Online Appendix: Do Women Respond Less to Performance Pay? Building Evidence from Multiple Experiments

Oriana Bandiera, Greg Fischer, Andrea Prat and Erina Ytsma

A Estimation

Our estimation of the Bayesian hierarchical models follows closely the procedures described in Gelman and Hill (2007) and Gelman et al. (2004). For clarity of exposition, we describe the univariate model here, which extends immediately to the full multivariate model. Following equation (3) in the main text, we assume that the site-specific effects, η_s, are drawn from a normal distribution with hyperparameters (η, τ):

$$p(\eta_1, \ldots, \eta_S | \eta, \tau^2) = \prod_{s=1}^{S} N(\eta_s | \eta, \tau^2).$$

Applying Bayes Rule, the posterior of the study effects and hyperparameters conditional on the observed effects can be expressed as:[1]

$$p(\{\eta_i\}_{i=1}^{S}, \eta, \tau^2 | y) = p(\tau^2 | y) p(\eta | \tau^2, y) p(\{\eta_i\}_{i=1}^{S} | \eta, \tau^2, y).$$

It is relatively straightforward to characterize this distribution, even for extensions to multiple parameters, using Markov Chain Monte Carlo (MCMC) methods to sample iteratively from the component distributions. Intuitively, in each step k, we first simulate $\tau^{(k)}$ from its distribution and then calculate $p(\tau^2|y)$, where $y = \{\hat{\eta}_i, \hat{\sigma}_j\}_{i=1}^{S}$ is our data. Using this draw of $\tau^{(k)}$ we then sample $p(\eta|\tau^2, y)$ from the normal distribution to obtain $\eta^{(k)}$. This is then used to sample $p(\{\eta_i\}_{i=1}^{S} | \eta, \tau^2, y)$, generating each $\eta_j^{(k)}$ independently. We update parameters subject to an acceptance rule and then repeat.

[1] The marginal posterior of the hyperparameters is typically written as $p(\eta, \tau^2|y) \propto p(\eta, \tau^2) \prod_{s=1}^{S} N(\hat{\eta}_S | \eta, \sigma_s^2 + \tau^2)$, however for the normal-normal model we can simplify by integrating over η leaving $p(\eta, \tau^2|y) = p(\eta|\tau^2, y)p(\tau^2|y)$. See Gelman et al. (2004) for details.

In practice, this is easily accomplished using the RStan package for the programming language R (Stan Development Team, 2020). We use the default HMC/NUTS sampler for Stan, which employs the Hamiltonian Monte Carlo algorithm (Betancourt and Girolami, 2015) with path lengths set adaptively using the no-U-turn sampler (NUTS; Hoffman and Gelman, 2014). Inference relies on the assumption that for large enough k, the simulated distribution of $\left\{\{\eta_i\}_{i=1}^{S}, \eta, \tau^2\right\}^{(k)}$ is close to the target distribution $p(\{\eta_i\}_{i=1}^{S}, \eta, \tau^2|y)$. We initialize four independent chains for the sampler with random draws from the prior density. We then let each chain run for 14,500 iterations, discarding the first 2,000 simulations as warm-up. These parallel chains are then tested for mixing—the between-chain and within-chain variances should be equal—and stationarity. After confirming that the chains are well behaved, we combine them to generate the simulated posterior distributions for both the hyperparameters, η and τ^2, as well as the true study-level effects, $\{\eta_i\}_{i=1}^{S}$.

B Comparison with pooling model

To motivate the Bayesian hierarchical model that we estimate, it is useful to consider the pooling model as an alternative approach to aggregating empirical evidence, where we focus on univariate models for ease of exposition. The pooling model (in statistics, often referred to as the classical fixed-effects model) assumes that each individual study is estimating a common effect, η. That is, observed differences in study results are solely due to idiosyncratic variation and not differences in the sample population, type of incentive, or outcomes studied. This model has the following form:

$$\hat{\eta}_s \sim N[\eta, \sigma_s^2] \quad s = 1, \ldots, S. \tag{5}$$

This approach is quite common and easy to estimate by what is often referred to as the inverse-variance method. The estimate of the common effect η is given by the precision-weighted average of the individual study effects,

$$\hat{\eta}^{Pool} = \sum w_s^{Pool} \eta_s / \sum w_s^{Pool}, \tag{6}$$

where the weight $w_s^{Pool} = \hat{\sigma}_s^{-2}$ is the precision of our estimate for $\hat{\eta}_s$. In the presence of cross-study heterogeneity, the estimated variance of $\hat{\eta}^{Pool}$ will be too small.

B.1 Pooling model results

The pooling estimate of the gender-incentive interaction hyperparameter is, with a mean of 0.077 (s.e.: 0.038), of similar magnitude as the BHM estimate. Not surprisingly therefore, the BHM estimate of cross-study heterogeneity is relatively low (median $\tau_\eta = 0.106$), which rationalizes the similarity of the BHM and pooling estimate. Yet, despite this similarity across studies, assuming away heterogeneity, as is done in the pooling model, leads to standard errors on $\hat{\eta}$ that are too small. While the pooling model therefore suggests there is a positive gender difference in the response to incentives, zero remains in the credible interval for the BHM, which allows for and estimates heterogeneity.

The pooling estimate of the incentive effect hyperparameter γ, in contrast, is smaller than the posterior BHM estimate. With a mean of 0.276 (s.e.: 0.031), the 75^{th} percentile is smaller than the 25^{th} percentile of the BHM estimate. This difference can be explained by substantial cross-study heterogeneity. Indeed, with a median estimate of τ_γ of 0.295 and no mass on values less than 0.098, we can easily reject the pooling hypothesis.

C Pooling Metrics

A natural question to ask when synthesizing findings from comparable studies is, should we believe that each is contributing to a common answer regarding the effect in the population ($\tau^2 = 0$) or should we treat each study as a stand-alone answer to a distinct question ($\tau^2 \to \infty$). Models that explicitly recognize and quantify heterogeneity allow for a potentially more realistic intermediate answer.

It may be intuitive to think about the degree of pooling in terms of effective sample size. That is, when estimating the population hyperparameters, do we have 24,060 observations or 17? Or, in the extreme case of no pooling, is the notion of a population mean not well-defined, leaving us with effectively no observations with which to estimate it?

A range of pooling diagnostics and metrics have been developed to quantify the degree of commonality across studies. If each study is estimating a common effect, then pooling the data across studies will produce a better estimate for the parameter in *each* experiment (Rubin, 1981). The classical test of the hypothesis that the studies are all estimating a common effect yields a χ^2-statistic $\sum_{s=1}^{S}\{(\hat{\eta}_s - \hat{\eta}^{Pool})^2/\hat{\sigma}_s^2\}$, which is distributed with $S-1$ degrees of freedom.

However, pooling need not be an all or nothing proposition. Our estimates of τ^2 and the observed $\hat{\sigma}_k$s can be combined to give some sense of the extent to which observed effects are site-specific versus representing a common effect. First, note that we can characterize the mean of the Bayesian posterior as a shrinkage estimator:

$$\hat{\eta}_s^{Post} = (1 - \lambda_s)\hat{\eta}_k + \lambda_s \eta, \tag{7}$$

where $\lambda_s \in [0, 1]$ can be thought of as a pooling factor that represents the degree to which the estimates are pooled towards the estimated population mean (η) rather than based on their observed value.[2] When τ^2 is large relative to σ_s^2, we are approaching the no pooling case in which our estimate for the effect in study s will be largely determined by its own separate estimate; λ_s will be close to zero. Intuitively, when λ_s is small there is little a study in one context can tell us about the expected effect in another. In contrast, if τ^2 is small relative to σ_s^2, λ_s will be close to 1 and the appropriate estimate will be close to the population mean irrespective of the site-specific estimate. The pooling model corresponds to $\tau^2 = 0$.

