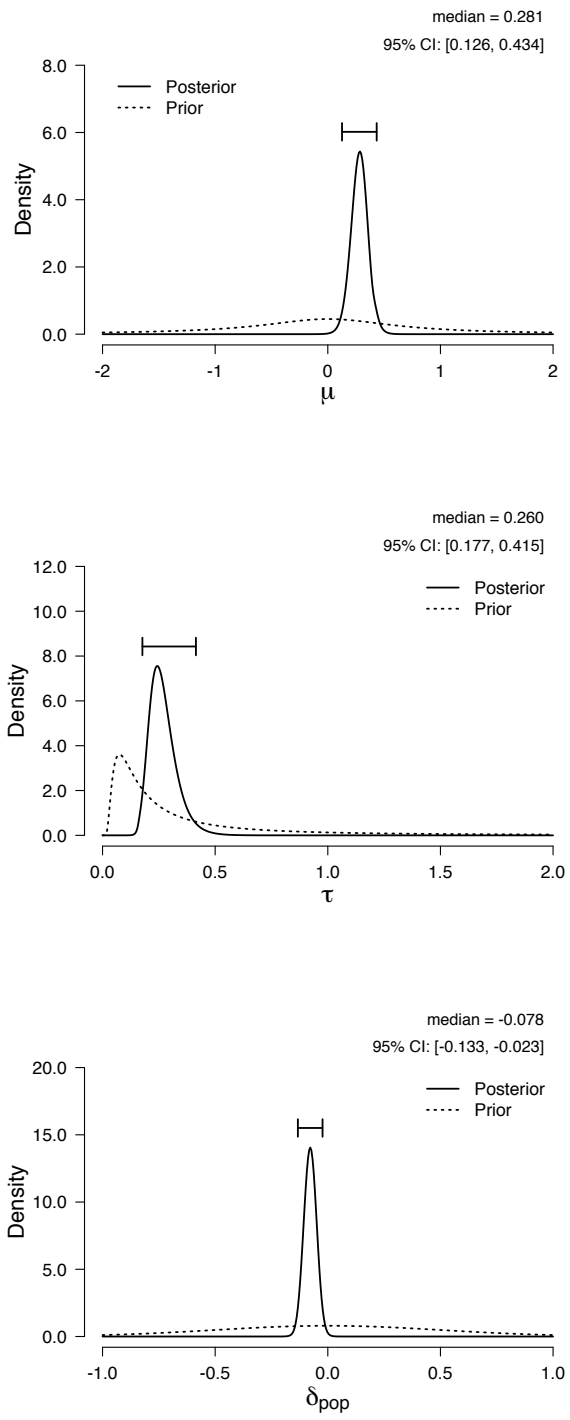

Figure 9. Estimation results for Q3 (*unfiltered* data). The upper panel displays the results for the group-level mean effect size μ_3, the middle panel displays the results for the across-team heterogeneity τ_3, and the lower panel displays the results for the difference $\delta_{\text{pop},3}$ between the MTurk and the PureProfile populations. Each panel shows the prior and posterior distribution, the posterior median, and a 95% posterior credible interval.

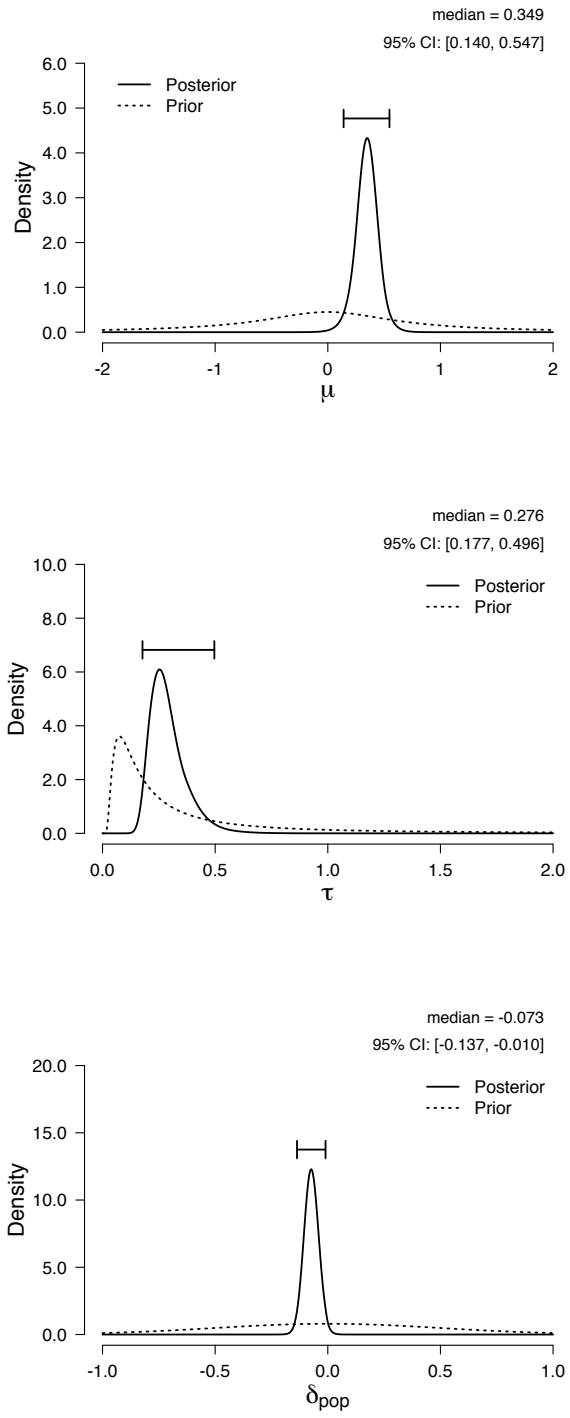

Figure 10. Estimation results for Q3 (*filtered* data). The upper panel displays the results for the group-level mean effect size μ_3, the middle panel displays the results for the across-team heterogeneity τ_3, and the lower panel displays the results for the difference $\delta_{\text{pop},3}$ between the MTurk and the PureProfile populations. Each panel shows the prior and posterior distribution, the posterior median, and a 95% posterior credible interval.

