Economica

Economica (2022) **89**, S38–S70
doi:10.1111/ecca.12407

Men are from Mars, and Women Too: A Bayesian Meta-analysis of Overconfidence Experiments

By Oriana Bandiera*, Nidhi Parekh†, Barbara Petrongolo‡ and Michelle Rao*

**LSE* †*J-PAL Africa* ‡*Oxford University*

Final version received 9 December 2021.

Gender differences in self-confidence could explain women's under-representation in high-income occupations and glass-ceiling effects. We draw lessons from the economic literature via a survey of experts and a Bayesian hierarchical model that aggregates experimental findings over the last 20 years. The experts' survey indicates beliefs that men are overconfident and women underconfident. Yet the literature reveals that both men and women are typically overconfident. Moreover, the model cannot reject the hypothesis that gender differences in self-confidence are equal to zero. In addition, the estimated pooling factor is low, implying that each study contains little information over a common phenomenon. The discordance can be reconciled if the experts overestimate the pooling factor or have priors that are biased and precise.

I. Introduction

Gender inequality in the labour market is rife. Women make systematically different education choices from men, are under-represented in high-earning careers, and bear the bulk of the earning penalty associated with parenthood (Kleven *et al*. 2019).

There are two, fundamentally different, explanations for this difference. The first is that men and women are equal in all relevant dimensions but face different opportunities or constraints. In this case, gender inequality can be a symptom of misallocation, and policies that promote gender equality can increase efficiency. The second is that men and women have different psychological traits that drive educational choices and labour market outcomes. In this case, gender inequality in labour outcomes is a manifestation of gender differences in traits. We contribute to this debate by aggregating the evidence on gender differences in traits, with a particular focus on overconfidence.

Several laboratory, and more recently field, experiments (surveyed, among others, by Croson and Gneezy 2009; Bertrand 2011, 2018; Azmat and Petrongolo 2014) have investigated gender differences in personality traits. Those that matter for labour outcomes can be grouped in three broad areas: attitudes towards risk, social preferences and confidence. These shape decisions at every stage of a person's career, from years of schooling to major choices, from job applications to the choice of sector and firms, and, once at work, on pay and promotions.[1] Knowing whether there are systematic gender differences in these traits is key to interpreting differences in outcomes, but drawing definitive conclusions from the large body of experiments is limited by the fact that due to differences in settings, stakes and design, findings cannot be aggregated easily.

This is what we attempt to do: aggregate findings. We begin by surveying experts—academic economists—who are the main 'consumers' of this literature, and draw lessons from it. Our sampling frame is the universe of research fellows and affiliates of the Centre for Economic Policy Research (CEPR)—1300 economists based mostly in Europe and the USA. Our sample of respondents (342, a 26% response rate) are asked to score men and women on risk aversion, confidence and altruism on a 0–100 scale, based on their

reading of the literature. Men come out more confident, risk-loving and selfish. The mode of the gender gap is, however, close to zero for risk and altruism, as experts who rank men highly also rank women highly on the same dimension. In contrast, the modal gap for overconfidence is positive, and the within-person correlation is negative, as experts who rank men highly overconfident typically rank women as very underconfident. On average, men are rated overconfident and women underconfident.

In the second part of the paper we assemble a dataset comprising all the experimental tests of confidence published in the last 20 years, and estimate a Bayesian hierarchical model (BHM) to aggregate the findings. The model tackles the main challenge of aggregating evidence from different settings and experimental designs head on by estimating, together with the gap, a 'pooling factor' that measures the extent to which each result is informative about a common phenomenon versus its own context-specific effects.

Why confidence? Self-confidence is important for understanding selection into certain education tracks, occupations and careers, as well as *ex post* payoffs to these choices, which depend on group composition whenever one's performance is remunerated against the performance of peers. When remuneration has a zero-sum, it pays to be realistic about one's chances of success. In other words, underconfident individuals compete too little, and overconfident individuals compete too much, relative to the choices that would maximize their expected payoffs. But in real-life circumstances, remuneration rarely has zero-sum, as the expected return from several competitive settings—such as the expected outcome of a job application or a promotion—is positive. This is because the worst-case scenario is typically one's status quo. By shying away from such situations, underconfident individuals forego positive chances of success. In addition, they may also miss out on feedback and experience that could be gained from participating, regardless of outcomes. Hence it is important to establish whether some groups are likely to be held back in the labour market by a tendency to underestimate their performance, whether absolute or relative to peers. For example, in the study by Niederle and Vesterlund (2007), men and women overstate own performance in a simple arithmetic task at which they are on average equally good, but men tend to be more overconfident about their performance than women. This gap in overconfidence explains part of the gender difference in the choice of tournament over piece-rate compensation for the same task.

Our sample includes papers that appear between 2000 to 2020 in peer-reviewed economic journals or widely circulating working paper series, that provide a measure of confidence for men and women on a real task, whether in laboratory or field settings. We identify 38 papers that satisfy these criteria, providing a pool of 90 paired results. In stark contrast to the experts' assessment, 72% of the estimates indicate that both genders are overconfident, while only 18% are consistent with the shared belief that men are overconfident and women underconfident.

Our Bayesian analysis can be performed on a subsample of 39 experiments that provide standard errors for their adopted overconfidence metric. Our main results deliver a difference in the (standardized) overconfidence of men and women of 0.094, with standard error 0.054. Overall, men are more confident than women, but we cannot reject that this difference is zero at the 5% level. This is in line with the findings of BHM estimates of gender gaps in altruism (Rao 2020) and response to incentives (Bandiera *et al*. 2021). In contrast to these studies, however, the pooling factor is low and the relatively large posterior variance implies that each individual study is poorly informative about a common phenomenon. Based on a standard pooling metric used in the Bayesian hierarchical literature (Rubin 1981; Gelman *et al*. 2013), we estimate that only 23% of

variation across studies is sampling variation, with more than three-quarters of the variation being driven by genuine differences across studies.[2] This implies that if we were to run a new experiment, we could not be reasonably confident that the resulting gender difference in overconfidence would be close to the posterior mean in the original sample. A naive meta-analysis that ignores this heterogeneity and assumes that each study contributes to identify a common effect yields an estimate of 0.114 (s.e. 0.013).

The full-pooling model can (qualitatively) match the profession's beliefs about gender gaps in overconfidence, but it cannot explain why the majority of economists believe that women are underconfident.

We consider two further explanations within a general BHM with partial pooling.

First, we acknowledge that different results may achieve varying degrees of visibility in the profession, implying different intensities of belief updating. We hypothesize that the salience of various results is proportional to the citations that they receive, which are in turn related to how long they have been in the public domain, the prestige of the outlet where they are published, and the authors' prominence. We estimate a modified BHM in which the precision of the study estimates is adjusted by the citations received, and obtain posterior estimates that are very close to those from the original model, with a posterior mean gap in confidence of 0.080, and standard error 0.068. The conclusion is that results delivering a larger gender gap in confidence do not systematically obtain more cites, hence the distribution of citations across papers may not explain the observed beliefs among economists.

Second, we explore the role of prior beliefs in shaping the updating process. Our BHM assumes a normally distributed prior for the average gender difference in overconfidence, with mean zero and unit variance. This reflects that we are *a priori* agnostic about which gender is more overconfident, but we allow for substantial variation around such a zero mean. We next consider a biased prior (e.g. one that is commensurate with gender differences in overconfidence observed in survey responses), while keeping a unit variance around it, and obtain nearly identical posterior estimates to those from the original BHM with unbiased priors. Only when we significantly increase the precision of the biased prior do we obtain a positive and precise posterior estimate of the gender gap in overconfidence. In other words, when priors are very precise, they are hardly updated when new information is received. If the prior is biased and very precise, then so will be the posterior.

In summary, we conjecture that biased beliefs in the profession could stem from extreme priors or lack of Bayesian updating. The full-pooling model described above is one special case of updating, which delivers much more precise results than the general BHM, but is not supported by the data. Other cases could involve selective or non-probabilistic updating, such that individuals interpret information to confirm what they believe in the first place, or form beliefs that are deterministic functions of their information sets (see, for example, Jackson *et al*. 2021). Lack of Bayesian updating could be rationalized in terms of its cognitive costs, whereby simplifying the updating process is a way to save on cognitive effort. As a consequence, gender stereotypes may arise as generalizations that help individuals to economize on cognitive resources when forming beliefs about the characteristics of groups members.[3]

Our findings are in line with the social psychology literature that benchmarks between-gender differences in traits to within-gender differences. The idea is to not just focus on the difference in means between men's and women's characteristics, but also take into account the overlap between the respective distributions. The findings in this literature indicate that for most relevant traits related to cognitive and non-cognitive

abilities—including self-confidence—within-gender differences are much larger than between-gender differences, so there is a substantial overlap in the gender-specific distributions (Hyde 2005, 2014; Bertrand 2020).

The rest of the paper is organized as follows. Section I describes the survey of economists' beliefs. Section describes the data, including the process of study selection and the Bayesian sample. Section III discusses the empirical approach. The results are described in Section IV. Section V highlights the knowledge gap between experts' beliefs and the evidence, and considers model extensions that could explain the differences between our meta-analysis and survey results. Section VI concludes.

II. Experts' Survey

In the autumn of 2019, we surveyed the universe of Research Fellows and Affiliates of CEPR, a leading research network in Europe. The survey was sent by email (see Figures A1 and A2 in the Appendix for screenshots of the full survey) to 1300 economists, and the response rate was 26%; that is, 342 experts responded. Respondents were asked to rate men and women on three traits—confidence, risk attitudes and altruism—*based on their reading of the literature*.

Ratings are given on a scale of 0 to 100. We set 50 as the neutral point in the risk question, realistic for the confidence question, and care equally about self and others for the altruism question. Of the 342 researchers who responded, 64% were male and 57% were full professors; Applied Micro is the modal field, accounting for 32% of the observations. Table A1 in the Appendix reports a full breakdown of the respondents' fields of specialization.

Figure 1 shows the mean answers as well as the gender gaps (men minus women) on all three measures. The figure shows positive gaps on all three dimensions—that is, men are more confident, more risk-loving, more selfish—but the confidence gap is much larger, as men are rated twice as confident as women, and the average expert rates women as underconfident (35 out of 100) and men as overconfident (69 out of 100), with 77% of the surveyed population rating women as underconfident and men overconfident.

Figure 2 shows the scatterplots of scores of men and women on the three dimensions. Each circle represents the grades given by one person. We find a positive correlation for both altruism and risk; that is, experts who score men high also score women high. In contrast, the correlation for confidence is negative; that is, experts who score men high also score women low. This partly explains why the gap is so much larger.

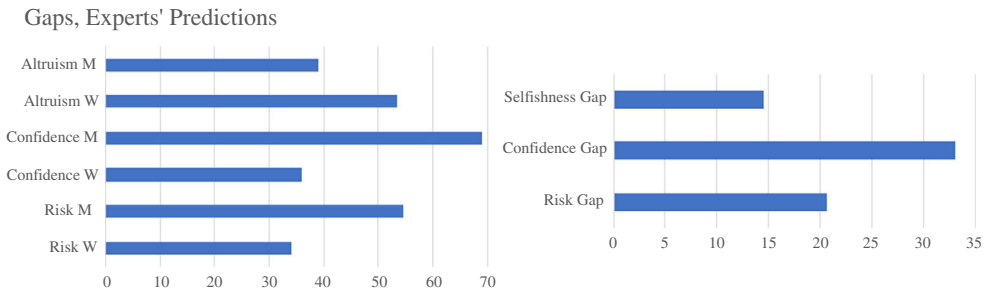

FIGURE 1. Experts' answers: means.
Notes: The panel on the left reports the mean of experts' answers on altruism, self-confidence and risk-taking. The scale used is described in Figure A1. The panel on the right reports mean gender gaps. $N = 342$.