Box and Tiao (1973) show that in the single parameter model when η and τ^2 are known, equation (7) characterizes the analytical mean of $\hat{\eta}_s$ with $\lambda_s = \frac{\sigma_i^2}{\sigma_i^2 + \tau^2}$. This suggests two alternative study-level pooling statistics: $\lambda_s^1 = \frac{\hat{\sigma}_i^2}{\hat{\sigma}_i^2 + \hat{\tau}^2}$, that is, the variance pooling metric calculated from the posterior means of the error terms, and $\lambda_s^2 = \frac{\hat{\eta}_k^{POST} - \hat{\eta}_k}{\eta - \hat{\eta}_k}$, a shrinkage metric that directly measures the extent to which the posterior mean of the study-level effect is determined by the posterior mean of the population effect. Note that in the multivariate model, λ_s^2 is not restricted to the interval $[0, 1]$. Correlation with other parameters makes it possible that the true effect in a study is outside the interval between the observed effect and the population mean.[3]

Gelman and Pardoe (2006) generalize this idea to develop a common pooling factor that summarizes the extent to which estimates at each level

[2] It is more common in the statistics literature to see this formulation expressed in terms of a shrinkage factor equal to $1 - \lambda_s$. Since we are primarily interested in the extent to which study-level results can be thought of as providing information about a population mean, we find it more natural to follow Gelman and Pardoe (2006) and focus on the degree of pooling.

[3] For example, suppose we observe a strong negative correlation between β and η, implying that women are relatively more responsive to incentives in settings when women's unincentivized performance is comparatively less. All else equal, when evaluating incentives for a task when women are at a comparative disadvantage, we will tend to have a higher posterior belief for the gender difference in the response to incentives.

of a hierarchical model are pooled together based on level-specific factors rather than based on lower-level or study-specific estimates. In the case of our two-level model, they define the pooling factor as

$$\lambda = 1 - \frac{V_{s=1}^{K} E(\epsilon_s)}{E(V_{s=1}^{K} \epsilon_s)}, \tag{8}$$

where E represents the posterior mean, V is the finite sample variance operator (i.e., $V_{i=1}^{n} = \frac{1}{n-1}\sum(x_i - \bar{x})^2$), and $\epsilon_s = \eta_s - \eta$. They suggest that the value of 0.5 provides a clear reference point. If $\lambda < 0.5$ there is more information at the study level than at the population level. At the extreme of $\lambda = 0$, there is no pooling and the broader population contributes no information to the true effect in a particular setting. When $\lambda > 0.5$, there is more information at the population-level, with local estimates being fully pulled toward the population mean at the extreme of $\lambda = 1$.

Finally, we can look directly at the marginal posterior density of the variance hyperparameter, $p(\tau|y)$. This is useful in that study-level posterior means can easily be calculated as functions of τ and the posterior uncertainty about τ and η_s displayed visually.

C.1 Estimates

Consistent with the posterior estimates for each of the τ parameters reported in Table 2 in the paper and depicted in Figure A1 in the Appendix, the pooling metrics (Appendix Table A4) demonstrate substantial commonality across studies for the gender-incentive interaction term (η). The common pooling factor of 0.806 means that with respect to any given study, there is relatively more information at the population level, that is, from the other $n - 1$ studies, than from the individual study itself. The average variance pooling factor across the studies is 0.440, suggesting that along this dimension the studies in our sample have reasonably high external validity. Results in one context have a substantial influence on our beliefs in another.

In contrast, the results for the incentive (γ) and gender (β) main effects exhibit more local-level than population-level information. The common pooling factors are 0.252 and 0.275, respectively, suggesting that while each experiment informs and is informed by beliefs about the population mean, most of the information about these effects must come from the context itself.

This is perhaps not surprising. The studies in our sample exhibit tremendous variation in both the type of task and the form of incentives. What is,

however, surprising is that men and women respond similarly to financial workplace incentives across such a diverse set of contexts.

D Posteriors

The Bayesian hierarchical model provides a precise and transparent method to incorporate data from other studies into our beliefs regarding the true effect in a particular setting. As noted in the main text, the best (i.e., lowest mean squared error) estimate for the true effect in a particular context is typically not equal to the mean estimate of a single, internally valid study in that context. Figures A2, A3, and A4 compare the posterior predicted distributions for each of the main parameters, η, γ, β, to the original estimates from the studies themselves. The posterior estimates are pulled towards the population mean to the extent the studies appear to be estimating a common parameter, as tempered by the precision of the study-specific, internally valid estimate and other available information such as the estimates of covarying parameters. The common and predictable pattern is that the posteriors for each study mostly lie between the original and the hyperparameter estimates. What is most surprising is that some of the gaps, that is, the degree of pooling, are quite large. This is most evident for the incentive-gender interaction (η), where the common pooling factor is large and some of the study-level estimates quite imprecise. However, there are still substantial differences between the posterior and the site-specific estimates for the other parameters in several studies.

Take, for example, the estimated effect of incentives (γ) in Bandiera et al. (2005). As shown in Figure A3, the parameter estimate in this study is large, $+0.86\sigma$, with a standard error of 0.16σ. However, with a 95%-credible interval spanning $[0.55, 1.17]$, there remains quite a bit of uncertainty about the magnitude of the effect. Furthermore, the estimates are substantially larger than the mean in all but four other studies. The mean of the posterior distribution for γ_s is $+0.74\sigma$, still a very large effect but pulled substantially towards the population mean of $+0.36\sigma$. The degree of pooling depends primarily on the uncertainty of the local parameter estimate and the estimated distribution of the population hyperparameter (γ, τ_γ).

Figure A5 demonstrates the relationship between the estimated standard deviation of the hyperparameter (τ_η) and the posterior mean of η_s, the study-specific effect. Here, we return to the gender-incentive interaction term, our primary outcome of interest. The upper half of the figure plots the posterior distribution of η_s for each study conditional on τ_η. If τ_η were 0, each study

would be estimating a common effect and the posterior for each η_s would be equal to our posterior estimate of the population mean. As τ_η increases, the extent to which the posterior for any study is pooled toward the population mean diminishes, and as $\tau_\eta \to \infty$ the posterior for each study tends towards the site-specific estimate.

Figure A5 shows that the posterior estimates for each η_s diverge rapidly as τ_η increases. For values of τ_η above 0.5 the posteriors for each study are very close to the site-specific estimate. The lower half of Figure A5 overlays the posterior distribution of τ_η, which has a mean estimate of 0.114. The substantial degree of observed pooling can be seen at the corresponding level of τ in the upper half of the figure.

E Model Checking

After computing the posterior distribution of all parameters, we can test how well the predictions of our model fit observed but unmodeled features of the data. It is, of course, possible alternative probability models could also fit our data but generate different posterior predictions. Therefore, we also test the sensitivity of our posterior predictions to alternative assumptions. Our aim is not so much to accept or reject the model, but to understand the limits of its applicability.

The key idea behind posterior predictive checking is that data replicated under our estimated model should look similar to the observed data (Gelman et al., 2004). We can construct test statistics, T, from any function of the data and then calculate the Bayesian p-value for each of these statistics:

$$p = Pr\left(T(y^{sim}, \theta) \geq T(y, \theta | \, y)\right).$$

These p-values can be directly interpreted as the probability that the test statistic in the posterior distribution, y^{sim}, is larger than in the observed data. For example, we calculate the share of draws from the simulated data for which the maximum $\hat{\eta}_s$ is greater than what was observed in the actual data. Thus, p-values near 0 or 1 indicate that the statistic observed in the data would be unlikely to be seen in simulations based on our specified probability model.

Figure A6 plots the observed order statistic for each of the model parameters against the mean from the simulated posterior distribution[4]. In the case of the gender-incentive interaction term, the posterior predictive

[4]Table A5 in the Appendix reports the associated Bayesian p-values.

distribution matches the observed data very well, including at the extremes. Although the settings for the included studies were certainly not chosen at random from the population of possible study sites, our hierarchical model that treats the study-level parameters as if they were normally distributed around a population mean does a remarkably good job of capturing important features of the data. The model also performs reasonably well for the gender (β) and incentive (γ) parameters, with the exception of slightly fatter tails in the distribution of γ.

