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

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

Existing empirical work raises the hypothesis that performance pay – whatever its output gains – may widen the gender earnings gap, because women may respond less to incentives. We evaluate this possibility by aggregating evidence from existing experiments on performance incentives with male and female subjects. Using a Bayesian hierarchical model we estimate both the average effect and heterogeneity across studies. We find that the gender response difference is close to zero and heterogeneity across studies small, while performance pay increases output by 0.36 standard deviations on average. The data thus support agency theory for men and women alike.
JEL: J16, J31, C11
Keywords: wage differentials, gender, econometrics, meta-analysis

Women may respond less to incentive pay for a number of cultural and psychological reasons, such as differences in risk aversion or self-confidence (Charness and Gneezy, 2012; Eckel and Grossman, 2008*b*; Reuben et al., 2012; Niederle and Vesterlund, 2007). Shurchkov and Eckel (2018); Blau and Kahn (2017); Azmat and Petrongolo (2014); Bertrand (2011) and Croson and Gneezy (2009) review this literature in detail, highlighting the lack of evidence on the impact of these differences on labor market outcomes. Importantly, if women respond less to performance incentives, this raises the possibility that performance pay – whatever its output gains – may increase gender inequality in earnings.

This paper tests whether women are less responsive to high-powered incentives commonly underlying performance pay in the workplace using a large, hitherto unexplored collection of laboratory and field experiments that identify the response to performance incentives. We use a Bayesian hierarchical model to estimate both the average gender differences as well as heterogeneity across studies. This method has recently been introduced to economics to aggregate evidence on a topic (Hsiang, Burke and Miguel (2013); Burke, Hsiang and Miguel (2015); Vivalt (2015) and Meager (2019; 2020)) but also to ask new questions Meager (2019). In particular, BHMs can be used to explore dimensions of heterogeneity that individual studies cannot, either because they lack statistical power or because it was not among their original stated goals. 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.[1] Second, the model uses the data itself to estimate the degree

* Bandiera: Department of Economics and STICERD, LSE and CEPR (o.bandiera@lse.ac.uk); Fischer: School of Public Policy and STICERD, LSE and CEPR (g.fischer@lse.ac.uk); Prat: Columbia University and CEPR (andrea.prat@columbia.edu); Ytsma: Tepper School of Business, Carnegie Mellon University (eytsma@andrew.cmu.edu). Rohini Pande was coeditor for this article. We thank Florian Blum and Szymon Sacher for excellent research assistance. We are grateful for helpful comments from Gharad Bryan, Ray Fisman, Andrew Gelman, Gerard Padró i Miquel, Jörn-Steffen Pischke, Bernard Salanié and seminar participants at UCL, Bocconi, Yale, Stanford, University of Washington, Columbia, the University of Manchester, LMU Munich, DFID and DIW.

[1] See Casey, Glennerster and Miguel (2012) and Olken (2015). We see this as a natural complement to Athey and Imbens (2015) and Dwork et al. (2015), which address this issue at the study-level.

to which each study is informative about a common phenomenon versus its own context-specific effect; thus it allows us to quantify how informative the findings of one study are for another.

Agency theory predicts that performance pay affects an individual's effort on the job, expected earnings and, through this, selection into jobs (see e.g. Lazear (2000)). Thus if women respond less to performance pay, they may also sort into jobs that do not offer performance pay.[2] Here we focus on the effort effect both because agency theory predicts it drives the selection effect and because experiments on selection are rare.

Women have been found to be more risk averse than men (Charness and Gneezy, 2012; Eckel and Grossman, 2008*b*), less confident (Reuben et al., 2012; Niederle and Vesterlund, 2007), more altruistic (Croson and Gneezy, 2009; Eckel and Grossman, 2008*a*) and more averse to competition (Niederle, 2016; Niederle and Vesterlund, 2011). Importantly for this paper, moral hazard theory would predict that these traits affect the expected utility of effort and thus the response to performance pay. Indeed, several experimental studies have found a weaker incentive response in risk-averse subjects (Cadsby et al., 2016; Zubanov, 2012; Cadsby, Song and Tapon, 2007); subjects with low self-confidence (Heite, Hoisl and Lakhani, 2019) or in pro-social tasks (Gneezy and Rustichini, 2000), provided financial incentives are low (Hossain and Li, 2014). Furthermore, Gneezy, Niederle and Rustichini (2003) show that men outperform women in tournaments, though only in mixed tournaments.

To proceed, we 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 studies that provide evidence relevant for understanding the effect of performance pay in the workplace, 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. We identified 29 studies satisfying the inclusion criteria and were able to obtain and use data from 17.

Our sample comprises 9 lab and 8 field experiments involving 8791 subjects, 50.5% of which are women. Tasks include uncovering curves or placing sliders, taking or grading exams, picking fruits or inspecting consumer electronics. The high-powered incentives range from tournament pay to bonuses, monitoring, commission or piece rates, while control conditions feature fixed pay or a lower prize, commission, piece rate or monitoring probability.

The Bayesian hierarchical model (BHM) posits that the observed estimate ($\hat{\eta_s}$) in 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 in men, σ. Our main finding is that the estimated distribution

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

of the gender-incentive coefficient (η) has a mean that is positive but close to zero ($+0.07\sigma$)—implying women are slightly *more* responsive to financial incentives—with little variance (0.11σ) across studies. That is, women and men respond similarly to different variants of performance pay across a wide range of contexts. If we were to run a new experiment, we would expect a similar response to steeper incentives in men and women, 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; the mean response to performance pay is positive and large ($+0.36\sigma$). Given the diversity of contexts and treatments, the estimated heterogeneity is also quite large, though it affects primarily the magnitude rather than the sign of the effect. Replicating the existing set of studies, a classical approach to inference is expected to yield a negative significant (at the 5%-level) effect of incentives, in fewer than 1% of cases.

The rest of this paper is organized as follows. Section I describes the study sample, Section II presents the methodology, and Section III the results. Section IV concludes.

I. Study sample

The first step in building evidence from multiple studies is to establish inclusion criteria[3] for study selection.

To maximize quality while minimizing subjective judgments, we restrict our sample to lab and field experiments published in refereed journals or the working paper series of the main research associations (CEPR, IZA, NBER). As experimental analyses of incentives have started relatively recently, we restrict our search to papers published between 1990 and 2012, when this study began.[4]

The second set of criteria serves to select studies that can be informative of gender differences in the response to financial incentives in the workplace. We therefore restrict our sample to studies where subjects choose effort that is (i) real, as opposed to hypothetical, and (ii) produces output. Furthermore, we only include studies with at least two treatments, one of which is unambiguously more high-powered than the other, such that the expected marginal effect on pay of an increase in performance is larger.