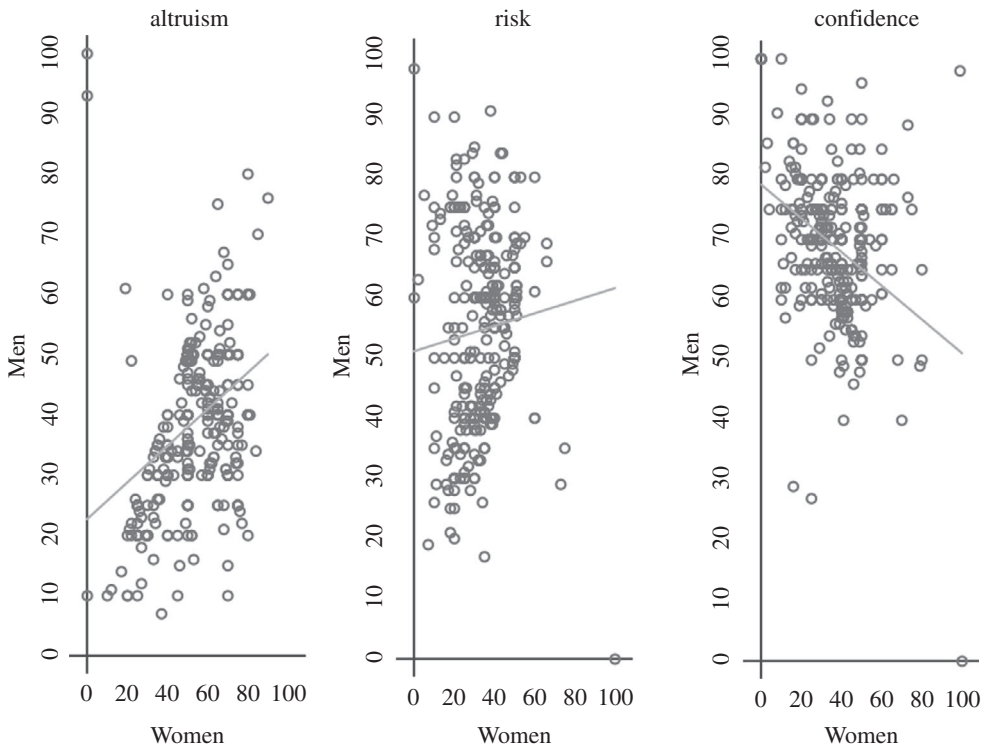

FIGURE 2. Experts' answers: correlations.
Notes: The graphs plot answers about men against answers about women for each respondent. $N = 342$.

Figure 3 shows the distributions of the three gaps. The confidence gap stands out both because there is much more variation across experts and because the mode is positive, while it is zero for the other two. Figures A3, A4 and A5 in the Appendix show scatterplots of experts' responses on the confidence question by gender, field of specialization and seniority, and reveal no evidence of systematic differences in responses along any of these dimensions.

Interestingly, respondents who report that men and women are very similar on altruism (within 10 points of one another) estimate that men are more confident than women by 23 points. And respondents who report that men and women are very similar on risk estimate that men are more confident than women by 22 points. Thus even respondents who believe that men and women are fairly similar on other traits—a belief in line with the meta-analyses of Bandiera *et al.* (2021) and Rao (2020)—believe that they differ in overconfidence. The next section will show how findings from the experimental literature compare to economists' beliefs about gender differences in overconfidence.

III. DATA

Paper selection and summary evidence

We select papers that appeared in the public domain during the past two decades, a very active period for the literature on gender differences in psychological traits. It is important to note that selection into this sample biases our estimate against the null of zero effects as publication is on average biased in favour of significant results (Kasy

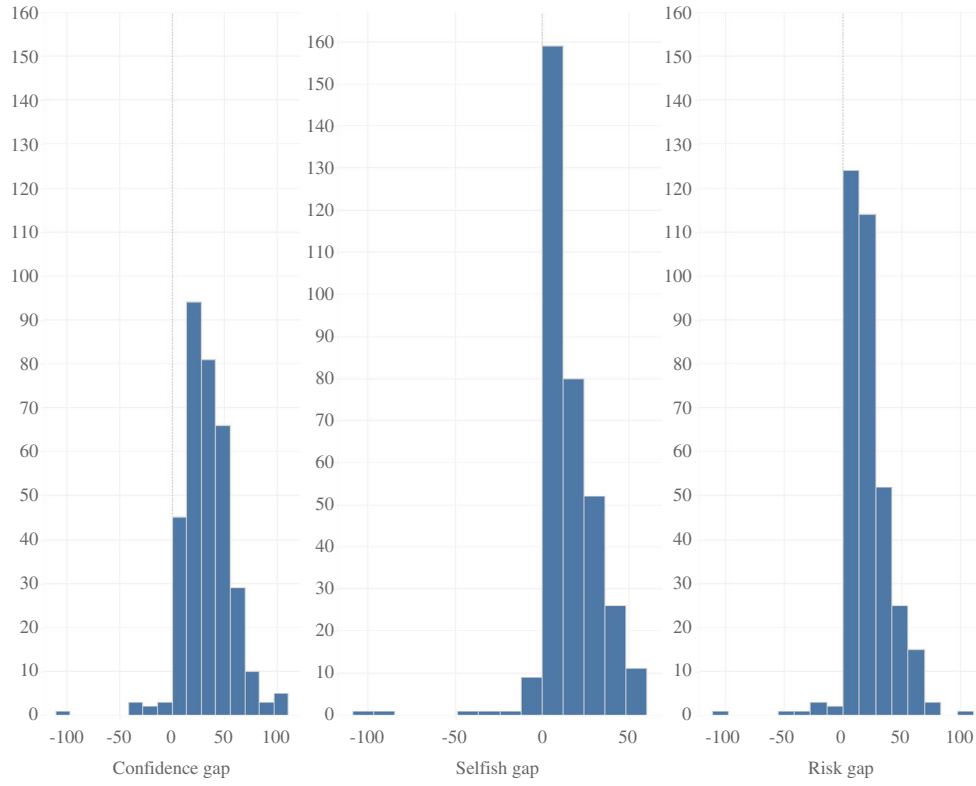

FIGURE 3. Experts' answers: distributions.
Notes: The histograms display the distributions of answers about gender gaps. $N = 342$.

2021). By searching Google Scholar and RePEC online repositories for papers with keywords 'confidence' and 'gender', we identified 474 such papers that appeared during 2000–20. We next selected those published in the top 100 economic journals (according to RePEC), or in the NBER, CEPR or IZA Working Paper series, yielding 140 papers. We finally checked each of these papers for relevant information on confidence by gender, selecting papers in which confidence was measured in relation to actual performance in a specific task. Specifically, we scouted for information on either of the following.

- The difference between self-assessed performance score or rank (according to the specific study setting), and the respective actual performance score or rank, by gender.[4] Based on this difference, we obtain the average degree (or intensity) of overconfidence by gender for each study. This is also referred to as the 'intensive margin' measure.[5]
- The difference in the share of men and women who are overconfident or underconfident. This is referred to as the 'extensive margin' measure. It is calculated based on one of the following two measures:

  – the share of men and women who overstate or understate their performance score or rank;[6]
  – the share of men and women who self-select into a tournament, believing that they will win the tournament, but do not, or the share of men and women who do not self-

select into a tournament, believing that they will not win, but would have won based on their performance.[7]

Our working sample consists of 90 studies, that is, paired observations on self-confidence for men and women, from 38 papers that meet the criteria laid out previously. The list of papers is provided in Table A2 in the Appendix. 71 observations are obtained in the laboratory, 9 in the field, and 10 from combinations of laboratory and field experiments. Most experiments are based on a student subject population in a high-income country.

Based on this sample, we build a dataset containing relevant measures of overconfidence by gender with the associated metric of statistical significance (whenever available) for each experiment included in the papers, as well as information on authors, publication outlet and impact factor. For working papers, the journal impact factor is imputed by assuming that the paper will eventually be published in the journal where the most cited author is most published.

Bayesian analysis sample

To aggregate evidence across studies using Bayesian hierarchical methods, we need estimates of the standard errors for each result included in the analysis. We therefore further select results for which standard errors are reported (or could be obtained from reported *p*-values or *t*-statistics) for the gender gap in overconfidence. 39 studies (from 16 papers) meet these criteria. Among these, 24 studies also report standard errors separately for overconfidence of men and women.

Figure 4 compares the distribution of the raw results across the different samples. The first bar refers to the whole sample of papers that provide paired observations on self-confidence for men and women; the second bar refers to the subsample that also provides significance levels for the gender difference in overconfidence; and the third bar refers to the subsample that provides significance levels separately for men and women. Irrespective of selection criteria, the vast majority of studies find that both men and women are overconfident, while only a minority find that men are overconfident and women are underconfident. The subsamples on which we perform the Bayesian analysis are therefore representative of the larger population of papers that measure gender differences in overconfidence.

Further details on the analysis samples are reported in Table 1. 26 out of 39 studies provide measures of the degree of overconfidence among men and women (the intensive margin sample). Of these, 17 studies report standard errors separately for men and women. Men and women on average overestimate their score (or underestimate their rank) by 4 and 2.7 percentage points, respectively, and the average gender gap is 2.9 percentage points. The remaining 13 studies report only shares of overconfident men and women (the extensive margin sample). Of these, only 7 report standard errors separately for men and women. The data reported imply that 52.2% of men and 46.2% of women overestimate their performance, and on average the share of men overestimating their performance exceeds the female share by 8.5 percentage points. Measures of overconfidence from the two subsamples can be combined in a standardized measure, given by the specific metric, divided by the within-sample standard deviation (Cohen 2013). Men and women overestimate their ability relative to their performance by 0.421 and 0.323 standard deviations, respectively; and men overestimate their ability by 0.115 standard deviations more than women.