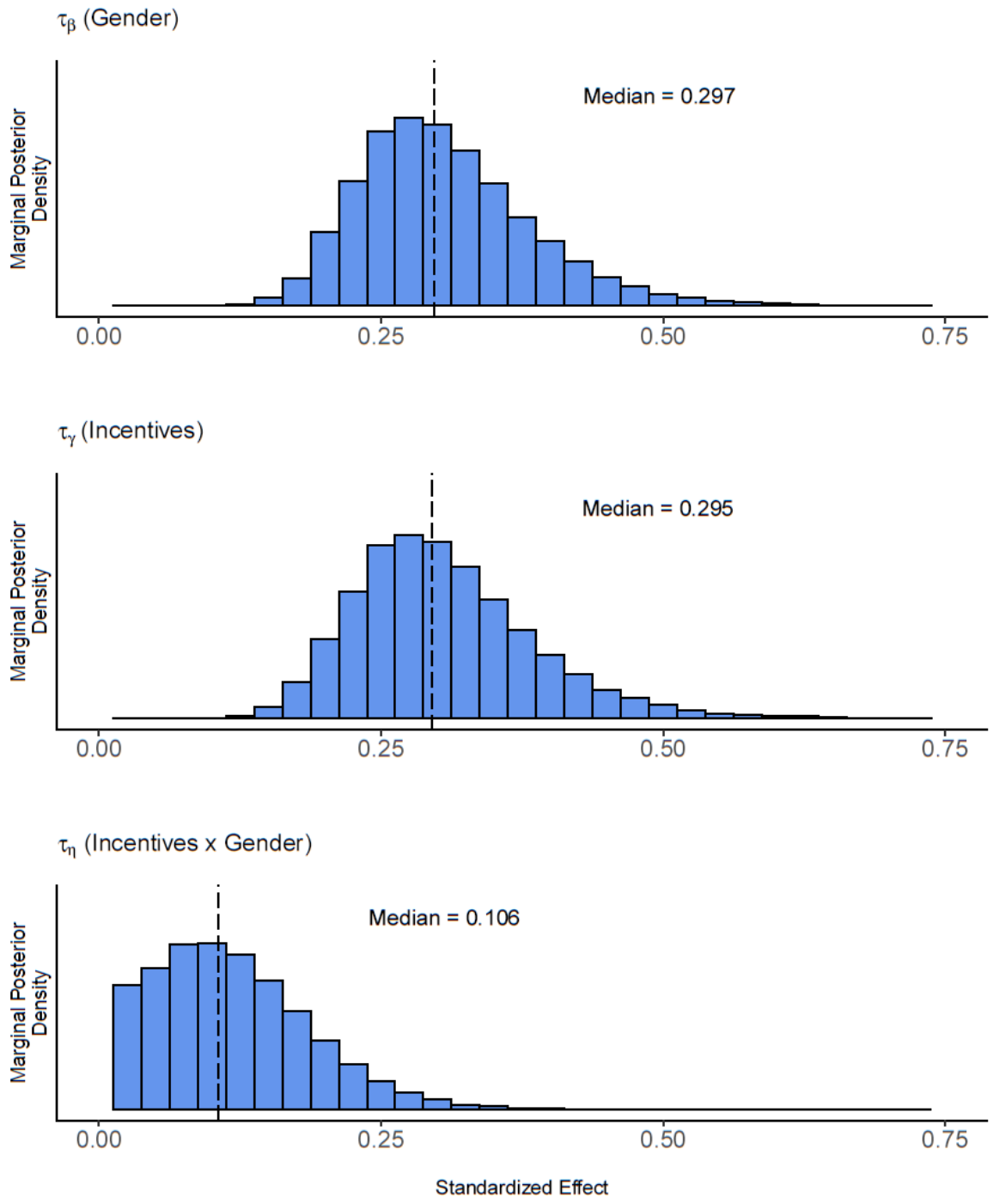

FIGURE A.1. POSTERIOR DISTRIBUTION OF τ (HYPERPARAMETER VARIANCE)

Notes: Figure shows the full posterior distribution of the hyperparameter variance. See section III.A for details.

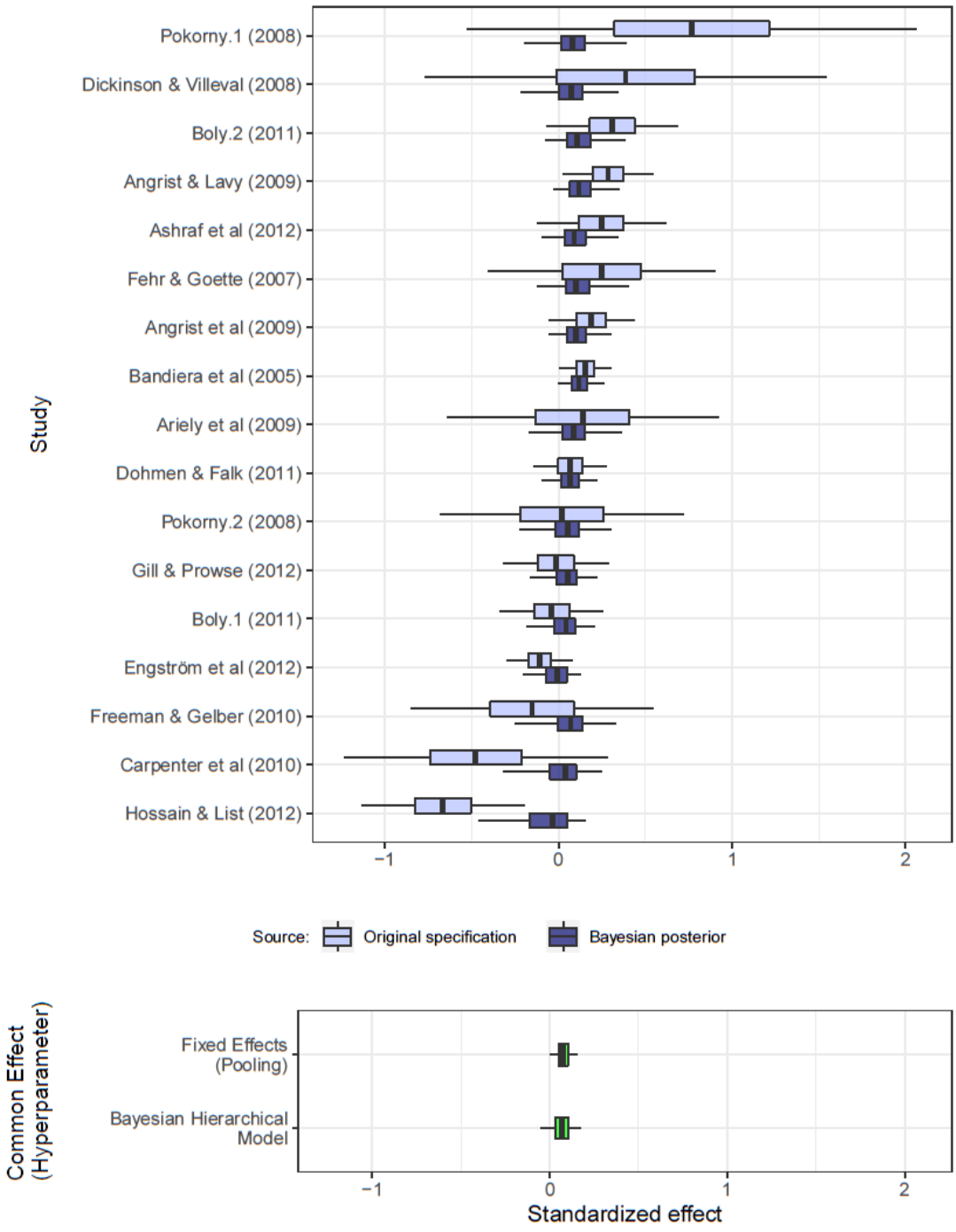

FIGURE A.2. ORIGINAL & POSTERIOR ESTIMATES FOR η (INCENTIVES X GENDER)

Notes: Outcome variable for each study is standardized using mean and standard deviation of men in control group. Vertical line indicates median estimate, box indicates 50%−interval and line indicates 95%−interval. Fixed effects model calculated using the metafor package for R (Viechtbauer, 2010). Bayesian Hierarchical model implemented in Rstan (Stan Development Team, 2020). See Appendix sections B and D for details.