Finally, since we focus on the effort response to incentives, we only include studies in which subjects cannot self-select into incentive schemes, to avoid confounding effects. We also exclude studies with externalities in production, such as team production and incentives, to avoid bringing in vastly different mechanisms like cooperation.

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 166 papers, of which 29 passed the inclusion criteria[5]. For 15 of these, the

[3]Summarized in Appendix Table A1.

[4]A small number of experimental studies have looked at gender differences in the response to performance incentives since, with mixed results. Tonin and Vlassopoulos (2015) find a larger positive performance response to piece rates in men, Erat and Gneezy (2016) find a larger negative response to competitive pay in women, while Mbiti et al. (2019) find no significant gender difference in the response to bonuses.

[5]Appendix Table A3 lists these 29 papers.

data was available online or shared with us by the authors. Among the rest, 7 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.[6] Of the 15 papers, two report two experiments – Boly (2011) and Pokorny (2008). These are included separately as they meet the inclusion criteria individually[7]. Table 1 summarizes all included studies.

For each study, we focus on the cleanest test of financial incentives meeting our selection criteria. In all but one case, this is the paper's primary analysis; for Dohmen et al. (2011) we use data from the first two preliminary rounds of the experiment as only these satisfy our no self-selection criterion.

There are 9 lab and 8 field experiments which, together, report on the behavior of 8,791 unique subjects, of which 50.5% are women. In the lab experiments, tasks range from pressing key pairs to uncovering a curve or placing sliders, grading exams, stuffing envelopes, solving multiplication problems or mazes, taking an IQ test or performing counting tasks. In the field experiments, tasks range from taking or grading exams to applying for jobs, selling condoms, picking fruits, making deliveries or inspecting consumer electronics. While the lab experiments generally employ university students in North America or Europe as subjects, locations and subjects in the field experiments range from high school and university students in Israel, Canada and Burkina Faso, to unemployed job seekers in Sweden, hair stylists in Zambia, fruit pickers in the UK, bike messengers in Switzerland and factory workers in China.

The incentives introduced also vary considerably. Three experiments feature tournament pay as the high-powered incentive scheme, three others feature bonuses, seven experiments introduce commission or piece rates and the remaining four introduce monitoring regimes. Control conditions range from fixed pay to a lower prize, commission, piece rate or monitoring probability.

The diversity in contexts and incentive schemes across studies is essential to identify a truly universal pattern in the response to workplace financial incentives. It also complicates comparing incentive power across studies, though we note that the highest monetary value rewards occur in field experiments. Importantly however, differences in incentive power should not matter for the primary objective of this paper - to assess whether men and women respond systematically differently to incentives. In each context, men and women face the same incentives. Moreover, we test for heterogeneity in the gender difference by incentive strength and context in sections (III.B) and (III.C) below.

A few included studies collect data on some of the traits in which men and women are thought to differ, namely risk preferences (Dickinson and Villeval, 2008; Carpenter, Matthews and Schirm, 2010; Dohmen et al., 2011) and social preferences (Ashraf, Bandiera and Jack, 2014; Dohmen and Falk, 2011). None of these papers evaluate whether such traits impact the effort response to incentives. Fehr, Klein and Schmidt (2007), however, find that loss averse subjects drive the effort response to incentives on the intensive margin. Only one of the studies reports a gender-incentive interaction term in the original paper; Freeman and Gelber (2010) mention that the interaction is not significant in the classical sense.

[6] We cannot include these studies, because the BHM requires the full variance-covariance matrix of any estimation and normalized outcome measures.

[7] In both papers, the two experiments have distinct control groups.

II. Methodology

A. Descriptive model of performance

In order to estimate the relative effect of incentives on the productivity of women versus men, we begin with a descriptive model of the performance of individual i on a task in 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 variable for women and T_{is} for the high-powered treatment. For instance, if one group is paid fixed wages and the other piece rates, we set $T_{is} = 1$ for the latter. Equation (1) is the non-parametric cell-means regression with respect to gender and incentives, so 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, the gender-incentive effect, which captures the differential effect of incentives on women relative to men in study s. If men and women respond similarly to incentives, η_s equals zero. Hence, even though the treatment dummy T_{is} does not differentiate between incentive strength of the high-powered treatment across experiments, this should not affect our core objective, to test whether η_s equals zero.

We aim to understand generalizable differences in the response to incentives, and doing so entails aggregating across disparate studies. For comparability, we therefore normalize the outcome variable as $\tilde{y}_{is} = (y_{is} - \bar{y_s})/\hat{\sigma}_s$, where $\bar{y_s}$ is the sample mean and $\hat{\sigma}_s$ the sample standard deviation for men in the control group. Such standardization is common in the education literature, for instance, to deal with variation in test scores across schools (Lavy, 2009; Glewwe, Ilias and Kremer, 2010; Duflo, Dupas and Kremer, 2015). Furthermore, standardization should not affect our central test, whether the gender difference in the incentive response is zero. Moreover, we find that results are robust to standardizing by the full control sample in a study rather than only men in the control sample.

For each study we then estimate the vector of parameters, $\theta_s = (\tilde{\beta}_s, \tilde{\gamma}_s, \tilde{\eta}_s)'$ 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 - an OLS regression with study-specific controls in most cases, only adding the gender-incentive interaction term where necessary[8]. Appendix Table A2 details the included specifications for each paper and Appendix Table A6 provides data citations. As a robustness check, we also estimate a common specification for each study, excluding covariates[9].

Table 1 shows that OLS estimation of (1) and (2) yields a positive and significant effort response to incentives in ten experiments, while the gender difference in the incentive response is significant - and positive - in only two. Without standardization, the effort response estimates range from -0.981 to 851.56, and from -0.148 to 1.006 only after standardization, while the gender-incentive effect ranges

[8] Accordingly, to replicate the specifications in Engström, Hesselius and Holmlund (2012); Freeman and Gelber (2010); Fehr and Goette (2007), we estimate OLS regressions, even though the outcome measure is a binary variable in the first and the data has a panel structure in the latter two studies.

[9] Results in Appendix Figure A7.

from -1.385 to 609.455 without standardization, and from -0.665 to 0.768 with standardization.

The vector of parameter estimates, $\hat{\theta}_s = (\hat{\theta_s}, \hat{\gamma_s}, \hat{\eta_s})$, and the associated covariance matrix, $\hat{\Sigma}_s$, for each study form the inputs in the Bayesian hierarchical model we describe below.