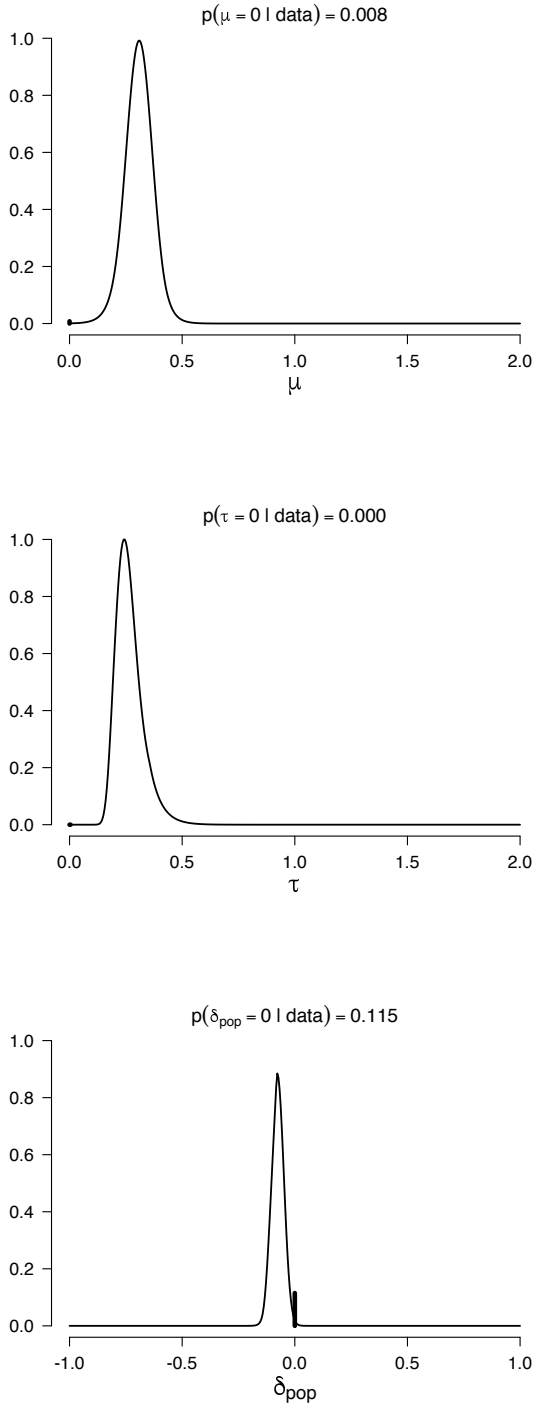

Figure 11. Model averaging results for Q3 (*unfiltered* data). The upper panel displays the results for the group-level mean effect size μ_3, the middle panel displays the results for the across-team heterogeneity τ_3, and the lower panel displays the results for the difference $\delta_{\text{pop},3}$ 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.

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 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.

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 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 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 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.

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 $\delta_{\text{pop},3}$ does not equal 0. The summed posterior probability for the models in which $\delta_{\text{pop},3} = 0$ equals 0.115. The lower panel of Figure 11 shows the model-averaged posterior distribution for $\delta_{\text{pop},3}$ across all eight models, where the height of the spike at zero corresponds to the summed posterior probability that $\delta_{\text{pop},3} = 0$. 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.

[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.

