Do women respond less to performance pay? Building evidence from multiple experiments

Oriana Bandiera, Greg Fischer, Andrea Prat and Erina Ytsma*

October 2017

Abstract

Performance pay increases productivity but also earnings inequality. Can it widen the gender gap because women are less responsive? We provide answers by aggregating evidence from existing experiments on performance pay that have both male and female subjects, regardless of whether they test for gender differences. We develop a Bayesian hierarchical model (BHM) that allows us to estimate both the average effect and the heterogeneity across studies. We find that the gender response difference is close to zero and heterogeneity across studies is small. We also find that the average effect of performance pay is positive, increasing output by 0.28 standard deviations. The data are thus strongly supportive of agency theory for men and women alike.

JEL: J16, J31, C11

Keywords: wage differentials, gender, econometrics, meta-analysis

*Bandiera: Department of Economics and STICERD, LSE and CEPR (o.bandiera@lse.ac.uk); Fischer: Department of Economics and STICERD, LSE and CEPR (g.fischer@lse.ac.uk); Prat: Columbia University and CEPR (andrea.prat@columbia.edu) ; Ytsma: MIT Sloan School of Management and IDE (eytsma@mit.edu). We thank Gharad Bryan, Ray Fisman, Andrew Gelman, Gerard Padró i Miquel, Jörn-Steffen Pischke, and Bernard Salanié. We are grateful for helpful comments from seminar participants at UCL, Bocconi, Yale, Stanford, University of Washington, Columbia, the University of Manchester, DFID and DIW.

1 Introduction

After almost a century of steady growth in female labor force participation the earnings gap between men and women remains large, especially for top earners (Manning and Swaffield, 2008; Bertrand et al., 2010). At the same time, performance pay, a cornerstone of good management practices (Bloom and Van Reenen, 2010; Bloom et al., 2012), has spread widely, fostering inequality: recent estimates suggest that it accounts for most of the recent top-end growth in wage dispersion in the US (Lemieux et al., 2009).

To the extent that women are less responsive to performance pay, its increase in popularity might have contributed to increasing the earnings gap. Women may respond less to incentive pay for a number of cultural and psychological reasons. For instance, laboratory evidence indicates that women are more risk averse (Croson and Gneezy, 2009); thus, faced with performance pay, they might take actions that reduce the variance but also the mean of performance. This effect can be reinforced by differences in confidence (Bertrand et al., 2010) or the ability to work under pressure (Azmat et al., Forthcoming).

This paper tests whether women are less responsive to performance pay using a large, hitherto unexplored collection of laboratory and field experiments that identify the response to performance pay, regardless of whether the studies themselves tested for gender differences. The goal of this paper is to aggregate this evidence and assess whether a clear gender pattern emerges.

We develop a Bayesian hierarchical model to estimate both the average gender differences as well as their heterogeneity across studies. This approach has two advantages. First, it leverages existing data to provide evidence on a new question while avoiding the pitfalls of ex post subgroup analysis. Second, the model uses the data itself to estimate the degree to which each study is informative about a common phenomenon versus its own context-specific effect; thus it allows us to quantify the extent to which the findings of one study are informative for another.

Agency theory makes precise that performance pay affects an individual's effort on the job, his or her expected earnings and, through this, selection into jobs. Thus if women respond less to performance pay, they may also sort into jobs that do not offer performance

pay.[1] Here we focus on the effort effect both because this drives the selection effect and because experiments on selection are rare. Importantly, agents in our sample are not selected on their responsiveness to performance pay because experiments are done in settings that offered no performance pay when agents signed up for the job. That is, this is the population where we should be most likely to find gender differences, if these exist.

To proceed we first identify a set of studies on performance pay and collate the data. To maximize the number of studies while ensuring quality and replicability of our aggregation process, we include only field and lab experiments published in peer-reviewed economics journals or a selected set of discussion paper series. To capture the parameters of interest, namely responses to performance pay in the workplace by men and women, we further require that (i) agents exert real and costly effort; (ii) performance is measured at the individual level; and (iii) the study includes at least two pay treatments, one of which is unambiguously more high-powered than the other (meaning that the marginal effect on pay of an increase in performance is always larger). We identified 37 studies satisfying the inclusion criteria and either obtained the published data or contacted the authors. After eliminating studies without gender variation or available data and those for which the authors did not reply, our final sample comprises data from 18 studies involving over 9,000 subjects across a wide range of contexts.

The Bayesian hierarchical model (BHM) posits that the observed estimate ($\hat{\eta}_s$) in a given study s is distributed normally conditional on certain parameters, most importantly η_s, the true average treatment effect in study s. These parameters are in turn distributed conditional on hyperparameters η and τ_η^2, which determine the mean and variance of study-level, average treatment effects in the population of potential studies. The BHM allows us to estimate both the average response by men and women as well as the heterogeneity of these responses across contexts.

Since different studies measure performance in different units, for comparability we rescale all outcomes in terms of each study's standard deviation of unincentivized performance, σ. Our main finding is that the estimated distribution of the gender-incentive

[1] For instance, Card et al. (2016) show that selection into firms that pay lower wage premia explains 15% of the gender earnings gap in Portugal.

coefficient (η) has a mean that is close to zero ($+0.07\sigma$)—implying women are slightly *more* responsive to financial incentives—with relatively little variance (0.13σ) across studies. That is, women and men respond similarly to different variants of performance pay across a wide range of contexts. This implies that, if we were to run a new experiment, we would expect men and women to respond to steeper incentives in a similar manner, and we would be quite confident in this expectation.

The model also allows us to estimate the common response to performance pay. Agency theory predicts this to be positive but psychological responses, such as intrinsic motivation crowding-out, might generate negative responses. The evidence favors agency theory as we find that the mean response to performance pay is positive and large ($+0.28\sigma$). Not surprisingly, given the diversity of contexts and of treatments, the estimated heterogeneity is also quite large. The heterogeneity, however, affects primarily the magnitude rather than the sign of the effect. The probability that the true effect of incentives is negative in a population split evenly between men and women is approximately 6%. In most of these cases, the effect would be indistinguishable from zero. Replicating the existing set of studies, we would expect to obtain a negative significant effect of incentives, significant at the 5%-level, in fewer than 1 out of 100 cases.

Finally, the model highlights the key insight, made previously by Efron and Morris (1977) and Rubin (1981), that the best estimate for the true effect in any particular context may not be the mean estimate of an internally valid study *in that very same context*. An internally-valid study provides an unbiased estimate of the treatment effect in that study. However, this object differs from the expected treatment effect if the study were repeated in the same context or in a different location. To the extent the studies are estimating a common effect, results from the other $n-1$ studies should be combined with our estimates from a particular study to inform our beliefs about the true effect *in that study*. The BHM makes this updating process transparent.

Our contribution to the literature is twofold. The first is substantive. We contribute to the literature on gender earnings gaps (Altonji and Blank, 1999; Olivetti and Petrongolo, 2016; Azmat and Ferrer, forthcoming; Bertrand et al., 2010; Goldin, 2014) by ruling out one possible cause of the gap: women do not respond differently to performance pay. If there

are indeed differences in risk aversion or other behavioral parameters, these are not strong enough to generate systematic differences in the response to incentives. Of course there are differences in the response of men and women in specific contexts, but such differences are not rationalizable as a manifestation of a consistent gender pattern.

The second contribution is methodological: we demonstrate how BHMs can harness the recent explosion in the number of field and laboratory experiments to answer new questions with existing data. While BHMs have been used for decades to aggregate information from multiple studies in other disciplines (see Rubin, 1981 for an early example and Gelman et al., 2004 for the textbook exposition), they have only recently begun to gain popularity in economics. Hsiang, Burke and Miguel (2013; 2015) analyze the link between climate change and conflict; Vivalt (2015) examines generalizability across a wide range of impact evaluations; and Meager (2015) looks carefully at the impact of microcredit.

The Bayesian approach allows to distill a common lesson from studies in diverse contexts; this is essential in economics where studies typically differ in terms of participants, interventions and outcomes and where this has fueled debate about what we can learn from experiments (Deaton, 2010; Banerjee and Duflo, 2009; Rodrik, 2010; Pritchett and Sandefur, 2015; Allcott, 2015).[2] We show that, in addition to aggregate answers to a given question, these methods can be used to ask new questions, in particular, to explore dimensions of heterogeneity that individual studies cannot either because they lack statistical power or because it was not part of their original stated goals.[3]

The rest of this paper is organized as follows. Section 2 reviews gender differences in personality traits that determine the response to incentives. Section 3 discusses study selection, Section 4 presents the methodology, and Section 5 the results. Section 6 concludes.

[2] In contrast, within the medical literature the Cochrane Handbook (2011) states that aggregating evidence "should only be considered when a group of studies is sufficiently homogeneous in terms of participants, interventions and outcomes." The Cochrane Handbook forms the basis for conducting and reporting systematic reviews as set forth by Cochrane, a group of 37,000 medical research professionals formed to systematically organize and evaluate medical randomized trials.

[3] Specifying ex ante the subgroup or subgroups along which heterogeneity will be analyzed and defining clear inclusion criteria for studies alleviates the cherry-picking concerns that normally plague ex post subgroup analysis. See Casey et al. (2012) and Olken (2015). We see this as a natural complement to the work of Athey and Imbens (2015) and Dwork et al. (2015), which addresses this issue at the study-level.

2 Foundations of the gender performance pay gap

Underpinning our research question is an extensive experimental literature that documents systematic differences in psychological traits between men and women. Bertrand (2011) and Azmat and Petrongolo (2014) review this literature in detail, highlighting the lack of evidence on the impact of these differences on labor market outcomes. Importantly for this paper, moral hazard theory would predict that three of the four traits that are found to broadly differ by gender—risk aversion, overconfidence and altruism—directly affect the expected utility of effort and thus should affect the response to performance pay. The fourth trait, attitudes towards competition, also affects the response to incentives but only if incentives have tournament structure. This will later motivate us to look for differential effects in tournament and non-tournament settings.

The consensus on risk attitudes is that women appear to be more risk averse than men. This result has emerged in many laboratory experiments (see reviews by Croson and Gneezy, 2009; Eckel and Grossman, 2008; and Charness and Gneezy, 2012) and in rarer surveys of the general population such as Dohmen et al. (2011). Risk attitudes play an important role in determining how individuals respond to pay for performance because, in the presence of production shocks, linking pay to performance transfers some of the production risk from the employer to the employee. Risk-averse employees will be more willing to take actions that reduce both the mean and the variance of performance, and hence of pay. For instance, in most jobs that offer performance pay, performance is measured by quality-adjusted quantity per unit of time. Working faster increases the expected quantity but also the probability of mistakes, thus increasing the variance of pay. Moreover, even if working harder does not increase the probability of mistakes, higher performance pay magnifies whatever variance was already present. How employees value this trade-off depends on their risk attitudes. In particular, risk averse individuals will suffer a bigger utility loss from the increase in variance and hence will be less likely to respond by increasing effort.[4]

[4]This is true in a model where, for instance, $y = \varepsilon x$, where: y is observed performance, x is the agent's effort, and ε is a productivity shock that is unknown to the agent at the time he or she selects effort (ε is a random variable that can only take non-negative values). Then, an agent with constant absolute risk aversion parameter γ will maximize $bE[\varepsilon]x - \gamma b^2 Var[\varepsilon]x^2 - c(x)$, where b is the power of the contract and $c(.)$ is the cost of effort. It is easy to see that an increase in γ will cause a decrease in the equilibrium value of effort x.

To the extent that women are more risk averse than men, they will respond less to increases in incentive power.

Overconfidence is the second trait over which men and women seem to differ, with women being less confident than men (Croson and Gneezy, 2009; Reuben et al., 2012). Overconfidence affects the response to incentives because agents rationally choose to try different strategies to improve their performance only if the expected return exceeds the cost, and the expected return depends on their own assessment of their probability of success. Underconfident individuals will underestimate this probability and hence will be less likely to improve their performance in response to performance pay.

Third, a large body of experimental research, reviewed in Croson and Gneezy (2009) and Eckel and Grossman (2008), tests whether women are more altruistic than men. Overall, evidence from dictator game experiments (for example, Bolton and Katok, 1995; Eckel and Grossman, 1998, and Andreoni and Vesterlund, 2001) suggests that women give away more than men, which can be interpreted as women being more altruistic, that is putting a larger weight on the utility of the other, typically anonymous, subject.

Differences in altruism affect the response to incentives because they determine performance when no incentives are offered. In particular, if women are more likely to internalize the effect of their effort on the welfare of the employer then their effort levels will be higher under fixed wages and the response to performance pay weaker.

Last but not least, a number of papers have shown that women are less responsive to competition (Croson and Gneezy, 2009). These differences could stem from the effect of competition per se. They may also reflect aforementioned differences in overconfidence and risk aversion because men and women seem to respond equally to competition when individuals know their relative ability (Bertrand, 2011).

Taken together, the experimental differences in risk aversion, overconfidence, altruism and competitive attitudes all pull in the same direction, making women potentially less responsive to performance pay. In this paper we test whether this is indeed the case in a variety of jobs and contexts.

3 Methodology

3.1 Study selection

The first step in building evidence from multiple studies is to establish the inclusion criteria for study selection. We focus on studies that evaluate the effect of performance pay on output either in the workplace or a comparable lab environment. As discussed above, different effects by gender might instigate a response also on the selection margin. If women expect to respond less and hence earn less under performance pay, they might avoid jobs with performance pay. Here we focus on the response to incentives both because it is the cause of this selection response and because selection experiments are rare.

We begin by tackling the issue of quality control. In this respect, our guiding principle is to minimize the use of subjective judgments. To this purpose, we use the established quality screening of the publication process and restrict our sample to articles published in refereed journals or the working paper series of one of the main research associations (CEPR, IZA, NBER). We thus refrain from the practice, common in other fields, to rank research output by quality and weight it accordingly.

For the same reason, we refrain from judging which studies successfully identify causal effects by focusing exclusively on papers that use experimental variation, either in the lab or in the field, in the exposure to performance pay. Our choice does not imply that credible results can only be obtained by randomizing incentives treatments. Rather we choose papers that use a common source of variation, orthogonal to the potential outcomes of interest, to avoid subjective judgments on the credibility of alternative identification strategies. Finally, as experimental analyses of incentives have started only recently, we restrict our search to papers published between 1990 and 2012, when this study began.

The second set of criteria aims to select studies that can be informative of gender differences in the response to incentives in the workplace. For this reason, we only include studies where there are at least two treatments that differ in the power of monetary incentive pay and can be ranked according to their power. This criterion excludes studies that compare different forms of incentives, e.g., piece rates vs. tournaments, that cannot be unambiguously ranked. It also excludes studies that use non-monetary performance rewards,

such as recognition, as these cannot be easily compared.