B. The Bayesian Hierarchical Model

Our analysis focuses on the Bayesian hierarchical model for the full parameter vector, $\theta = (\beta, \gamma, \eta)$, to allow us to explore heterogeneity across studies along the dimension of potentially correlated parameters. We use the canonical multivariate BHM for aggregating across studies as described in (Rubin, 1981). The BHM assumes that each observed study result, $\hat{\theta}_s$, is estimating its own study-specific effect, θ_s. These study-specific θ_s's are in turn distributed in the population with mean θ and covariance Σ, where the population hyperparameters θ and Σ are themselves random variables. Formally:

$$\begin{aligned} \hat{\theta}_s &\sim N[\theta_s, \Sigma_s] \quad s = 1, \ldots, S \\ \theta_s &\sim N[\theta, \Sigma], \end{aligned} \tag{3}$$

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:

$$\begin{aligned} \theta &\sim N[0,\ 1000^2 * I_3] \\ \Sigma &\sim diag(\sigma)\ \Omega\ diag(\sigma) \\ \sigma_k &\sim Cauchy(0, 2.5),\ \text{for } k \in \{\beta,\gamma,\eta\} \text{ and } \sigma_k > 0 \\ \Omega &\sim LKJcorr(2) \end{aligned} \tag{4}$$

where N denotes a multi-variate normal distribution, Ω is a correlation matrix and σ is the vector of coefficient scales (Gelman, 2006).The LKJ distribution (Lewandowski, Kurowicka and Joe, 2009) is a distribution over correlation matrices, i.e., positive semi-definite matrices with unit diagonals.

The second line embodies a critical assumption: the study-level effects $(\theta_1, \ldots, \theta_S)$ are themselves normally distributed in the population with mean θ and covariance Σ. We assume a normal distribution because it aids tractability and has been shown to perform well in various applications (McCulloch and Neuhaus, 2011; Gelman et al., 2004). We test the appropriateness of this assumption in Appendix E and find that the data conform quite well. Even so, our results are best interpreted as the distribution of incentive effects in the population of contexts in which economists have been willing to run experiments. The extent to which these settings represent the broader population points to further questions regarding the placement of experiments and the representativeness of empirical work more generally (see e.g., Cartwright and Deaton, 2016 and Allcott, 2015).

The key assumption to estimate the joint probability model is exchangeability. Technically, this means that the joint distribution of $(\theta_1, \ldots, \theta_S)$ is invariant to

permutations of the indices $(1, \ldots, S)$. It allows us to write the joint distribution of the θ_s's as i.i.d. given hyperparameters θ and Σ. Intuitively, it means there is no information other than the data, y, to distinguish one study from another. In practice, this assumption is not very restrictive and can easily be relaxed with partial or conditional exchangeability. If there are study-level characteristics that one expects to be informative about the parameters of interest, one could group data together with an additional level of hierarchy or add additional parameters to the analysis (e.g., expanding the parameter space by including interactions with study type), as we do below.

Finally, (4) indicates prior distributions for the hyperparameters. We 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. Our posterior predictions are largely insensitive to alternative priors, suggesting that there is sufficient information contained in our data indeed.[10,11]

Our estimation of BHMs follows closely the procedures described in Gelman and Hill (2007) and Gelman et al. (2004) (see Appendix A for details[12]). The key outputs from this estimation are the simulated posterior distributions for the hyperparameters, θ and Σ, and the true study-level effects, $\{\theta_i\}_{i=1}^{S}$. 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. In addition to calculating means and posterior intervals (the Bayesian analog to confidence intervals), we can also use these simulated distributions to test other functions of the parameters. For instance, we can calculate cross-correlations of parameter values drawn from these simulated distributions, to evaluate whether the gender-incentive effect, η, is greater in contexts with a stronger incentive effect, γ (see section (III.B)).

The simulated posterior is a joint distribution over not only the population hyperparameters—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 the mean estimate of an internally valid study *in that very same context*—is first attributed to Stein (Efron and Morris, 1977). The Bayesian hierarchical model serves to make this belief-updating process transparent and precise.

III. Results

A. *The response to incentives for men and women*

Table 2 summarizes the posterior distribution of the hyperparameters (γ, η, and β, and the corresponding elements of τ).[13] Given the available data and our speci-

[10]Reducing the variance of the prior on θ from $1000^2 * I_3$ to $0.1^2 * I_3$ changes the median of the posterior for η by less than 0.001. Even a strongly informed uncentered prior for η ($N(-0.1, 0.1^2)$) only reduces the posterior median from 0.068 to 0.049.

[11]For the LKJ distribution too, the choice of prior has little impact on the posterior distributions. For example, changing the scale parameter for the LKJ prior from 2 to 1—making correlations across parameters more likely—does not change the median of the posterior on η (within rounding errors) and moves the correlations of the posterior predictive distribution on e.g. β and η from -0.37 to only -0.40.

[12]Please see Bandiera et al. (2020) for data and code.

[13]Appendix D discusses posteriors of the true study-level effects.

fied (uninformative) prior beliefs, it describes the population distribution of (i) the gender difference in the response to incentives, (ii) men's response to incentives and (iii) the gender difference in unincentivized productivity, as well as the estimated standard deviation of each of these parameters. Because the data are standardized, the unit of measure for the parameters is the standard deviation of productivity for unincentivized men in each setting.

The table shows that $\eta = 0$, embodying the idea that men and women respond equally, is well within the credible interval. The median and mean of the BHM estimates for the gender-incentive interaction hyperparameter, η, are 0.068 and 0.066, with a 95%-interval of $[-0.050,\ 0.173]$. The sign of the estimate is positive, suggesting that, contrary to the implications of gender differences in traits like risk aversion, women respond slightly more to incentives than men do. Results are robust to standardizing by the full control sample in a study rather than only men in the control sample.[14]

Table 2 also shows that the estimated cross-study heterogeneity is relatively low (median $\tau_\eta = 0.106$). Moreover, there is considerable mass in the posterior distribution at $\tau_\eta \approx 0$[15]. This implies that the estimated gender response difference in study n is highly predictive of the same in study $n + 1$. That is, despite substantial variation in context, including task, location, and incentives, 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; knowing that the gender differential is zero implies that the next, hypothetical study is also very likely to find a zero effect. A further assessment of the heterogeneity and commonality across contexts is provided in Appendix C, which discusses pooling metrics.

Having established that women and men respond similarly, we are interested in assessing whether they both respond positively. 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. Men increase productivity by about one-third of one standard deviation in response to high-powered incentives. As shown in Table 2, the median and mean for the posterior estimate of γ are 0.356 and 0.357, with a 95%-interval of $[0.188,\ 0.532]$. This is consistent with the main prediction of agency theory and casts doubt on the practical relevance of crowd-out.

There is substantial cross-study heterogeneity in γ; the median estimate of τ_γ is 0.295 and values below 0.098 have no mass. This 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.

For completeness, Table 2 also reports the estimates of β, 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 -0.061 and -0.062. Not surprisingly, given the diversity of contexts covered by the sample studies, the distribution is quite spread out. The 95%-interval spans $[-0.240,\ 0.113]$, and the median for τ_β is 0.297.