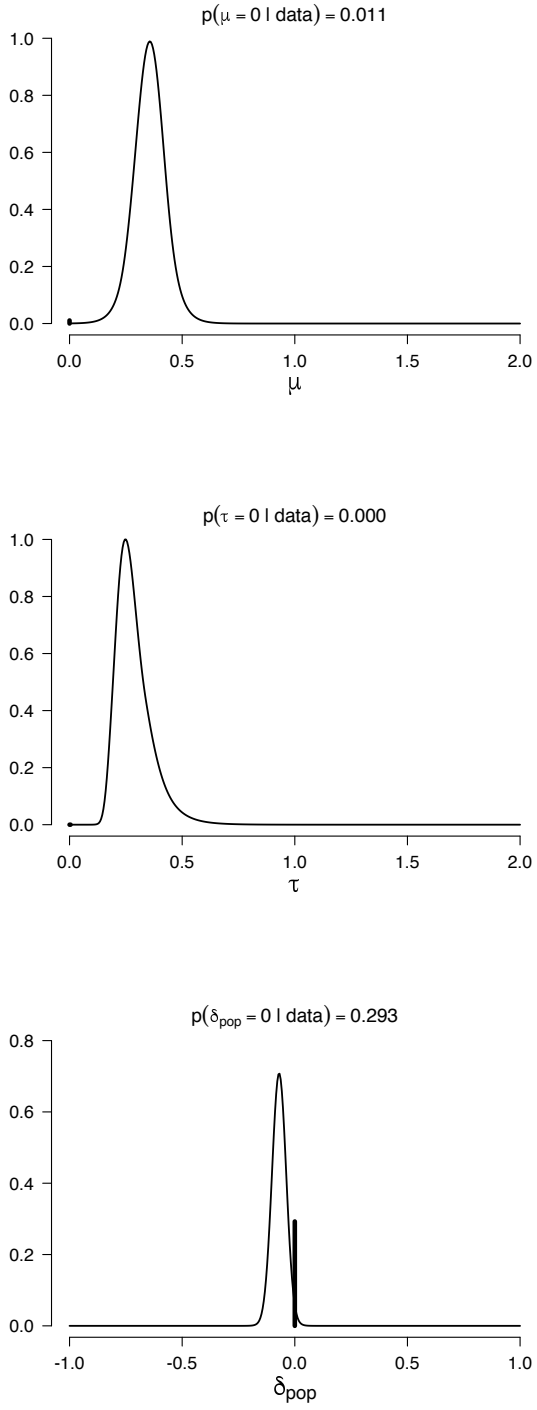

Figure 12. Model averaging results for Q3 (*filtered* data). The upper panel displays the results for the group-level mean effect size μ_3, the middle panel displays the results for the across-team heterogeneity τ_3, and the lower panel displays the results for the difference $\delta_{\text{pop},3}$ 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 2.416 in favor of the proposition that $\delta_{\text{pop},3}$ does not equal 0. The summed posterior probability for the models in which $\delta_{\text{pop},3} = 0$ equals 0.293. The lower panel of Figure 12 shows the model-averaged posterior distribution for $\delta_{\text{pop},3}$ across all eight models, where the height of the spike at zero corresponds to the summed posterior probability that $\delta_{\text{pop},3} = 0$. 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.

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 $\delta_{\text{pop},4}$ between the MTurk and the PureProfile populations.

First we present the results of the *unfiltered* data. Figure 13 shows the prior and posterior distributions from the model with all three parameters free to vary. The top panel of Figure 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 14 shows the prior and posterior distributions from the model with all three parameters free to vary. The top panel of Figure 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 $\delta_{\text{pop},4} = 0$.

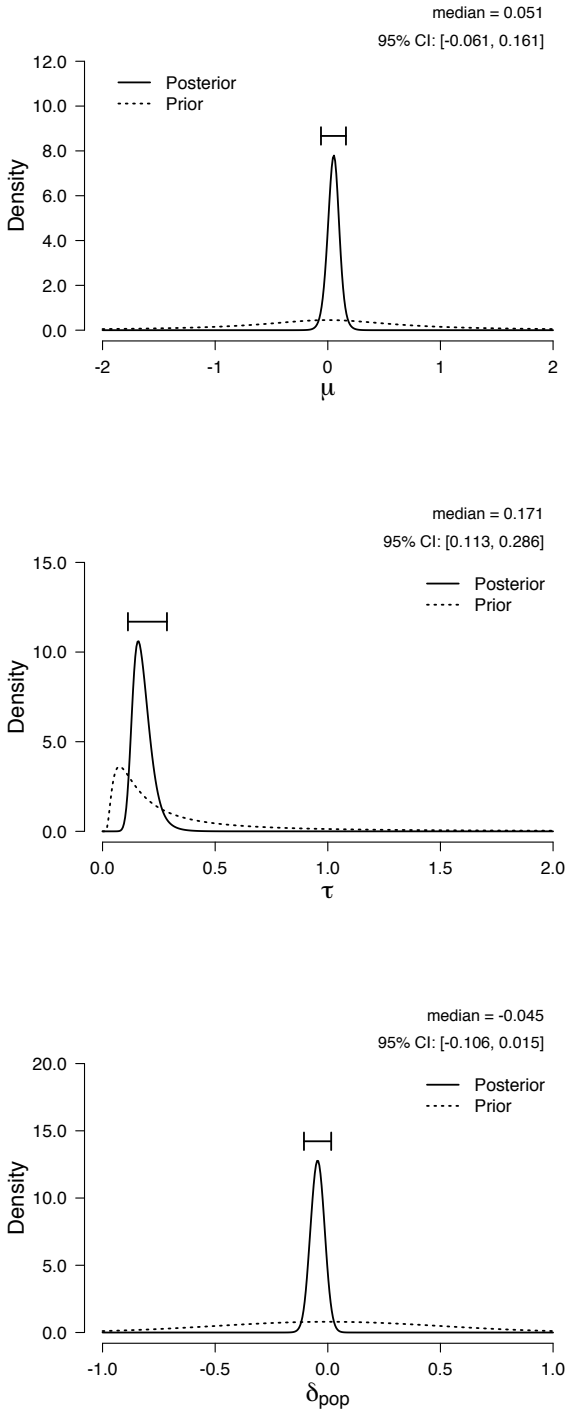

Figure 13. Estimation results for Q4 (*unfiltered* data). The upper panel displays the results for the group-level mean effect size μ_4, the middle panel displays the results for the across-team heterogeneity τ_4, and the lower panel displays the results for the difference $\delta_{\text{pop},4}$ between the MTurk and the PureProfile populations. Each panel shows the prior and posterior distribution, the posterior median, and a 95% posterior credible interval.

