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

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

November 2020

Abstract

Performance pay increases productivity but also earnings inequality. Can it contribute to the gender gap because women are less responsive? We provide answers by aggregating evidence from existing experiments on performance incentives with male and female subjects, regardless of whether they test for gender differences. 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 is 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

*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). 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 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 (**??**). At the same time, performance pay, a cornerstone of good management practices (**??**), has spread widely. **?**, for instance, show that the incidence of performance pay increased from 38% in the 1970s to 45% in the 1990s in the US, **?** document a rise from 16.3% in 1998 to 32% in 2004 in the UK, and **?** finds an increase from 15.4% in 1984 to 39.4% in 2009 in Germany.

To the extent that women are less responsive to performance pay, its increase in popularity might have contributed to the earnings gap. Indeed **?** document a gender gap in performance pay ranging from 14% in India, to 17% in the US, Germany and Canada, and 18% in the UK and The Netherlands. **?** find that piece rates and reward rates increase gender wage differentials in Finland, though bonuses decrease it, while **?** show that performance pay accounts for a small but potentially increasing share of the gender wage gap in the US.

This paper tests whether women are less responsive to high-powered performance 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, regardless of whether the studies themselves tested for gender differences - which most do not. The goal of this paper is to aggregate this evidence and assess whether a clear gender pattern emerges.

We use 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, do women and men respond differently to performance pay, 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 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. **?**). 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 agency theory predicts it drives the selection effect and because experiments on selection are rare.

Women may respond less to incentive pay for a number of cultural and psychological reasons. **?????** review this literature in detail, highlighting the lack of evidence on the impact of these differences on labor market outcomes. Women have been found to be more risk averse than men (**??**), less confident (**??**), more altruistic (**??**) and more averse to competition (**??**). 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 (**???**); subjects with low self-confidence **?** or in pro-social tasks **?**, provided financial incentives are low **?**. Furthermore, **?** 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..

[1] For instance, **?** show that selection into firms that pay lower wage premia explains 15% of the gender earnings gap in Portugal.

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 of the gender-incentive coefficient (η) has a mean that is 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.

Our paper thus shows that men and women respond equally positively to high-powered incentives. If there are differences in risk aversion or other behavioral parameters, these are not strong enough to generate systematic differences in the response to incentives. The absence of a gender difference in the incentive response suggests it is unlikely that incentives underlying performance pay in the workplace contribute to the gender earnings gap directly. As such, we contribute to the literature on gender earnings gaps (**??????**).

Our paper also contributes to a small but growing literature in economics which uses BHMs to distill a common lesson from studies in diverse contexts. Hsiang, Burke and Miguel (**??**) analyze the link between climate change and conflict; **?** examines generalizability across a wide range of impact evaluations; and Meager (**??**) looks at the impact of microcredit. Like **?**, we show that, in addition to aggregating answers to a given question, these methods can be used to ask new questions. 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.

The rest of this paper is organized as follows. Section 2 describes the study sample, Section 3 presents the methodology, and Section 4 the results. Section 5 concludes.

2 Study sample

The first step in building evidence from multiple studies is to establish inclusion criteria[2] 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.[3]

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

[2] Summarized in Appendix Table A1.

[3] A small number of experimental studies have looked at gender differences in the response to performance incentives since, with mixed results. **?** find a larger positive performance response to piece rates in men, **?** find a larger negative response to competitive pay in women, while **?** find no significant gender difference in the response to bonuses.

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.might generate different responses due to gender differences in competitiveness or cooperation, hence bringing in radically different mechanisms.

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[4]. For 15 of these, the 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.[5] Of the 15 papers, two report two experiments – **?** and **?**. These are included separately as they meet the inclusion criteria individually[6]. 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 **?** 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 uni-

[4] Appendix Table A3 lists these 29 papers.

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

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

versity 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 (4.2) and (4.3) below.

A few included studies collect data on some of the traits in which men and women are thought to differ, namely risk preferences (**???**) and social preferences (**??**). None of these papers evaluate whether such traits impact the effort response to incentives. **?**, 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; **?** mention that the interaction is not significant in the classical sense.

3 Methodology

3.1 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 $s \in \{1, \dots, S\}$:

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

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 (**???**). Furthermore, standardization should not affect our central test, whether the gender difference in the incentive response is zero. Nevertheless, we provide a robustness check with alternative standardization below.

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[7]. Appendix Table A2 details the in-

[7] Accordingly, to replicate the specifications in **???**, 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.

cluded specifications for each paper. As a robustness check, we also estimate a common specification for each study, excluding covariates[8].

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 -0.98 to 851.56, and from -0.15 to 1.01 only after 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.

3.2 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 (**?**). 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

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

[8] Results in Appendix Figure A7.

We use the following priors for the hyperparameters:

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

where N denotes a multi-variate normal distribution, Ω is a correlation matrix and σ is the vector of coefficient scales (**?**).The LKJ distribution (**?**) 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 (**??**). 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., **?** and **?**).

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 noninformative (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.[9],[10]

Our estimation of BHMs follows closely the procedures described in **?** and **?** (see Appendix A for details). 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 (4.2)).

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 (**?**). The Bayesian hierarchical model serves to make this belief-updating process transparent and precise.

[9] Reducing the variance of the prior on θ from $1000^2 \times I_3$ to $0.1^2 \times I_3$ 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.

[10] 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.

4 Results

4.1 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).[11] Given the available data and our specified (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.[12]

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 $\tau_\eta \approx 0$[13]. 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 assess-

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

[12] Appendix B compares BHM estimates with pooling model estimates.

[13] Full posterior distribution in Appendix Figure A1.

ing 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.