[14] Appendix B compares BHM estimates with pooling model estimates.
[15] Full posterior distribution in Appendix Figure A1.

1. Predictions

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 1 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 0.36σ (with an interquartile range from 0.30σ to 0.41σ) and for women by an average of 0.42σ (with an interquartile range from 0.37σ to 0.48σ). Comparing the two distributions, the median of the posterior predictive distribution for women is at the 79^{th} 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 ($\eta_{S+1} < -0.18$) in 4.7% of studies and less than the mean effect for men in about 1% of studies. Alternatively, if we could rerun the 17 experiments included in this study, maintaining all the design features including sample size, classical inference would expect to find a negative and statistically significant (at the 5%-level) gender difference in 2.7% of the replications and a positive and statistically significant difference in 10%. In other words, 87% of replications would not be able to statistically distinguish the responses of women and men. In contrast, replicating the existing set of studies, classical inference would expect to find a negative and significant effect of incentives in fewer than 1% of cases and a positive and significant effect in 53%.

B. Cross-correlations

As noted above, it is difficult to compare incentive power across experiments because studies differ in incentive structure and strength as well as context. Accordingly, our descriptive model of performance features only an indicator variable for higher-powered incentives. We would, however, like to assess whether the gender-incentive interaction varies with incentive power, and in particular, whether the gender difference in incentive responses grows with incentive power. To do so, we draw values for γ, men's responsiveness to incentives, and η, the gender difference in responsiveness, from the posterior predictive distribution, then plot pairwise combinations in bivariate scatter plots and calculate correlations.

Figure 2 shows that the estimated correlation between γ and η is -0.253, and the estimated average gender-incentive effect is consistently positive, albeit small. To the extent that the incentive response (γ) is stronger when incentive power is greater, as agency theory predicts, the correlation suggests that the incentive response of men and women becomes more similar, rather than more divergent, as incentives grow stronger.

A similar test can be implemented with respect to β, the gender productivity gap. The estimated correlation between β and η is -0.371, with η large and positive when β is small and negative. Hence, when women perform worse than men with low-powered incentives, women respond more strongly to high-powered incentives than men, thus closing the productivity gap. Whatever causes women's productivity to be less than men's under low-powered incentives (e.g. distaste for a task, less complementary inputs), this result suggests that stronger incentives drive women to make up for this difference with extra effort.

Finally, the bottom panel of Figure 2 shows that the correlation between β and γ is close to 0. Thus there is no discernible relationship between the gender productivity gap and the effect of financial incentives for men.

C. Study-level heterogeneity

As a final test, we assess heterogeneity in the distribution of treatment effects with respect to two study-level characteristics: (1) whether the study was a field or lab experiment and (2) whether the incentives were tournament-based. To do so, we expand the parameter space for θ in (3) to allow both the main incentive effect, γ, and the gender-incentive interaction, η, to vary according to study type by including interaction terms.

Some of the gender differences in behavioral traits have been found to be context dependent, for instance overconfidence (Bordalo et al., 2019) and altruism (Andreoni and Vesterlund, 2001; Eckel and Grossman, 1998; Bolton and Katok, 1995). The literature on gender norms suggests a possible explanation; differences in behaviors might reflect norm-conforming behavior rather than innate traits (Akerlof and Kranton, 2000; Alesina, Giuliano and Nunn, 2013; Bertrand, Kamenica and Pan, 2015). D'Acunto (2019); Cadsby, Servátka and Song (2013) and Boschini, Muren and Persson (2012) for instance show evidence of gender differences in aversion to competition, altruism, risk aversion and overconfidence when gender roles are made more salient. But then, differences in the salience of gender norms between lab and field studies could give rise to different gender-incentive effects. Furthermore, if the power of incentives is higher in field experiments, comparing the gender-incentive effect across lab and field studies may provide a further test of its sensitivity to incentive power. Field experiments may also expose subjects to more production risk. If women are more risk averse, and if this risk increases with effort, we may then expect to find a weaker incentive response in women compared to men in field experiments.

As shown in Figure 3, we find no evidence of systematic differences between field and lab experiments. While the incentive-gender interaction term is 0.13σ higher for field experiments, the 95%-interval includes 0 and spans $[-0.12\sigma, 0.38\sigma]$. This suggests that there are no substantial differences in the salience of norms or the exposure to risk, or that any differences are too small to bring about a noticeable divergence in the incentive response of men and women. Any differences in incentive strength between lab and field experiments are also either too small or not causing the gender difference in incentive responses to bifurcate substantially.

We also analyze heterogeneity between tournament and non-tournament incentives, motivated by potential differences in women's and men's attitudes towards competition. We find that the incentive-gender interaction term is 0.22σ lower for tournaments than for non-competitive incentives, with a 95%-interval of $[-1.02\sigma, 0.56\sigma]$. Our sample only contains three tournaments and the parameters are only weakly identifiable, so the results should be interpreted with caution, but they suggest that further experimentation along this dimension could be fruitful.

IV. Discussion

Performance pay is at the core of agency theory and management practices. Not surprisingly, given this popularity with theorists and practitioners, the effectiveness of various performance incentives has been tested in several lab and field experiments. In this paper we use a Bayesian Hierarchical Model to aggregate this

evidence to test whether incentives increase performance to the same extent for men and women. We find that incentives commonly underlying performance pay schemes in the workplace increase performance for men and women alike across a variety of contexts and for a variety of incentive designs. This finding suggests that the widespread use of performance pay[16] is unlikely to contribute to the gender earnings gap directly.

To the extent that women differ in risk aversion, confidence and altruism, our finding suggests that these differences are not strong enough to generate different responses. One possible explanation could be that women do not differ in behavioral traits so much as they engage in norm-appropriate behavior. If the experiments did not activate gender norms, the resulting absence of norm-appropriate behavior may have given rise to gender-neutral responses. In a similar vein, if the link between risk and higher effort is either weak or absent in experiments, we may fail to find gender differences in the response to incentives even if women are more risk averse. More research on whether gender norms or risk exposure give rise to gender differences in the response to incentives would therefore be valuable.

Another reason for the gender-neutral result could be the absence of the selection channel in the included experiments. Although we assume, following e.g. Lazear (2000), that the effort effect drives the selection effect, it may be that other factors influence selection in the labor market. Women might have a distaste for competition (Niederle and Vesterlund, 2007; Gneezy, Niederle and Rustichini, 2003), or a greater preference for flexible work hours which may intersect with household composition (Heywood and Parent, 2017; Goldin, 2014; Bertrand, Goldin and Katz, 2010) for example. Furthermore, men and women may optimally negotiate different compensation contracts in the labor market if they differ on behavioral traits (Albanesi, Olivetti and Prados, 2015). Here too, more research would be valuable.