To ensure that the setting is informative of workplace behavior, we restrict our sample to studies where subjects choose effort that is (i) real and (ii) produces output. Criterion (i) excludes all experiments that use hypothetical effort—for instance, those that require subjects to choose from a list of numbers that represent effort. Criterion (ii) excludes all experiments that aim to affect behavior outside the workplace, for instance those that pay people to stop smoking, to go to the gym, etc.

Finally, we only include studies with no externalities in production, that is, we exclude all forms of team production and team incentives. The rationale for this latter criterion is that team work might generate different responses due to gender differences in competitiveness or cooperation, hence bringing in a radically different mechanism. Table 1 summarizes our selection criteria.

We search EconLit, Google Scholar, and the working paper series of CEPR, IZA and NBER for the following combinations of keywords "incentive, productivity, experiment", "incentive, effort, experiment", "performance, pay, experiment" as well as "incentives", "performance", "pay", "effort", and "productivity". The search yields 169 papers, of which 37 passed the inclusion criteria. Of these, 16 either had their data available online or the authors shared the data with us. Among the rest, 14 were not usable either because the authors no longer had the data or because they did not record gender, and 7 sent us regression results but not the underlying data.[5] Of the 16 papers, two report two experiments and these are included separately as they meet the inclusion criteria individually.

Table 2 summarizes the 18 experiments we use for the analysis. For each study, we focus on the specification with the cleanest test of workplace financial incentives meeting our selection criteria. In most cases, this is the paper's primary specification; however, in some cases this may be preliminary rounds of an experiment where the comparison between high- and low-powered incentives is most direct. The table describes the included specifications for each experiment. In all but 4, approximately half of the subjects are women. In aggregate, the studies report on the behavior of 9,968 unique subjects, of which 50.1%

[5]Because our aggregation method requires the full variance-covariance matrix of any estimation and normalized outcome measures, we do not include these studies in our main results.

were women. The pool of studies is equally split between field and lab experiments. As the table makes clear, there is a fair amount of heterogeneity in the size of the subject pool, the exact type of performance pay used, and the context. Real outcomes in the included studies range from teacher attendance in Kenya to mazes solved in an experimental lab in Israel and condoms sold in Zambia. This diversity across studies is essential if we are to identify a truly universal pattern in the response to workplace financial incentives. However, it also brings to the fore the main challenge that our methodology will need to address: how to distill a common lesson from diverse experimental contexts. We turn to this next.

3.2 Descriptive model of performance

In order to estimate the relative effect of financial incentives on the productivity of women versus men, we begin with a descriptive model of the performance of individual i on a productive task in experiment or study $s \in \{1,\ldots,S\}$:

$$y_{is} = \alpha_s + \beta_s G_{is} + \gamma_s T_{is} + \eta_s G_{is} \times T_{is} + \varepsilon_{is}, \tag{1}$$

where G_{is} is an indicator that is equal to 1 if subject i in experiment s is a woman and T_{is} is an indicator that is equal to 1 if subject i in experiment s receives the higher powered treatment. For instance, if one group is paid fixed wages and the other piece rates, we set $T_{is} = 1$ for the piece rate group. Note that equation (1) is the non-parametric cell-means regression with respect to gender and incentives. That is, α_s equals the average productivity of unincentivized men in experiment s; $\alpha_s + \beta_s$ equals the average productivity of unincentivized women; etc. Our primary parameter of interest is η_s, which captures the differential effect of incentives on women relative to men in study s. Under the null that women and men respond equivalently to incentives, η_s equals zero.

Our aim is to understand generalizable differences in the response to incentives, and doing so entails aggregating across disparate studies. As such it is necessary to normalize the outcome measure across tasks. We do so by converting the outcome variable in each study to its standardized value, $\tilde{y}_{is} = (y_{is} - \bar{y}_s)/\hat{\sigma}_s$, where $\bar{y}_s$ is the sample mean for men in the control group and $\hat{\sigma}_s$ is the sample standard deviation, again for men in the control

group.

For each study we then estimate the vector of parameters, $\theta_s = (\tilde{\beta}_s, \tilde{\gamma}_s, \tilde{\eta}_s)'$, following specification in equation (1) on the transformed data:

$$\tilde{y}_{is} = \tilde{\alpha}_s + \tilde{\beta}_s G_{is} + \tilde{\gamma}_s T_{is} + \tilde{\eta}_s G_{is} \times T_{is} + f(X_{is}) + \tilde{\varepsilon}_{is}, \tag{2}$$

where $f(X_{is})$ are study-specific controls. We aim to replicate each study's preferred specification, only adding the gender-incentive interaction term where necessary.[6] The vector of parameter estimates, $\hat{\theta}_s$, and the associated covariance matrix, $\hat{\Sigma}_s$, for each study form the inputs in the Bayesian hierarchical model as described below.

As a robustness check, we also consider an alternative formulation in which we estimate a common specification for each study, excluding covariates. Where multiple observations are available for each individual, we collapse the data to subject-treatment-level means, and estimate equation 1. In both cases, we implement the estimation using the error structure assumed in the original paper, e.g., clustering at the session level.

3.3 The Bayesian Hierarchical Model

To motivate the Bayesian hierarchical model that we estimate, it is useful to consider two alternative approaches to aggregating empirical evidence. 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:

$$\hat{\eta}_S \sim N[\eta, \sigma_s^2] \quad s = 1, \ldots, S. \tag{3}$$

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-

[6] For the two studies that employed panel data (Bandiera et al, 2005 and Fehr & Goette, 2007), we collapse the data to the individual level in order to estimate the main gender effect and associated elements of the covariance matrices. This has little effect on the estimates of incentive and incentive-gender effects, and all results are robust to excluding these studies.

weighted average of the individual study effects,

$$\hat{\eta}^{Pool} = \sum w_s^{Pool} \eta_s / \sum w_s^{Pool}, \tag{4}$$

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.

In contrast, the random-effects model assumes that each observed study result, $\hat{\eta}_S$, is estimating its own study-specific effect, η_S. These study-specific η_S's are in turn distributed around a common population mean, η. The random-effects model thus explicitly allows for variation in the true treatment effect across studies as well as within-study idiosyncratic variation. The model has the following form:

$$\begin{aligned} \hat{\eta}_s &\sim N[\eta_S, \sigma_s^2] \quad s = 1, \ldots, S \\ \eta_s &\sim N[\eta, \tau^2]. \end{aligned} \tag{5}$$

We can decompose the difference between study-level parameter estimates, $\hat{\eta}_s$, and the population mean, η, into two components. First, there is statistical variation in the local parameter estimate, $\hat{\eta}_s - \eta_s$, that is, the difference between the estimated and true effects in that particular context. This would include both the idiosyncratic variation in our estimates as well as any potential bias. Second, there are context-specific factors, $\eta_s - \eta$, that would include variation across studies in the types of incentives, sample populations, study characteristics, and implementation.

The estimate of η is again the weighted average of the individual study effects as in (4); however, the weights are now given by $w_i^{RE} = (\hat{\sigma}_i + \hat{\tau}^2)^{-1}$, where $\hat{\tau}^2$ is an estimate of the between-study variance:

$$\hat{\eta}^{RE} = \sum w_s^{RE} \eta_s / \sum w_s^{RE}. \tag{6}$$

The random-effects model reduces the relative weight on more precise studies and reduces the effective precision for all studies. It thus generates more conservative estimates for the variance of the estimate of the common effect, η. However, while classical random effects explicitly accounts for potential heterogeneity across studies, it treats τ^2 as a constant once

estimated rather than a random variable. In ignoring this level of uncertainty, akin to the generated regressors problem (Pagan, 1984), it underestimates cross-study heterogeneity (Rubin, 1981).

The Bayesian hierarchical model mirrors the classical approach in (5), but treats the hyperparameters, η and τ^2 , themselves as random variables :[7]

$$\begin{aligned} \hat{\eta}_s &\sim N[\eta_S, \sigma_s^2] \quad s = 1, \ldots, S \\ \eta_s &\sim N[\eta, \tau^2] \\ \eta &\sim [-,-] \\ \tau^2 &\sim [-,-], \end{aligned} \tag{7}$$

where $[-,-]$ indicates a prior distribution that must be specified.

The first line of (7) is uncontroversial and proceeds immediately from the assumption of each study's internal validity and the fact that sample sizes are sufficiently large that the central limit theorem justifies using the normal distribution. The assumption of known variances is also reasonable. Given the sample sizes of the included studies, the study-level parameter variances are precisely estimated and modeling the uncertainty would add little to the analysis.

The second line embodies a critical assumption: the parameters $(\eta_1, \ldots, \eta_s)$, the study-level effects, are themselves normally distributed in the population with mean η and variance τ^2. It is perhaps unreasonable to assume that studies were placed at random, with contexts chosen from a large population with approximately normally distributed effects.

Previously, the normality assumption for study-level effects was required for analytical tractability (Stein et al., 1956; James and Stein, 1961; Efron and Morris, 1971); however, modern estimation techniques have freed us from that necessity. We begin by estimating

[7]Since at least Lindley (1971) and Lindley and Smith (1972), this structure, sometimes called a one-way normal random-effects model with known data variance, has been commonly employed in hierarchical modeling, at least in part owing to its analytical tractability. This means we have no prior knowledge with which to distinguish η_j from $\eta_k, \forall j \neq k$. For example, before seeing the data, there is no reason to believe that the effect size in study j should be larger than the effect size in study k. The assumption that the study-specific effects, η_S, are i.i.d. is implied by a stronger underlying assumption that the S values of y_i are exchangeable, i.e., the joint probability density of the data, $p(y_1, \ldots, y_S)$ is invariant to permutations of the indices. If in addition to the data we have access to additional information that would allow us to distinguish the studies *a priori*, then our model can be expanded to include this information so the model is exchangeable in the data and the covariates.

the normal-normal model as a convenient starting point. We then test the appropriateness of this assumption. As described in Section 4.4, the data conform quite well to the normality assumption. Although studies were not selected at random, the net effect of differences in location, incentives, subject pools and other study characteristics can be characterized quite well by the hierarchical model we estimate and the normal distribution. That is, the study-level results are distributed as if the studies were "placed at random", with the true effects in each context distributed approximately normally around a population mean. Nevertheless, we check the robustness of our results to alternative distribution assumptions; Section 4.4 describes the results. Despite the accuracy of the model we estimate, it remains important to note that our results are best interpreted as the distribution of incentive effects from the population of contexts in which economists have been willing to run experiments. The extent to which these settings are representative of the broader population points to further questions regarding the placement of experiments and the representativeness of empirical work more generally (see, for example, Cartwright and Deaton, 2016 and Allcott, 2015).

The key assumption required for us to estimate the joint probability model is exchangeability. Technically, this means that the joint distribution of $(\eta_1, ..., \eta_S)$ is invariant to permutations of the indexes $(1, \ldots, S)$. It allows us to write the joint distribution of the η_s's as i.i.d. given hyperparameters η and τ that also have prior distributions. Intuitively, it means there is no information, other than the data, y, to distinguish one study from other. For example, before seeing the results from the studies, there is no reason to believe that the results from, say, experiment 1 should be larger than those in experiment 2 or more similar to those in 3 than in 4.

In practice, this assumption is less restrictive than it appears and can easily be relaxed with partial or conditional exchangeability. If there is information available that distinguishes one study from another, it can simply be included in the data. For example, if there were study-level characteristics that we thought were informative about the parameters of interest, we could group data together with an additional level of hierarchy (e.g., grouping field and lab experiments separately) or add additional variables to the analysis (e.g., estimating a three-parameter model that allows for correlation across the parameters or including study date or location as covariates).

Finally, lines 3 and 4 of equation (7) indicate some prior distribution for the hyperparameters. For both, we will focus on non-informative (reference) priors, motivated by the notion that the information we have about the response to incentives is contained in the data themselves. As with the normality assumption, we can easily test the robustness of our choice of priors, which we discuss further in Section 4.4.

The univariate model provides a clear setting to demonstrate the intuitive appeal of the Bayesian hierarchical model. The pooling model described in (3) above is a natural reference point for thinking about analyzing data across studies. It assumes that each individual study is estimating a common effect, η. Observed differences in parameter estimates arise only due to idiosyncratic experimental variation.

This model performs well when parameter estimates across studies are similar. Figure 1a illustrates the posterior distribution of η that would be generated by two similar signals, e.g., parameter estimates from two different studies, and an uninformative prior: $\hat{\eta}_1 = -0.50(1.00)$; $\hat{\eta}_2 = 0.50(1.00)$. The mean of this posterior is the precision-weighted mean of the signals, and the posterior precision is the sum of the precision in the signals: $\hat{\eta}^{post}_{pool} = 0.00(0.71)$. While this posterior and, in particular, its increasing precision appear plausible when studies are relatively homogeneous, the implications of the pooling model are less reasonable in the presence of heterogeneity.

Figure 1b again illustrates the posterior distribution of η that would be generated by two signals with the same precision as before, but now the means of the two signals differ substantially: $\hat{\eta}_1 = -2.00(1.00)$; $\hat{\eta}_2 = 2.00(1.00)$. In this case, the pooling model generates the identical posterior distribution. This is unrealistic. Note that the posterior for the population hyperparameter has little overlap with the sampling distributions of either study estimate. Our natural intuition would be to conclude that the two studies in Figure 1b are measuring different effects.

The Bayesian hierarchical model not only captures this intuition but quantifies the cross-study heterogeneity. Figures 1c and 1d display the BHM's posterior estimates for the population parameter, η, based on the same signals as shown in 1a and 1b, respectively. When the signals are similar, as in 1c, the posterior distribution η is similar to that generated by the pooling model. In fact, in this example, they are identical: $\hat{\eta}^{post}_{BHM} = 0.00(0.71)$. The

BHM explicitly allows for the possibility of heterogeneity in the true effect across studies, but when the estimated cross-study variation is zero, as it is here, the BHM recovers the pooling model. However, when cross-study heterogeneity is large, as in Figure 1d, the BHM rejects the pooling model, and the posterior distribution on the population quantifies the variation we observe across studies: $\hat{\eta}_{BHM}^{post} = 0.00(2.00)$.

A simple but unappealing alternative to using the BHM in the presence of cross-study heterogeneity would be to disregard any commonality and treat each study as an estimate of its own unique, context-specific parameter. This would be equivalent to fixing τ^2 in (7) at ∞. In essence, there is no population parameter. While this avoids the troubling implications of the pooling model, it comes at the cost of precluding any attempt to discern universal patterns in fundamental economic questions such as the response to incentives.

By allowing for intermediate degrees of pooling, the BHM makes it possible to aggregate evidence across similar but not necessarily identical studies. Rather than assuming the extent to which one study is informative about another, one uses the data itself to quantify both the degree to which studies are estimating a common parameter and uncertainty about this variation. One could also incorporate beliefs about the degree of heterogeneity across studies—for example, confidence but not certainty that a set of studies is estimating a common effect—in the form of the prior distribution of τ. Aggregating evidence with the BHM also provides a foundation for the inductive exploration into the sources of heterogeneity, that is, to what extent can characteristics at the level of the study or the individual participants explain the variation in parameter estimates.