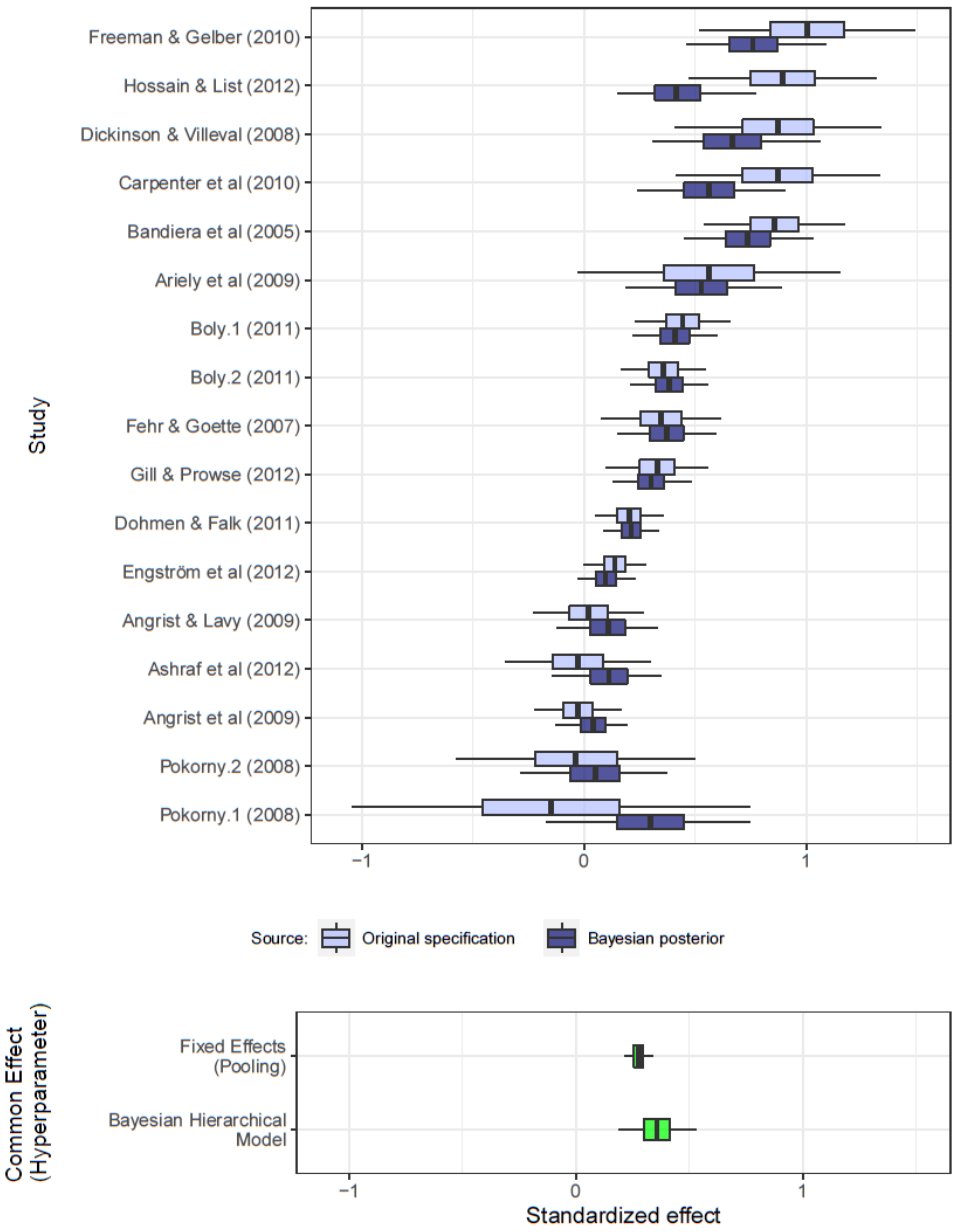

FIGURE A.3. ORIGINAL & POSTERIOR ESTIMATES FOR γ (INCENTIVES)

Notes: Outcome variable for each study is standardized using mean and standard deviation of men in control group. Vertical line indicates median estimate, box indicates 50%−interval and line indicates 95%−interval. Fixed effects model calculated using the metafor package for R (Viechtbauer, 2010). Bayesian Hierarchical model implemented in Rstan (Stan Development Team, 2020). See Appendix sections B and D for details.

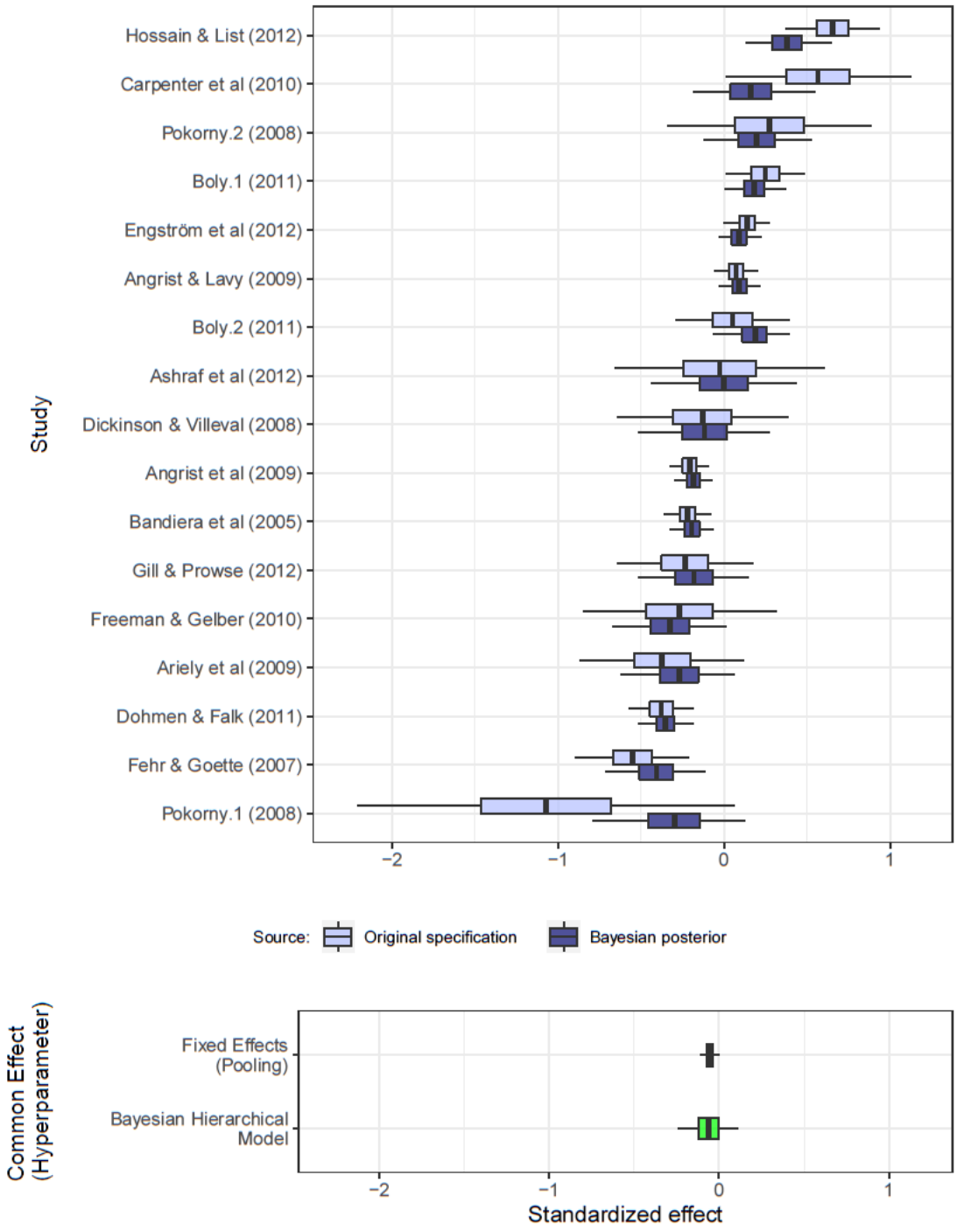

FIGURE A.4. ORIGINAL & POSTERIOR ESTIMATES FOR β (GENDER)

Notes: Outcome variable for each study is standardized using mean and standard deviation of men in control group. Vertical line indicates median estimate, box indicates 50%−interval and line indicates 95%−interval. Fixed effects model calculated using the metafor package for R (Viechtbauer, 2010). Bayesian Hierarchical model implemented in Rstan (Stan Development Team, 2020). See Appendix sections B and D for details.