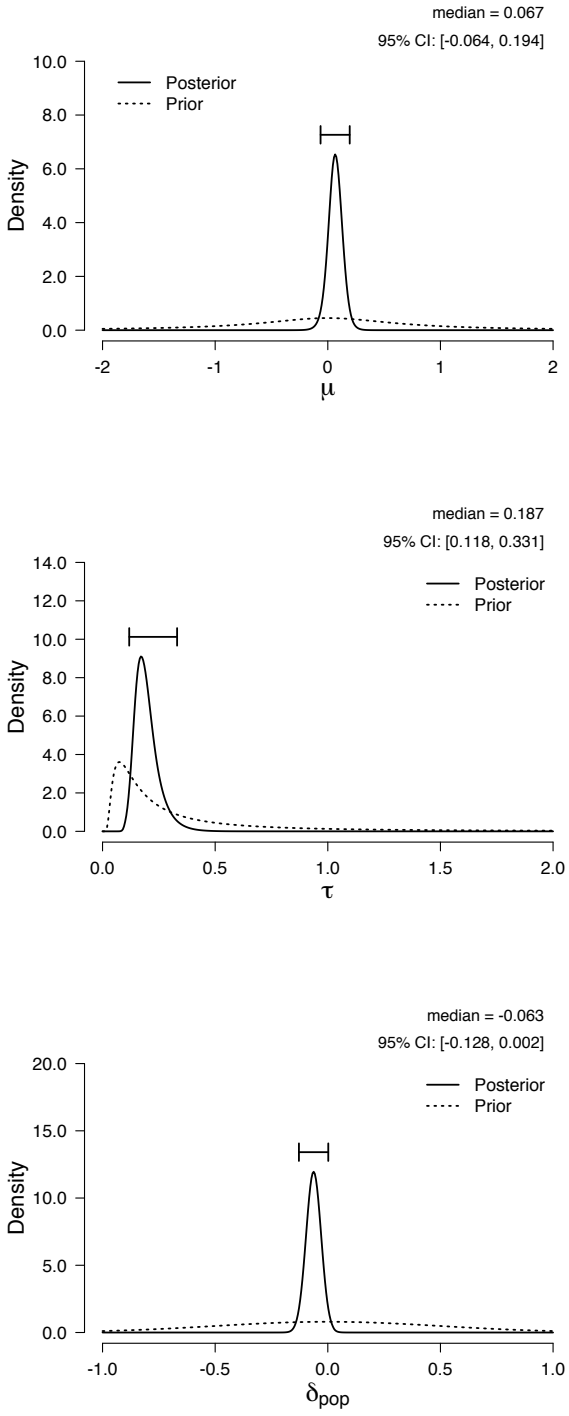

Figure 14. Estimation results for Q4 (*filtered* data). The upper panel displays the results for the group-level mean effect size μ_4, the middle panel displays the results for the across-team heterogeneity τ_4, and the lower panel displays the results for the difference $\delta_{\text{pop},4}$ between the MTurk and the PureProfile populations. Each panel shows the prior and posterior distribution, the posterior median, and a 95% posterior credible interval.

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 1 presents the posterior model probabilities for Q4 based on the *unfiltered* data. The fourth column of Table 2 presents the posterior model probabilities for Q4 based on the *filtered* data.

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 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 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 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.

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 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 $\delta_{\text{pop},4}$ equals 0. The summed posterior probability for the models in which $\delta_{\text{pop},4} = 0$ equals 0.639. The lower panel of Figure 15 shows the model-averaged posterior distribution for $\delta_{\text{pop},4}$ across all eight models, where

[6]The true Bayes factor is so large that it exceeds the available numerical precision.

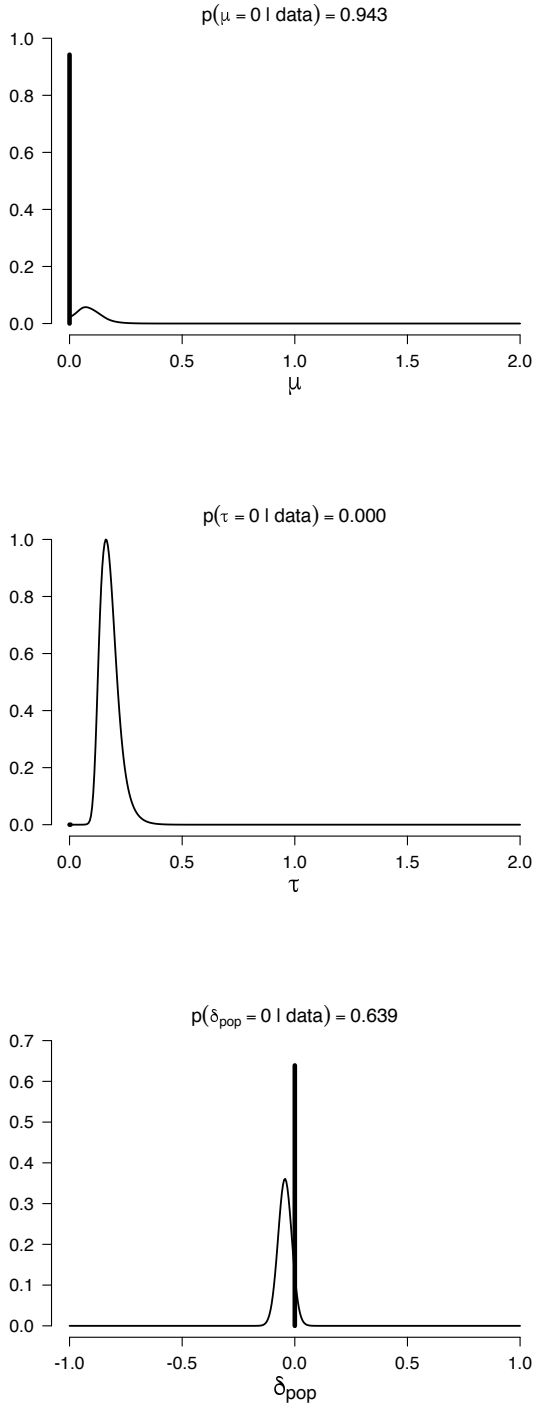

Figure 15. Model averaging results for Q4 (*unfiltered* data). The upper panel displays the results for the group-level mean effect size μ_4, the middle panel displays the results for the across-team heterogeneity τ_4, and the lower panel displays the results for the difference $\delta_{\text{pop},4}$ 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.