The core of our analysis focuses on the Bayesian hierarchical model for the full parameter vector, $\theta = (\beta, \gamma, \eta)$, which will allow us to explore heterogeneity across studies along the dimension of potentially correlated parameters. For example, the gender-incentive interaction may be increasing in the size of the main incentive effect or in the relative performance of women vs. men. The structure again parallels those above:

$$\begin{aligned} \hat{\theta}_s &\sim N[\theta_S, \Sigma_s] \quad s = 1, \ldots, S \\ \theta_s &\sim N[\theta, \Sigma], \end{aligned} \tag{8}$$

where

$$\Sigma = \begin{bmatrix} \tau_\beta^2 & \tau_{\beta\gamma} & \tau_{\beta\eta} \\ \tau_{\beta\gamma} & \tau_\gamma^2 & \tau_{\gamma\eta} \\ \tau_{\beta\eta} & \tau_{\gamma\eta} & \tau_\eta^2 \end{bmatrix}.$$

We use the following priors for the hyperparameters:

$$\begin{aligned} \theta &\sim N[0,\ 100^2] \\ \Sigma &\sim diag(\sigma)\ \Omega\ diag(\sigma) \\ \sigma_k &\sim Cauchy(0, 2.5),\ \text{for } k \in \{\beta, \gamma, \eta\} \text{ and } \sigma_k > 0 \\ \Omega &\sim LKJcorr(2) \end{aligned} \tag{9}$$

where Ω is a correlation matrix and σ is the vector of coefficient scales (Gelman, 2006).[8] The vector of parameter estimates, $\hat{\theta}_s = (\hat{\theta}_s, \hat{\gamma}_s, \hat{\eta}_s)$, from each study and the associated estimated covariance matrix, $\hat{\Sigma}_s$, are the key inputs in the hierarchical model on which we focus.

Our estimation of the Bayesian hierarchical models follows closely the procedures described in Gelman and Hill (2007) and Gelman et al. (2004). Appendix A describes the estimation procedure in more detail. The key outputs from this estimation are the simulated posterior distributions for both the hyperparameters, θ and Σ, as well as the true study-level effects, $\{\theta_i\}_{i=1}^S$. In addition to calculating standard measures such as means and posterior intervals (the Bayesian analog to classical confidence intervals), we can also use these simulated distributions to test any other functions of the parameters that may be of interest. The bulk of our analysis focuses on the simulated marginal posterior distributions of these parameters. We define y^{sim} as the simulated parameters that could have been observed if the studies in our sample were replicated and the parameter estimates were distributed according to our specified probability model.

Note that the simulated posterior is a joint distribution over not only the population

[8]The LKJ distribution (Lewandowski et al., 2009) is a distribution over correlation matrices, i.e., positive semi-definite matrices with unit diagonals. When the shape parameter is equal to one, the density is uniform over all correlation matrices of a given order. When the shape parameter is greater than one, the modal correlation matrix is the identity, with the distribution becoming more concentrated about the identity matrix as the shape parameter increases (Gelman et al., 2004).

hyperparameters, that is, the average effect of monetary incentives and its dispersion, but also each study-level effect. That is, our beliefs about the effect of incentives in any given setting are based not only on the results obtained in that setting but on the results in the other $n-1$ similar settings.

This insight—the seeming paradox that in the presence of other information the best (i.e., lowest mean squared error) estimate of the true effect in any particular context may not be simply mean estimate of an internally valid study *in that very same context*—is first attributed to Stein (Efron and Morris, 1977). Intuitively, this is the same process most of us engage in when evaluating new empirical evidence. When presented with the estimates from a carefully executed experiment, we shrink (or pool) our beliefs about what would happen if the exact same experiment were re-run in the same context. The degree of shrinkage is based on the precision of the estimates and our beliefs about how similar the particular study is to others in the population. The Bayesian hierarchical model serves to make that process transparent and precise.

3.4 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 60,892 observations or 18? 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:

$$\hat{\eta}_s^{Post} = (1-\lambda_s)\hat{\eta}_k + \lambda_s \eta, \tag{10}$$

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.[9] 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 (10) 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 pooling metric calculated from the posterior means of the error terms, and $\lambda_s^2 = \frac{\hat{\eta}_k^{POST}-\hat{\eta}_k}{\eta-\hat{\eta}_k}$, which 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.[10]

Gelman and Pardoe (2006) generalize this idea to develop a common pooling factor that

[9] 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.

[10] 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.

summarizes the extent to which estimates at each level 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

$$\lambda = 1 - \frac{V_{s=1}^{K} E(\varepsilon_s)}{E(V_{s=1}^{K} \varepsilon_s)}, \tag{11}$$

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 $\varepsilon_s = \eta_s - \eta$. 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.

4 Results

4.1 The response to incentives for men and women

Table 3 summarizes the posterior distribution of the hyperparameters (γ, η, and β, and the corresponding elements of τ). That is, given the available data and our specified (uninformative) prior beliefs, it describes the population distribution of (i) men's response to incentives, (ii) the gender difference in response to incentives and (iii) the gender difference in unincentivized productivity as well as the estimated standard deviation of each of these parameters. As described in Section 3.2, the data are standardized so the unit of measure for the parameters is the standard deviation of productivity for unincentivized men in each setting. For each parameter we report both the BHM and the pooling estimate for comparison.

The table shows that we cannot reject the null that men and women respond equally.

The median and mean of the BHM estimates for the gender-incentive interaction hyperparameter, η, are 0.069 and 0.066, with a 95%-interval of $[-0.026,\ 0.176]$. The sign of the estimate is positive, suggesting that, contrary to the implications of gender differences in risk aversion and overconfidence, women respond slightly more to incentives than men do. The similarity between men and women is not an artifact of main incentive effects close to zero. As described below, the effect of incentives is consistently large and positive. Moreover, the responses of men and women are more similar in settings where incentives have larger effects on men.

The pooling estimate of the gender-incentive interaction hyperparameter is of similar magnitude, with a mean of 0.052 (s.e.: 0.020). Recall that the difference between the two estimates is that the pooling model assumes no heterogeneity across studies while the BHM allows for heterogeneity and estimates it. Figure 5 shows that the estimated cross-study heterogeneity is relatively low (median $\tau_\eta = 0.121$). Moreover, there is significant mass in the posterior distribution at $\tau_\eta \approx 0$, which rationalizes the similarity of the BHM and pooling estimates.

Despite significant variation in context, including task, location, and the structure of pay for performance, the differences between men and women in the response to incentives appear to be relatively consistent and consistently close to zero. This implies that these studies have external validity, that is knowing that the gender differential is zero implies that the next, hypothetical study is also very likely to find a zero effect.

Despite this similarity across studies, assuming away heterogeneity for the pooling model, leads to standard errors on $\hat{\eta}$ that are approximately one-third those of the more conservative BHM. More importantly, the low estimates of cross-study heterogeneity imply that the estimated gender response difference in study n is highly predictive of the same in study $n+1$.

Having established that women and men respond similarly, we are interested in assessing whether they both respond positively, as predicted by agency theory, or negatively, as predicted by crowding-out. Because our estimate of gender differences is essentially zero, we will focus on the distribution of γ, the estimated effect of incentives on male subjects. Increasing the power of incentives leads men to increase productivity by about one-fourth

of one standard deviation. As shown in Table 3, the median and mean for the posterior estimate of γ are 0.277 and 0.275, with a 95%-interval of $[0.131,\ 0.431]$. This is consistent with the main prediction of agency theory and casts doubt on the practical relevance of crowd-out. We note that the pooling estimate is much smaller, with a mean of 0.097 (s.e.: 0.016), or roughly the 3^{rd} percentile of the BHM estimate. In line with this, the middle panel of Figure 5 shows a substantial amount of heterogeneity across studies. The median estimate of τ_γ is 0.268 and there is no mass on values less than 0.10. These results explain the difference between the BHM and pooling estimates of γ. Indeed, we can easily reject the pooling hypothesis.

That the magnitudes of the incentive effect are heterogeneous is to be expected because the different studies use different incentive schemes in different contexts; more studies with the same incentive scheme are needed to assess whether there is indeed a common response across contexts. Despite studies in different contexts estimating incentive effects of very different magnitudes, incentives unambiguously increase productivity across the sample.

A key advantage of our method is that the findings can be used to predict the response to incentives in a potential new study (γ_{S+1} and η_{S+1}). Figure 6 does so by combining the estimates of γ and η to generate a predictive distribution for men and women. As shown in the figure, if we were to run another study drawn from the same population of potential studies and knowing nothing more about the contextual details, we would expect incentives to increase performance for men by an average of 0.28σ (with an interquartile range from 0.23σ to 0.32σ) and for women by an average of 0.34σ (with an interquartile range from 0.29σ to 0.40σ). Comparing the two distributions, the median of the posterior predictive distribution for women is at the 82^{nd} percentile for men.

We expect the true, context-specific gender difference in the response to incentives to be negative and at least half as large as the estimated mean effect for men ($\eta_{S+1} < -0.14$) in about 5% of studies and less than the mean effect for men in only 7 out of 1000 studies. Alternatively, one could think about what would happen if we could simply rerun the 18 experiments included in this study, maintaining all the design features including sample size. Then, we would expect to find a negative and statistically significant (at the 5%-level) gender difference in 3.7% of the replications and a positive and statistically significant

difference in 12.6%. In other words, 84% of replications would not be able to statistically distinguish the responses of women and men. In contrast, the probability that the true effect of incentives is negative in a population split evenly between men and women is 6.3%. In most of these cases, the effect would be indistinguishable from zero. Replicating the existing set of studies, we would expect to obtain a negative and significant incentives coefficient ($\hat{\gamma}_s$) in fewer than 1 out of 100 cases.

For completeness, Table 3 also reports the estimates of β, that is the productivity difference between men and women in the absence of incentives. On average in the population of experimental settings, women are somewhat less productive. The median and mean estimates for β are -0.082 and -0.083. Not surprisingly, given the diversity of contexts covered by the sample studies, the distribution is quite spread out. The 95%-interval spans $[-0.246,\ 0.071]$, and the median for τ_β is 0.284.

Consistent with the posterior estimates for each of the τ parameters, the pooling metrics (Table 4) demonstrate significant commonality across studies for the gender-incentive interaction term (η). The common pooling factor of 0.684 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 0.394, 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 0.234 and 0.254, 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.

4.2 Cross-correlations

The fact that cross-study heterogeneity of men's response to incentives γ is large raises the possibility that gender differences in responsiveness might depend on the strength of incentives. For instance, men and women might have similar responses when incentives are weak but men might respond more when incentives are strong. To assess whether this is indeed the case, we estimate the correlation between γ and η. If any gender differences in responsiveness are large when incentives matter more, we would expect a positive correlation.

Figure 7 displays the correlations and bivariate scatter plots for each pairwise combination of the three regression parameters as drawn from the posterior predictive distribution. The estimated correlation between η and γ is -0.123, suggesting that when incentives are most effective for men (large γ) the difference between men and women is smallest. The figure also illustrates that the estimated average difference is consistently positive albeit small.

A similar test can be implemented with respect to β, the gender productivity gap. To the extent that this reflects gender differences in intrinsic motivation for the task at hand we expect incentives to be more effective for women when they are less motivated to start with. Figure 7 provides support for this hypothesis: η is large and positive when β is large and negative. Thus, when women perform worse than men with low powered incentives, an increase in incentive power closes the gap. This implies that, at least to some extent, gender differences in productivity may reflect different tastes rather than insurmountable gaps due, e.g., to differences in physical strength.

Finally, the bottom panel of Figure 7 shows the correlation between β, the gender productivity gap, and γ, men's responsiveness to incentives. There is no discernible relationship between the gender specificity of a task and the effect of financial incentives for men.

4.3 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 above, 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 2, 3, and 4 compare the posterior predicted distributions for each of the main parameters, γ, η, 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 in Bandiera et al. (2005). As shown in Figure 3, the parameter estimate in this study is large, $+0.87\sigma$, and with a t-statistic of 5.33 significantly different from zero at any conventional level. However, the estimates are substantially larger than the mean in all but one of the other studies. With a standard error of 0.16σ there remains quite a bit of uncertainty as to the magnitude of the effect even though one can say with near certainty that the effect is positive and economically significant. The mean of the posterior distribution for γ_s is $+0.70\sigma$, still a very large effect but pulled substantially towards to population mean of $+0.28\sigma$. 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 8 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 8 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 8 overlays the posterior distribution of τ_η, which has a mean estimate of 0.126. The substantial degree of observed pooling can be seen at the corresponding level of τ in the upper half of the figure.

4.4 Model Checking

After computing the posterior distribution of all parameters, it is essential to assess the fit of our model to the observed data. Using the posterior distributions, 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 will 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. 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 9 plots the observed order statistic for each of the model parameters against the mean from the simulated posterior distribution. Table 5 reports the associated p-values. In the case of the gender-incentive interaction term, the posterior predictive distribution matches the observed data very well, including at the extremes. Similarly, the symmetry test

statistics for the observed data closely match those from the simulated posterior. 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 performs reasonably well for the gender (β) and incentive (γ) parameters as well. Across all 18 order statistics, the minimum tail probabilities are 0.039 and 0.095, respectively.

In addition to directly evaluating the order statistics, we also test for asymmetry in the center of the distribution by constructing $T(y,\theta) = \left|\theta_{(15)} - \theta\right| - \left|\theta_{(2)} - \theta\right|$ for each of the parameters. The 2^{nd} and 15^{th} order statistics represent approximately the 10^{th} and 90^{th} percentiles of the distribution. For a symmetric distribution, this test quantity should be distributed around zero.[11] Figure 10 presents these results. There is some skewness in the observed data, leftwards for β and η and rightwards for γ, such that the normal distribution does not fully capture the asymmetry in the center of the distribution. In the case of the gender-incentive interaction term (η), this is only present at the particular points in the distribution we selected for the test. None of the departures from normality are significant.

4.5 Study-level heterogeneity

As noted above, it is straightforward to assess heterogeneity in the distribution of treatment effects with respect to study or context characteristics by adding another level to the hierarchical model described in equation (8). Motivated by potential differences in women's and men's attitudes towards competition, we investigate differences between their responses to tournament and non-tournament incentives. Figure 11 plots the differences.

Our sample contains only four tournaments, but the results are suggestive. In three of the studies, women are substantially less responsive to tournament-based incentives than men. Although the differences are economically meaningful, none are significant at standard thresholds. When aggregating within the hierarchical model, the incentive-gender interaction term is 0.17σ lower for tournaments than for individual incentives, with 0 falling

[11] For the observed data, each draw of the test statistic is calculated from the observed order statistics, which are fixed throughout, and draws from the posterior distribution of the relevant hyperparameter.

just inside the 95%-interval (the classical p-value is 0.12). While these results are not dispositive, they suggest that further experimentation along this dimension would be fruitful.