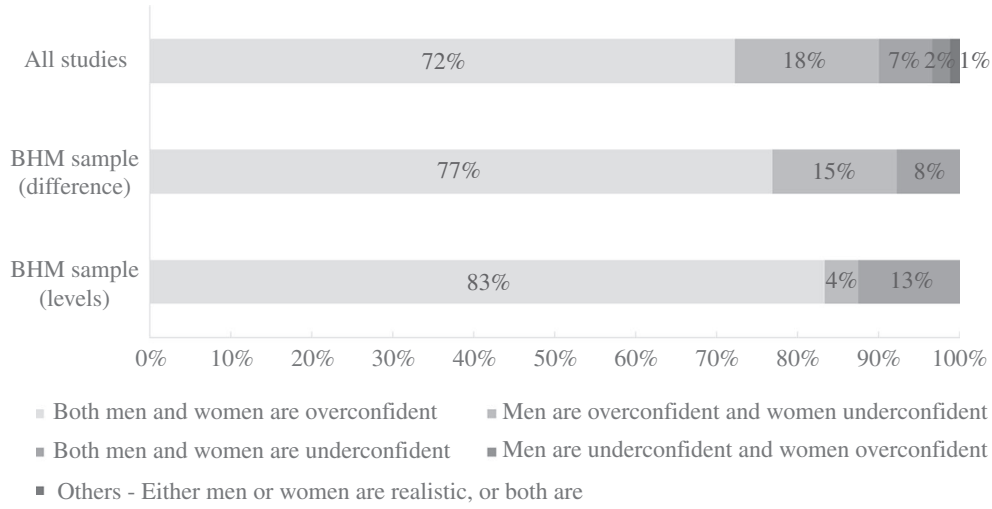

FIGURE 4. Distribution of results on self-confidence.
Notes: This figure compares the distribution of raw (non-Bayesian aggregated) results across the three samples of the literature on confidence. All studies ($N = 90$) include results with paired observations on confidence for men and women that meet the criteria laid out in Section 2. 'BHM sample (difference)' ($N = 39$) refers to the subsample of results that report standard errors for the gender gap in overconfidence. 'BHM sample (levels)' ($N = 24$) refers to the subsample of results that report standard errors separately for overconfidence of men and women.

TABLE 1
SUMMARY EVIDENCE ON OVERCONFIDENCE

	N	Mean	S.D.
Intensive margin sample			
Overconfidence, men	17	0.040	0.073
Overconfidence, women	17	0.027	0.076
Difference (men − women)	26	0.029	0.067
Extensive margin sample			
Overconfidence, men	7	0.522	0.186
Overconfidence, women	7	0.462	0.213
Difference (men − women)	13	0.085	0.125
Full sample			
Overconfidence, men	24	0.421	0.412
Overconfidence, women	24	0.323	0.372
Difference (men − women)	39	0.115	0.245

Notes
The sample includes 39 studies that report standard errors for the overconfidence metric adopted. The intensive margin subsample includes studies that report the shares of men and women who overstate (understate) their performance score (rank). The extensive margin subsample includes studies that report the shares of men and women who are overconfident, based on the shares of men and women who overstate (understate) their performance score (rank) or the shares of men and women who self-select into a tournament, believing that they will win the tournament, but do not, or the share of men and women who do not self-select into a tournament, believing that they will not win, but would have won based on their performance.

In the empirical analysis that follows, we will provide posterior overconfidence estimates for each subsample, as well as the full sample, for men and women separately and for the corresponding gender gap.

IV. Empirical Approach

By combining information from several data sources that are potentially interrelated, meta-analysis naturally lends itself to hierarchical modelling. The Bayesian hierarchical model (BHM) provides a versatile framework to aggregate findings from comparable studies and disentangle genuine variation across studies, resulting from cross-study differences in the respective empirical contexts, from sampling variation in the study-level estimates.

Consider S studies, with associated estimates for the parameter of interest $\hat{\beta}_s$, $s = 1, \ldots, S$, which may denote, for example, the estimated degree of overconfidence for either gender, or the gap in overconfidence between genders. The difference between each study-level estimate $\hat{\beta}_s$ and the population mean β can be decomposed into two components. The first component, $\hat{\beta}_s - \beta_s$, represents the difference between study-specific estimates and the respective true values, and reflects sampling (i.e. idiosyncratic) variation, as well as potential biases. The second component, $\beta_s - \beta$, represents the difference between study-specific values and the population value, stemming from systematic differences in the subject population, treatment or outcomes studied, among other factors.

The above decomposition has two extreme cases. At one extreme, each study identifies a common population effect ($\beta_s = \beta$), and variation across studies is purely idiosyncratic. This is known as the full-pooling (or fixed-effect) model, and has the form

$$\hat{\beta}_s \sim N(\beta,\ \sigma^2), \qquad s = 1,\ \ldots,\ S.$$

The estimate of the population mean β is given by the precision-weighted average of the study-level effects:

$$\hat{\beta}^{\text{pool}} = \frac{\sum \hat{\beta}_s / \hat{\sigma}_s^2}{\sum 1/\hat{\sigma}_s^2},$$

where $\hat{\sigma}_s^2$ denotes the variance of each study-level estimate.

Alternatively, in the random effects (RE) model, each study-level estimate $\hat{\beta}_s$ identifies its own study-specific effect β_s, and the study-specific effects are in turn distributed around the population mean β:

$$\hat{\beta}_s \sim N(\beta_s,\ \sigma_s^2), \qquad \beta_s \sim N(\beta,\ \sigma^2), \qquad s = 1,\ \ldots,\ S. \tag{1}$$

The estimate of the population parameter β is again a weighted average of the study estimates, in which weights now factor in the individual study variances $\hat{\sigma}_s^2$ as well as the between-study variance $\hat{\sigma}^2$:

$$\hat{\beta}^{\text{RE}} = \frac{\sum \hat{\beta}_s / (\hat{\sigma}_s^2 + \hat{\sigma}^2)}{\sum 1/(\hat{\sigma}_s^2 + \hat{\sigma}^2)}.$$

In the estimation of the population effect, the random effects model reduces the precision on all estimates, and relatively more so for more precisely estimated parameters.

The BHM lies between the two extremes. Its formulation resembles that of the random effects model in (1), but—unlike the random effects model—it treats the 'hyperparameters' β and σ^2 as random variables to be estimated:

$$\hat{\beta}_s \sim N(\beta_s,\ \sigma_s^2), \tag{2}$$

$$\beta_s \sim N(\beta,\ \sigma^2), \quad s = 1,\ \ldots,\ S, \tag{3}$$

$$\beta \sim N(-,\ -), \tag{4}$$

$$\sigma^2 \sim N(-,\ -), \tag{5}$$

where (−,−) indicates a prior distribution that needs to be specified. The clear advantage of the BHM is that estimation of σ^2 effectively allows for varying degrees of pooling, where $\sigma^2 = 0$ corresponds to full pooling and $\sigma^2 \to \infty$ corresponds to no pooling.

In the BHM (2)–(5), we are making a few assumptions. First, condition (2) assumes normality of the study effects $\hat{\beta}_s$, which follows from the assumption of internal validity of study-level estimates and the fact that the respective sample sizes are sufficiently large that the central limit theorem can be invoked. Condition (3) assumes that the study-level effects are distributed normally around the population mean β. While there is no obvious justification for this assumption, McCulloch and Neuhaus (2011) provide reasonable conditions under which inference on β and σ^2 under the normality assumption is reliable even when the underlying distribution is not normal.

One key assumption in the model above is exchangeability, imposing that the joint distribution of $(\beta_1,\ \ldots,\ \beta_S)$ is invariant to permutations of the indices $1, \ldots, S$, allowing us to write the joint distribution of the β_s as i.i.d. The interpretation of the exchangeability assumption is that studies should be indistinguishable from each other, except for the estimate that they provide, such that, for example, there is no reason *ex ante* to believe that the estimate from study 1 should be closer to the estimate from study 2 than to the estimate of study 3. This assumption is likely to be violated whenever there are study characteristics that would naturally make some studies more similar to one another than to other studies in the sample. This is clearly the case when multiple estimates are provided within the same paper and are plausibly subject to experimenter effects (Rosenthal 1976).

To address this potential violation of the exchangeability assumption, we introduce an additional layer in the hierarchical model. Our estimation procedure has two steps. In the first step, for each multi-study paper, we estimate a BHM to aggregate information from multiple estimates $k = 1, \ldots, K$ within each paper s:

$$\hat{\beta}_{ks} \sim N(\beta_{ks},\ \sigma_{ks}^2), \quad \beta_{ks} \sim N(\beta_s,\ \sigma_s^2), \quad k = 1,\ \ldots,\ K.$$

In the second step, we use posterior means $\hat{\beta}_s$ and $\hat{\sigma}_s^2$ for multi-study papers, as well as the original estimates from single-study papers, as inputs to the model in (2)–(5). In the resulting two-step framework, the assumption of exchangeability is imposed within each step.

Finally, as in all Bayesian models, we need to specify a prior distribution on the hyperparameters. In the context of the two-step model, we specify prior distributions for $\hat{\beta}_s$ and $\hat{\sigma}_s^2$ in the first step, and for β and σ^2 in the second step. We assume the following prior distributions:

$$\beta \sim N(0,\ 1), \qquad \sigma \sim N(0,\ 1), \qquad \beta_s \sim N(0,\ 0.2^2), \qquad \sigma_s \sim N(0,\ 0.2^2).$$

In all four cases we choose weakly informative priors, which means that we prefer for our posterior distribution (and hence inference) to be driven by information from the data rather than any prior beliefs on overconfidence. The posterior distribution, a function of the prior and the likelihood, is a probability distribution on our parameters of interest β and σ. In particular, for our second-stage priors, we assume that priors for β and σ^2 are normally distributed with zero mean and unit variance. This reflects the fact that absent seeing the data: (i) we have no reason to expect men or women to be fully realistic, or overestimate or underestimate their ability relative to their performance by greater than 1 standard deviation; (ii) we have no reason to expect men to be more overconfident than women, or to be more overconfident than women by greater than 1 standard deviation, or vice versa. The assumption of a zero mean is also consistent with a standard frequentist approach to hypothesis testing, in which the null hypothesis is zero.

Similarly, we assume that our first-stage priors β_s and σ_s are normally distributed with zero-mean and standard deviation 0.2. Given that we have a smaller number of experiments within each study,[8] we need more precise priors to regularize the estimates and to prevent over-fitting. This choice may also be justified by noting that there should be smaller heterogeneity in experiments within the same paper than across papers. The prior standard deviation 0.2 is similar to the mean standard deviation in the levels of estimated overconfidence within papers (0.19 for women, 0.25 for men), and over twice as large as the mean standard deviation in the corresponding gender gap within multi-study papers (0.097). In the Appendix we show that our results are invariant to different choices of scale and location parameters, and functional forms. We also show that our results remain robust to fitting a standard one-stage specification, as in Rubin (1981), in which we treat each experimental observation in the sample, $\hat{\beta}_{ks}$, as exchangeable.

The posterior distribution of the model is proportional to the likelihood and the prior distributions specified above. While we cannot solve for a closed-form solution of the posterior distribution, in practice, we characterize the posterior distribution via simulation using Hamiltonian Monte Carlo (HMC), a subset of Markov Chain Monte Carlo (MCMC). HMC methods use derivatives of the density function to construct Markov transitions that sample from the posterior distribution. They do so by introducing auxiliary momentum variables and sampling from a joint density that depends on the auxiliary and posterior distributions. HMC methods are more efficient and better suited for estimating hierarchical models than other common MCMC algorithms, including Random Walk Metropolis and Gibbs Sampler (Betancourt and Girolami 2015; Neal 2011).[9]

Note that the estimated posterior is a joint distribution over not just the population hyperparameters but also each study-level effect. In other words, the best belief about the true effect in a setting is not simply the study-specific estimate. One can, in fact, improve on the study-specific estimate by factoring in information from $S-1$ comparable studies. This seemingly paradoxical result was first attributed to Charles Stein (Efron and Morris 1977). The intuition behind it is as follows.[10] Consider results from S studies obtained in

S specific settings, $\hat{\beta}_s$, $s = 1, \ldots, S$, and the overall average $\hat{\beta}$. Imagine next replicating study s in the exact same context. The best prediction for the associated effect is not simply $\hat{\beta}_s$, but indeed it will 'shrink' towards the overall average. More generally, all estimates that are above the overall average would be adjusted downwards, and vice versa. The degree of shrinkage (or pooling) depends on the informative content of each study s about the population of interest. By distinguishing between genuine and sampling variation across studies, the BHM makes this process rigorous and transparent.