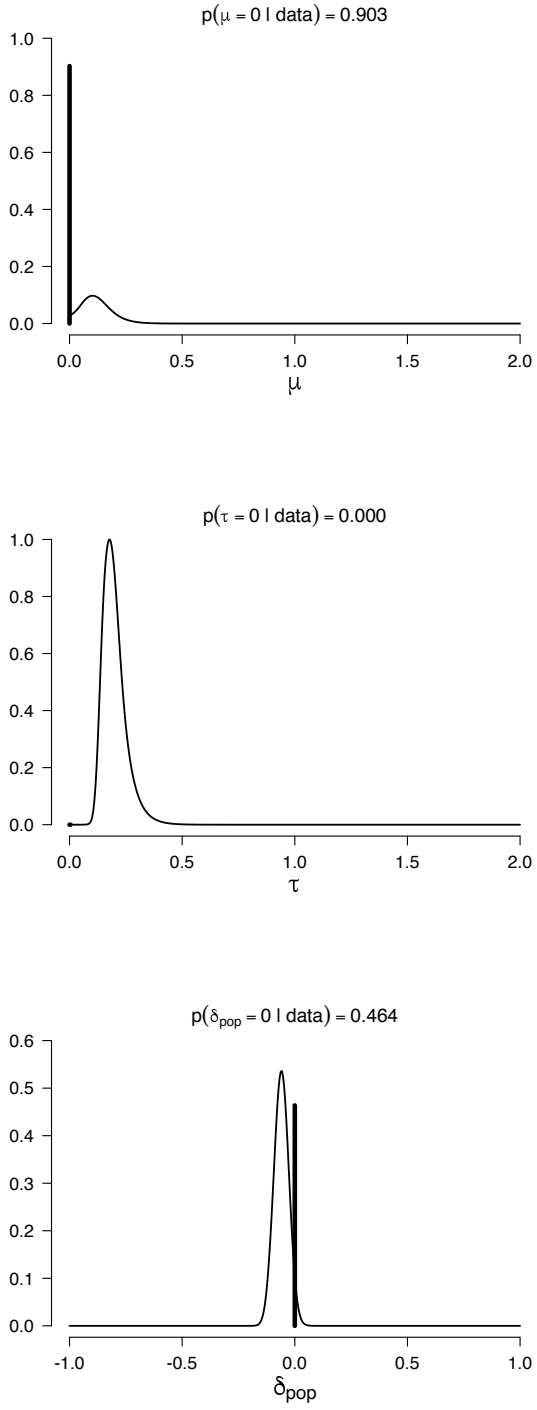

Figure 16. Model averaging results for Q4 (*filtered* data). The upper panel displays the results for the group-level mean effect size μ_4, the middle panel displays the results for the across-team heterogeneity τ_4, and the lower panel displays the results for the difference $\delta_{\text{pop},4}$ 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.

the height of the spike at zero corresponds to the summed posterior probability that $\delta_{\text{pop},4} = 0$. 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.

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 $\delta_{\text{pop},4}$ does not equal 0. The summed posterior probability for the models in which $\delta_{\text{pop},4} = 0$ equals 0.464. The lower panel of Figure 16 shows the model-averaged posterior distribution for $\delta_{\text{pop},4}$ across all eight models, where the height of the spike at zero corresponds to the summed posterior probability that $\delta_{\text{pop},4} = 0$. 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 $\delta_{\text{pop},5}$ between the MTurk and the PureProfile populations.

First we present the results of the *unfiltered* data. Figure 17 shows the prior and posterior distributions from the model with all three parameters free to vary. The top panel of Figure 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 18 shows the prior and posterior distributions from the model with all three parameters free to vary. The top panel of Figure 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 $\delta_{\text{pop},5} = 0$. Each model is assigned equal prior probability; hence, each restriction is a priori

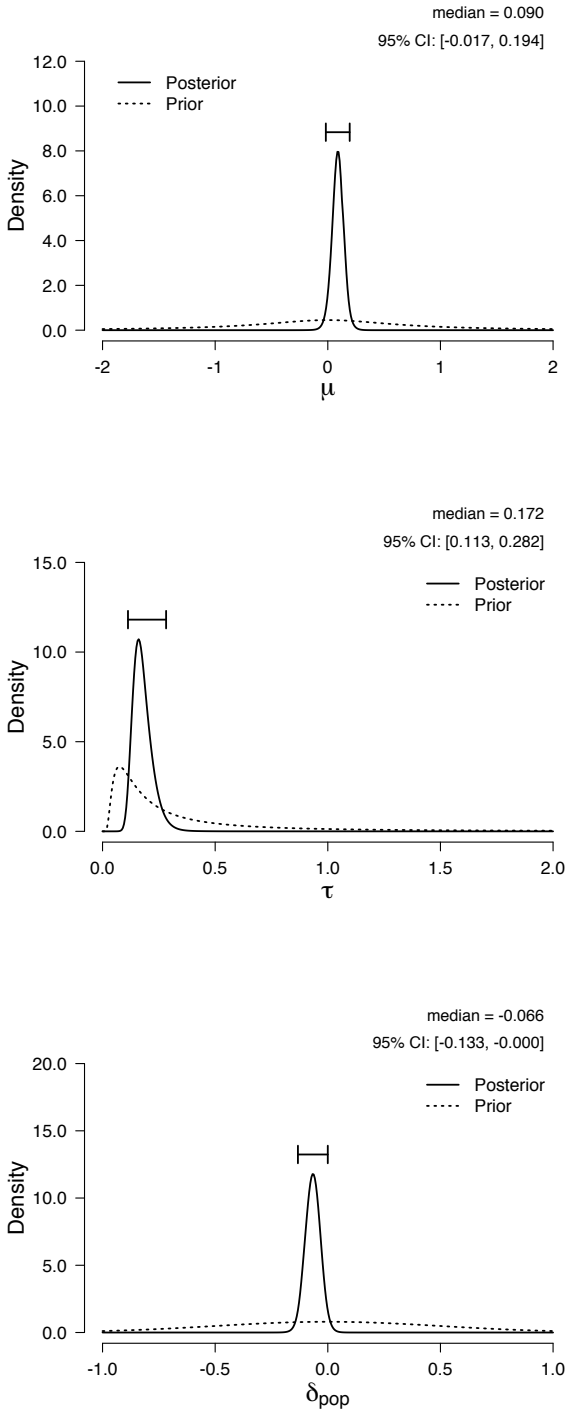

Figure 17. Estimation results for Q5 (*unfiltered* data). The upper panel displays the results for the group-level mean effect size μ_5, the middle panel displays the results for the across-team heterogeneity τ_5, and the lower panel displays the results for the difference $\delta_{\text{pop},5}$ between the MTurk and the PureProfile populations. Each panel shows the prior and posterior distribution, the posterior median, and a 95% posterior credible interval.