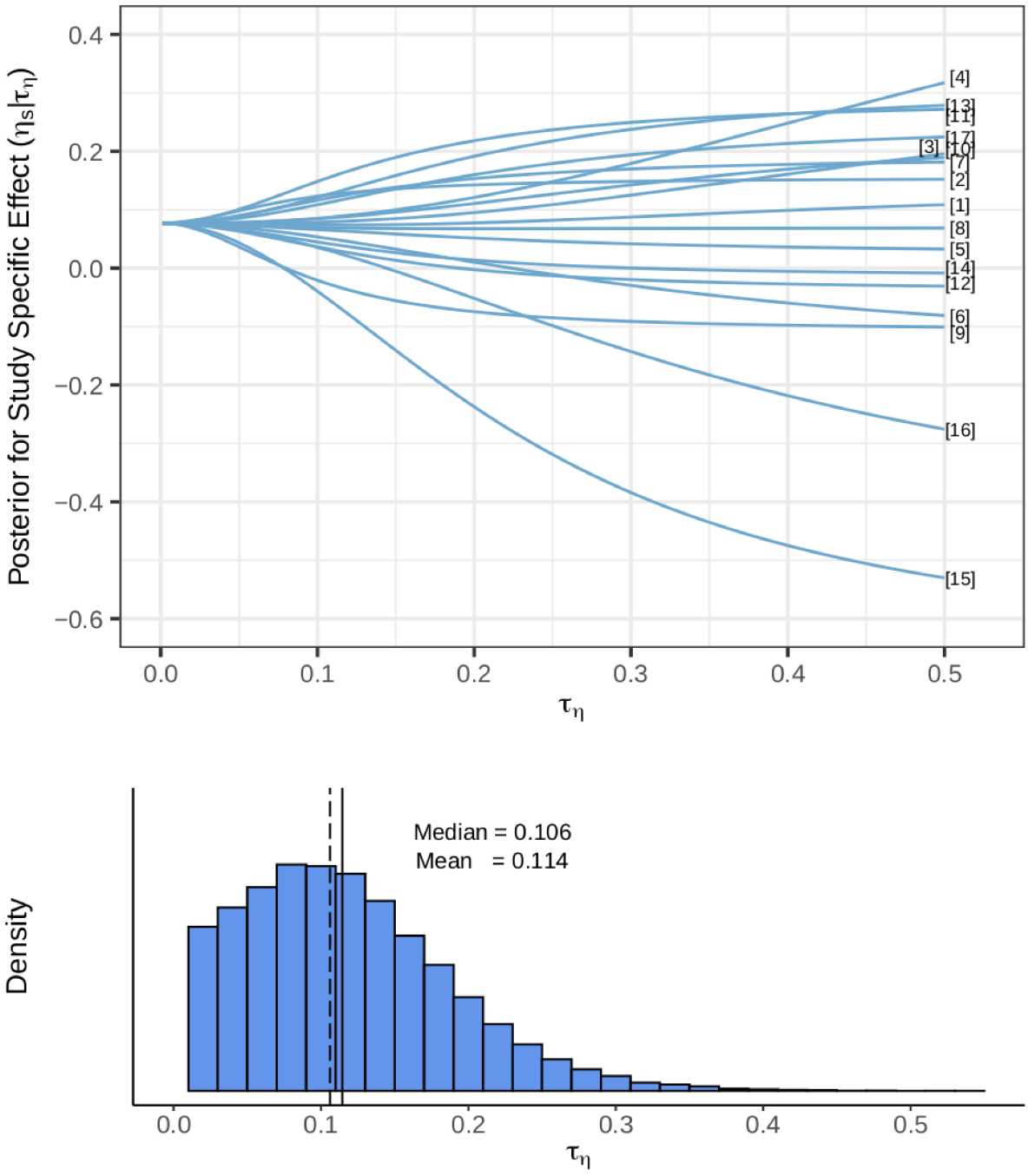

FIGURE A.5. POSTERIOR MEAN OF η_s (GENDER X INCENTIVES) CONDITIONAL ON τ_η

Notes: Conditional posterior for: [1] Ariely et al (2009); [2] Bandiera et al (2005); [3] Fehr & Goette (2007); [4] Pokorny.1 (2008); [5] Pokorny.2 (2008); [6] Freeman & Gelber (2010); [7] Angrist et al (2009); [8] Dohmen & Falk (2011); [9] Engstrom et al (2012); [10] Dickinson & Villeval (2008); [11] Angrist & Lavy (2009); [12] Boly.1 (2011); [13] Boly.2 (2011); [14] Gill & Prowse (2012); [15] Hossain & List (2012); [16] Carpenter et al (2010); and [17] Ashraf et al (2012). See Appendix section D for details.

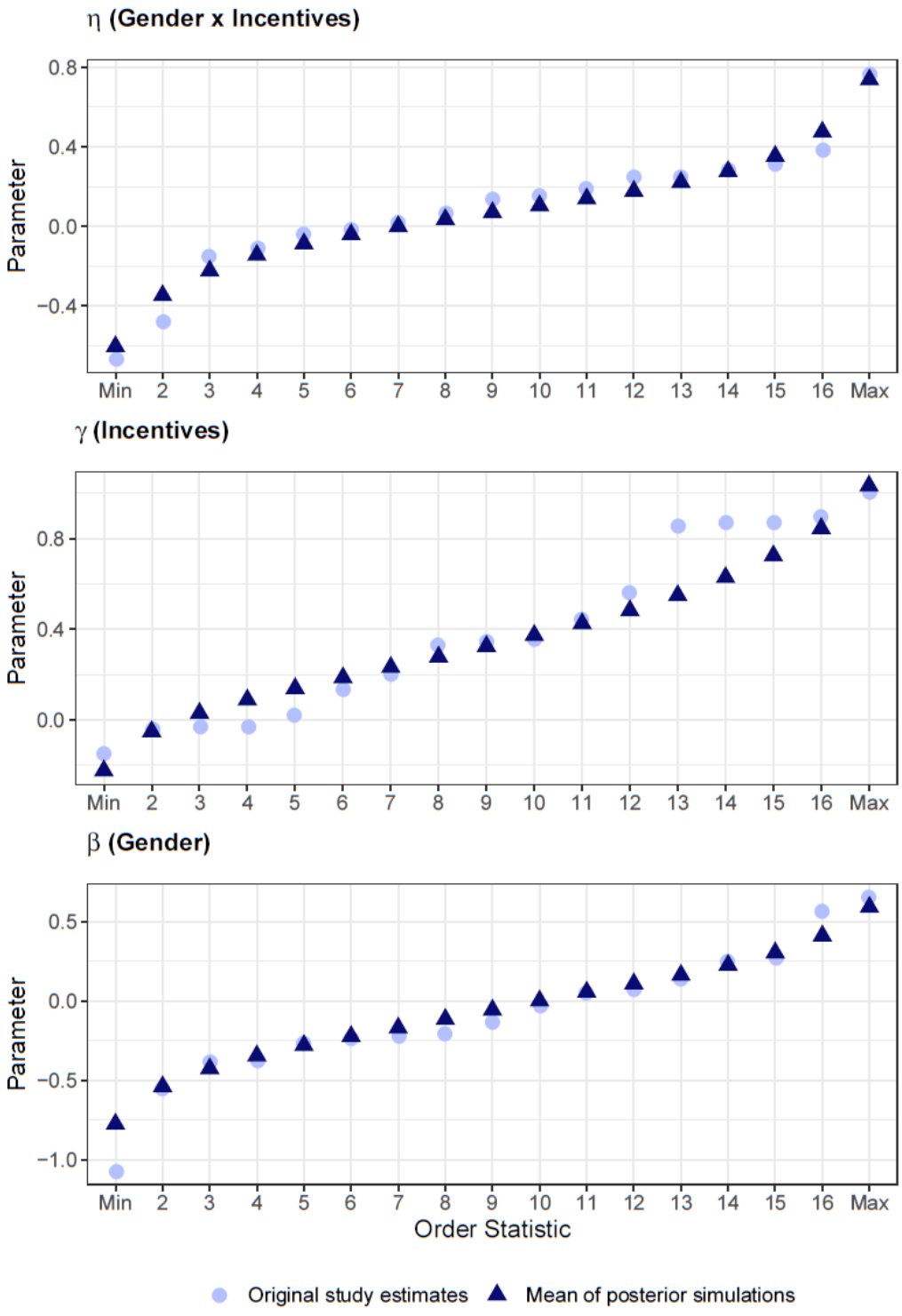

FIGURE A.6. POSTERIOR PREDICTIVE CHECKS, ORDER STATISTICS

Notes: Each plot compares the order statistic for observed parameter estimates to the analogous mean in posterior simulations. See Appendix section E for a further discussion of posterior predictive checks.