To see this more formally, note that in a normal-normal hierarchical model specified by equations (3) and (4), where population parameters β and σ are known, the estimate of the parameter β_s for each study s can be characterized as a shrinkage estimator

$$\hat{\beta}_s^p = (1 - \lambda_s)\hat{\beta}_s + \lambda_s \hat{\beta}^p,$$

where the superscript p denotes posterior estimates, and the pooling factor $\lambda_s \in [0, \ 1]$ captures the degree to which the posterior estimates are shrunk towards the posterior mean.

Following this intuition, Rubin (1981) and Gelman *et al.* (2013) suggest a pooling metric given by

$$\hat{\lambda}_s = \frac{\hat{\sigma}_s^{2,p}}{\hat{\sigma}_s^{2,p} + \hat{\sigma}^{2,p}}, \tag{6}$$

where $\hat{\sigma}_s^{2,p}$ is the posterior standard error estimate at the study level, and $\hat{\sigma}^{2,p}$ is the corresponding population estimate. In a two-step model, for multi-study papers $\hat{\sigma}_s^{2,p}$ is the posterior estimate from the first stage. For each s, $\lambda_s = 0$ corresponds to full pooling, while $\lambda_s = 1$ corresponds to no pooling. To obtain an indicator for the degree of pooling at the population level, we estimate $\hat{\lambda}$, which is an arithmetic mean of the pooling metric per study, $\hat{\lambda}_s$.

V. Results

Table 2 summarizes the posterior distribution of the hyperparameters β and σ for overconfidence of men and women, and for the gender gap in overconfidence. In all samples, both men and women are found to be overconfident, although only in the full sample does the 95% interval not include zero. In this case, men and women overestimate their performance by about 0.39 and 0.35 standard deviations relative to their ability, respectively. We find little evidence to suggest that men are more overconfident than women. While the estimated gender difference in overconfidence is positive across all three subsamples, its magnitude is small relative to gender-specific means, and in all samples its 95% interval includes zero.

Critically, the results from the BHM suggest that there is a high degree of heterogeneity in the levels and differences in overconfidence across studies. Figure 5 compares posterior β estimates and their 95% and 90% posterior intervals from each sample to the corresponding full-pooling estimates, which one would obtain under the assumption that each study identifies a common effect. Clearly, the posterior intervals are much wider for the BHM estimates than under the full-pooling model, reflecting a high degree of genuine heterogeneity across settings.

TABLE 2
POSTERIOR ESTIMATES FOR HYPERPARAMETERS

	N	J	$\hat{\beta}^p$	$\hat{\sigma}^p$	Percentiles				
					2.5	25	50	75	97.5
Intensive margin sample									
Overconfidence, men	17	7	0.076	0.051	−0.015	0.047	0.072	0.098	0.229
Overconfidence, women	17	7	0.065	0.044	−0.026	0.040	0.065	0.090	0.154
Difference (men – women)	26	13	0.027	0.019	−0.011	0.015	0.027	0.039	0.064
Extensive margin sample									
Overconfidence, men	7	2	0.339	0.374	−0.527	0.171	0.363	0.522	1.106
Overconfidence, women	7	2	0.272	0.382	−0.631	0.096	0.301	0.474	1.027
Difference (men – women)	13	4	0.059	0.088	−0.124	0.018	0.058	0.100	0.248
Full sample									
Overconfidence, men	24	9	0.392	0.150	0.086	0.300	0.394	0.485	0.687
Overconfidence, women	24	9	0.352	0.154	0.038	0.259	0.354	0.448	0.657
Difference (men – women)	39	16	0.094	0.054	−0.015	0.059	0.094	0.129	0.200

Notes
The table reports estimates of a two-stage BHM for gender-specific overconfidence and for the associated gender gap. The first stage aggregates estimates within each paper. The second stage aggregates the paper-level estimates (from the first stage) across papers. N denotes the number of results included, and J denotes the number of papers that these come from.

This in turn implies that the available body of evidence from the S studies would not be highly-informative about the likely result from the next study, $\hat{\beta}_{S+1}$. Figure 6 plots the posterior *predictive* distribution for $\hat{\beta}_{S+1}$ for the gender gap in overconfidence in the full sample. Indeed, there is 63.8% probability that the next study would find a gender gap in overconfidence ranging (widely) between −0.2 and 0.2.

We next present a more detailed breakdown of results for the full sample,[11] plotting posterior estimates for each study in Figure 7 for gender-specific overconfidence and Figure 8 for the gender gap in overconfidence. Compared to the original study estimates, the posterior $\hat{\beta}_s^p$ estimates 'shrink' closer to the hyperparameter $\hat{\beta}^p$, plotted at the bottom of each graph. But, as suggested by the comparison between the BHM and full-pooling models in Figure 5, the degree of shrinkage or pooling is quite limited. Table 3 reports pooling factors, obtained as sample averages of expression (6). These imply that only 8% and 6.9% of the variation in estimated overconfidence for men and women, respectively, is explained by sampling variation. For the associated gender gap, the degree of pooling is somewhat higher, at 23%. Overall, the reported pooling factors imply that the differences in estimates across studies are largely explained by genuine heterogeneity across settings. Thus each additional study on overconfidence tells us little about the overall population mean. In other words, each individual study has limited external validity.

VI. EXPLAINING THE KNOWLEDGE GAP

The BHM estimates indicate that there is no significant gender gap in self-confidence. This, however, implies a knowledge gap, as experts' opinions are at odds with the BHM analysis of available evidence. The discrepancy is stark. As we have seen, most experts' interpretation of the literature is that there is a positive confidence gap for men, whereas the BHM estimates cannot reject a zero gap. On a simple count, Figure 9 shows that 72%

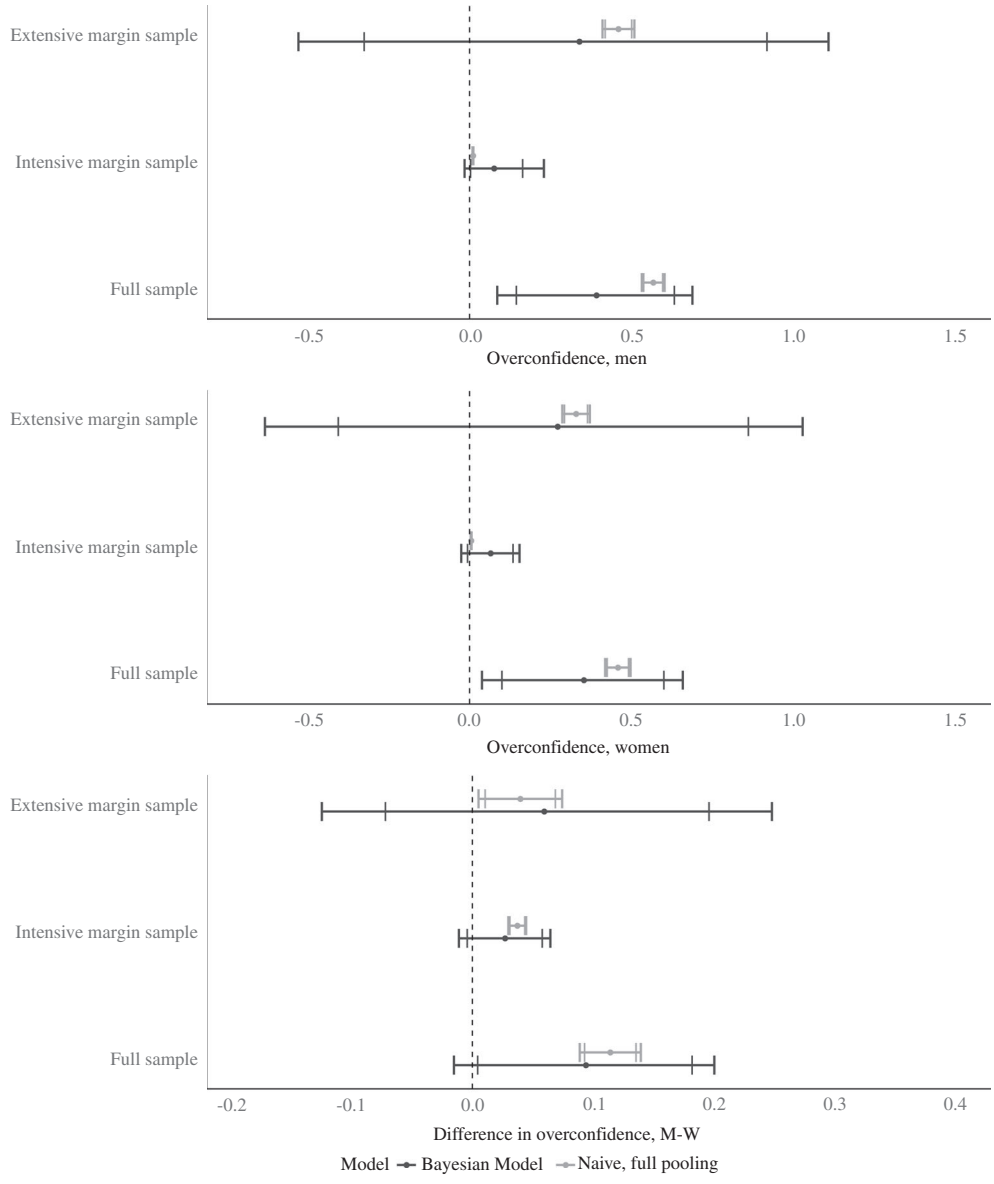

FIGURE 5. Overconfidence of men and women, by sample.
Notes: The figure reports posterior estimates of a two-stage BHM for gender-specific overconfidence and for the corresponding gender difference. The first stage aggregates estimates within each paper. The second stage aggregates the paper-level estimates (from the first stage) across papers.

of the findings indicate that both men and women are overconfident, yet only 8%—26 of 342 respondents—had this interpretation. On the other hand, 77%—265 respondents—believed that men are overconfident and women underconfident, while only 18% of the findings are in line with this interpretation.

Why are expert economists' beliefs starkly different to the economics literature on confidence? We explore two possible explanations.