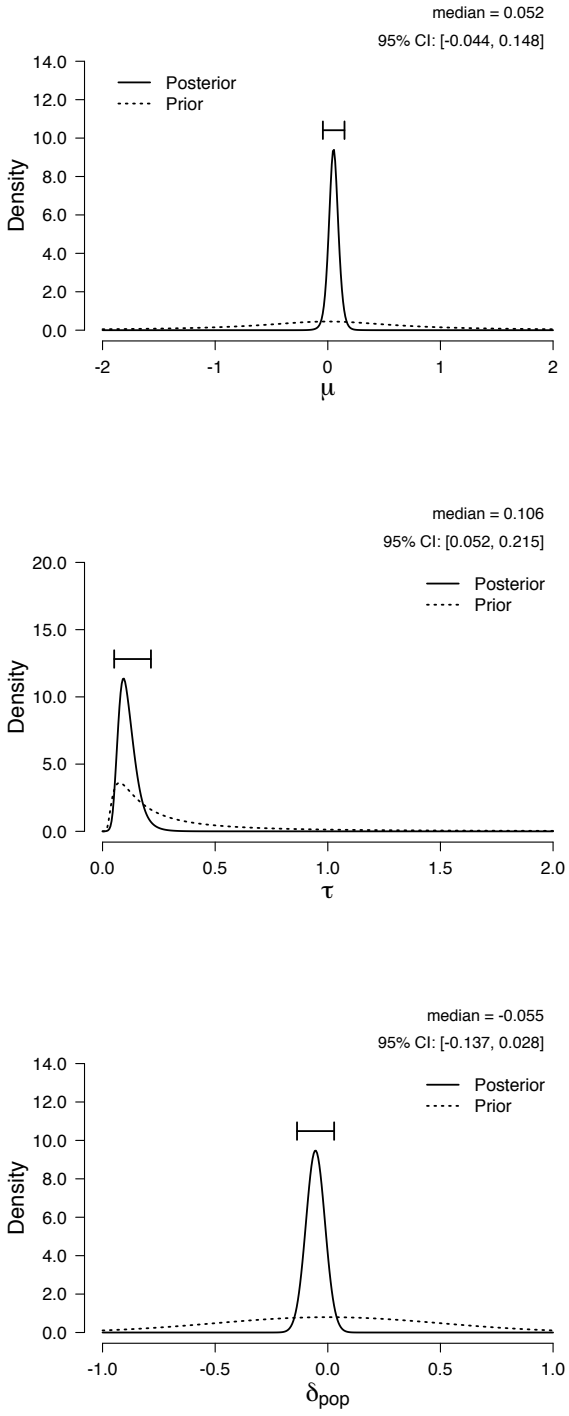

Figure 18. Estimation results for Q5 (*filtered* data). The upper panel displays the results for the group-level mean effect size μ_5, the middle panel displays the results for the across-team heterogeneity τ_5, and the lower panel displays the results for the difference $\delta_{\text{pop},5}$ between the MTurk and the PureProfile populations. Each panel shows the prior and posterior distribution, the posterior median, and a 95% posterior credible interval.

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 1 presents the posterior model probabilities for Q5 based on the *unfiltered* data. The fifth column of Table 2 presents the posterior model probabilities for Q5 based on the *filtered* data.

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 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 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 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 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 $\delta_{\text{pop},5}$ does not equal 0. The summed posterior probability for the models in which $\delta_{\text{pop},5} = 0$ equals 0.406. The lower panel of Figure 19 shows the model-averaged posterior distribution for $\delta_{\text{pop},5}$ across all eight models, where the height of the spike at zero corresponds to the summed posterior probability that $\delta_{\text{pop},5} = 0$. In sum, for Q5 the *unfiltered* data provide weak evidence