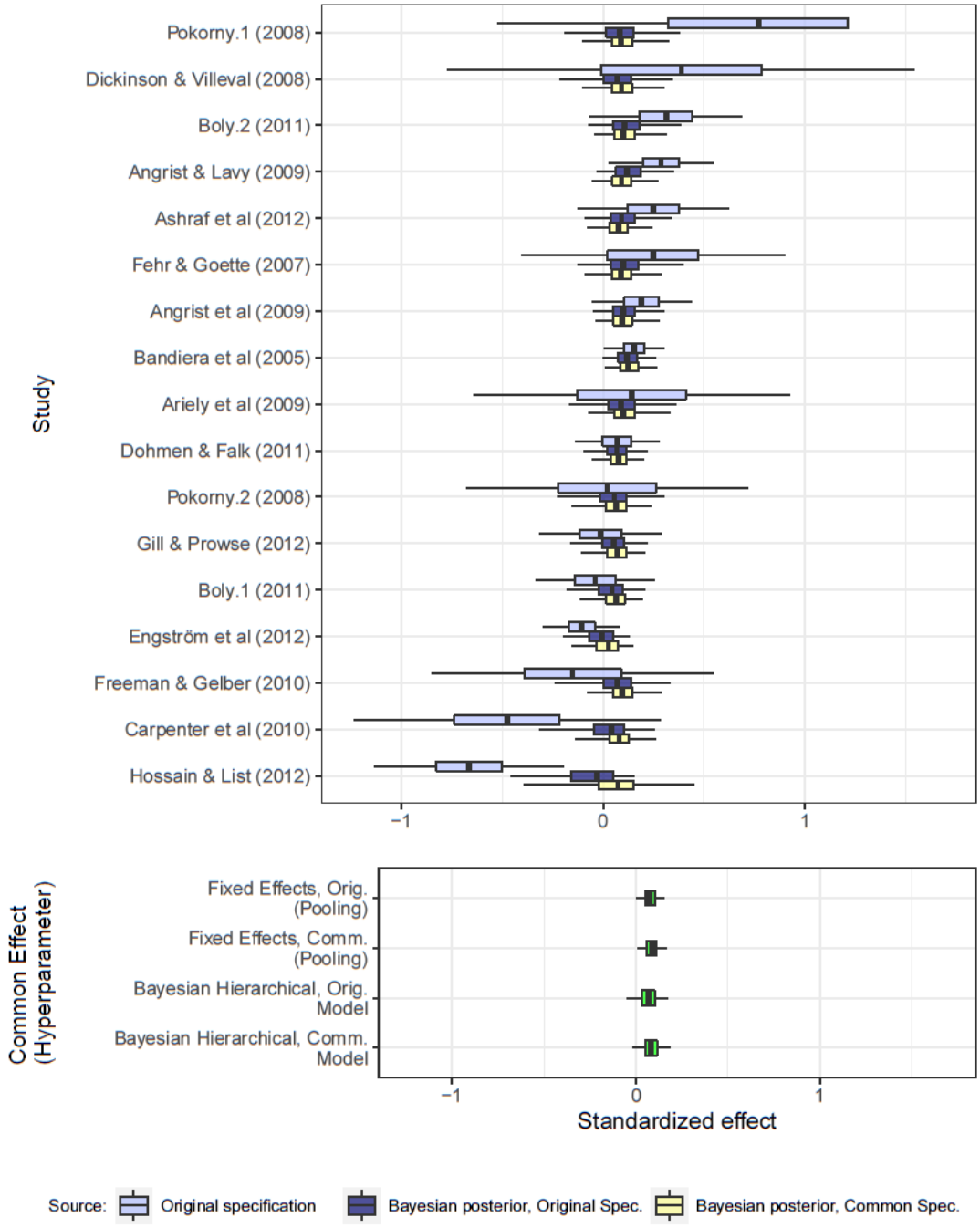

FIGURE A.7. POSTERIOR ESTIMATES FOR INCENTIVES X GENDER ORIGINAL & COMMON SPECIFICATIONS

Notes: Outcome variable for each study is standardized using mean and standard deviation of men in control group. Vertical line indicates median estimate, box indicates 50%−interval and line indicates 95%−interval. Fixed effects model calculated using the metafor package for R (Viechtbauer, 2010). Bayesian Hierarchical model implemented in Rstan (Stan Development Team, 2020). See section II and Appendix section B for details.

TABLE A.1 — Selection Criteria

Criterion	Requirement
Quality control	- Papers published in peer reviewed journals or renowned working paper series
Comparable identification	- Variation in incentive power generated randomly (either lab or field)
	- Only monetary performance rewards
Workplace relevance	- At least two treatments that can be ranked according to their power
	- Real, costly effort
	- Higher effort leads to higher output
Confounding mechanisms	- No externalities in production
	- No self-selection according to incentives

TABLE A.2 — Summary of Included Studies, Specification Detail

Study	Closest Specification in Paper	Specification in Meta-Analysis
Angrist & Lavy (2009)	Table 2, Panel A, column 1, row 2 (page 1394)	OLS of Bagrut status on a treatment dummy, a dummy for Arab schools and a dummy for Jewish religious schools, a female dummy and female X treatment interaction effect, with pair-randomization fixed effects. BRL (Biased Reduced Linearisation) standard errors, clustered at the school level, are estimated.
Angrist et al (2009)	Table 5, Panel B, column 4 (page 149)	OLS of first year GPA on a dummy for combined bonus treatments (SFP (any)), a female dummy and female X treatment interaction effect, a dummy for the peer advising treatment (SSP), as well as a full set of controls (mother tongue dummies, high school quartile dummies, dummies for the number of courses enrolled in, and dummies for responses to survey questions on procrastination and parents' education). Sample restricted to students with fall grades in year 1 and excluding no-shows. Standard errors are robust.
Ariely et al (2009)	Table 1, Panel A, column 1 (page 551)	OLS of key press pairs on a dummy for private monetary incentives, a female dummy and female X private monetary incentives interaction, controlling for individual perceptions of the majority view of the specific cause (on full scale: -5 to +5), for subjects in the private condition only. Robust standard errors.
Ashraf et al (2012)	Table 1, column 2 (page 42)	OLS of condoms sold on a dummy for large financial reward, a female dummy and female X large financial reward interaction, dummies for whether the shop is a barbershop or barbershop and hairdresssers, a dummy for whether the shop is near a bar, log number of employees, number of trained salons in same area (cell), whether stylist sells other products, whether stylist is in bottom quartile of asset disribution, whether stylist's socio-economic status is low, whether stylist's donation in dictator game is above the median, whether stylist's self-reported motivation is social, whether stylist's religion is roman catholic. Restricted to subjects in the voluntary condition and high financial reward treatment. Standard errors clustered at cell (area) level.
Bandiera et al (2005)	Table 2, column 4 (page 934)	Linear regression of log productivity on a piece rate dummy, female dummy, female X piece rate interaction, a time trend, field life cycle and worker experience, with field fixed effects and standard errors clustered at worker and field-day level.
Boly (2011)	Table 4, column 1 (page 249) Table 4, column 2 (page 249)	Random effects GLS of the negative of the absolute deviation between the number of mistakes in an exam reported by subject and the number of actual mistakes, controlling for age, a female dummy, a treatment dummy for any monitoring (low or high), a female X any monitoring interaction, paper ranking (1 through 10) and paper ranking * treatment dummy interaction effects, with exam paper fixed effects. Standard errors clustered at individual marker (= subject) level. Restricted to normal wage, low and high monitoring treatments. Restricted to lab or field environment, respectively
Carpenter et al (2010)	Table 2, column 4 (page 511)	OLS of quality adjusted number of enveloppes produced on a tournament dummy, a female dummy and female X tournament interaction, dummies for international student, being in the top 10% of risk taking, and expecting teammates to correctly report output, controlling for GPA, birth order, number of siblings, employment status, the number of other participants known, and proxies for family wealth, for non-sabotage treatments only. Robust standard errors.