The results also illustrate the usefulness of Bayesian hierarchical models as a tool to build evidence from existing studies and assess external validity and, in doing so, we contribute to a growing literature in economics (Meager, 2019, 2020; Vivalt, 2015; Burke, Hsiang and Miguel, 2015; Hsiang, Burke and Miguel, 2013). Moreover, like (Meager, 2019), we show that building evidence from existing studies allows researchers to test for heterogeneity across subgroups for which individual studies might be underpowered, and to capitalize on the recent explosion in field and laboratory experiments to answer new questions with existing data. As such, we see BHMs as a powerful tool to build on existing knowledge and give directions on what experiments to run next.

[16]Lemieux, Macleod and Parent (2009), for instance, show that the incidence of performance pay increased from 38% in the 1970s to 45% in the 1990s in the US, Manning and Saidi (2010) document a rise from 16.3% in 1998 to 32% in 2004 in the UK, and Sommerfeld (2013) finds an increase from 15.4% in 1984 to 39.4% in 2009 in Germany.

References

Akerlof, George A, and Rachel E Kranton. 2000. "Economics and identity." *The quarterly journal of economics*, 115(3): 715–753.

Albanesi, Stefania, Claudia Olivetti, and María José Prados. 2015. *Gender and Dynamic Agency: Theory and Evidence on the Compensation of Top Executives.* Emerald Group Publishing Limited.

Alesina, Alberto, Paola Giuliano, and Nathan Nunn. 2013. "On the origins of gender roles: Women and the plough." *The Quarterly Journal of Economics*, 128(2): 469–530.

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

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

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

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

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

Bandiera, Oriana, Greg Fischer, Andrea Prat, and Erina Ytsma. 2020. "Replication data for: Do Women Respond Less to Performance Pay? Building Evidence from Multiple Experiments." *American Economic Association, Inter-university Consortium for Political and Social Research, https://doi.org/10.3886/E129821V1.*

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

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

Bertrand, Marianne, Emir Kamenica, and Jessica Pan. 2015. "Gender identity and relative income within households." *The Quarterly Journal of Economics*, 130(2): 571–614.

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

Blau, Francine D, and Lawrence M Kahn. 2017. "The gender wage gap: Extent, trends, and explanations." *Journal of Economic Literature*, 55(3): 789–865.

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

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

Bordalo, Pedro, Katherine Coffman, Nicola Gennaioli, and Andrei Shleifer. 2019. "Beliefs about gender." *American Economic Review*, 109(3): 739–73.

Boschini, Anne, Astri Muren, and Mats Persson. 2012. "Constructing gender differences in the economics lab." *Journal of Economic Behavior & Organization*, 84(3): 741–752.

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

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

Cadsby, C Bram, Fei Song, and Francis Tapon. 2007. "Sorting and incentive effects of pay for performance: An experimental investigation." *Academy of management journal*, 50(2): 387–405.

Cadsby, C Bram, Fei Song, Francis Tapon, et al. 2016. "The impact of risk-aversion and stress on the incentive effect of performance-pay." *Experiments in Organizational Economics*, 19: 189–227.

Cadsby, C Bram, Maroš Servátka, and Fei Song. 2013. "How competitive are female professionals? A tale of identity conflict." *Journal of Economic Behavior & Organization*, 92: 284–303.

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

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

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

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

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

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

D'Acunto, Francesco. 2019. "Identity and choice under risk." *Available at SSRN 3466626*.

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

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

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

Duflo, Esther, Pascaline Dupas, and Michael Kremer. 2015. "School Governance, Teacher Incentives, and Pupil–teacher Ratios: Experimental Evidence from Kenyan Primary Schools." *Journal of Public Economics*, 123: 92–110.

Dwork, Cynthia, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Roth. 2015. "The reusable holdout: Preserving validity in adaptive data analysis." *Science*, 349(6248): 636–638.

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

Eckel, Catherine C, and Philip J Grossman. 2008*a*. "Differences in the Economic Decisions of Men and Women: Experimental Evidence." *Handbook of Experimental Economics Results*, 1: 509–519.

Eckel, Catherine C, and Philip J Grossman. 2008*b*. "Men, women and risk aversion: Experimental evidence." *Handbook of experimental economics results*, 1: 1061–1073.

Efron, Bradley, and Carl Morris. 1977. "Stein's Paradox in Statistics." *Scientific American*, 236: 119–127.

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

Erat, Sanjiv, and Uri Gneezy. 2016. "Incentives for creativity." *Experimental Economics*, 19(2): 269–280.

Fehr, Ernst, Alexander Klein, and Klaus M Schmidt. 2007. "Fairness and Contract Design." *Econometrica*, 75(1): 121–154.

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

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

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

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

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

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

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

Gneezy, Uri, and Aldo Rustichini. 2000. "Pay enough or don't pay at all." *The Quarterly journal of economics*, 115(3): 791–810.

Gneezy, Uri, Muriel Niederle, and Aldo Rustichini. 2003. "Performance in competitive environments: Gender differences." *The quarterly journal of economics*, 118(3): 1049–1074.

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

Heite, Jonas, Karin Hoisl, and Karim R Lakhani. 2019. "Performance in Contests: The Role of Risk and Confidence."

Heywood, John S, and Daniel Parent. 2017. "Performance pay, the gender gap, and specialization within marriage." *Journal of Labor Research*, 38(4): 387–427.

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

Hossain, Tanjim, and King King Li. 2014. "Crowding out in the labor market: A prosocial setting is necessary." *Management Science*, 60(5): 1148–1160.

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

Lavy, Victor. 2009. "Performance Pay and Teachers' Effort, Productivity, and Grading Ethics." *The American Economic Review*, 99(5): 1979–2021.

Lazear, Edward P. 2000. "Performance Pay and Productivity." *American Economic Review*, 90(5): 1346–1361.

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

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

Manning, Alan, and Farzad Saidi. 2010. "Understanding the gender pay gap: what's competition got to do with it?" *ILR Review*, 63(4): 681–698.

Mbiti, Isaac, Karthik Muralidharan, Mauricio Romero, Youdi Schipper, Constantine Manda, and Rakesh Rajani. 2019. "Inputs, incentives, and complementarities in education: Experimental evidence from Tanzania." *The Quarterly Journal of Economics*, 134(3): 1627–1673.

McCulloch, Charles E, and John M Neuhaus. 2011. "Misspecifying the shape of a random effects distribution: why getting it wrong may not matter." *Statistical science*, 388–402.

Meager, Rachael. 2019. "Understanding the average impact of microcredit expansions: A Bayesian hierarchical analysis of seven randomized experiments." *American Economic Journal: Applied Economics*, 11(1): 57–91.

Meager, Rachael. 2020. "Aggregating Distributional Treatment Effects: A Bayesian Hierarchical Analysis of the Microcredit Literature." *mimeo*.