We also look for differences between field and lab experiments. As visible in Figure 11, we find no evidence of any systematic variation with respect to the setting.

5 Discussion

Performance pay is at the core of agency theory and management practices. Not surprisingly, given this popularity with theorists and practitioners, its effectiveness has been tested in several lab and field experiments. In this paper we propose a methodology to aggregate this evidence to test whether performance pay increases performance to the same extent for men and women. The answer provides evidence on whether differential responses to performance pay might underpin some of the gender earnings gap.

The results vindicate agency theory: across a variety of contexts and for a variety of incentive designs, we find that performance pay increases performance for men and women alike. To the extent that women differ in risk aversion, confidence, cooperation these differences are not strong enough to generate different responses. The question that remains open is why women avoid sorting into organizations with a strong performance pay component if their response does not differ. A likely explanation is that these organizations have other traits, e.g., a culture of long hours, that are not attractive to women (Bertrand et al., 2010; Goldin, 2014).

The results also illustrate the usefulness of Bayesian hierarchical models as a tool to build evidence from existing studies. BHMs are especially well suited to aggregate evidence in economics because economists run experiments in different contexts and BHMs allow to estimate the level of heterogeneity. This, in turn, is useful to assess external validity as aggregating studies by means of a BHM allows researchers to quantify the extent to which existing studies are informative about an hypothetical next study. In addition, building evidence from existing studies allows researchers to test for heterogeneity across subgroups for which which individual studies might be underpowered and to capitalize on the recent explosion in evidence from field and laboratory experiments to answer new questions with

existing data.

While we believe that this methodology is very promising, it faces several hurdles because of the way economists classify their studies and present their results. The data availability requirement recently imposed by many journals is a first step in the right direction but this should be accompanied by the requirement to report a common set of statistics such as means and standard deviations in treatment and control groups and the requirement to state the main specification upfront.[12]

A limitation that the BHM cannot overcome is that, by definition, it can only aggregate evidence from settings where experiments were run. This is particularly relevant for field experiments where partners would not normally allow researchers to test interventions they have a strong prior on. For instance, organizations are generally reluctant to test whether pay-off equivalent rewards and punishments have the same effect. A more disturbing selection comes from the researchers themselves both in terms of where to run experiments and which results to publish. As we have demonstrated above, subgroup analysis that is not the goal of the experiments is less prone to these biases.

Despite these challenges, we see BHMs as a powerful tool to build on existing knowledge and giving directions on which is the most useful experiment to run next. Most importantly it is a powerful tool to test the relevance of theory across different settings.

[12]Some of our sample studies did, others reported pooled means and standard deviations, others reported none.

References

Allcott, Hunt, "Site Selection Bias in Program Evaluation," *The Quarterly Journal of Economics*, 2015, *130* (3), 1117–1165.

Altonji, Joseph G and Rebecca M Blank, "Race and Gender in the Labor Market," *Handbook of Labor Economics*, 1999, *3*, 3143–3259.

Andreoni, James and Lise Vesterlund, "Which is the Fair Sex? Gender Differences in Altruism," *Quarterly Journal of Economics*, 2001, pp. 293–312.

Angrist, Joshua and Victor Lavy, "The Effects of High Stakes High School Achievement Awards: Evidence from a Randomized Trial," *The American Economic Review*, 2009, *99* (4), 1384–1414.

_ **, Daniel Lang, and Philip Oreopoulos**, "Incentives and Services for College Achievement: Evidence from a Randomized Trial," *American Economic Journal: Applied Economics*, 2009, *1* (1), 136–163.

Ariely, Dan, Anat Bracha, and Stephan Meier, "Doing Good or Doing Well? Image Motivation and Monetary Incentives in Behaving Prosocially," *The American Economic Review*, 2009, *99* (1), 544–555.

Ashraf, Nava, Oriana Bandiera, and B. Kelsey Jack, "No Margin, No Mission? a Field Experiment on Incentives for Public Service Delivery," *Journal of Public Economics*, 2014, *120*, 1 – 17.

Athey, Susan and Guido W Imbens, "Machine Learning Methods for Estimating Heterogeneous Causal Effects," *stat*, 2015, *1050*, 5.

Azmat, Ghazala and Barbara Petrongolo, "Gender and the Labor Market: What Have We Learned from Field and Lab Experiments?," *Labour Economics*, 2014, *30*, 32–40.

_ **and Rosa Ferrer**, "Gender Gaps in Performance: Evidence from Young Lawyers," *Journal of Political Economy*, forthcoming.

_ **, Caterina Calsamiglia, and Nagore Iriberri**, "Gender Difference in Response to Big Stakes," *Journal of the European Economic Association*, Forthcoming.

Bandiera, Oriana, Iwan Barankay, and Imran Rasul, "Social Preferences and the Response to Incentives: Evidence from Personnel Data," *The Quarterly Journal of Economics*, 2005, pp. 917–962.

Banerjee, Abhijit V and Esther Duflo, "The Experimental Approach to Development Economics," *Annu. Rev. Econ*, 2009, *1*, 151–78.

Bertrand, Marianne, "New Perspectives on Gender," *Handbook of Labor Economics*, 2011, *4*, 1543–1590.

_ **, Claudia Goldin, and Lawrence F Katz**, "Dynamics of the Gender Gap for Young Professionals in the Financial and Corporate Sectors," *American Economic Journal: Applied Economics*, 2010, *2* (3), 228–255.

Betancourt, Michael and Mark Girolami, "Hamiltonian Monte Carlo for Hierarchical Models," *Current Trends in Bayesian Methodology with Applications*, 2015, *79*.

Bloom, Nicholas and John Van Reenen, "Why Do Management Practices Differ Across Firms and Countries?," *The Journal of Economic Perspectives*, 2010, *24* (1), 203–224.

_ , Christos Genakos, Raffaella Sadun, and John Van Reenen, "Management Practices Across Firms and Countries," *The Academy of Management Perspectives*, 2012, *26* (1), 12–33.

Bolton, Gary E and Elena Katok, "An Experimental Test for Gender Differences in Beneficent Behavior," *Economics Letters*, 1995, *48* (3), 287–292.

Boly, Amadou, "On the Incentive Effects of Monitoring: Evidence from the Lab and the Field," *Experimental Economics*, 2011, *14* (2), 241–253.

Box, George E.P. and George C. Tiao, *Bayesian Inference in Statistical Analysis*, Wiley Classics, 1973.

Burke, Marshall, Solomon M. Hsiang, and Edward Miguel, "Climate and Conflict," *Annual Review of Economics*, 2015, *7*, 577–617.

Card, David, Ana Rute Cardoso, and Patrick Kline, "Bargaining, Sorting, and the Gender Wage Gap: Quantifying the Impact of Firms on the Relative Pay of Women," *The Quarterly Journal of Economics*, 2016, *131* (2), 633–686.

Carpenter, Jeffrey, Peter Hans Matthews, and John Schirm, "Tournaments and Office Politics: Evidence from a Real Effort Experiment," *The American Economic Review*, 2010, *100* (1), 504–517.

Cartwright, N and A Deaton, "Understanding and Misunderstanding Randomized Controlled Trials.," 2016, (w22595).

Casey, Katherine, Rachel Glennerster, and Edward Miguel, "Reshaping Institutions: Evidence on Aid Impacts Using a Preanalysis Plan*.," *Quarterly Journal of Economics*, 2012, *127* (4).

Charness, Gary and Uri Gneezy, "Strong Evidence for Gender Differences in Risk Taking," *Journal of Economic Behavior & Organization*, 2012, *83* (1), 50–58.

Croson, Rachel and Uri Gneezy, "Gender Differences in Preferences," *Journal of Economic Literature*, 2009, *47* (2), 448–474.

Deaton, Angus, "Instruments, Randomization, and Learning about Development," *Journal of Economic Literature*, 2010, *48* (2), 424–455.

Dickinson, David and Marie claire Villeval, "Does Monitoring Decrease Work Effort?: The Complementarity between Agency and Crowding-out Theories," *Games and Economic Behavior*, 2008, *63* (1), 56–76.

Dohmen, Thomas and Armin Falk, "Performance Pay and Multidimensional Sorting: Productivity, Preferences, and Gender," *The American Economic Review*, 2011, *101* (2), 556–590.

_ , _ , **David Huffman, Uwe Sunde, Jürgen Schupp, and Gert G Wagner**, "Individual Risk Attitudes: Measurement, Determinants, and Behavioral Consequences," *Journal of the European Economic Association*, 2011, *9* (3), 522–550.

Dwork, Cynthia, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Roth, "The Reusable Holdout: Preserving Validity in Adaptive Data Analysis," *Science*, 2015, *349* (6248), 636–638.

Eckel, Catherine C and Philip J Grossman, "Are Women Less Selfish than Men?: Evidence from Dictator Experiments," *The Economic Journal*, 1998, *108* (448), 726–735.

_ **and** _ , "Differences in the Economic Decisions of Men and Women: Experimental Evidence," *Handbook of Experimental Economics Results*, 2008, *1*, 509–519.

Efron, Bradley and Carl Morris, "Limiting the Risk of Bayes and Empirical Bayes Estimators–Part I: the Bayes Case," *Journal of the American Statistical Association*, 1971, *66* (336), 807–815.

_ **and** _ , "Stein's Paradox in Statistics," *Scientific American*, 1977, *236*, 119–127.

Engström, Per, Patrik Hesselius, and Bertil Holmlund, "Vacancy Referrals, Job Search, and the Duration of Unemployment: A Randomized Experiment," *Labour*, 2012, *26* (4), 419–435.

Fehr, Ernst and Lorenz Goette, "Do Workers Work More If Wages Are High? Evidence from a Randomized Field Experiment," *The American Economic Review*, 2007, *97* (1), 298–317.

Freeman, Richard B. and Alexander Gelber, "Prize Structure and Information in Tournaments: Experimental Evidence," *American Economic Journal: Applied Economics, 2:1*, 2010, *2* (1), 149–164.

Gelman, Andrew, "Prior Distributions for Variance Parameters in Hierarchical Models (comment on article by Browne and Draper)," *Bayesian Analysis*, 2006, *1* (3), 515–534.

_ **and Iain Pardoe**, "Bayesian Measures of Explained Variance and Pooling in Multilevel (Hierarchical) Models," *Technometrics*, 2006, *48* (2), 241–251.

_ **and Jennifer Hill**, *Data Analysis Using Regression and Multilevel/Hierarchical Models*, Cambridge University Press, 2007.

_ **, John B. Carlin, Hal S. Stern, , and Donald B. Rubin**, *Bayesian Data Analysis*, second edition ed., Vol. 2, Boca Raton, FL: Chapman & Hall/CRC Press, 2004.

Gill, David and Victoria Prowse, "A Structural Analysis of Disappointment Aversion in a Real Effort Competition," *The American Economic Review*, 2012, *102* (1), 469–503.

Glewwe, Paul, Nauman Ilias, and Michael Kremer, "Teacher Incentives," *American Economic Journal: Applied Economics*, 2010, *2* (3), 205–227.

Goldin, Claudia, "A Grand Gender Convergence: Its Last Chapter," *The American Economic Review*, 2014, *104* (4), 1091–1119.

Higgins, Julian PT and Sally Green, eds, *Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0*, The Cochrane Collaboration, 2011.

Hoffman, Matthew D and Andrew Gelman, "The No-U-turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo.," *Journal of Machine Learning Research*, 2014, *15* (1), 1593–1623.

Hossain, Tanjim and John A List, "The Behavioralist Visits the Factory: Increasing Productivity Using Simple Framing Manipulations," *Management Science*, 2012, *58* (12), 2151–2167.

Hsiang, Solomon M., Marshall Burke, and Edward Miguel, "Quantifying the Influence of Climate on Human Conflict," *Science*, 2013, *341* (6151).

James, William and Charles Stein, "Estimation with Quadratic Loss," in "Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability," Vol. 1 1961, pp. 361–379.

Lemieux, Thomas, W Bentley Macleod, and Daniel Parent, "Performance Pay and Wage Inequality," *The Quarterly Journal of Economics*, 2009, *74* (1), 1–49.

Lewandowski, Daniel, Dorota Kurowicka, and Harry Joe, "Generating Random Correlation Matrices Based on Vines and Extended Onion Method," *Journal of Multivariate Analysis*, 2009, *100* (9), 1989–2001.

Lindley, Dennis V., *Bayesian Statistics, A Review*, Society for Industrial and Applied Mathematics, 1971.

_ **and Adrian F.M. Smith**, "Bayes Estimates for the Linear Model," *Journal of the Royal Statistical Society. Series B (Methodological)*, 1972, pp. 1–41.

Manning, Alan and Joanna Swaffield, "The Gender Gap in Early-Career Wage Growth," *The Economic Journal*, 2008, *118* (530), 983–1024.

Meager, Rachael, "Understanding the Impact of Microcredit Expansions: A Bayesian Hierarchical Analysis of 7 Randomised Experiments," 2015. Working paper.

Olivetti, Claudia and Barbara Petrongolo, "The Evolution of the Gender Gap in Industrialized Countries," *Annual Review of Economics*, 2016, *8* (1).

Olken, Benjamin A, "Promises and Perils of Pre-Analysis Plans," *The Journal of Economic Perspectives*, 2015, *29* (3), 61–80.

Pagan, Adrian, "Econometric Issues in the Analysis of Regressions with Generated Regressors," *International Economic Review*, 1984, *25* (1), 221–247.

Pokorny, Kathrin, "Pay But Do Not Pay Too Much: An Experimental Study on the Impact of Incentives," *Journal of Economic Behavior & Organization*, 2008, *66* (2), 251–264.

Pritchett, Lant and Justin Sandefur, "Learning from Experiments when Context Matters," *American Economic Review*, 2015, *105* (5), 471–75.

Reuben, Ernesto, Pedro Rey-Biel, Paola Sapienza, and Luigi Zingales, "The Emergence of Male Leadership in Competitive Environments," *Journal of Economic Behavior & Organization*, 2012, *83* (1), 111–117.

Rodrik, Dani, "Diagnostics Before Prescription," *The Journal of Economic Perspectives*, 2010, *24* (3), 33–44.

Rubin, Donald B, "Estimation in Parallel Randomized Experiments," *Journal of educational and behavioral statistics*, 1981, *6* (4), 377–401.

Stein, Charles et al., "Inadmissibility of the usual estimator for the mean of a multivariate normal distribution," in "Proceedings of the Third Berkeley symposium on mathematical statistics and probability," Vol. 1 1956, pp. 197–206.

Vivalt, Eva, "Heterogeneous Treatment Effects in Impact Evaluation," *American Economic Review*, 2015, *105* (5), 467–70.