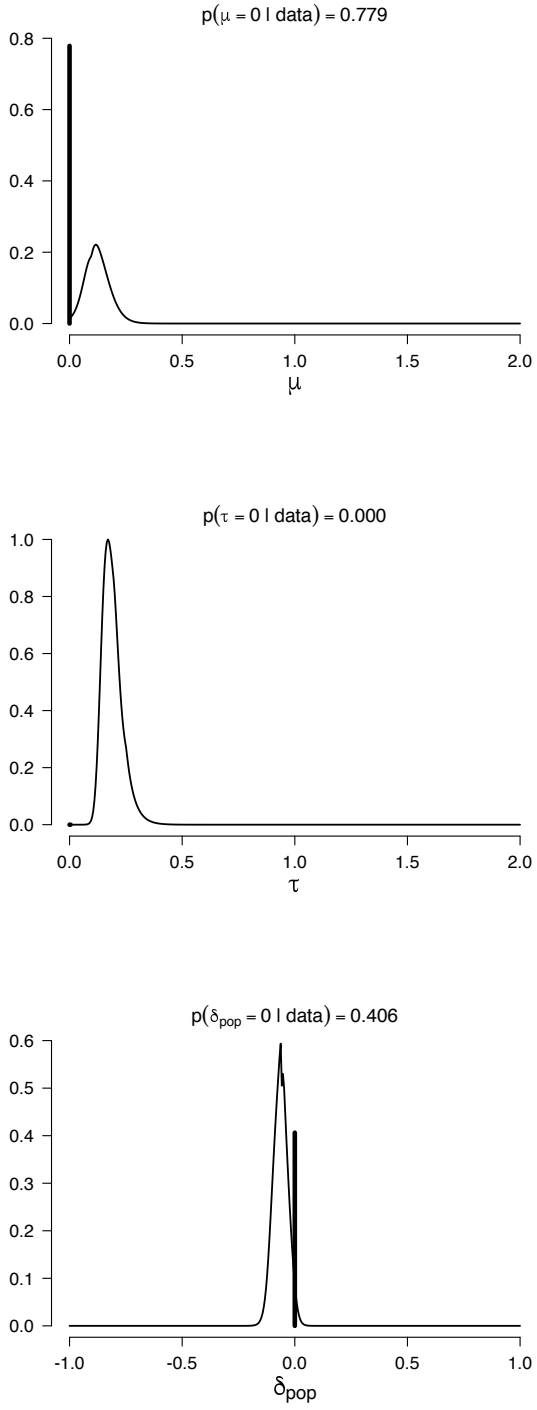

Figure 19. Model averaging results for Q5 (*unfiltered* data). The upper panel displays the results for the group-level mean effect size μ_5, the middle panel displays the results for the across-team heterogeneity τ_5, and the lower panel displays the results for the difference $\delta_{\text{pop},5}$ 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.

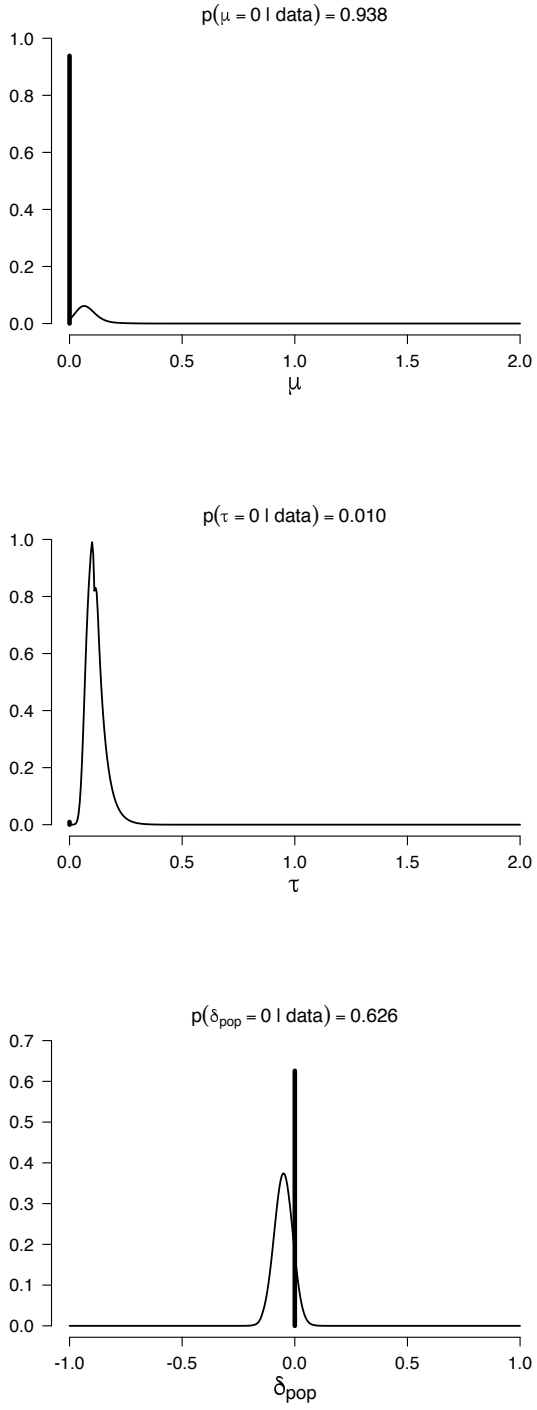

Figure 20. Model averaging results for Q5 (*filtered* data). The upper panel displays the results for the group-level mean effect size μ_5, the middle panel displays the results for the across-team heterogeneity τ_5, and the lower panel displays the results for the difference $\delta_{\text{pop},5}$ 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.

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 1.673 in favor of the proposition that $\delta_{\text{pop},5}$ equals 0. The summed posterior probability for the models in which $\delta_{\text{pop},5} = 0$ equals 0.626. The lower panel of Figure 20 shows the model-averaged posterior distribution for $\delta_{\text{pop},5}$ across all eight models, where the height of the spike at zero corresponds to the summed posterior probability that $\delta_{\text{pop},5} = 0$. 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 an 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 where included in the null model, under the "Model" tab. The results of this analysis is summarized in Table 3. The Bayes factor of $\text{BF}_{01} = 12.03$ (2.69 % error) indicates evidence for absence over presence of an lab effect. A similar conclusion can be drawn from the descriptive plot of Fig. 21,

Table 3
Model Comparison - standardisedAcrossAll

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

which plots the latent abilities of each lab separated by question with a 95% credible interval.

In addition to the analysis summarized in Table 3, we also performed the same analysis based on the unstandardized effect sizes. The results are provided by Table 4 and note that the evidence in favor of absence over presence of lab effect increases: $\text{BF}_{01} = 38.26$ (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

[7]For the reported two-sample tests the effective sample sizes were used.