Niederle, Muriel. 2016. "Gender." In *Handbook of Experimental Economics*. . 2 ed., , ed. John Kagel and Alvin E. Roth, 481–553. Princeton University Press.

Niederle, Muriel, and Lise Vesterlund. 2007. "Do Women Shy Away from Competition? Do Men Compete Too Much?" *The Quarterly Journal of Economics*, 1067–1101.

Niederle, Muriel, and Lise Vesterlund. 2011. "Gender and competition." *Annu. Rev. Econ.*, 3(1): 601–630.

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

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

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

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

Shurchkov, Olga, and Catherine C. Eckel. 2018. "Gender Differences in Behavioral Traits and Labor Market Outcomes." In *The Oxford Handbook of Women and the Economy*. 481. Oxford University Press.

Sommerfeld, Katrin. 2013. "Higher and higher? Performance pay and wage inequality in Germany." *Applied Economics*, 45(30): 4236–4247.

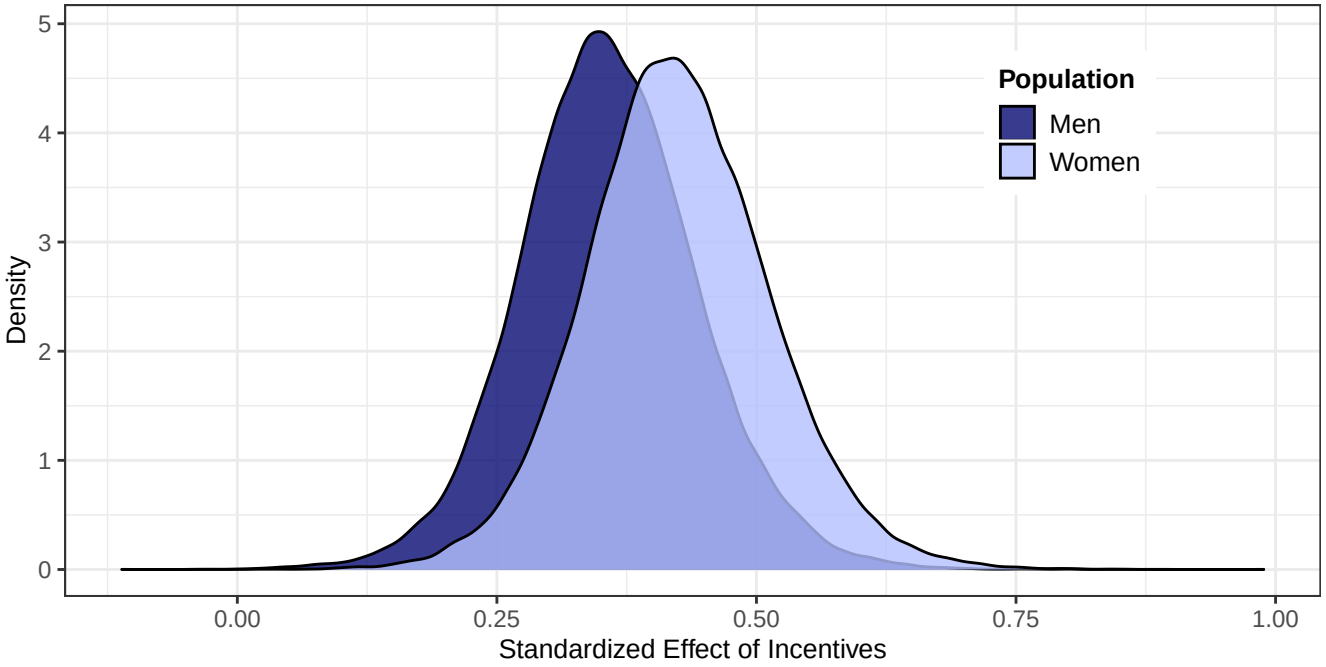
Stan Development Team. 2020. "RStan: the R interface to Stan." R package version 2.21.2.

Tonin, Mirco, and Michael Vlassopoulos. 2015. "Corporate philanthropy and productivity: Evidence from an online real effort experiment." *Management Science*, 61(8): 1795–1811.

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

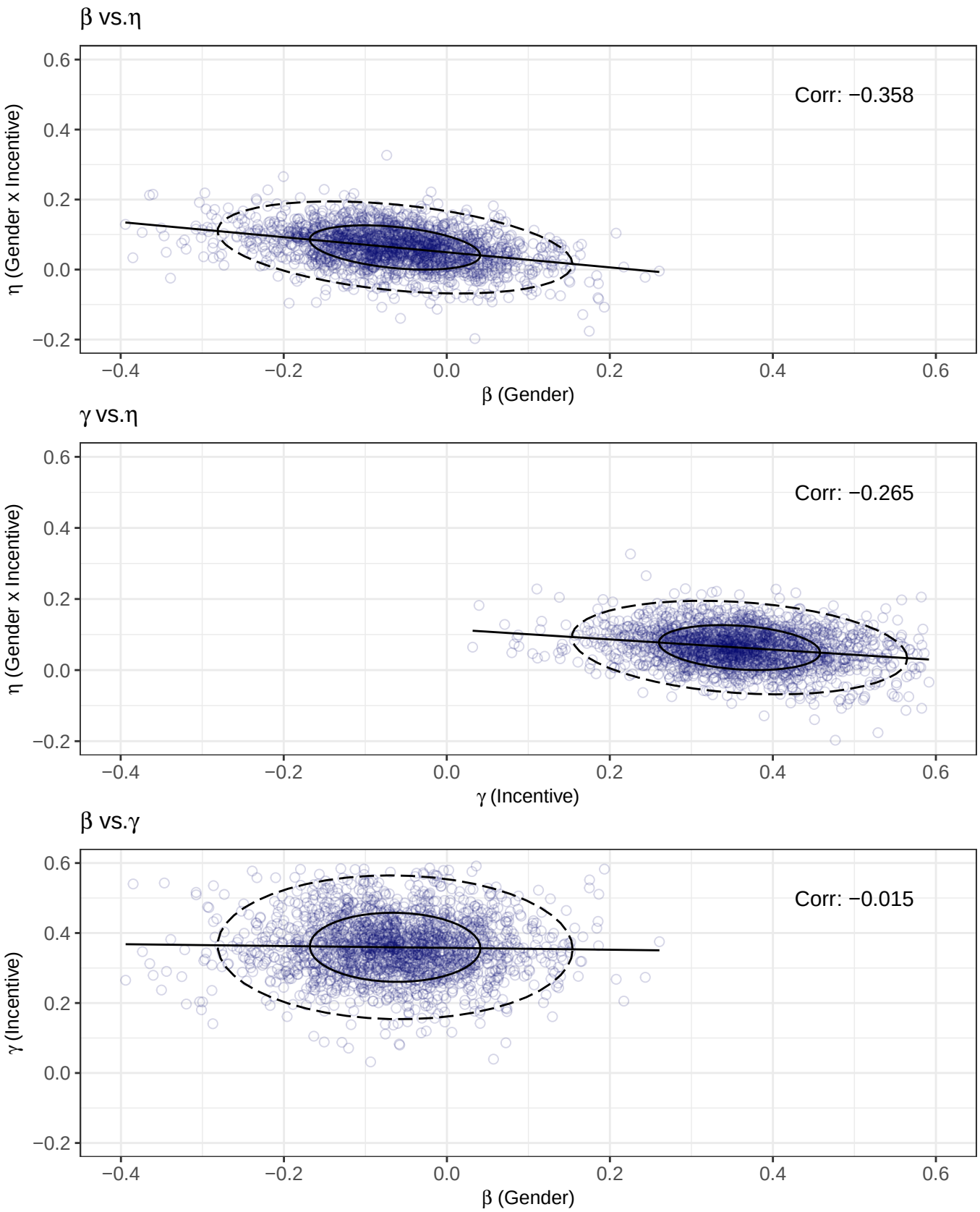
Zubanov, Nick. 2012. "Risk aversion and effort in an incentive pay scheme with multiplicative noise: theory and experimental evidence." *ERIM Report Series Reference No. ERS-2012-005-STR.*