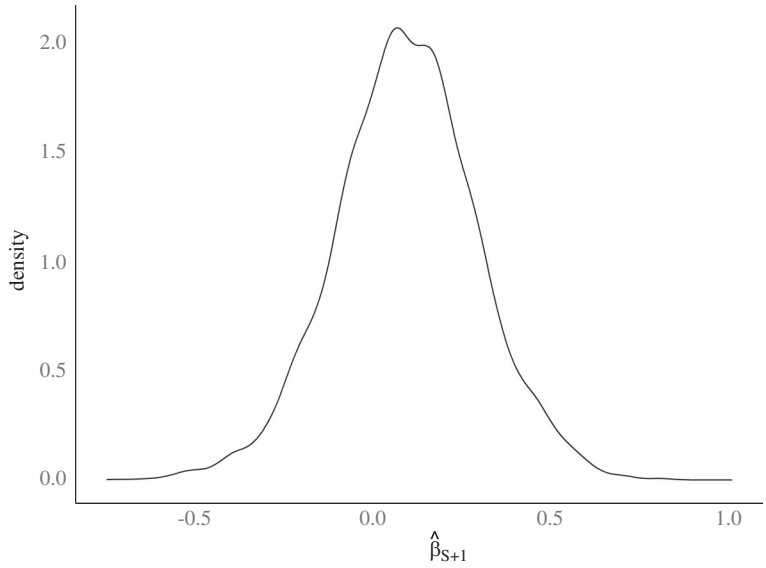

FIGURE 6. Gender difference in overconfidence, posterior predictive distribution of $\hat{\beta}_{S+1}$.

First, it is reasonable to hypothesize that highly cited papers play a relatively stronger role in shaping beliefs in the profession. Below, we take on board the role of citations by adjusting the estimated precision of each study-level estimate according to its citations. To do so, we estimate a BHM in which we inflate the precision of each study-level estimate by its citations relative to the median number of citations in the sample:

$$\text{citation} - \text{adjusted s.e.} \equiv \tilde{\sigma}_s = \hat{\sigma}_s \times \frac{\text{med}(citations_s)}{citations_s}. \tag{7}$$

When estimating the posterior mean $\hat{\beta}^p$, this procedure revises upwards the precision of studies with higher than median citations, and vice versa.

Table 4 reports the results obtained, as well as those based on the original standard errors for reference. Comparing estimates in the two rows, the adjustment does little to change our main results, and if anything, the posterior mean of 0.080 is slightly smaller than the 0.094 estimate obtained on the original standard errors. Furthermore, the 95% interval is now wider. The interpretation is that papers finding a larger gender gap in confidence do not systematically attract more citations.

Can the differences between experts' beliefs and the literature be explained by biased and/or strong prior beliefs on the gender differences in overconfidence? We explore this hypothesis by considering how our estimates change with different assumptions on the moments of the hyper-prior of β, using the standardized mean (1.46) and standard deviation (0.078) of survey responses as a proxy for prior beliefs on the gender differences in overconfidence.

As can be seen from Table 5, the gap between the beliefs and the literature can be largely accounted for by an extreme hyper-prior of $\beta \sim N(1.46,\ 0.078^2)$, wherein the prior belief is not only non-zero but also very strongly held. When we move from our standard model with hyper-priors $\beta \sim N(0, 1)$ to one with just a change in the hyper-prior mean ($\beta \sim N(1.46, 1)$), the estimated posterior mean and distribution is largely

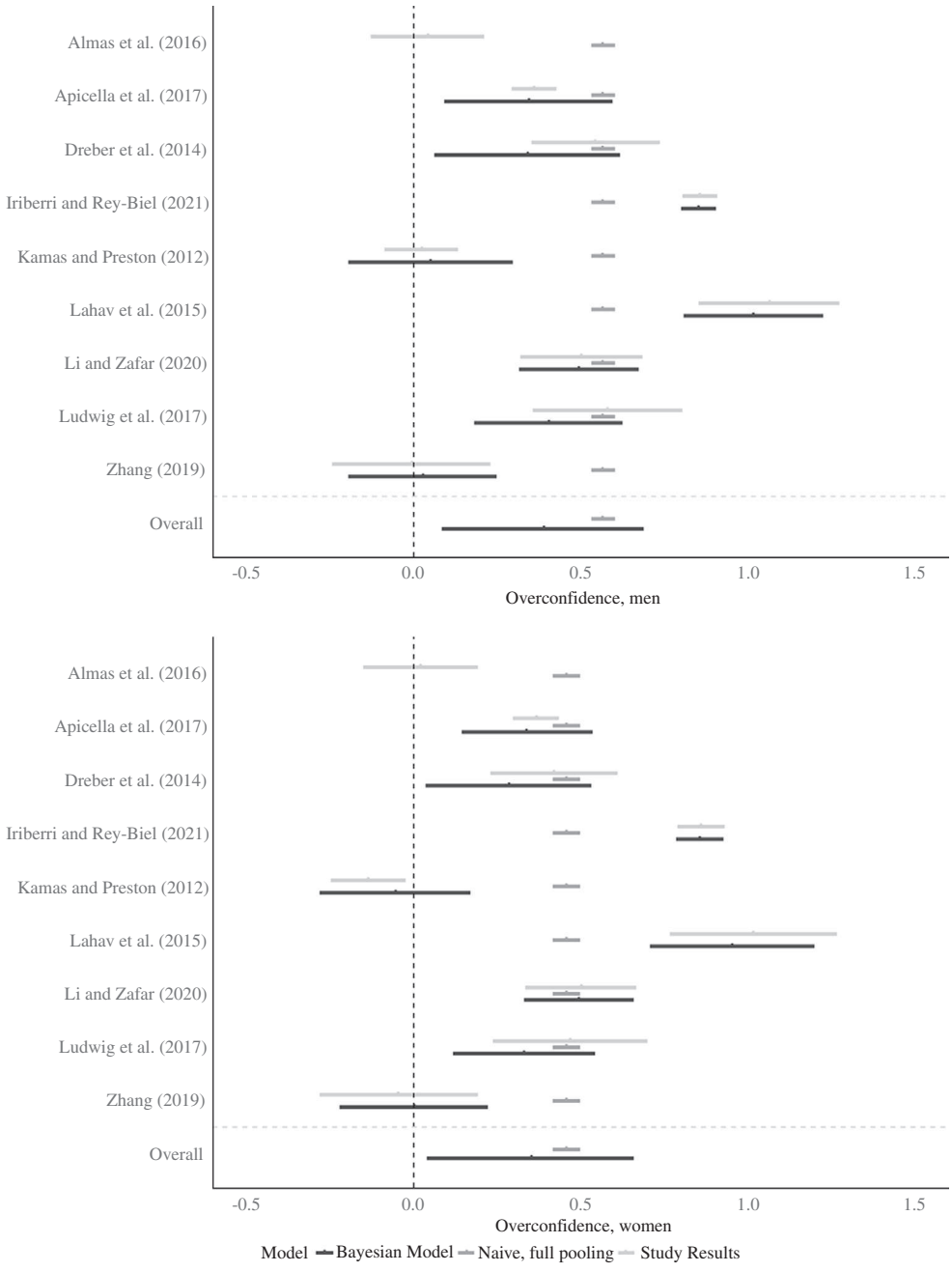

FIGURE 7. Model comparison—overconfidence, by gender and paper.
Notes: The figure reports posterior estimates of a two-stage BHM for gender-specific overconfidence. The first stage aggregates estimates within each paper. The second stage aggregates the paper-level estimates (from the first stage) across papers.

unchanged when compared to our baseline model. However, once we also increase the confidence around the beliefs on the mean, the posterior mean on the average differences in overconfidence almost perfectly coincides with the survey beliefs. This result is

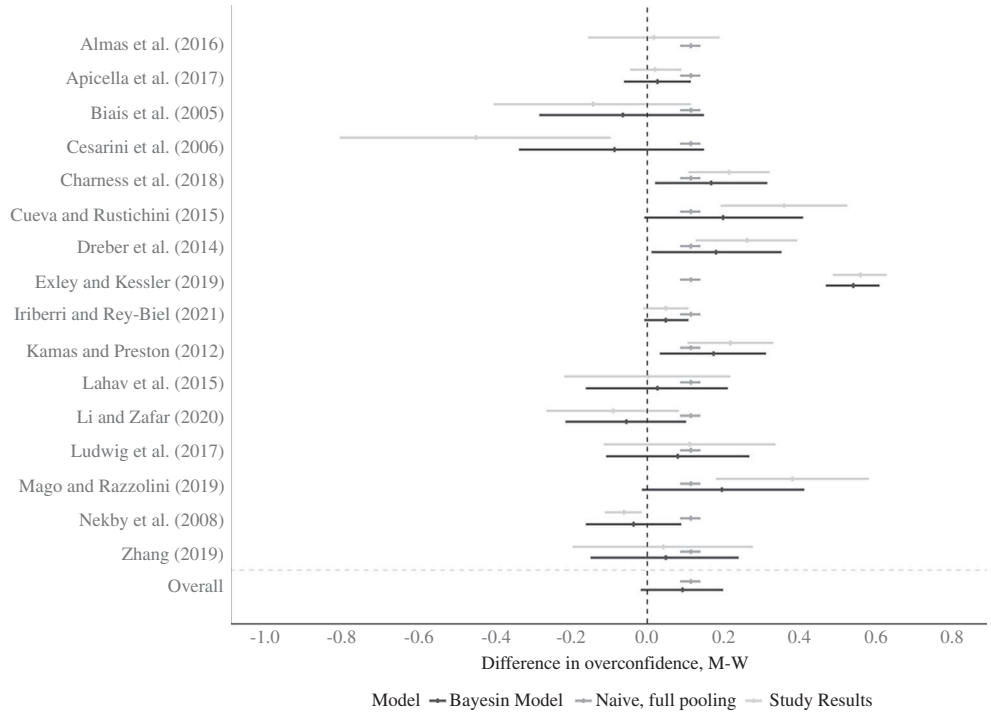

FIGURE 8. Model comparison—gender differences in overconfidence, by paper.
Notes: The figure reports posterior estimates of a two-stage BHM for gender gaps in overconfidence. The first stage aggregates estimates within each paper. The second stage aggregates the paper-level estimates (from the first stage) across papers.

TABLE 3
POOLING FACTORS BY METRIC

	$\hat{\lambda}$
Overconfidence of men	0.080
Overconfidence of women	0.069
Gender difference in overconfidence	0.23

Notes
The overall pooling factor is obtained as a sample average of expression (6).

intuitive and follows almost mechanically from the setup of the Bayesian model: given very precise priors, there will be hardly any updating on the posterior mean, regardless of what is found in the literature.[12]

VII. DISCUSSION

Our analysis yields two main lessons. The first is that the literature in economics provides little support to the hypothesis that differences in self-confidence can explain differences in labour market outcomes because, against popular stereotypes, if men are from Mars, then so are women. This is important because if men and women do not differ on traits such as confidence, then it may be that the barriers/opportunities that they face are different, and

FIGURE 9. Experts' beliefs versus results on overconfidence.
Notes: This figure compares the distribution of experts' beliefs collected from the survey (first bar) to the raw evidence (non-Bayesian aggregated) from the literature (bars 2–4). All studies ($N = 90$) include results with paired observations on confidence for men and women that meet the criteria laid out in Section II. 'BHM sample (difference)' ($N = 39$) refers to the subsample of results that report standard errors for the gender gap in overconfidence. 'BHM sample (levels)' ($N = 24$) refers to the subsample of results that report standard errors separately for overconfidence of men and women.