Figure 1: Examples of Aggregation Models

Figure 1a: Pooling Model
Low variation across studies

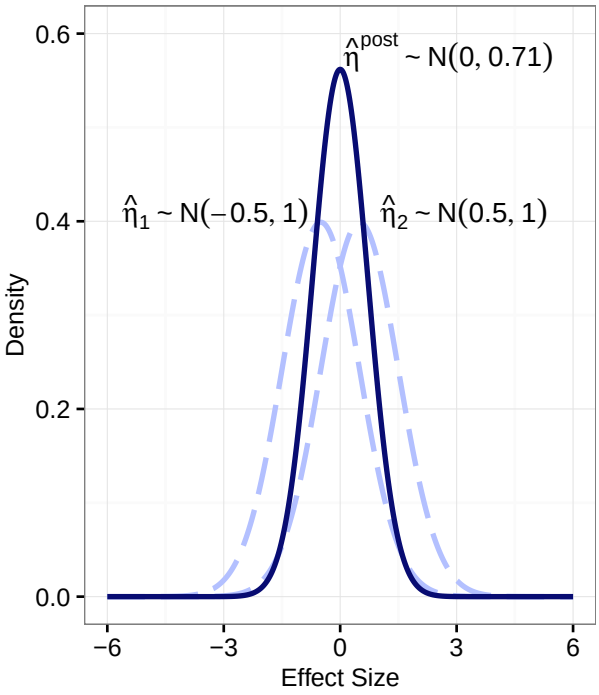


Figure 1b: Pooling Model
High variation across studies

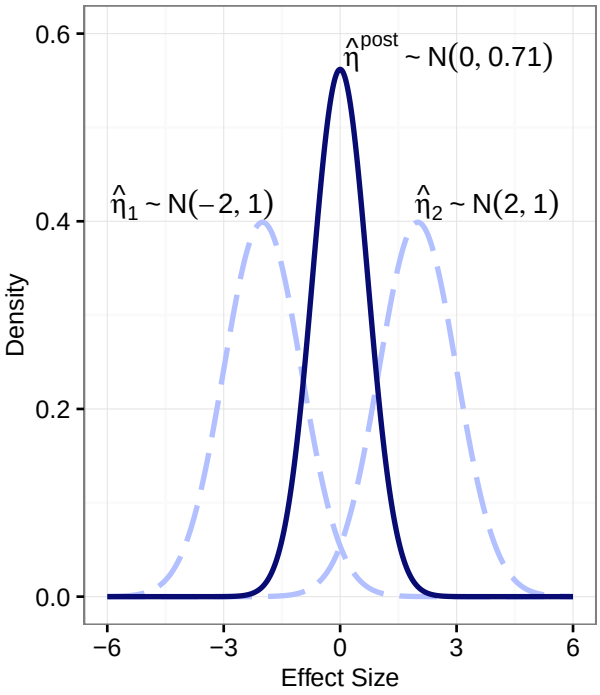


Figure 1c: Bayesian Hierarchical Model
Low variation across studies

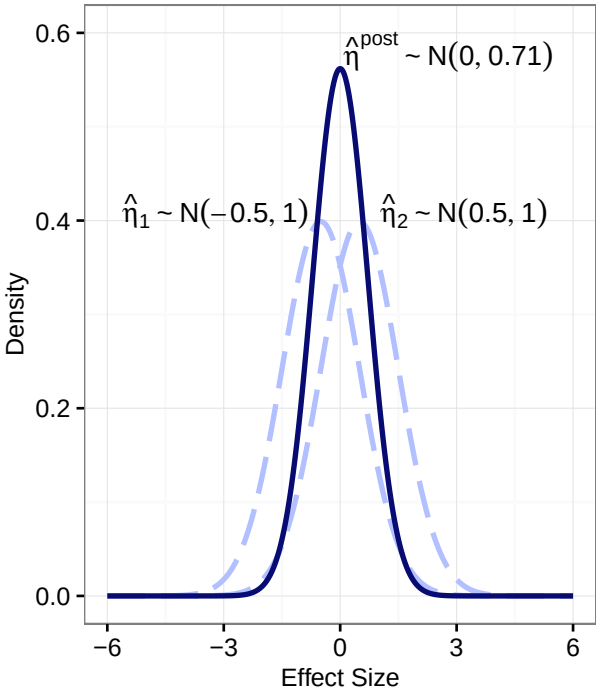
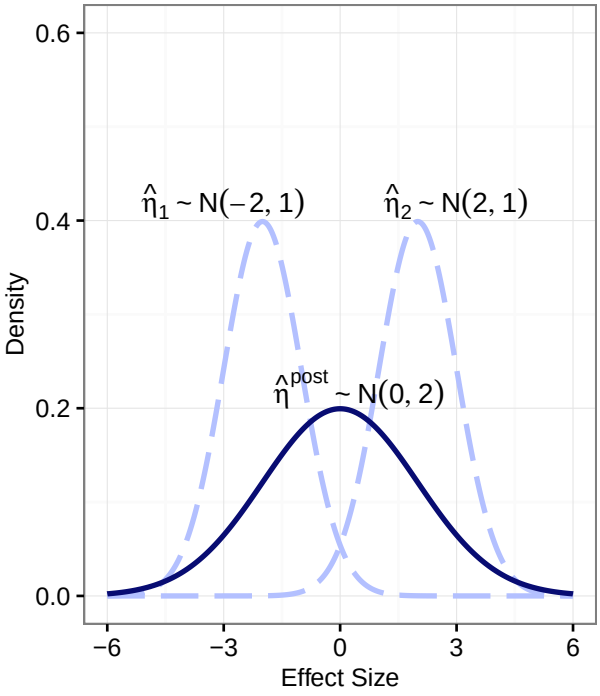
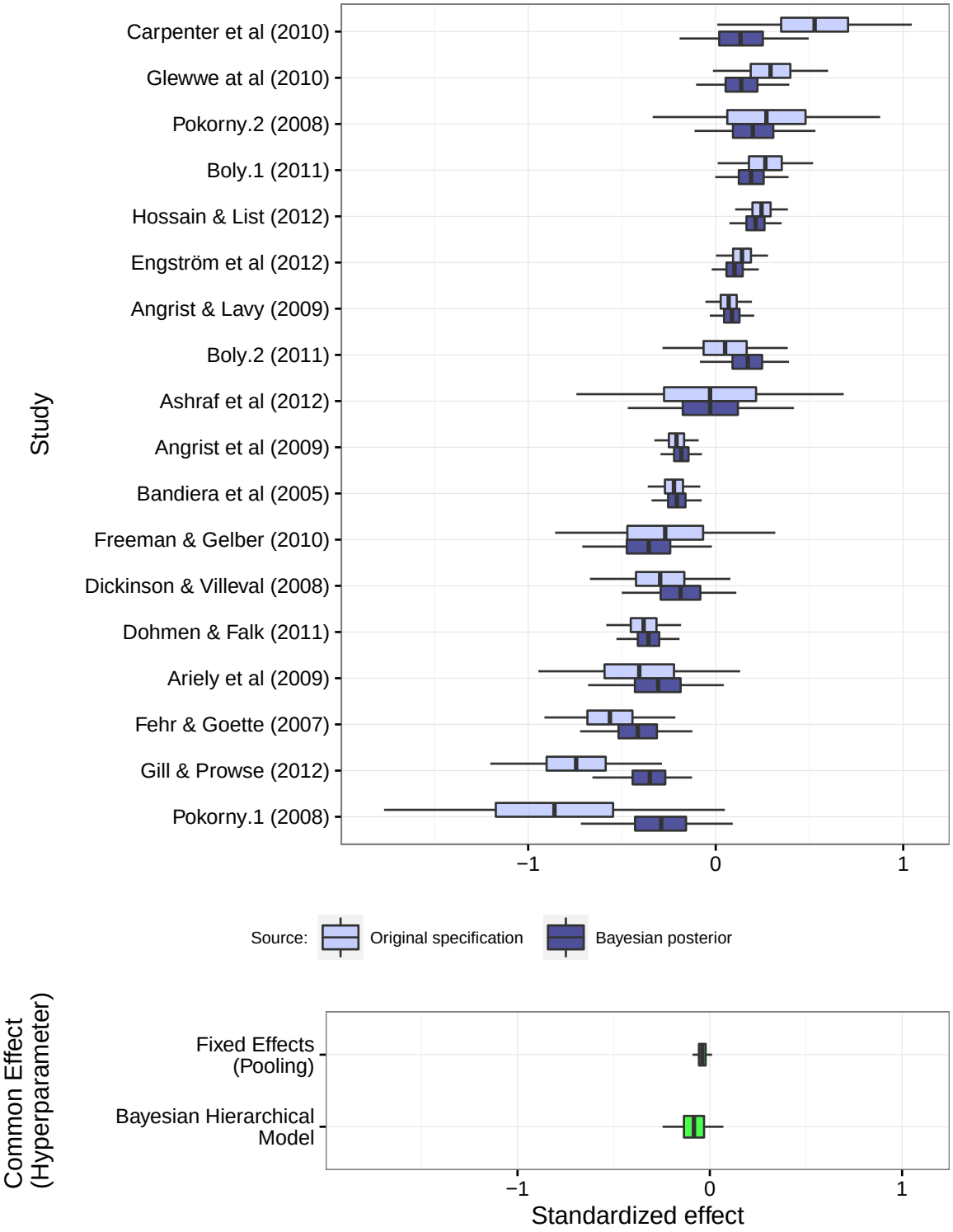


Figure 1d: Bayesian Hierarchical Model
High variation across studies



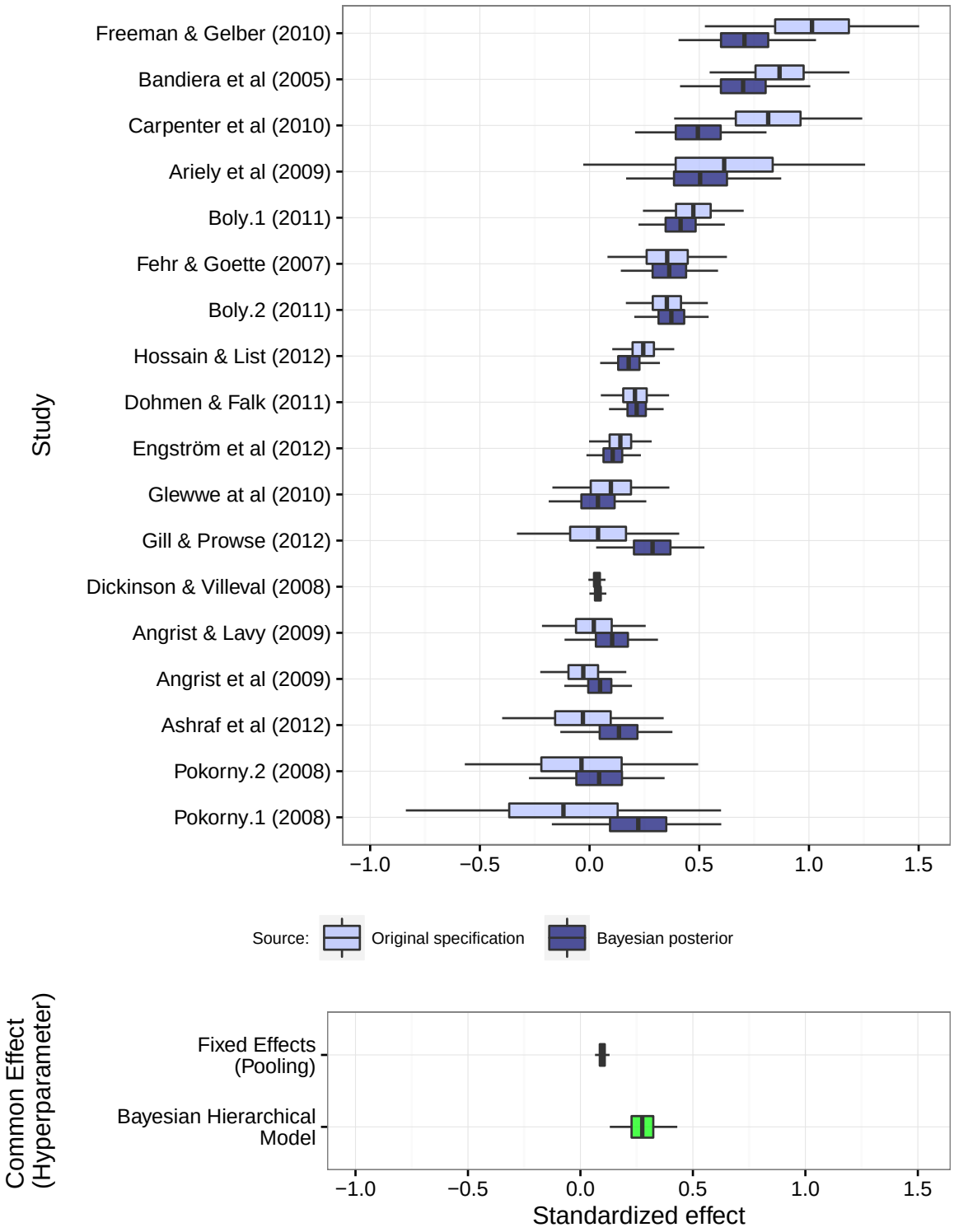
Note: Dashed lines represent distribution of study−level parameter estimates. Solid lines represent posterior distribution of population parameter, assuming uninformative prior. See Section 3.3 for discussion.