4.1.1 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%.

4.2 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.

4.3 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 (**?**) and altriusm (**???**). The literature on gender norms suggests a possible explanation; differences in behaviors might reflect norm-conforming behavior rather than innate traits (**???**). **???** 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 substanstial differences in the salience of norms or the exposure to risk, or that any differences are too small to bring about a noticable 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.

5 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 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. **?**, 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 (**??**), or a greater preference for flexible work hours which may intersect with household composition (**???**) for example. Furthermore, men and women may optimally negotiate different compensation contracts in the labor market if they differ on behavioral traits (**?**). 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 (**?????**). Moreover, like (**?**), 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.

References

Online Appendix

A Estimation

Our estimation of the Bayesian hierarchical models follows closely the procedures described in **?** and **?**. For clarity of exposition, we describe the univariate model here, which extends immediately to the full multivariate model. Following (3) 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:[14]

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

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 (**?**) with path lengths set adaptively using the no-U-turn sampler (NUTS; **?**). 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.

[14] 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 **?** for details.

We then let each chain run for 14,500 iterations, discarding the first 2,000 simulations as warm-up. These parallel chains are then tested for mixing—the between-chain and within-chain variances should be equal—and stationarity. After confirming that the chains are well behaved, we combine them to generate the simulated posterior distributions for both the hyperparameters, η and τ^2, as well as the true study-level effects, $\{\eta_i\}_{i=1}^{S}$.

B Comparison with pooling model

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

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

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

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

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

B.1 Pooling model results

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

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

C Pooling Metrics

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

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

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

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

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

where $\lambda_s \in [0,1]$ can be thought of as a pooling factor that represents the degree to which the estimates are pooled towards the estimated population mean (η) rather than based on their observed value.[15] 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 $\tau^2 = 0$.

? show that in the single parameter model when η and τ^2 are known, equation (7) characterizes the analytical mean of $\hat{\eta}_s$ with $\lambda_s = \frac{\sigma_i^2}{\sigma_i^2 + \tau^2}$. This suggests two alternative study-level pooling statistics: $\lambda_s^1 = \frac{\hat{\sigma}_i^2}{\hat{\sigma}_i^2 + \hat{\tau}^2}$, that is, the variance pooling metric calculated from the posterior means of the error terms, and $\lambda_s^2 = \frac{\hat{\eta}_k^{POST} - \hat{\eta}_k}{\eta - \hat{\eta}_k}$, a shrinkage metric that directly measures the extent to which the posterior mean of the study-level effect is determined by the posterior mean of the population effect. Note that in the multivariate model, λ_s^2 is not restricted to the interval $[0,1]$. Correlation with other parameters makes it possible that the true effect in a study is outside the interval between the observed effect and the population mean.[16]

? generalize this idea to develop a common pooling factor that 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

[15] 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 **?** and focus on the degree of pooling.

[16] 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.

our two-level model, they define the pooling factor as

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

where E represents the posterior mean, V is the finite sample variance operator (i.e., $V_{i=1}^{n} = \frac{1}{n-1}\sum(x_i - \bar{x})^2$), and $\varepsilon_s = \eta_s - \eta$. 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.

C.1 Estimates

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

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

This is perhaps not surprising. The studies in our sample exhibit tremendous variation in both the type of task and the form of incentives. What is, however, surprising is that men and women respond similarly to financial workplace incentives across such a diverse set of contexts.

D Posteriors

The Bayesian hierarchical model provides a precise and transparent method to incorporate data from other studies into our beliefs regarding the true effect in a particular setting. As noted in the main text, the best (i.e., lowest mean squared error) estimate for the true effect in a particular context is typically not equal to the mean estimate of a single, internally valid study in that context. Figures A2, A3, and A4 compare the posterior predicted distributions for each of the main parameters, 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 (γ) in Bandiera et al. (2005). As shown in Figure A3, the parameter estimate in this study is large, $+0.86\sigma$, with a standard error of 0.16σ. However, with a 95%-credible interval spanning $[0.55, 1.17]$, there remains quite a bit of uncertainty about the magnitude of the effect. Furthermore, the estimates are substantially larger than the mean in all but four other studies. The mean of the posterior distribution for γ_s is $+0.74\sigma$, still a very large effect but pulled substantially towards the population mean of $+0.36\sigma$. The degree of pooling depends primarily on the uncertainty of the local parameter estimate and the estimated distribution of the population hyperparameter (γ, τ_γ).

Figure A5 demonstrates the relationship between the estimated standard deviation of the hyperparameter (τ_η) and the posterior mean of η_S, the study-specific effect. Here, we return to the gender-incentive interaction term, our primary outcome of interest. The upper half of the figure plots the posterior distribution of η_s for each study conditional on τ_η. If τ_η were 0, each study would be estimating a common effect and the posterior for each η_s would be equal to our posterior estimate of the population mean. As τ_η increases, the extent to which the posterior for any study is pooled toward the population mean diminishes, and as $\tau_\eta \rightarrow \infty$ the posterior for each study tends towards the site-specific estimate.

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

E Model Checking

After computing the posterior distribution of all parameters, 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 (**?**). 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 A6 plots the observed order statistic for each of the model parameters against the mean from the simulated posterior distribution[17]. In the case of the gender-incentive interaction term, the posterior predictive distribution matches the observed data very well, including at the extremes. Although the settings for the included studies were certainly not chosen at random from the population of possible study sites, our hierarchical model that treats the study-level parameters as if they were normally distributed around a population mean does a remarkably good job of capturing important features of the data. The model also performs reasonably well for the gender (β) and incentive (γ) parameters, with the exception of slightly fatter tails in the distribution of γ.

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