TABLE 4
POSTERIOR ESTIMATES FOR HYPERPARAMETER-BASED CITATION-ADJUSTED STANDARD ERRORS

	N	$\hat{\beta}^p$	$\hat{\sigma}^p$	Percentiles				
				2.5	25	50	75	97.5
Original s.e.	39	0.094	0.054	−0.016	0.059	0.094	0.129	0.200
Citation-adjusted s.e.	39	0.080	0.068	−0.056	0.037	0.081	0.124	0.213

Notes
The table reports posterior estimates of a two-stage BHM for gender gaps in overconfidence. The first stage aggregates estimates within each paper. The second stage aggregates the paper-level estimates (from the first stage) across papers. The standard errors used in estimates in the second row have been adjusted for cites received, according to expression (7).

that is what needs to be addressed. However, there is no doubt that in some settings women are less confident than men, but in many others they are not. Indeed, the BHM estimate of the pooling factor is quite low, implying that self-confidence is context-specific.

The second, intriguing, finding is that the experts' interpretation of the literature is close to naive pooling and at odds with Bayesian learning. This is especially surprising because for other traits—especially altruism and risk attitudes—the experts' opinions are more in line with BHM estimates. One way to reconcile this is to note that in these domains the pooling factor is high, so that the naive pooling estimate is close to the Bayesian posterior.

TABLE 5
POSTERIOR ESTIMATES FOR HYPERPARAMETERS FOR ALTERNATIVE PRIOR DISTRIBUTIONS

Prior on β	N	$\hat{\beta}^p$	$\hat{\sigma}^p$	Percentiles				
				2.5	25	50	75	97.5
$\beta \sim N(0,1)$	39	0.094	0.054	−0.015	0.059	0.094	0.129	0.200
$\beta \sim N(1.46,1)$	39	0.098	0.054	−0.010	0.063	0.098	0.133	0.205
$\beta \sim N(1.46,\ 0.078^2)$	39	1.463	0.006	1.451	1.459	1.463	1.467	1.475

Notes
The table reports posterior estimates of a two-stage BHM for gender gaps in overconfidence. The first stage aggregates estimates within each paper. The second stage aggregates the paper-level estimates (from the first stage) across papers.

This raises the question of how experts learn, because, ultimately, this is what determines the advancement of science.

ACKNOWLEDGMENTS

This research was undertaken separately from Parekh's role at J-PAL Africa.

We are grateful to participants to the *Economica* Centenary Conference for helpful comments.

NOTES

1. High-income careers typically develop in competitive environments, in which winners may be disproportionately rewarded, and are characterized by a relatively high variability of earnings. Individuals who are unwilling to compete or are particularly risk-averse may thus simply not embark on those careers. Likewise, pro-social preferences may lead to choices that do not maximize own monetary payoffs. These are only some of the channels whereby gender differences in these traits may interfere with women's labour market success.
2. One possible interpretation is that the context matters more than individual preferences for confidence experiments, i.e. the same individual might be overconfident or underconfident depending on the circumstances.
3. See the related discussion in Bertrand (2020). See also Bordalo *et al.* (2019) for evidence on the role of stereotypes in shaping beliefs about gender skills.
4. Measures based on performance scores provide an estimate of absolute overconfidence, while measures based on performance rank in a tournament provide an estimate of relative overconfidence, both of which are used in this paper.
5. For example, Kamas and Preston (2012) compare a participant's actual score to their estimated score in a maths task to measure their self-confidence. In another experiment, they compare a participant's actual ranking in a group competition to their estimated ranking.
6. For example, in Reuben *et al.* (2017), students perform addition tasks under tournament and piece-rate compensation. After completing the tasks, students are asked to rank their beliefs on their performance within a group of four. The authors measure confidence by comparing the percentage of men and women who think that they would rank first versus the percentage who would have come first, based on their performance.
7. For example, Dreber *et al.* (2014) consider the share of boys and girls who choose to compete in a verbal task tournament. As participants are compensated only if they win a tournament, the participation rate is used to measure one's beliefs about outperforming other participants. The authors find that 33% of boys choose to compete in the verbal task, compared to 28% of girls. However, based on performance in the tournament, the probability of winning is similar for boys and girls, implying that as many girls as boys should have chosen to compete. In this context, the authors measure confidence as the difference between the share of boys and girls who choose to compete versus those who *should* compete, as proxied by their actual performance.
8. Across all studies in our analysis, we have a range of 1–6 experiments per study.
9. We implement this using Stan, a C++ programme that is commonly used for estimating Bayesian models. For each model and metric of interest, we use 8 chains and 100,000 iterations per chain.
10. For the sake of this simple argument, we discuss a one-stage BHM framework.

11. Similar breakdowns for the intensive and extensive margin samples can be found in Figures A6–A9 in the Appendix.
12. In Tables A3, A4 and A5 of the Appendix, we show that our main findings remain robust to changes on the functional form on priors for β and σ. Tables A6–A9 show robustness analysis based on a one-step BHM.

REFERENCES

Almås, I., Cappelen, A. W., Salvanes, K. G., Sørensen, E. Ø. and Tungodden, B. (2016). Willingness to compete: family matters. *Management Science*, **62**(8), 2149–62.

Apicella, C. L., Demiral, E. E. and Mollerstrom, J. (2017). No gender difference in willingness to compete when competing against self. *American Economic Review*, **107**(5), 136–40.

Azmat, G. and Petrongolo, B. (2014). Gender and the labor market: what have we learned from field and lab experiments?*Labour Economics*, **30**, 32–40.

Balafoutas, L. and Sutter, M. (2010). Gender, competition and the efficiency of policy interventions. IZA Discussion Paper.

Baldiga, N. R. and Coffman, K. B. (2018). Laboratory evidence on the effects of sponsorship on the competitive preferences of men and women. *Management Science*, **64**(2), 888–901.

Bandiera, O., Fischer, G., Prat, A. and Ytsma, E. (2021). Do women respond less to performance pay? Building evidence from multiple experiments. *American Economic Review: Insights*, forthcoming.

Beaurain, G. and Masclet, D. (2016). Does affirmative action reduce gender discrimination and enhance efficiency? New experimental evidence. *European Economic Review*, **90**, 350–62.

Bertrand, M. (2011). New perspectives on gender. In D.Card and O.Ashenfelter (eds), *Handbook of Labor Economics*, Vol. **4**. Amsterdam: Elsevier, pp. 1543–90.

Bertrand, M. (2018). Coase Lecture—The Glass Ceiling. *Economica*, **85**(338), 205–31.

Bertrand, M. (2020). Gender in the twenty-first century. *AEA Papers and Proceedings*, **110**, 1–24.

Betancourt, M. and Girolami, M. (2015). Hamiltonian Monte Carlo for hierarchical models: current trends in Bayesian methodology with applications, **79**(30), 2–4.

Biais, B., Hilton, D., Mazurier, K. and Pouget, S. (2005). Judgemental overconfidence, self-monitoring, and trading performance in an experimental financial market. *Review of Economic Studies*, **72**(2), 287–312.

Bordalo, P., Coffman, K., Gennaioli, N. and Shleifer, A. (2019). Beliefs about gender. *American Economic Review*, **109**(3), 739–73.

Buser, T., Niederle, M. and Oosterbeek, H. (2014). Gender, competitiveness, and career choices. *Quarterly Journal of Economics*, **129**(3), 1409–47.

Buser, T. and Yuan, H. (2019). Do women give up competing more easily? Evidence from the lab and the Dutch Math Olympiad. *American Economic Journal: Applied Economics*, **11**(3), 225–52.

Cesarini, D., Sandewall, Ö. and Johannesson, M. (2006). Confidence interval estimation tasks and the economics of overconfidence. *Journal of Economic Behavior & Organization*, **61**(3), 453–70.

Charness, G., Rustichini, A. and Van de Ven, J. (2018). Self-confidence and strategic behavior. *Experimental Economics*, **21**(1), 72–98.

Coffman, K. B. (2014). Evidence on self-stereotyping and the contribution of ideas. *Quarterly Journal of Economics*, **129**(4), 1625–60.

Cohen, J. (2013). *Statistical Power Analysis for the Behavioral Sciences*. Cambridge, MA: Academic Press.

Croson, R. and Gneezy, U. (2009). Gender differences in preferences. *Journal of Economic Literature*, **47**(2), 448–74.

Cueva, C. and Rustichini, A. (2015). Is financial instability male-driven? Gender and cognitive skills in experimental asset markets. *Journal of Economic Behavior & Organization*, **119**, 330–44.

Dreber, A., Von Essen, E. and Ranehill, E. (2014). Gender and competition in adolescence: task matters. *Experimental Economics*, **17**(1), 154–72.

Efron, B. and Morris, C. (1977). Stein's paradox in statistics. *Scientific American*, **236**(5), 119–27.

Ewers, M. and Zimmermann, F. (2015). Image and misreporting. *Journal of the European Economic Association*, **13**(2), 363–80.

Exley, C. L. and Kessler, J. B. (2019). The gender gap in self-promotion. NBER Technical Report.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A. and Rubin, D. B.(2013). *Bayesian Data Analysis*. Boca Raton, FL: CRC Press.

Gupta, N. D., Poulsen, A. and Villeval, M. C. (2005). Male and female competitive behavior—experimental evidence. IZA Discussion Paper.

Hardies, K., Breesch, D. and Branson, J. (2013). Gender differences in overconfidence and risk taking: do self-selection and socialization matter?*Economics Letters*, **118**(3), 442–4.