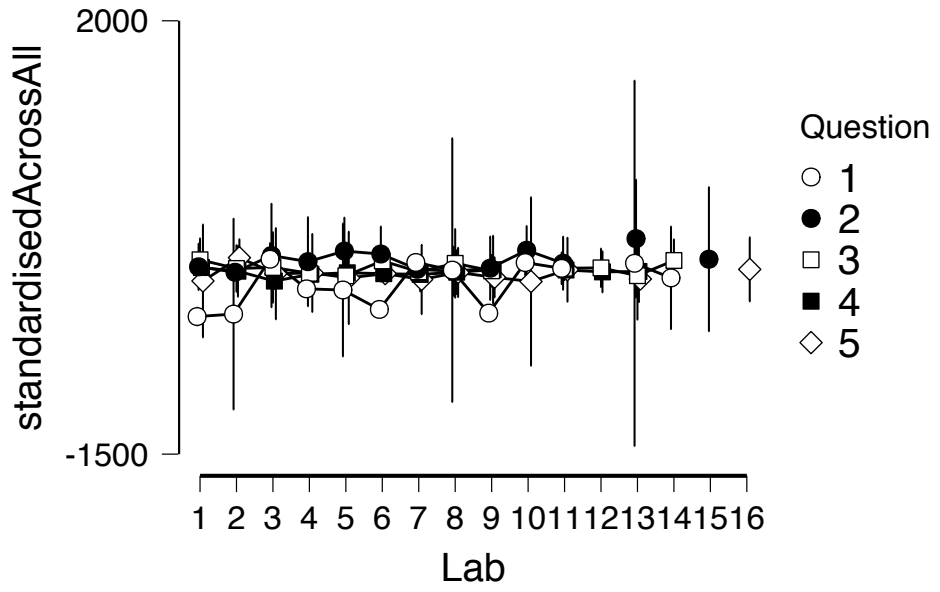

Figure 21. Descriptives plot with 95% credible interval separated by questions and lab on the horizontal axis. Note that not all labs designed studies for all five questions.

Table 4
Model Comparison - effectSize

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

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 5. Each of these ten models were given a prior probability of $P(\mathrm{M}) = 0.10$, 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 $P(\mathrm{M} \mid \mathrm{data}) = 0.56$ 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 $\mathrm{BF}_{10} = 7.60e+6$ 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 $\mathrm{BF}_M = 11.25$, that is, $P(\mathrm{M} \mid \mathrm{data})/(1 - P(\mathrm{M} \mid \mathrm{data})) = 0.56/0.44$ divided by $P(\mathrm{M})/(1 - P(\mathrm{M})) = 0.10/0.90$.

Note that the factor Question appears in several models and one might be in-

Table 5
Model Comparison - standardisedAcrossAll

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

terested to study how effective it is to include this factor across all these models. Table 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 6
Analysis of Effects - standardisedAcrossAll

Effects	P(incl)	P(incl|data)	$BF_{Inclusion}$
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 $BF_{01} \approx 1$. The Bayes factor of $BF_{01} = 1.40$ with 2.67% error indicates neither evidence for or against an lab effect, see Table 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 Fig. 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

Table 7
Model Comparison - standardisedAcrossBetter

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

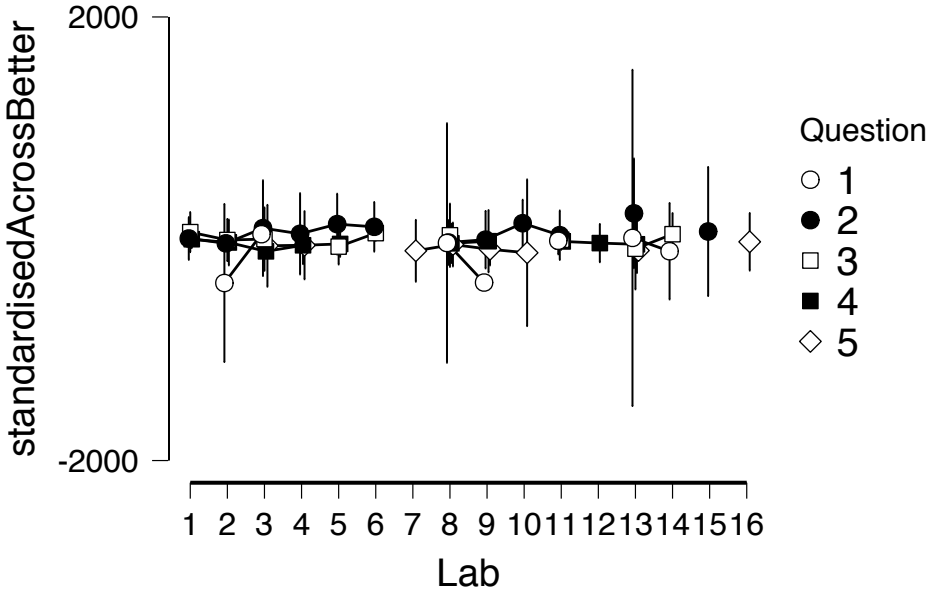

Figure 22. Descriptives plot with 95% credible interval separated by questions and lab on the horizontal axis based on studies that were rated five or higher.

into the null model. The results are summarized in Table 8, which shows that the "Question"-model is $BF_{10} = 238,598.75$ times more likely than the intercept only model. Similarly, Table 8 shows that after seeing the data, the evidence in favor of including the factor Question in the model went up, i.e., $BF_{\text{Inclusion}} = 349,323.04$, whereas the inclusion Bayes factors $BF_{\text{Inclusion}} = 0.91$ and $BF_{\text{Inclusion}} = 0.57$ 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.

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

Table 8
Analysis of Effects - standardisedAcrossBetter

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

change the results.

- 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.

- We did not calculate the replication Bayes factors, because the summary statistics data lead to effect size estimates and standard errors that differed slightly from the data set given to us. Hence, this would introduce new inconsistencies.

- For the transformation used for the ANOVA I used the effect sample sizes instead of the sample sizes of the two groups.

References

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 0.9.11-1.* 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