Figure 1: Predictive Distribution by Gender



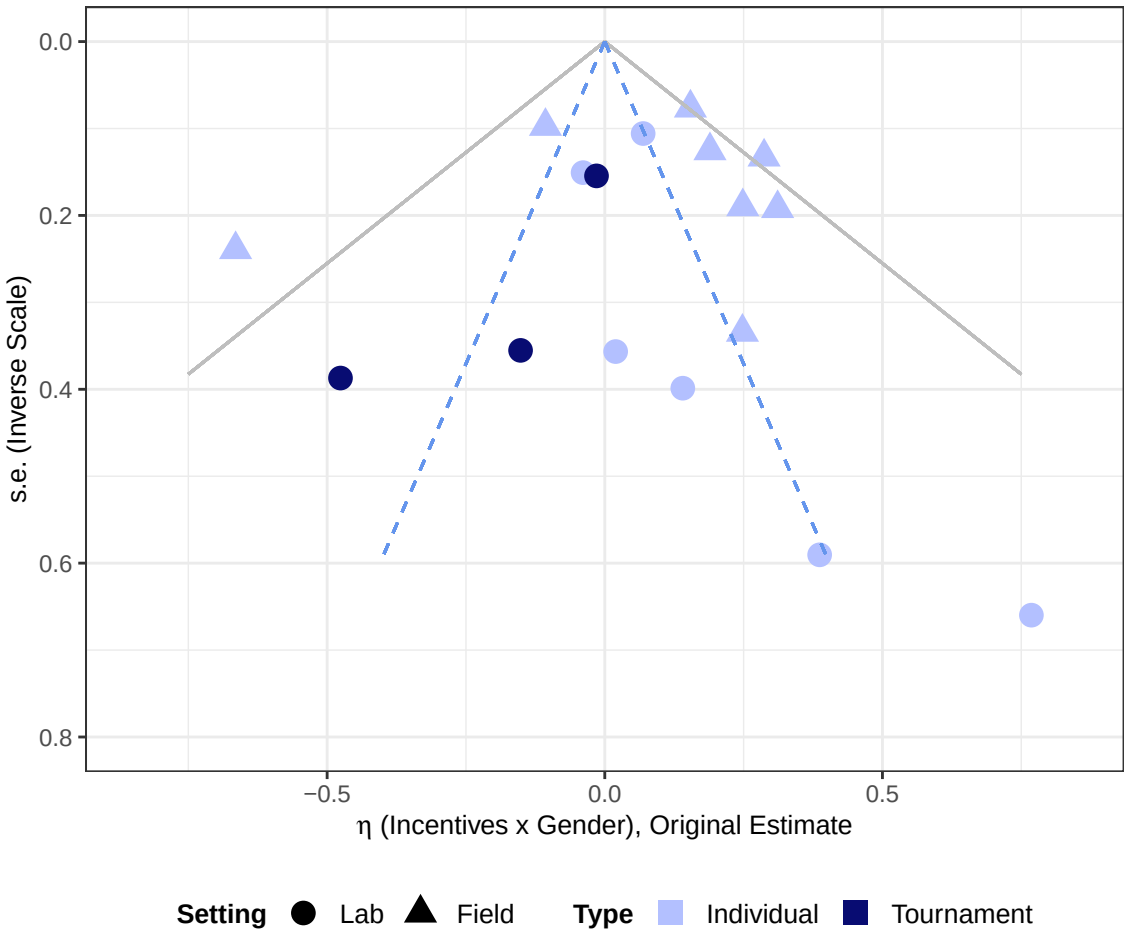
Notes: Outcome variable for each study is standardized using mean and standard deviation of men in control group. Vertical line indicates median estimate, box indicates 50% interval and line indicates 95% interval. Fixed effects model calculated using the metafor package for R (Viechtbauer, 2010). Bayesian Hierarchical model implemented in Rstan (Stan Development Team, 2020). See section III.A.1 for details.

Figure 2: Bivariate Correlations of Hyperparameters



Notes: Joint parameter estimates from simulated posterior distribution. Ellipses represent 50% and 95% intervals. Line displays linear best fit. See section III.B for details.

Figure 3: Incentive x Gender Effect and Study Type



Note: Lines represent 95% (solid) and 50% (dashed) intervals for the original estimates of the incentive−gender effect in the 17 included studies. See section III.C for details.