HYDE, J. S. (2005). The gender similarities hypothesis. *American Psychologist*, **60**(6), 581–92.
HYDE, J. S. (2014). Gender similarities and differences. *Annual Review of Psychology*, **65**(1), 373–98.
IRIBERRI, N. and REY-BIEL, P. (2021). Brave boys and play-it-safe girls: gender differences in willingness to guess in a large scale natural field experiment. *European Economic Review*, **131**, 103603.
JACKSON, M. O., HAGHTALAB, N. and PROCACCIA, A. D. (2021). Belief polarization in a complex world: a learning theory perspective. *PNAS*, **118**.
JAKOBSSON, N. (2012). Gender and confidence: are women underconfident?*Applied Economics Letters*, **19**(11), 1057–9.
KAMAS, L. and PRESTON, A. (2012). The importance of being confident; gender, career choice, and willingness to compete. *Journal of Economic Behavior & Organization*, **83**(1), 82–97.
KAMAS, L. and PRESTON, A. (2018). Competing with confidence: the ticket to labor market success for college-educated women. *Journal of Economic Behavior & Organization*, **155**, 231–52.
KASY, M. (2021). Of forking paths and tied hands: selective publication of findings, and what economists should do about it. *Journal of Economic Perspectives*, **35**(3), 175–92.
KLEVEN, H., LANDAIS, C. and SØGAARD, J. E. (2019). Children and gender inequality: evidence from Denmark. *American Economic Journal: Applied Economics*, **11**(4), 181–209.
LAHAV, E., NIR, A. and SINIVER, E. (2015). Do differing pay schemes help close the gender gap in overconfidence? *Economics Bulletin,***35**(1), 30–6.
LI, C. H. and ZAFAR, B. (2020). Ask and you shall receive? Gender differences in regrades in college. IZA Discussion Paper.
LUDWIG, S., FELLNER-RÖHLING, G. and THOMA, C. (2017). Do women have more shame than men? An experiment on self-assessment and the shame of overestimating oneself. *European Economic Review*, **92**, 31–46.
MAGO, S. D. and RAZZOLINI, L. (2019). Best-of-five contest: an experiment on gender differences. *Journal of Economic Behavior & Organization*, **162**, 164–87.
MCCULLOCH, C. E. and NEUHAUS, J. M. (2011). Misspecifying the shape of a random effects distribution: why getting it wrong may not matter. *Statistical Science*, **26**(3), 388–402.
NEAL, R. M. (2011). MCMC using Hamiltonian dynamics. In S.Brooks, A.Gelman, G. L.Jones and X.-L.Meng (eds), *Handbook of Markov Chain Monte Carlo*. Boca Raton, FL: Chapman & Hall/CRC Press, pp. 113–62.
NEKBY, L., THOURSIE, P. S. and VAHTRIK, L. (2008). Gender and self-selection into a competitive environment: are women more overconfident than men?*Economics Letters*, **100**(3), 405–7.
NIEDERLE, M., SEGAL, C. and VESTERLUND, L. (2013). How costly is diversity? Affirmative action in light of gender differences in competitiveness. *Management Science*, **59**(1), 1–16.
NIEDERLE, M. and VESTERLUND, L. (2007). Do women shy away from competition? Do men compete too much?*Quarterly Journal of Economics*, **122**(3), 1067–101.
NIEDERLE, M. and YESTRUMSKAS, A. H. (2008). Gender differences in seeking challenges: the role of institutions. NBER Technical Report.
PIKULINA, E., RENNEBOOG, L. and TOBLER, P. N. (2017). Overconfidence and investment: an experimental approach. *Journal of Corporate Finance*, **43**, 175–92.
PROEGER, T. and MEUB, L. (2014). Overconfidence as a social bias: experimental evidence. *Economics Letters*, **122**(2), 203–7.
RAO, M. (2020). Gender differences in altruism: a Bayesian hierarchical analysis of dictator games. Working Paper.
REUBEN, E., REY-BIEL, P., SAPIENZA, P. and ZINGALES, L. (2012). The emergence of male leadership in competitive environments. *Journal of Economic Behavior & Organization*, **83**(1), 111–17.
REUBEN, E., WISWALL, M. and ZAFAR, B. (2017). Preferences and biases in educational choices and labour market expectations: shrinking the black box of gender. *Economic Journal*, **127**(604), 2153–86.
ROSENTHAL, R. (1976). *Experimenter Effects in Behavioral Research*. New York: Irvington Publishers.
RUBIN, D. B. (1981). Estimation in parallel randomized experiments. *Journal of Educational Statistics*, **6**(4), 377–401.
SAMAK, A. C. (2013). Is there a gender gap in preschoolers' competitiveness? An experiment in the US. *Journal of Economic Behavior & Organization*, **92**, 22–31.
SUTTER, M., GLÄTZLE-RÜTZLER, D., BALAFOUTAS, L. and CZERMAK, S. (2016). Cancelling out early age gender differences in competition: an analysis of policy interventions. *Experimental Economics*, **19**(2), 412–32.
WOZNIAK, D., HARBAUGH, W. T. and MAYR, U. (2014). The menstrual cycle and performance feedback alter gender differences in competitive choices. *Journal of Labor Economics*, **32**(1), 161–98.
ZHANG, Y. J. (2019). Culture, institutions and the gender gap in competitive inclination: evidence from the communist experiment in China. *Economic Journal*, **129**(617), 509–52.

APPENDIX

ADDITIONAL STATISTICS: EXPERTS' SURVEY

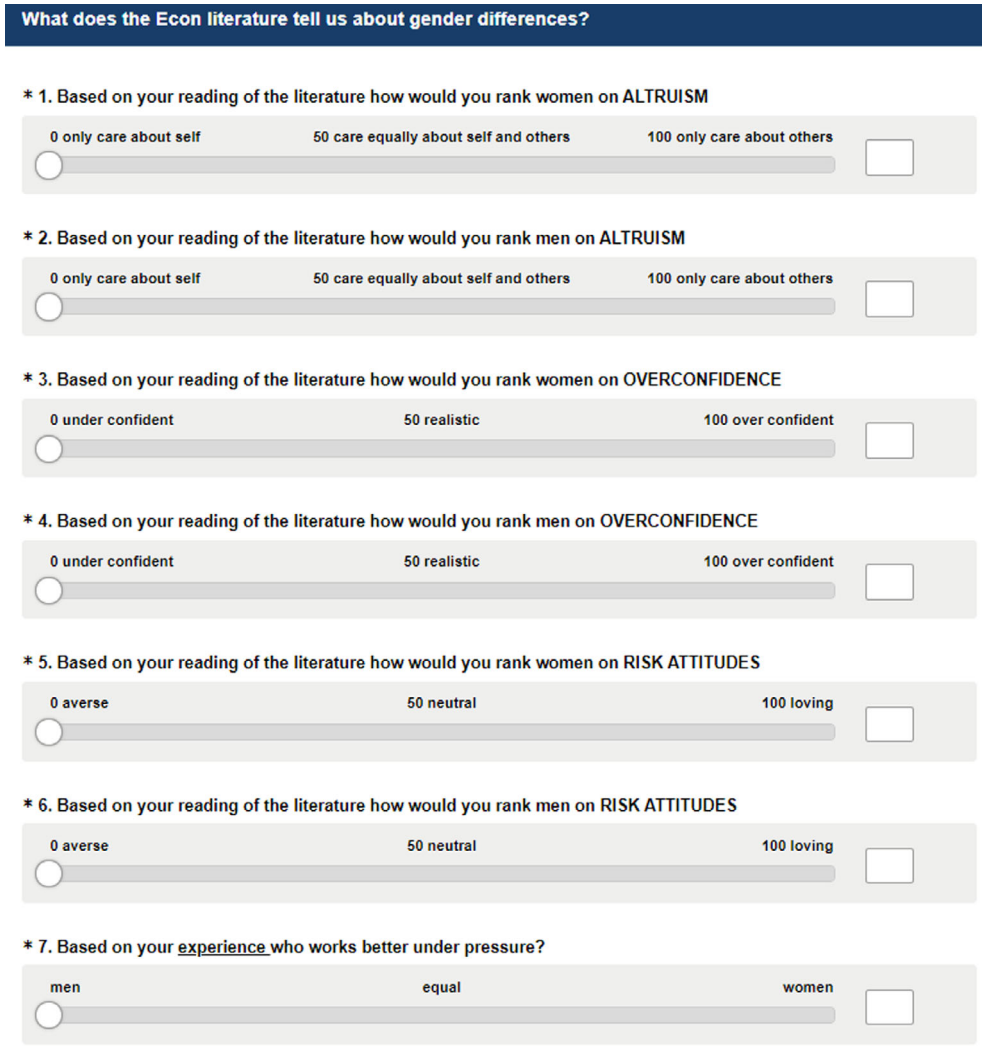

FIGURE A1. Expert survey questions 1.

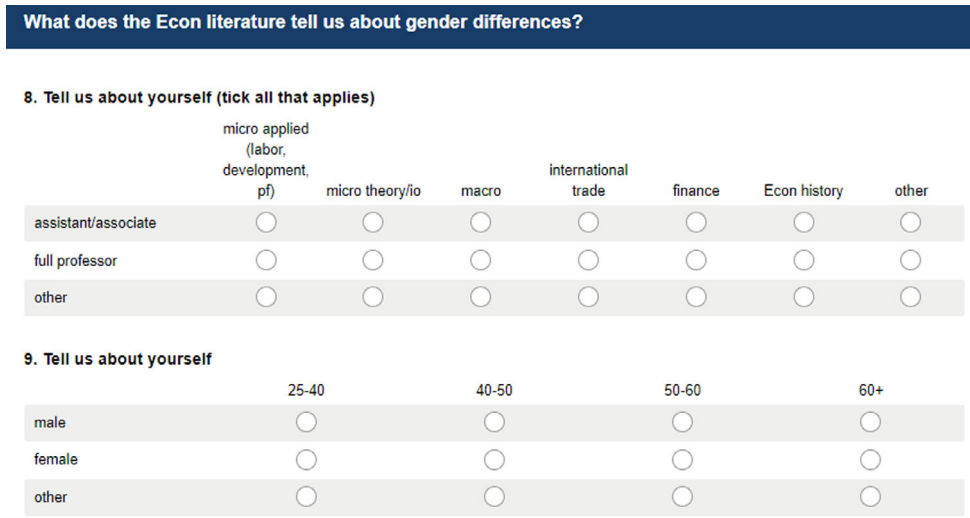

FIGURE A2. Expert survey questions 2.

TABLE A1
FIELDS OF SPECIALIZATION OF SURVEY RESPONDENTS

Specialization	No.
Econ. History	10
Finance	39
International Trade	33
Macro	70
Micro Applied	108
Micro Theory or IO	55
Other	16

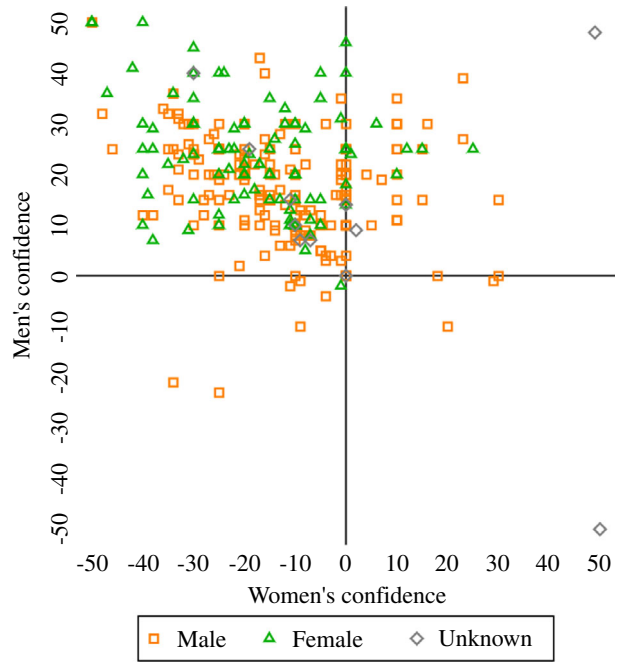

FIGURE A3. Survey results on confidence levels by gender.
Notes: Men ($N = 220$), women ($N = 111$) and unknown ($N = 11$). No respondents chose 'other' as their gender. Refer to Figures A1 and A2 for the survey questions.

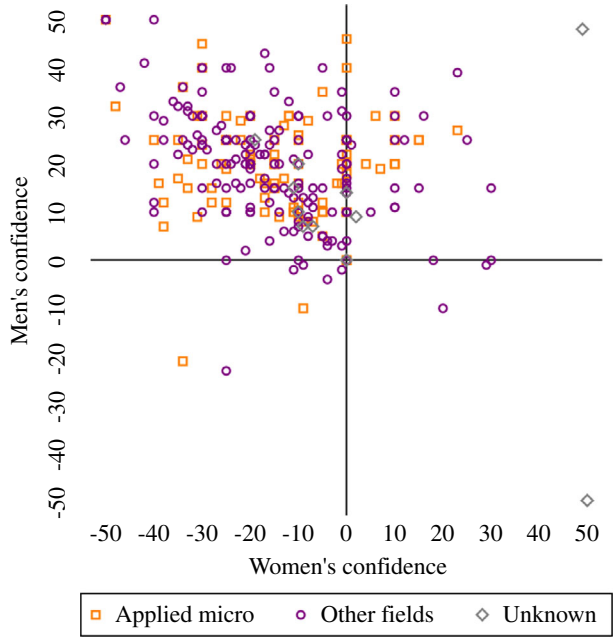

FIGURE A4. Survey results on confidence levels by field.
Notes: Applied micro ($N = 108$), other fields ($N = 234$), unknown ($N = 11$). Refer to Figures A1 and A2 for the survey questions.