Figure 2: Original & Posterior Estimates for β (Gender)



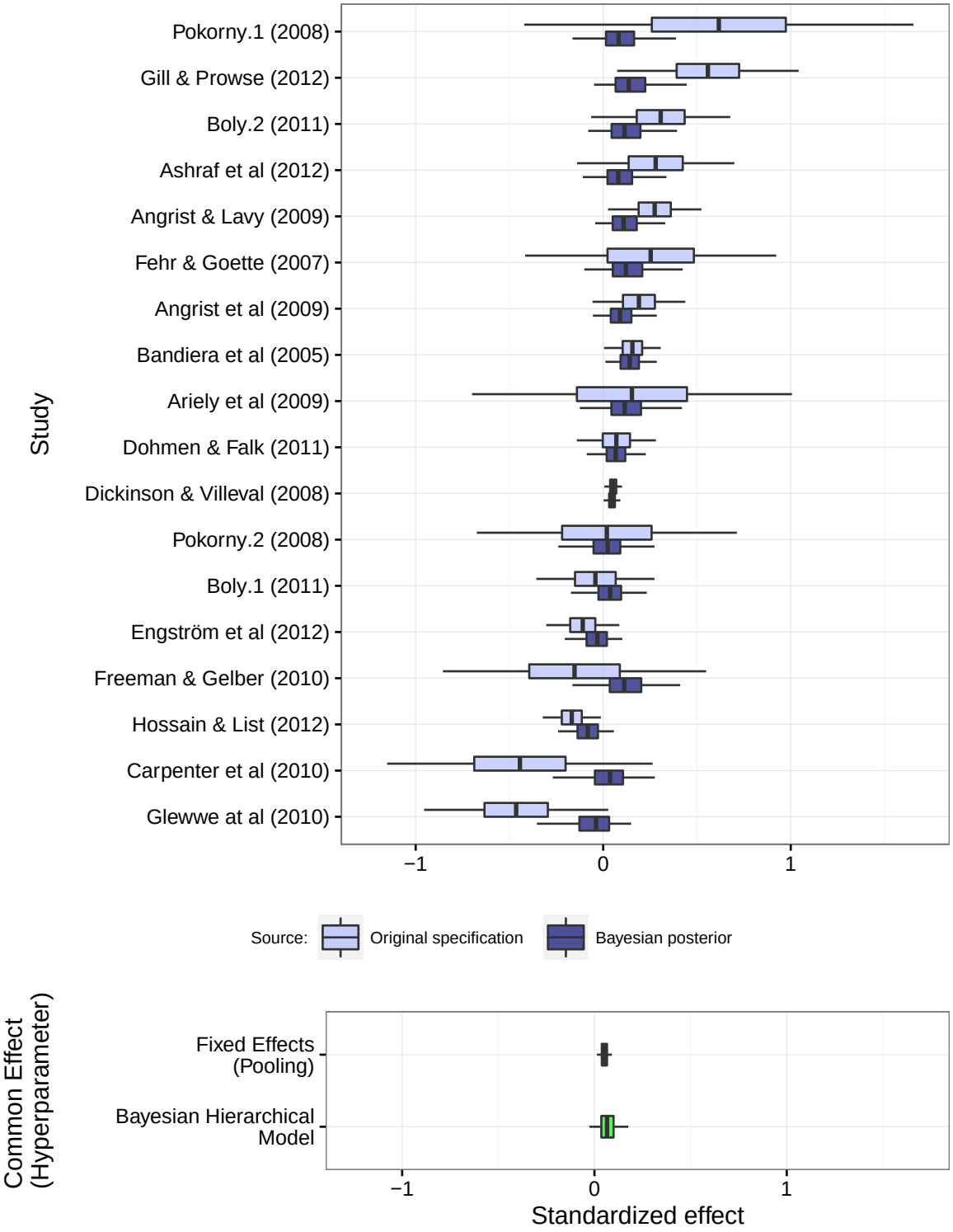
Note: Outcome variable for each study is standardized based on control group mean and standard deviation. Vertical line indicates median estimate, box indicates 50%−interval and line indicates 95%−interval. Fixed effects model calculated using the metafor package for R (Viechtbauer, 2010). Bayesian Hierarchical model implemented in Rstan.

Figure 3: Original & Posterior Estimates for γ (Incentives)



Note: Outcome variable for each study is standardized based on control group mean and standard deviation. 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.

Figure 4: Original & Posterior Estimates for η (Incentives x Gender)



Note: Outcome variable for each study is standardized based on control group mean and standard deviation. 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.

Figure 5: Posterior Distribution of τ (Hyperparameter Variance)

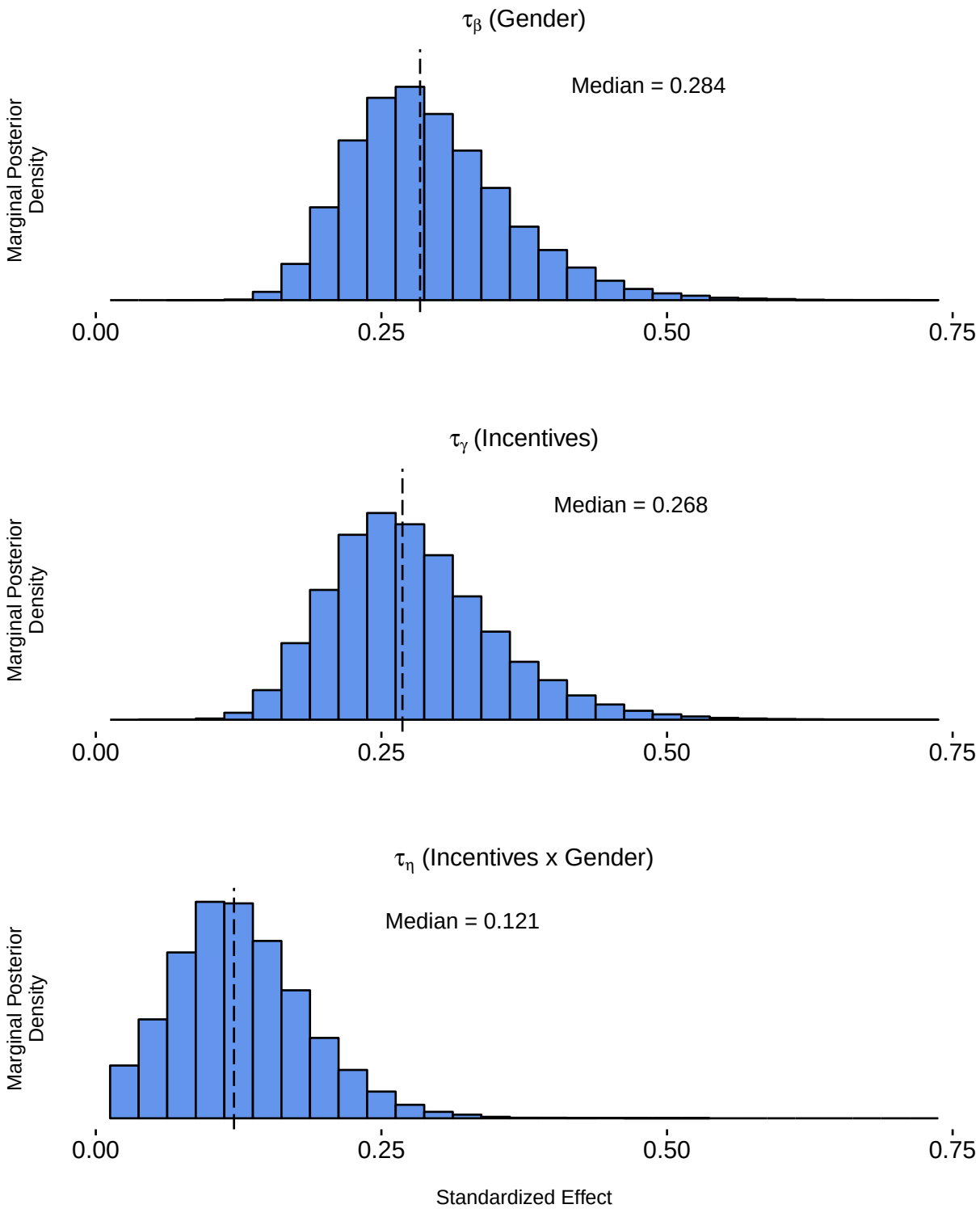
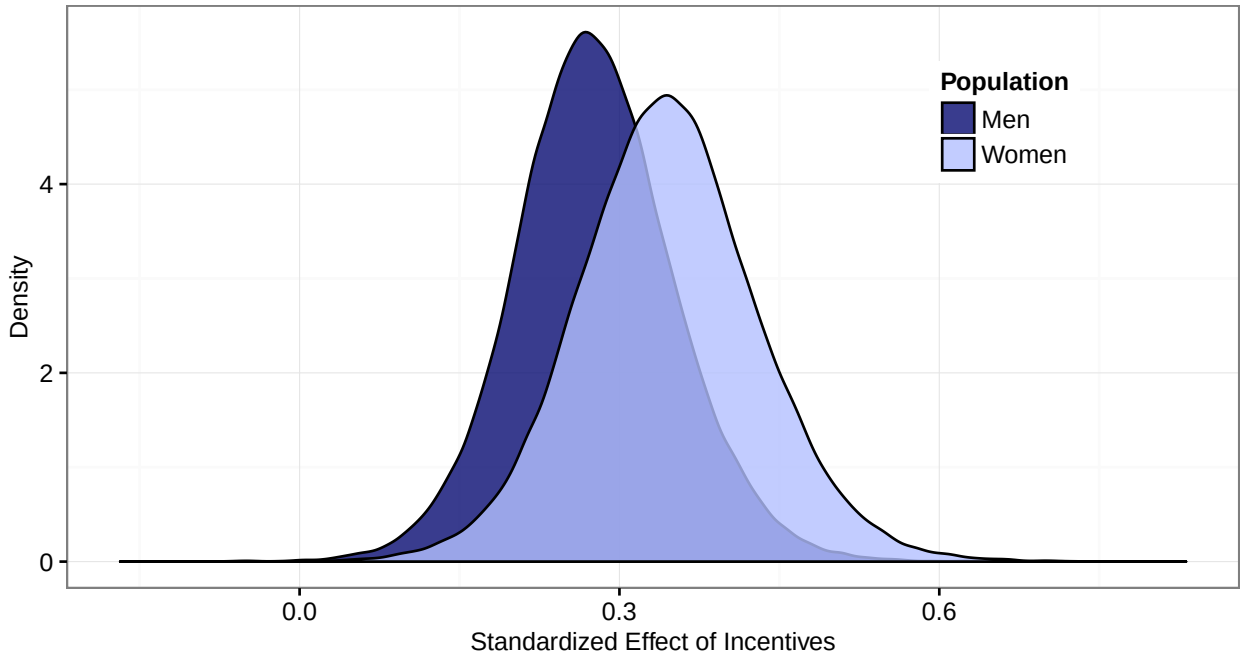
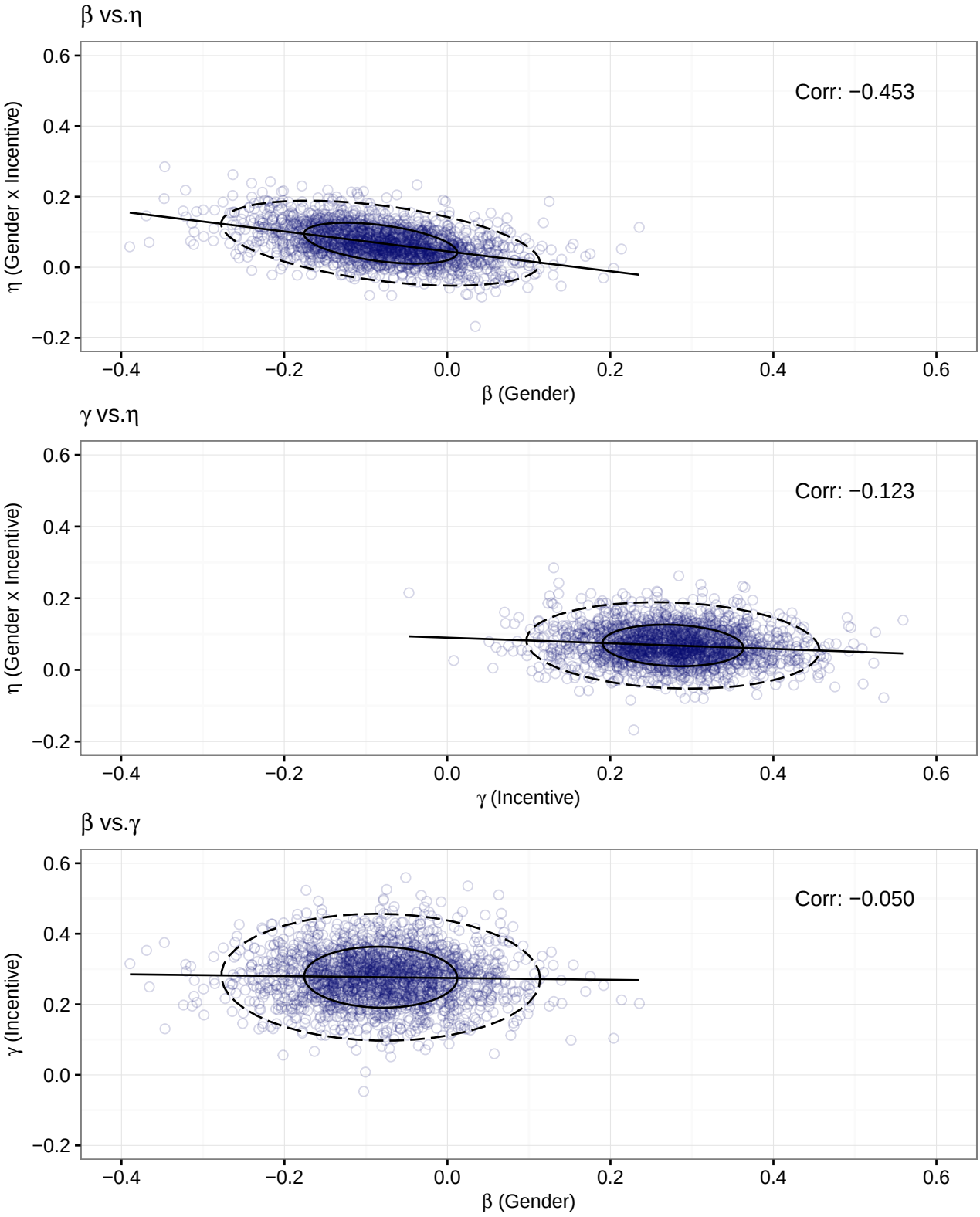


Figure 6: Predictive distribution by gender



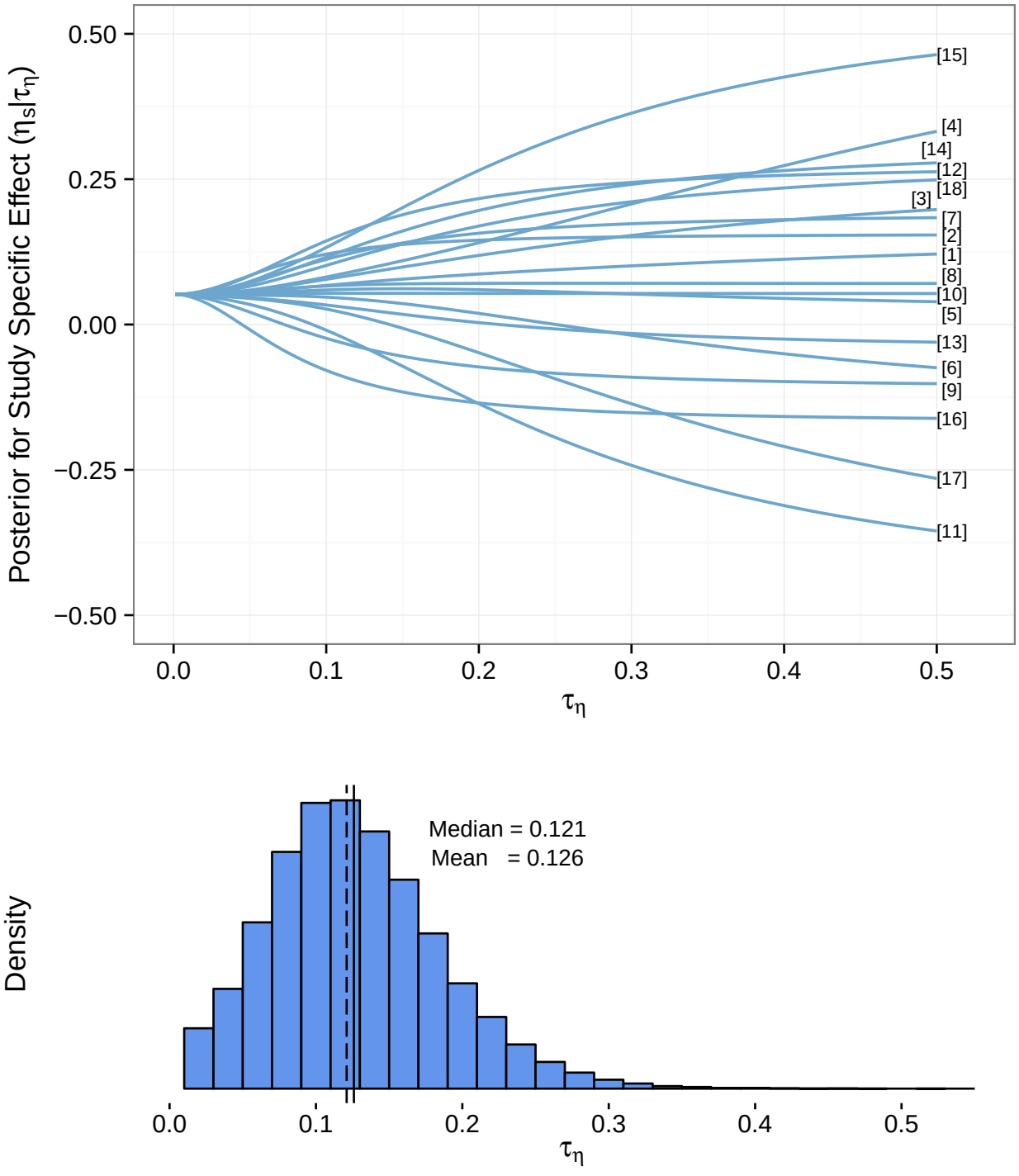
Note: Distributions based on posterior predictive distribution. The median for women is at the 82.8th percentile for men.