TABLE A.2 — Summary of Included Studies, Specification Detail

Study	Closest Specification in Paper	Specification in Meta-Analysis
Dickinson & Villeval (2008)	Table 4, column 1 (page 69)	Random effects GLS of agent score on task on monitoring probability (=treatment) variable, female dummy (1 if agent female), a female * monitoring probability interaction, a partner protocol dummy, partner protocol * monitoring probability interaction term, round number, first round of protocol dummy (1 for round 1 and 11), dummy for having partner protocol before stranger protocol, agent's risk aversion variable, same sex * partner protocol dummy interaction term, task difficulty index and task difficulty index squared, for the variable pay treatment only (in which the principal's pay increases with agent's output; monitoring probability above 0.2 is costly for principal). For normalization purposes, the control group is defined by below-median monitoring probability.
Dohmen & Falk (2011)	N/A: paper only shows graphical evidence of output responses and presents regression analyses only of sorting.	OLS of the negative of log stacked productivity indicators from step 1 and 2 (time needed to answer multiplication problem in step 1 (no pay, no time limit) and step 2 (piece rate pay, 30 sec. time limit), failures top-coded) on a dummy for step 2 (treatment dummy), a female dummy and a female X step 2 interaction effect. Robust standard errors, clustered at session level.
Engström et al (2012)	Table 4, column 2 (page 427)	OLS regression of binary variable of whether an individual applied for a job he/she was referred to on a female dummy, a treatment dummy (1 if in group A; group B is benchmark), a female X treatment interaction term, controlling for age, education level, number of prior referrals, days of UB receipt; conditional on the person receiving unemployment benefit during the referral period. Hetero-skedasticity robust standard errors.
Fehr & Goette (2007)	Table 3, column 1 (page 309)	OLS of revenues per four-week period on a treatment dummy, female dummy, a female X treatment interaction, with dummies for treatment period. Restricted to messengers that participate in the experiment only. Standard errors clustered at the individual level.
Freeman & Gelber (2010)	Table 2, column 2 (page 155)	OLS of the difference in mazes solved between round 1 (all piece rate pay) and round 2 (different pay schemes) on a dummy for either of the tournament pay schemes in round 2 (single prize or multiple prize), a female dummy and a female X tournament interaction. Restricted to the "no info" group to abstract from informational effects. Robust standard errors, clustered by group.
Gill & Prowse (2012)	Table 2, column 1 (page 15))	Random effects GLS of number of sliders correctly placed on a prize variable, a female dummy and a female X prize interaction, with dummies for rounds. Restricted to first movers. For normalization purposes, the control group is defined by prizes below the median prize level (0.5).
Hossain & List (2012)	Table 4, column 5 (page 2159)	OLS of the log of average hourly productivity in a week on three treatment dummies (one for the reward, one for the punishment and one for the gift treatment), a female dummy and female X treatment interaction terms, with set (group) fixed effects and set*week fixed effects. Restricted to individual workers. The meta-analysis uses as inputs only the estimate for the reward treatment dummy and female X reward interaction effect.
Pokorny (2008)	No regressions in paper, only graphical evidence and two-sided t-tests of performance differences across treatments (for IQ questions and Numbers Counting task separately)	OLS of score on task on a treatment dummy (1 for high incentive, low incentive and very low incentive), a female dummy and a female X treatment dummy. Restricted to IQ task and number counting task, respectively

TABLE A.3 — Overview of Studies that Meet Inclusion Criteria

Author(s)	Year	Title	Journal	Data obtained
Bettinger, Eric P.	2012	Paying to Learn: The Effect of Financial Incentives on Elementary School Test Scores	The Review of Economics and Statistics (94; 3)	Regression results
Duflo, Esther ; Hanna, Rema; Ryan, Stephen	2012	Incentives Work: Getting Teachers to Come to School	American Economic Review (102; 4)	No gender variation
Hossain, Tanjim; List, John A.	2012	The Behavioralist Visits the Factory: Increasing Productivity Using Simple Framing Manipulations	Management Science (58; 12)	Data set
Ashraf, Nava; Bandiera, Oriana; Jack, Kelsey	2012	No margin, no mission? A Field Experiment on Incentives for Pro-Social Tasks	CEPR discussion paper 8834	Data set
Engström, Per; Hesselius, Patrik; Holmlund, Bertil	2012	Vacancy Referrals, Job Search, and the Duration of Unemployment: A Randomized Experiment	Labour (26; 4), p. 419-435	Data set
Gill, David; Prowse, Victoria	2012	A Structural Analysis of Disappointment Aversion in a Real Effort Competition	American Economic Review (102; 1)	Data set
Fryer, Roland G. Jr.; Holden, Richard T.	2012	Aligning Student, Parent, and Teacher Incentives: Evidence from Houston Public Schools	NBER working paper 17752	Data cannot be obtained
Leuven, Edwin; Oosterbeek, Hessel; Sonnemans, Joep; Van der Klaauw, Bas	2011	Incentives Versus Sorting in Tournaments: Evidence from a Field Experiment	Journal of Labor Economics (29; 3)	Regression results
Dohmen, Thomas, J.; Falk, Armin	2011	Performance Pay and Multi-Dimensional Sorting: Productivity, Preferences and Gender	American Economic Review (101; 2)	Data set
Boly, Amadou	2011	On the incentive effects of monitoring: evidence from the lab and the field	Experimental Economics (14; 2)	Data set
Fryer, Roland G. Jr.	2011	Financial Incentives and Student Achievement: Evidence from Randomized Trials	Quarterly Journal of Economics (126)	Data cannot be obtained
Bellamare, Charles; Lepage, Patrick; Shearer, Bruce	2010	Peer pressure, incentives, and gender: An experimental analysis of motivation in the workplace	Labour Economics (17; 1)	Regression results
Shi, Lan	2010	Incentive Effect of Piece Rate Contracts: Evidence from Two Small Field Experiments	B.E. Journal of Economic Analysis and Policy (10; 1)	Regression results
Freeman, Richard B.; Gelber, Alexander M.	2010	Prize Structure and Information in Tournaments: Experimental Evidence	American Economic Journal: Applied Economics (2; 1)	Data set
Carpenter, Jeffrey P.; Matthews, Peter Hans; Schirm, John	2010	Tournaments and Office Politics: Evidence from a Real Effort Experiment	American Economic Review (100; 1)	Data set

Leuven, Edwin; Oosterbeek, Hessel; Van der Klaauw, Bas	2010	The effect of financial rewards on students' achievement: Evidence from a randomized experiment	Journal of the European Economic Association (8; 6)	Gender not recorded
Muralidharan, Karthik; Sundararaman, Venkatesh	2011	Teacher Performance Pay: Experimental Evidence from India	Journal of political Economy (119; 1)	Regression results
Ariely, Dan; Bracha, Anat; Meier, Stephan	2009	Doing Good or Doing Well? Image Motivation and Monetary Incentives in Behaving Prosocially	American Economic Review (99; 1)	Data set
Angrist, Joshua; Lang, Daniel; Oreopoulos, Philip	2009	Incentives and Services for College Achievement: Evidence from a Randomized Trial	American Economic Journal: Applied Economics (1; 1)	Data set
Angrist, Joshua; Lavy, Victor	2009	The Effects of High Stakes School Achievement Awards: Evidence from a Randomized Trial	American Economic Review (99; 4)	Data set
Dickinson, D.; Villeval, Marie Claire	2008	Does Monitoring Decrease Work Effort? The Complementarity between Agency and Crowding-out Theories	Games and Economic Behavior 63	Data set
Manthei-Pokorny, Kathrin	2008	Pay—but do not pay too much: An experimental study on the impact of incentives	Journal of Economic Behavior and Organization (66; 2)	Data set
Paarsch, Harry J.; Shearer, Bruce S.	2007	Do Women React Differently to Incentives? Evidence from Experimental Data and Payroll Records	European Economic Review (51)	Regression results
Fehr, Ernst; Goette, Lorenz	2007	Do Workers Work More if Wages Are High? Evidence from a Randomized Field Experiment	American Economic Review (97; 1)	Data set
Pozo, Susan; Stull, Charles A.	2006	Requiring a Math Skills Unit: Results of a Randomized Experiment	American Economic Review (96; 2)	No response
Bandiera, Oriana; Barankay, Iwan; Rasul, Imran	2005	Social Preferences and the Response to Incentives: Evidence from Personnel Data	Quarterly Journal of Economics (120; 3)	Data set
Nagin, Daniel S.; Rebitzer, James B.; Sanders, Seth; Taylor, Lowell J.	2002	Monitoring, Motivation, and Management: The Determinants of Opportunistic Behavior in a Field Experiment	American Economic Review (92; 4)	Data cannot be obtained
Gneezy, U.; Rustichini, A.	2000	Pay Enough or Don't Pay at All	Quarterly Journal of Economics (115; 3)	Gender not recorded
Dickinson, David L.; Villeval, Marie Claire	1999	An Experimental Examination of Labor Supply and Work Intensities	Journal of Labor Economics (17; 4)	Regression results