TABLE 1A — Summary of Included Studies

Study	Subjects	Treatment	Outcome Measure
Panel A: Field Experiments			
Angrist & Lavy (2009)	High school students	Treatment: increasing monetary bonuses (up to ca. $1500) for taking any matriculation test, passing any matriculation test and completing all matriculation requirements. Control: no bonuses.	Matriculation exam performance
Angrist et al (2009)	Undergraduate students	Treatment: monetary bonuses (scholarships) for meeting GPA targets. Higher bonus (up to $5000) for higher targets. Control: no bonuses.	1st year GPA
Ashraf et al (2012)	Hairstylists	Treatment: piece rate commission equal to 90% margin over retail price. Control: no commission or reward.	Number of packs of condoms sold
Bandiera et al (2005)	Fruit pickers	Treatment: constant piece rate. Control: piece rate decreases with average productivity of group of workers. (Rates are confidential)	Kilograms of fruit picked per hour
Fehr & Goette (2007)	Bicycle messengers	Treatment: 25% increase in commission rate. Control: standard commission rate.	Revenues per four week period
Hossain & List (2012)	Factory workers	Treatment: roughly 20% increase in pay for meeting productivity target. Control: fixed pay.	Log of units inspected per hour
Engström et al (2012)	Unemployed job seekers	Treatment: increased monitoring to check if job seeker applied for referred job, plus subjects are informed of this monitoring. Failure to apply for referred jobs can result in UI benefit sanctions, e.g. 25% benefit cut for 10 days. Control: increased monitoring but subjects are not informed.	Whether job seeker applies for referred jobs
Panel B: Lab Experiments			
Ariely et al (2009)	Lab subjects	Treatment: commission per unit of output, rate decreases with output ($0.01 for first 200 units, $0.005 for next 200, etc.). Control: no commission.	Number of key pairs pressed
Dickinson & Villeval (2008)	Lab subjects	Treatment: monitoring probability ranging from 0 to 1 at 0.1 intervals. If audited and output is low, pay is €0.27 in a round, else pay is €0.67. For normalization of variables only: control group defined by below-median monitoring probability.	Height of curve uncovered
Dohmen & Falk (2011)	Lab subjects	Treatment: commission decreases by €0.17 for every 2 seconds taken to solve problem. Control: no commission.	Negative of log time needed to solve multiplication problem
Pokorny (2008)	Lab subjects Lab subjects	Treatment: €0.50 (High Incentive), €0.05 (Low Incentive) or €0.01 (Very Low Incentive) per point score on top of show-up fee. Control: only show-up fee.	Score on IQ test Score on number counting task
Panel C: Lab and Field			
Boly (2011)	Lab subjects High school students	Treatment 1: monitoring at 1/20 rate, monetary penalties for mistakes follow flat schedule. Treatment 2: monitoring at 1/4 rate, monetary penalties for mistakes follow schedule that is up to twice as steep. Control: fixed wages.	Exam grading accuracy
Panel D: Lab Tournaments			
Carpenter et al (2010)	Lab subjects	Treatment: piece rate of $1 per unit of output plus $25 bonus for highest producer. Control: only piece rate of $1 per unit of output.	Quality adjusted envelopes produced
Freeman & Gelber (2010)	Lab subjects	Treatment: single prize ($30) or multiple prize ($15, $7, $5, $2, $1) tournament. Control: flat pay ($5).	Number of mazes solved
Gill & Prowse (2012)	Lab subjects	Treatment: continuous tournament prize variable ranging from £0.10 to £3.90 at £0.10 intervals, which we rescale to a unit scale. For normalization only: control group defined by below-median prize.	Number of sliders correctly placed

TABLE 1B — Summary of Included Studies

Study	N_{Men} N_{Women}	Raw		Normalized	
		Incentive Effect	Gender X Incentive Effect	Incentive Effect	Gender X Incentive Effect
Panel A: Field Experiments					
Angrist & Lavy (2009)	1,960	0.008	0.114	0.020	0.287
	1,861	(0.050)	(0.053)	(0.126)	(0.133)
Angrist et al (2009)	526	-0.026	0.172	-0.028	0.189
	729	(0.090)	(0.114)	(0.099)	(0.125)
Ashraf et al (2012)	187	-0.380	3.514	-0.027	0.249
	214	(2.358)	(2.685)	(0.167)	(0.190)
Bandiera et al (2005)	66	0.557	0.100	0.857	0.154
	76	(0.105)	(0.050)	(0.161)	(0.076)
Fehr & Goette (2007)	37	851.564	609.455	0.346	0.248
	5	(334.257)	(821.772)	(0.136)	(0.334)
Hossain & List (2012)	5	0.120	-0.089	0.896	-0.665
	16	(0.029)	(0.032)	(0.215)	(0.239)
Engström et al (2012)	752	0.068	-0.052	0.138	-0.107
	829	(0.035)	(0.048)	(0.071)	(0.097)
Panel B: Lab Experiments					
Ariely et al (2009)	41	222.050	55.478	0.562	0.140
	41	(118.610)	(157.465)	(0.300)	(0.399)
Dickinson & Villeval (2008)	45	10.929	4.842	0.873	0.387
	46	(2.947)	(7.388)	(0.235)	(0.590)
Dohmen & Falk (2011)	178	0.261	0.089	0.203	0.069
	181	(0.100)	(0.136)	(0.078)	(0.106)
Pokorny (2008)	56	-0.981	5.084	-0.148	0.768
	51	(3.022)	(4.366)	(0.457)	(0.66)
	64	-0.378	0.200	-0.037	0.020
	66	(2.793)	(3.641)	(0.274)	(0.357)
Panel C: Lab and Field					
Boly (2011)	87	1.38	-0.122	0.444	-0.039
	60	(0.342)	(0.469)	(0.110)	(0.151)
	176	1.208	1.051	0.358	0.311
	32	(0.327)	(0.648)	(0.097)	(0.192)
Panel D: Lab Tournaments					
Carpenter et al (2010)	51	2.539	-1.385	0.873	-0.476
	60	(0.682)	(1.126)	(0.234)	(0.387)
Freeman & Gelber (2010)	93	4.359	-0.657	1.006	-0.152
	141	(1.071)	(1.539)	(0.247)	(0.355)
Gill & Prowse (2012)	26	2.483	-0.112	0.329	-0.015
	33	(0.877)	(1.165)	(0.116)	(0.155)

Notes: Only Ashraf et al. (2012), was a working paper at time of inclusion. Coefficient estimates and standard errors of the incentive effect (γ) and gender-incentive effect (η) from regressions with outcome measures normalized by the mean and standard deviation of men in the control group (equation 4) in the "Normalized" columns, and non-normalized outcome measures as dependent variable (equation 3) in the "Raw" columns. See Section II.A for details and Appendix Table A2 for regression specifications.

TABLE 2 — Summary of Hyperparameter Estimates

			Quantiles				
	Mean	**S.E.**	**2.5%**	**25%**	**50%**	**75%**	**97.5%**
Gender x Incentives							
η (effect hyperparameter)	0.066	0.056	-0.050	0.032	0.068	0.102	0.173
τ_η (variance hyperparameter)	0.114	0.072	0.007	0.060	0.106	0.158	0.278
Incentives							
γ (effect hyperparameter)	0.357	0.086	0.188	0.302	0.356	0.412	0.532
τ_γ (variance hyperparameter)	0.305	0.079	0.179	0.249	0.295	0.350	0.490
Gender							
β (effect hyperparameter)	-0.062	0.089	-0.240	-0.119	-0.061	-0.005	0.113
τ_β (variance hyperparameter)	0.307	0.078	0.186	0.252	0.297	0.350	0.488

Note: Hyperparameter estimates from Bayesian hierarchical model based on empirical distribution from posterior simulations. See Section II.B for details.