Figure 7: Bivariate Correlations of Hyperparameters



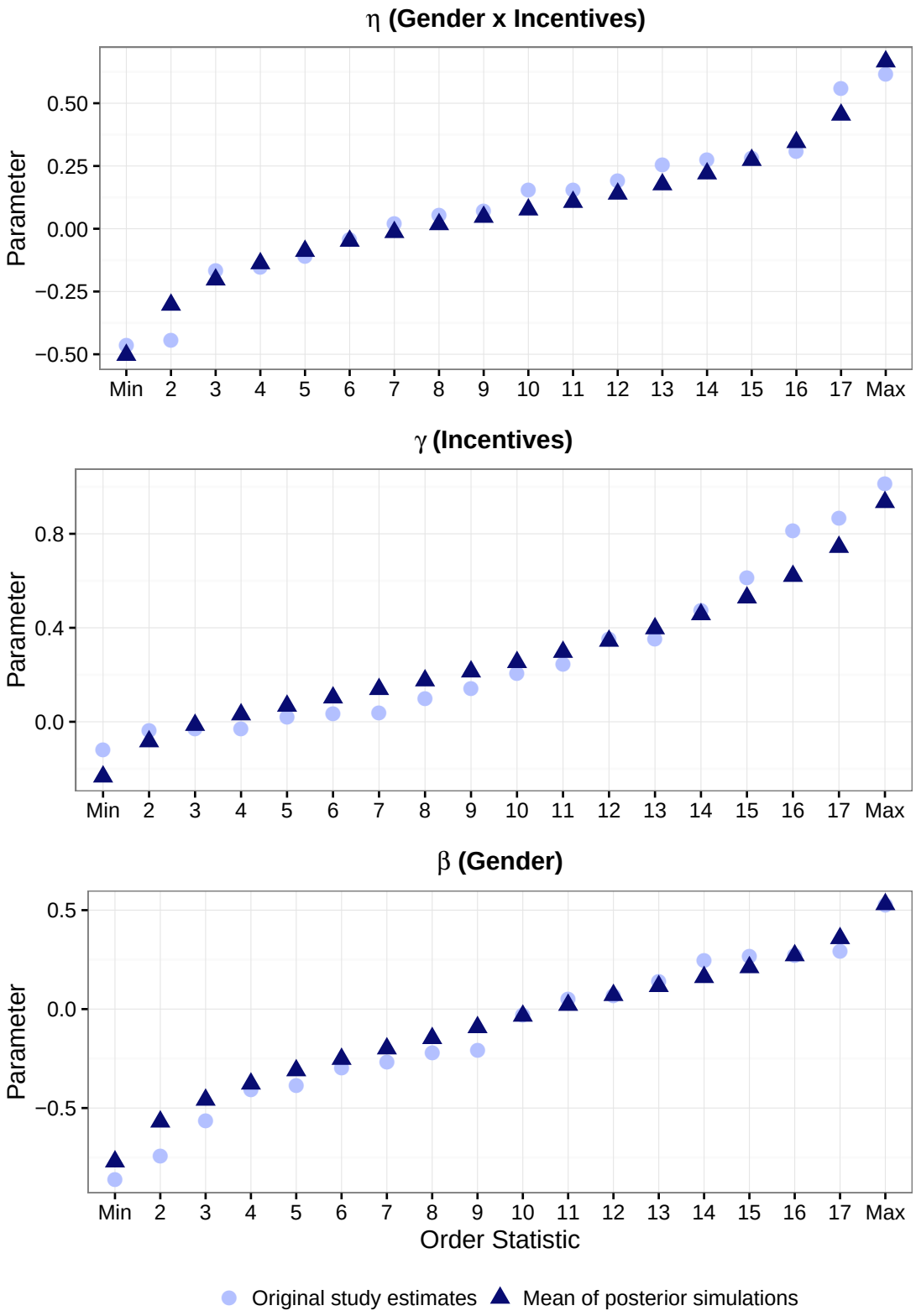
Note: Joint parameter estimates from simulated posterior distribution. Ellipses represent 50% and 95% intervals. Line displays linear best fit.

Figure 8: Posterior Mean of η_s (gender x incentives) conditional on τ_η



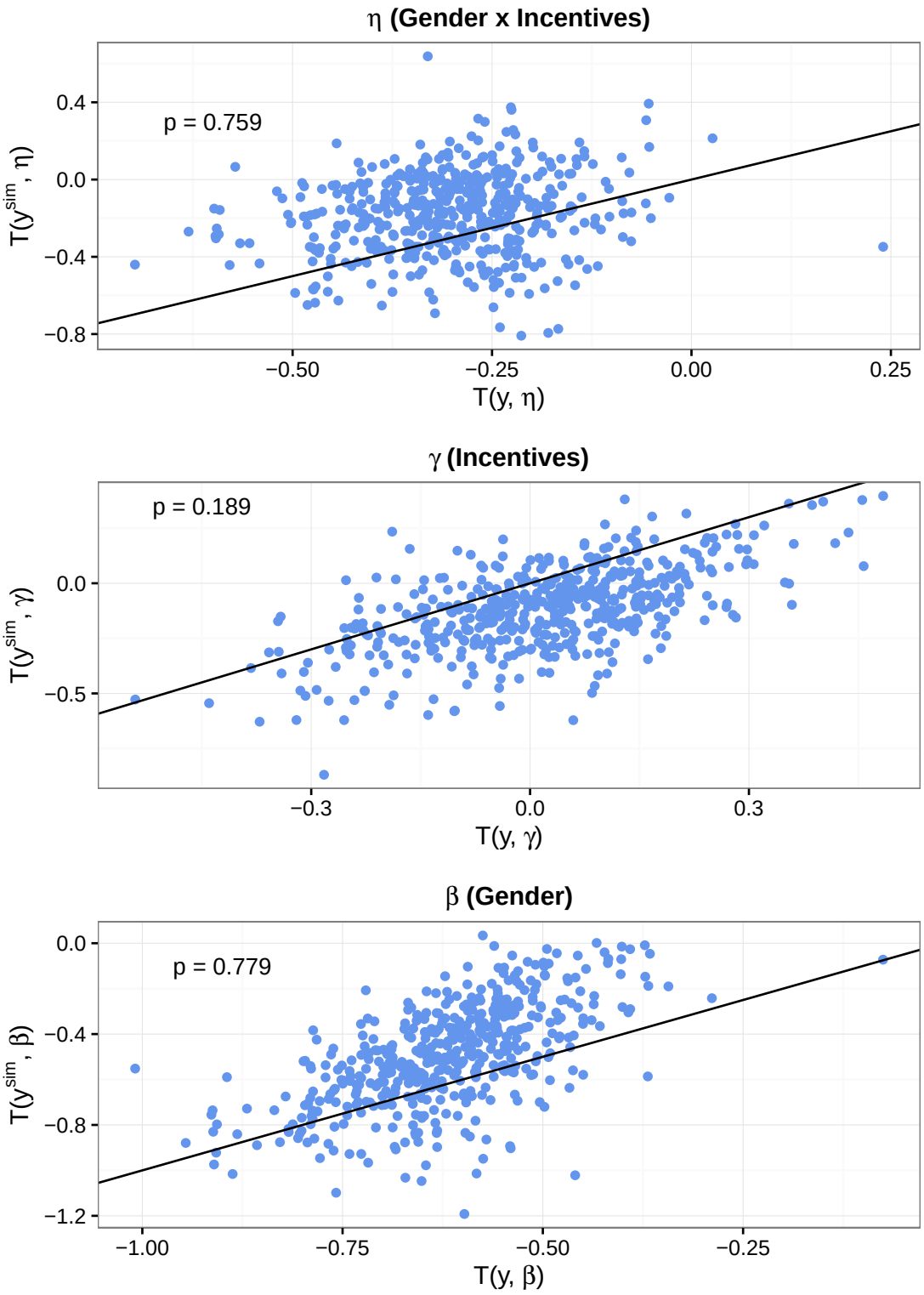
Note: Conditional posterior for: [1] Ariely et al (2009); [2] Bandiera et al (2005); [3] Fehr & Goette (2007); [4] Pokorny.1 (2008); [5] Pokorny.2 (2008); [6] Freeman & Gelber (2010); [7] Angrist et al (2009); [8] Dohmen & Falk (2011); [9] Engström et al (2012); [10] Dickinson & Villeval (2008); [11] Glewwe at al (2010); [12] Angrist & Lavy (2009); [13] Boly.1 (2011); [14] Boly.2 (2011); [15] Gill & Prowse (2012); [16] Hossain & List (2012); [17] Carpenter et al (2010); and [18] Ashraf et al (2012).

Figure 9: Posterior Predictive Checks, order statistics



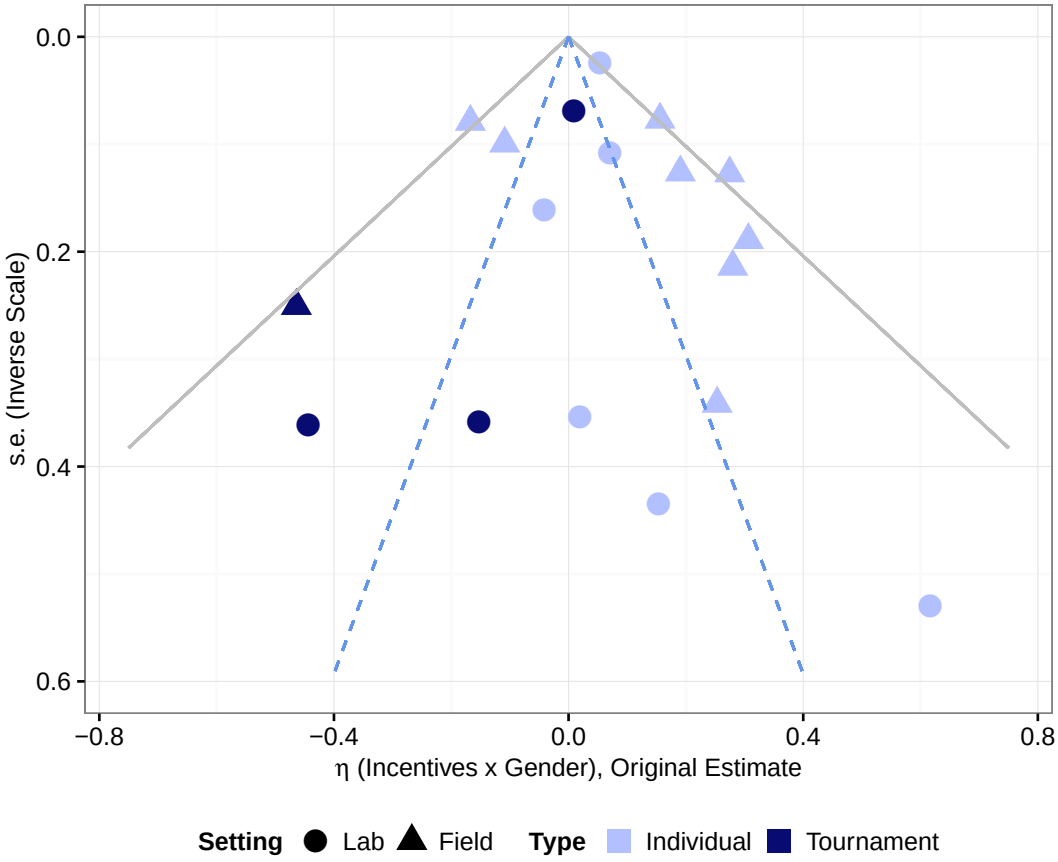
Note: Each plot compares the order statistic for observed parameter estimates to the analogous mean in posterior simulations. See section 4.4 for further discussion of posterior predictive checks.