TABLE A.4 — Pooling Metrics

	η (Gender x Incentives)		γ (Incentives)		β (Gender)	
Common pooling factor	0.806		0.252		0.275	
By study	**Variance**	**Shrinkage**	**Variance**	**Shrinkage**	**Variance**	**Shrinkage**
Angrist & Lavy (2009)	0.410	0.704	0.200	0.256	0.126	-0.146
Angrist et al (2009)	0.366	0.664	0.147	0.177	0.117	0.165
Ariely et al (2009)	0.484	0.690	0.307	0.160	0.295	0.323
Ashraf et al (2012)	0.423	0.814	0.216	0.355	0.385	-0.701
Bandiera et al (2005)	0.277	0.371	0.252	0.240	0.127	0.162
Boly.1 (2011)	0.365	0.693	0.166	0.404	0.157	0.208
Boly.2 (2011)	0.456	0.774	0.157	-86.377	0.195	-1.150
Carpenter et al (2010)	0.523	0.909	0.291	0.596	0.320	0.638
Dickinson & Villeval (2008)	0.503	0.989	0.336	0.389	0.346	0.176
Dohmen & Falk (2011)	0.305	0.979	0.120	0.061	0.144	0.079
Engström et al (2012)	0.346	0.528	0.124	-0.184	0.124	0.239
Fehr & Goette (2007)	0.496	0.740	0.192	2.331	0.254	0.292
Freeman & Gelber (2010)	0.517	0.985	0.272	0.373	0.290	-0.280
Gill & Prowse (2012)	0.358	0.743	0.154	-0.967	0.282	0.300
Hossain & List (2012)	0.615	0.813	0.270	0.867	0.225	0.379
Pokorny.1 (2008)	0.531	0.971	0.412	0.881	0.404	0.759
Pokorny.2 (2008)	0.495	0.633	0.295	0.216	0.279	0.224

Notes: See Appendix Section C for a discussion of pooling factor calculations. The common pooling factor λ is defined in equation 8. The variance pooling metric λ1 and the shrinkage metric λ2 are both defined in the text in the same section.

TABLE A.5 — Posterior Predictive Checks, Order Statistics

	p-value		
	η (Gender x Incentives)	**γ (Incentives)**	**β (Gender)**
Order Statistic			
Min	0.643	0.438	0.866
$\theta_{(2)}$	0.776	0.502	0.578
$\theta_{(3)}$	0.335	0.770	0.379
$\theta_{(4)}$	0.403	0.931	0.635
$\theta_{(5)}$	0.328	0.949	0.478
$\theta_{(6)}$	0.401	0.758	0.577
$\theta_{(7)}$	0.415	0.658	0.739
$\theta_{(8)}$	0.335	0.239	0.863
$\theta_{(9)}$	0.169	0.380	0.791
$\theta_{(10)}$	0.248	0.570	0.638
$\theta_{(11)}$	0.248	0.402	0.536
$\theta_{(12)}$	0.187	0.190	0.674
$\theta_{(13)}$	0.354	0.006	0.598
$\theta_{(14)}$	0.417	0.033	0.377
$\theta_{(15)}$	0.580	0.145	0.574
$\theta_{(16)}$	0.642	0.357	0.138
Max	0.388	0.518	0.329

Notes: See Appendix Section E for discussion of Bayesian p-values for posterior predictive model checking. These p-values can be directly interpreted as the probability that the test statistic in the simulated posterior distribution is larger than that in the observed data. p-values near either 0 or 1 indicate that the observed data would be unlikely to be seen in simulations based on our specified probability distribution.

TABLE A.6 — Data Citations

Study	Data Citation
Panel A: Study data sets available online	
Angrist & Lavy (2009)	Angrist, Joshua, and Lavy, Victor. Replication data for: The Effects of High Stakes High School Achievement Awards: Evidence from a Randomized Trial. Nashville, TN: American Economic Association, 2009. Ann Arbor, MI: Inter-university Consortium for Political and Social Research, 2019-10-12. https://doi.org/10.3886/E113319V1
Angrist et al (2009)	Angrist, Joshua, Lang, Daniel, and Oreopoulos, Philip. Replication data for: Incentives and Services for College Achievement: Evidence from a Randomized Trial. Nashville, TN: American Economic Association, 2009. Ann Arbor, MI: Inter-university Consortium for Political and Social Research, 2019-12-07. https://doi.org/10.3886/E116327V1
Ariely et al (2009)	Ariely, Dan, Bracha, Anat, and Meier, Stephan. Replication data for: Doing Good or Doing Well? Image Motivation and Monetary Incentives in Behaving Prosocially. Nashville, TN: American Economic Association, 2009. Ann Arbor, MI: Inter-university Consortium for Political and Social Research, 2019-10-12. https://doi.org/10.3886/E113292V1
Carpenter et al (2010)	Carpenter, Jeffrey, Matthews, Peter Hans, and Schirm, John. Replication data for: Tournaments and Office Politics: Evidence from a Real Effort Experiment. Nashville, TN: American Economic Association, 2010. Ann Arbor, MI: Inter-university Consortium for Political and Social Research, 2019-10-11. https://doi.org/10.3886/E112333V1
Dohmen & Falk (2011)	Dohmen, Thomas, and Falk, Armin. Replication data for: Performance Pay and Multidimensional Sorting: Productivity, Preferences, and Gender. Nashville, TN: American Economic Association, 2011. Ann Arbor, MI: Inter-university Consortium for Political and Social Research, 2019-10-11. https://doi.org/10.3886/E112408V1
Fehr & Goette (2007)	Fehr, Ernst, and Goette, Lorenz. Replication data for: Do Workers Work More if Wages Are High? Evidence from a Randomized Field Experiment. Nashville, TN: American Economic Association, 2007. Ann Arbor, MI: Inter-university Consortium for Political and Social Research, 2019-12-07. https://doi.org/10.3886/E116259V1
Freeman & Gelber (2010)	Freeman, Richard B., and Gelber, Alexander M. Replication data for: Prize Structure and Information in Tournaments: Experimental Evidence. Nashville, TN: American Economic Association, 2010. Ann Arbor, MI: Inter-university Consortium for Political and Social Research, 2019-10-12. https://doi.org/10.3886/E113737V1
Panel B: Study data sets available online and received from study authors	
Gill & Prowse (2012)	Gill, David, and Prowse, Victoria. Replication data for: A Structural Analysis of Disappointment Aversion in a Real Effort Competition. Nashville, TN: American Economic Association, 2012. Ann Arbor, MI: Inter-university Consortium for Political and Social Research, 2019-10-11. https://doi.org/10.3886/E112499V1 Gill, David, and Prowse, Victoria n.d.: "Data for: Gender Differences and Dynamics in Competition: The Role of Luck", unpublished data. Accessed 7 September 2014. (This data set is used for the gender-incentive regressions instead of the data set provided for the AER 2012 paper by the same authors, since the AER 2012 data set does not include gender data)

TABLE A.6 — Data Citations

Study	Data Citation
Panel C: Study data received from study authors	
Ashraf et al (2012)	Ashraf, Nava; Bandiera, Oriana; Jack, B. Kelsey n.d.: "Data for: No Margin, No Mission? A Field Experiment on Incentives for Pro-Social Tasks", unpublished data. Accessed 8 May 2014.
Bandiera et al (2005)	Bandiera, Oriana; Barankay, Iwan; Rasul, Imran n.d.: "Confidential data for: Social Preferences and the Response to Incentives: Evidence from Personnel Data", unpublished data. Accessed 15 April 2015.
Boly (2011)	Boly, Amadou n.d.: "Data for: On the incentive effects of monitoring: evidence from the lab and the field", unpublished data. Accessed 13 February 2013.
Dickinson & Villeval (2008)	Dickinson, David; Villeval, Marie Claire n.d.: "Data for: Does Monitoring Decrease Work Effort? The Complementarity between Agency and Crowding-out Theories", unpublished data. Accessed 13 February 2013.
Engström et al (2012)	Engström, Per; Hesselius, Patrik; Holmlund, Bertil n.d.: "Data for: Vacancy Referrals, Job Search, and the Duration of Unemployment: A Randomized Experiment", unpublished data. Accessed 20 March 2013.
Hossain & List (2012)	Hossain, Tanjim; List, John A. n.d.: "Data for: The Behavioralist Visits the Factory: Increasing Productivity Using Simple Framing Manipulations", unpublished data. Accessed 13 February 2013.
Pokorny (2008)	Pokorny, Kathrin n.d.: "Data for: Pay—but do not pay too much: An experimental study on the impact of incentives", unpublished data. Accessed 28 May 2014.