Figure 10: Posterior Predictive Checks, skewness



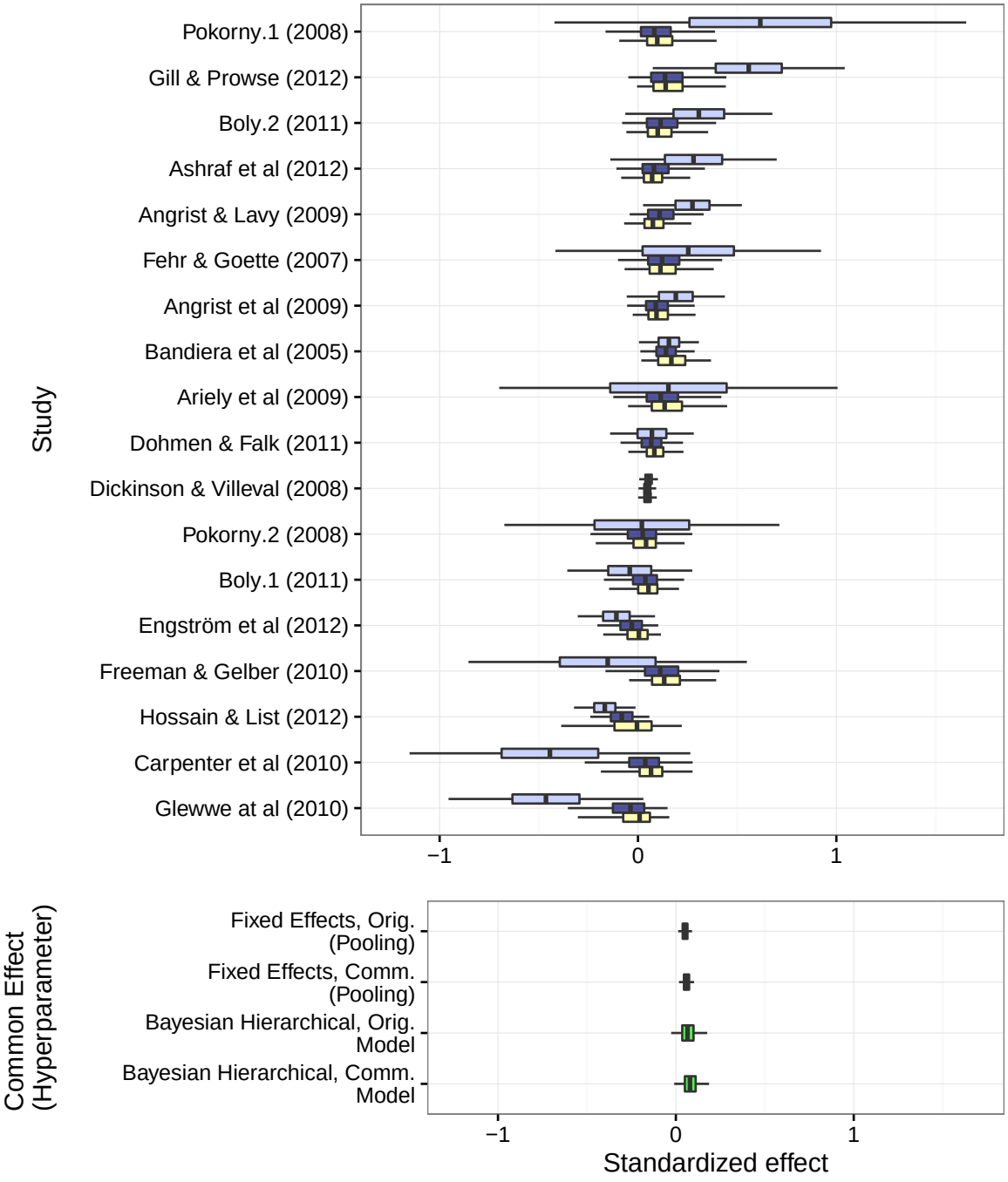
Note: Each point represents test statistic for on draw from posterior distribution. x−axis represents skewness measure, T, for observed data based on draw of parameter; y−axis represents skewness measure for simulated data, i.e., 18 study−level parameter estimates based on estimated model, and parameter draw. Bayesian p−values reported. See section 4.4 for detailed discussion.

Figure 11: Incentive x Gender Effect and Study Type



Note: Lines represent 95% (solid) and 50% (dashed) intervals for originally estimated values.

Figure A1: Posterior Estimates for Incentives x Gender Original & Common Specifications



Note: Outcome variable for each study is standardized based on control group mean and standard deviation. 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. See Section 3.2 for details.

TABLE 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 - No self-selection according to incentives

TABLE 2A: Summary of Included Studies

Study	Number of Subjects	Share Women	Field	Lab	Tournament
Angrist & Lavy (2009): The Effects of High Stakes School Achievement Awards: Evidence from a Randomized Trial	3,821	48.7%	x	-	-
Angrist et al (2009): Incentives And Services For College Achievement: Evidence From A Randomized Trial	1,255	58.1%	x	-	-
Ariely et al (2009): Doing Good or Doing Well? Image Motivation and Monetary Incentives in Behaving Prosocially	82	50.0%	-	x	-
Ashraf et al (2012): No Margin, No Mission? A Field Experiment on Incentives for Pro-Social Tasks	401	53.4%	x	-	-
Bandiera et al (2005): Social Preferences and the Response to Incentives: Evidence from Personnel Data	142	53.5%	x	-	-
Boly (2011): On the incentive effects of monitoring: evidence from the lab and the field	147 Lab	40.8%	-	x	-
	208 Field	15.4%	x	-	-
Carpenter et al (2010): Tournaments and Office Politics: Evidence from a Real Effort Experiment	111	54.1%	-	x	x
Dickinson & Villeval (2008): Does Monitoring Decrease Work Effort? The Complementarity between Agency and Crowding-out Theories	91	50.5%	-	x	-
Dohmen & Falk (2011): Performance Pay and Multi-Dimensional Sorting: Productivity, Preferences and Gender	359	50.4%	-	x	-
Engström et al (2012): Vacancy Referrals, Job Search, and the Duration of Unemployment: A Randomized Experiment	1,581	52.4%	x	-	-
Fehr & Goette (2007): Do Workers Work More if Wages Are High? Evidence from a Randomized Field Experiment	111	13.5%	x	-	-
Freeman & Gelber (2010): Prize Structure and Information in Tournaments: Experimental Evidence	234	60.3%	-	x	x
Gill & Prowse (2012): A Structural Analysis of Disappointment Aversion in a Real Effort Competition	590	55.9%	-	x	x
Glewwe at al (2010): Teacher Incentives	349	19.2%	x	-	x
Hossain & List (2012): The Behavioralist Visits the Factory: Increasing Productivity Using Simple Framing Manipulations	249	78.7%	x	-	-
Pokorny (2008): Pay—but do not pay too much: An experimental study on the impact of incentives	107 IQ	47.7%	-	x	-
	130 Counting	50.8%	-	x	-
Totals	**9,968**	**50.1%**	**9**	**9**	**4**

TABLE 2B: Summary of Included Studies, Treatment Detail

Study	Productivity Measure	Unit of Randomization	Description of Treatment	Description of Control
Angrist & Lavy (2009)	Bagrut (matriculation exam) performance.	Schools	Students are paid increasing monetary bonuses to: take any Bagrut (matriculation) test, pass any Bagrut test, complete all Bagrut requirements.	No financial incentives
Angrist et al (2009)	1st year GPA	Students	Students are paid monetary bonuses for improving their GPA.The bonus is higher for weaker students.	No bonuses
Ariely et al (2009)	Number of key press pairs (subjects had to hit the X and Z key sequentially)	Lab subjects	Subjects are paid a commission per output (key pair pressed), the commission rate is decreasing in output.	Subjects earn the same commission as in treatment but their earnings go to charity
Ashraf et al (2012)	Number of packs of condoms sold	Geographical clusters	Agents are paid a commission on each pack sold.	No commission
Bandiera et al (2005)	Log of worker's productivity (kilogram picked per hour per field-day)	Time periods	Workers are paid an exogenously fixed piece rate per kg of fruit picked.	Workers are paid a piece rate, decreasing with the average productivity of the group, per kg of fruit picked.
Boly (2011)	Exam grading accuracy	Lab subjects Students	Subjects' pay is negatively related to the number of grading mistakes in a randomly audited sample of exams. Treatment 1: randomly audit 1/20 exams, punishment schedule is flat. Treatment 2: randomly audit 5/20 exams, punishment schedule is steep.	Fixed wages
Carpenter et al (2010)	Quality adjusted envelopes produced (quality as assessed by outside expert)	Lab subjects	Piece rate per envelope produced plus bonus for the most productive agent in a group of eight.	Piece rate
Dickinson & Villeval (2008)	Curve height (subjects click a button to uncover a curve on a screen, the faster they click the higher the curve they uncover)	Lab subjects	Subjects' pay equals a fixed wage minus a penalty for low output if audited. Treatment equals 1 for high audit probability.	Same pay as in treatment but audit probability is low.
Dohmen & Falk (2011)	Negative of the time (in sec.) needed to solve multiplication problem in steps 1 and 2 of the experiment	Lab subjects	Subjects are paid a commission on output (number of multiplications solved in 5 minutes).	No commission.

TABLE 2B: Summary of Included Studies, Treatment Detail

Study	Productivity Measure	Unit of Randomization	Description of Treatment	Description of Control
Engström et al (2012)	Whether job seekers apply for jobs	Job seekers	Subjects' job applications are monitored more intensively.	Same monitoring as in treatment but unknown to workers.
Fehr & Goette (2007)	(Log) revenue per shift during fixed shifts (pure intensive margin response)	Bycicle messengers	Higher commission rate on deliveries.	Normal commission rate on deliveries.
Freeman & Gelber (2010)	Number of mazes solved	Lab subjects	Tournaments (single and multiple prizes).	Fixed wages.
Gill & Prowse (2012)	Number of slides correctly placed	Lab subjects	30 prizes of different magnitude; we scale continuous treatment dose pro rata between 0 (no prize) and 1.0 (the maximum prize).	
Glewwe at al (2010)	Teacher attendance in second year of programme (1999)	Schools	Tournaments with prizes for "Top-scoring schools" and "Most improved schools" based on exam performance of all grade 4-8 students.	Fixed wages.
Hossain & List (2012)	Log of per-hour productivity in a given week	Employees of tech firm	Monetary bonus awarded if productivity is above a set treshold.	Fixed wages.
Pokorny (2008)	Score on IQ test Score on number counting task	Lab subjects	Three treatments: high incentive, low incentive and very low incentive	No incentives.

TABLE 3: Summary of Hyperparameter Estimates

			Quantiles				
	Mean	**s.e.**	**2.5%**	**25%**	**50%**	**75%**	**97.5%**
Gender x Incentives							
η (effect hyperparameter)							
BHM	0.069	0.051	-0.026	0.036	0.066	0.099	0.176
Pooling (FE)	0.052	0.020	0.013	0.038	0.052	0.065	0.091
τ_η (variance hyperparameter)							
BHM	0.126	0.060	0.021	0.085	0.121	0.162	0.257
Pooling (FE)[1]	- -	- -	- -	- -	- -	- -	- -
Incentives							
γ (effect hyperparameter)							
BHM	0.277	0.075	0.131	0.227	0.275	0.324	0.431
Pooling (FE)	0.097	0.016	0.065	0.086	0.097	0.108	0.129
τ_γ (variance hyperparameter)							
BHM	0.277	0.072	0.161	0.226	0.268	0.318	0.442
Pooling (FE)[1]	- -	- -	- -	- -	- -	- -	- -
Gender							
β (effect hyperparameter)							
BHM	-0.083	0.081	-0.246	-0.134	-0.082	-0.030	0.071
Pooling (FE)	-0.039	0.026	-0.090	-0.056	-0.039	-0.021	0.012
τ_β (variance hyperparameter)							
BHM	0.293	0.070	0.182	0.243	0.284	0.333	0.453
Pooling (FE)[1]	- -	- -	- -	- -	- -	- -	- -

Hyperparameter estimates from Bayesian hierarchical model based on empirical distribution from posterior simulations. Fixed effects estimates calculate theoretical quantiles from estimated mean and standard error in pooling (classical fixed-effects) model. See Section 3 for details. [1]Note that pooling model assumes variance hyperparameter (τ) is zero, i.e., true study-level effects are everywhere identical

TABLE 4: Pooling Metrics

	η (Gender x Incentives)		γ (Incentives)		β (Gender)	
Common pooling factor	0.684		0.234		0.254	
By study	**Variance**	**Shrinkage**	**Variance**	**Shrinkage**	**Variance**	**Shrinkage**
Angrist & Lavy (2009)	0.356	0.755	0.202	0.319	0.113	-0.109
Angrist et al (2009)	0.321	0.757	0.146	0.243	0.108	0.203
Ariely et al (2009)	0.491	0.323	0.340	0.313	0.316	0.300
Ashraf et al (2012)	0.410	0.888	0.238	0.525	0.399	-0.033
Bandiera et al (2005)	0.254	0.151	0.279	0.279	0.120	0.108
Boly.1 (2011)	0.378	0.691	0.182	0.291	0.165	0.215
Boly.2 (2011)	0.443	0.754	0.157	-0.271	0.196	-0.870
Carpenter et al (2010)	0.504	0.919	0.285	0.589	0.306	0.639
Dickinson & Villeval (2008)	0.151	-0.411	0.086	0.018	0.269	0.501
Dohmen & Falk (2011)	0.299	1.220	0.121	0.108	0.142	0.085
Engström et al (2012)	0.328	0.406	0.126	-0.247	0.120	0.175
Fehr & Goette (2007)	0.477	0.644	0.205	-0.126	0.259	0.305
Freeman & Gelber (2010)	0.517	1.227	0.296	0.414	0.299	-0.488
Gill & Prowse (2012)	0.463	0.826	0.228	1.033	0.218	0.579
Glewwe at al (2010)	0.512	0.766	0.214	-0.330	0.217	0.410
Hossain & List (2012)	0.334	0.355	0.136	-2.033	0.129	0.098
Pokorny.1 (2008)	0.497	0.956	0.376	0.858	0.361	0.725
Pokorny.2 (2008)	0.480	0.015	0.302	0.250	0.285	0.196

See Section 3.4 for discussion of pooling factor calculations.

TABLE 5: Posterior Predictive Checks, Order Statistics

	p-value		
	η (Gender x Incentives)	**γ (Incentives)**	**β (Gender)**
Order Statistic			
Min	0.501	0.293	0.695
$\theta_{(2)}$	0.833	0.381	0.868
$\theta_{(3)}$	0.416	0.656	0.807
$\theta_{(4)}$	0.600	0.865	0.637
$\theta_{(5)}$	0.627	0.830	0.790
$\theta_{(6)}$	0.492	0.895	0.716
$\theta_{(7)}$	0.334	0.959	0.811
$\theta_{(8)}$	0.279	0.895	0.822
$\theta_{(9)}$	0.333	0.877	0.917
$\theta_{(10)}$	0.117	0.755	0.487
$\theta_{(11)}$	0.223	0.766	0.365
$\theta_{(12)}$	0.232	0.441	0.512
$\theta_{(13)}$	0.169	0.682	0.359
$\theta_{(14)}$	0.251	0.412	0.130
$\theta_{(15)}$	0.434	0.205	0.235
$\theta_{(16)}$	0.579	0.077	0.461
$\theta_{(17)}$	0.238	0.210	0.669
Max	0.520	0.321	0.429

See Section 4.4 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.

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, which extends immediately to the full multivariate model. Following (7) above, 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:[13]

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

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 $y = \left\{\hat{\eta}_i, \hat{\sigma}_j\right\}_{i=1}^S$ 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.

In practice, this is easily accomplished using the RStan package for the programming language R. 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

[13] 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.

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, $\{\eta_i\}_{i=1}^{S}$.