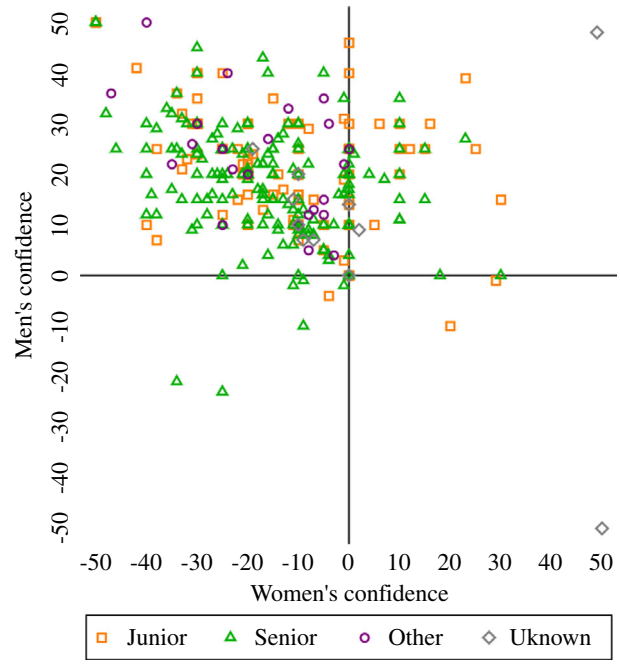

FIGURE A5. Survey results on confidence levels by seniority.
Notes: Junior, i.e. assistant/associate professor ($N = 111$); senior, i.e. full professor ($N = 196$); other ($N = 24$); unknown ($N = 11$). Refer to Figures A1 and A2 for the survey questions.

PAPERS USED IN ANALYSES

TABLE A2
LIST OF PAPERS USED

Title	Authors and year	Journal	Country	Exp.[a]	BHM diff[b]	BHM levels[c]
Ask and you shall receive? Gender differences in regrades in college	Li and Zafar (2020)	IZA Discussion Paper	USA	L	Y	Y
Beliefs about gender	Bordalo *et al.* (2019)	*American Economic Review*	USA	L	N	N
Best-of-five contest: an experiment on gender differences	Mago and Razzolini (2019)	*Journal of Economic Behavior & Organization*	USA	L	Y	N
Brave boys and play-it-safe girls: gender differences in willingness to guess in a large scale natural field experiment	Iriberri and Rey-Biel (2021)	*European Economic Review*	Spain	F	Y	Y
Cancelling out early age gender differences in competition: an analysis of policy interventions	Sutter *et al.* (2016)	*Experimental Economics*	Austria	H	N	N
Competing with confidence: the ticket to labor market success for college-educated women	Kamas and Preston (2018)	*Journal of Economic Behavior & Organization*	USA	L	N	N
Confidence interval estimation tasks and the economics of overconfidence	Cesarini *et al.* (2006)	*Journal of Economic Behavior & Organization*	Sweden	L	Y	N
Culture, institutions, and the gender gap in competitive inclination: evidence from the communist experiment in China	Zhang (2019)	*Economic Journal*	China	L	Y	Y
Do differing pay schemes help close the gender gap in overconfidence?	Lahav *et al.* (2015)	*Economics Bulletin*	Israel	L	Y	Y
Do women give up competing more easily? Evidence from the lab and the Dutch Math Olympiad	Buser and Yuan (2019)	*American Economic Journal: Applied Economics*	Netherlands	M	N	N
Do women have more shame than men? An experiment on self-assessment and the shame of overestimating oneself	Ludwig *et al.* (2017)	*European Economic Review*	Germany and Austria	L	Y	Y
Do women shy away from competition? Do men compete too much?	Niederle and Vesterlund (2007)	*Quarterly Journal of Economics*	USA	L	N	N

TABLE A2
CONTINUED

Title	Authors and year	Journal	Country	Exp.[a]	BHM diff[b]	BHM levels[c]
Does affirmative action reduce gender discrimination and enhance efficiency? New experimental evidence	Beaurain and Masclet (2016)	*European Economic Review*	France	L	N	N
Evidence on self-stereotyping and the contribution of ideas	Coffman (2014)	*Quarterly Journal of Economics*	USA	L	N	N
Gender and competition in adolescence: task matters	Dreber *et al.* (2014)	*Experimental Economics*	Sweden	L	Y	Y
Gender and confidence: are women underconfident	Jakobsson (2012)	*Applied Economics Letters*	Sweden	F	N	N
Gender and self-selection into a competitive environment: are women more overconfident than men?	Nekby *et al.* (2008)	*Economics Letters*	Sweden	F	Y	N
Gender differences in overconfidence and risk taking: do self-selection and socialization matter?	Hardies *et al.* (2013)	*Economics Letters*	Belgium	L	N	N
Gender differences in seeking challenges: the role of institutions	Niederle and Yestrumskas (2008)	NBER Working Paper	USA	L	N	N
Gender, Competition and the efficiency of policy interventions	Balafoutas and Sutter (2010)	IZA Discussion Paper	Austria	L	N	N
Gender, competitiveness, and career choices	Buser *et al.* (2014)	*Quarterly Journal of Economics*	Netherlands	L	N	N
How costly is diversity? Affirmative action in light of gender differences in competitiveness	Niederle *et al.* (2013)	*Management Science*	USA	L	N	N
Image and misreporting	Ewers and Zimmermann (2015)	*Journal of the European Economic Association*	Germany	L	N	N
Is financial instability male driven? Gender and cognitive skills in experimental asset markets	Cueva and Rustichini (2015)	*Journal of Economic Behavior & Organization*	UK	L	Y	N
Is there a gender gap in preschoolers' competitiveness? An experiment in the US	Samak (2013)	*Journal of Economic Behavior & Organization*	USA	L	N	N
Judgemental overconfidence, self-monitoring, and trading performance in an experimental financial market	Biais *et al.* (2005)	*Review of Economic Studies*	France and UK	L	Y	N

TABLE A2
CONTINUED

Title	Authors and year	Journal	Country	Exp.[a]	BHM diff[b]	BHM levels[c]
Laboratory evidence on the effects of sponsorship on the competitive preferences of men and women	Baldiga and Coffman (2018)	*Management Science*	USA	L	N	N
Male and female competitive behavior—experimental evidence	Gupta *et al.* (2005)	IZA Discussion Paper	France	L	N	N
No gender difference in willingness to compete when competing against self	Apicella *et al.* (2017)	*American Economic Review*	USA, Online	M	Y	Y
Overconfidence as a social bias: experimental evidence	Proeger and Meub (2014)	*Economics Letters*	Germany	L	N	N
Overconfidence and investment: an experimental approach	Pikulina *et al.* (2017)	*Journal of Corporate Finance*	Netherlands	L	N	N
Preferences and biases in educational choices and labour market expectations: shrinking the black box of gender	Reuben *et al.* (2017)	*Economic Journal*	USA	L	N	N
Self-confidence and strategic behavior	Charness *et al.* (2018)	*Experimental Economics*	Netherlands	L	Y	N
The emergence of male leadership in competitive environments	Reuben *et al.* (2012)	*Journal of Economic Behavior & Organization*	USA	L	N	N
The gender gap in self-promotion	Exley and Kessler (2019)	NBER Working Paper	USA	L	Y	N
The importance of being confident; gender, career choice, and willingness to compete	Kamas and Preston (2012)	*Journal of Economic Behavior & Organization*	USA	L	Y	Y
The menstrual cycle and performance feedback alter gender differences in competitive choices	Wozniak *et al.* (2014)	*Journal of Labor Economics*	USA	L	N	N
Willingness to compete: family matters	Almås *et al.* (2016)	*Management Science*	Norway	L	Y	Y

Notes
[a] L = laboratory experiment; F = field experiment; H = hybrid experiment; M = mix of experiments in a single paper. [b] Whether the results from the paper are used in the BHM for gender differences in confidence (Y = Yes, N = No). [c] Whether the results from the paper are used in the BHM for gender-specific overconfidence (Y = Yes, N = No).

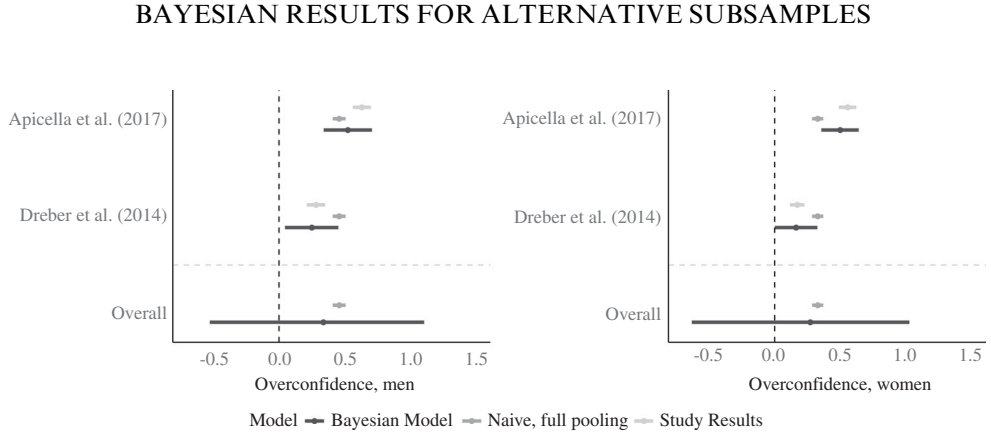

FIGURE A6. Model comparison: overconfidence by gender, extensive margin sample.

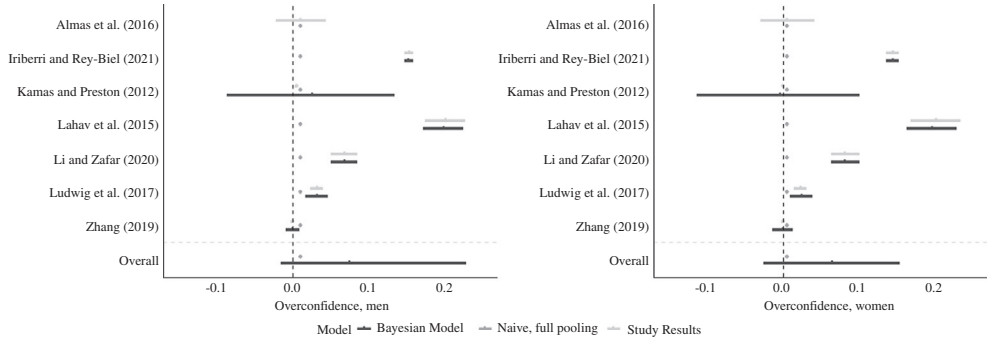

FIGURE A7. Model comparison: overconfidence by gender, intensive margin sample.

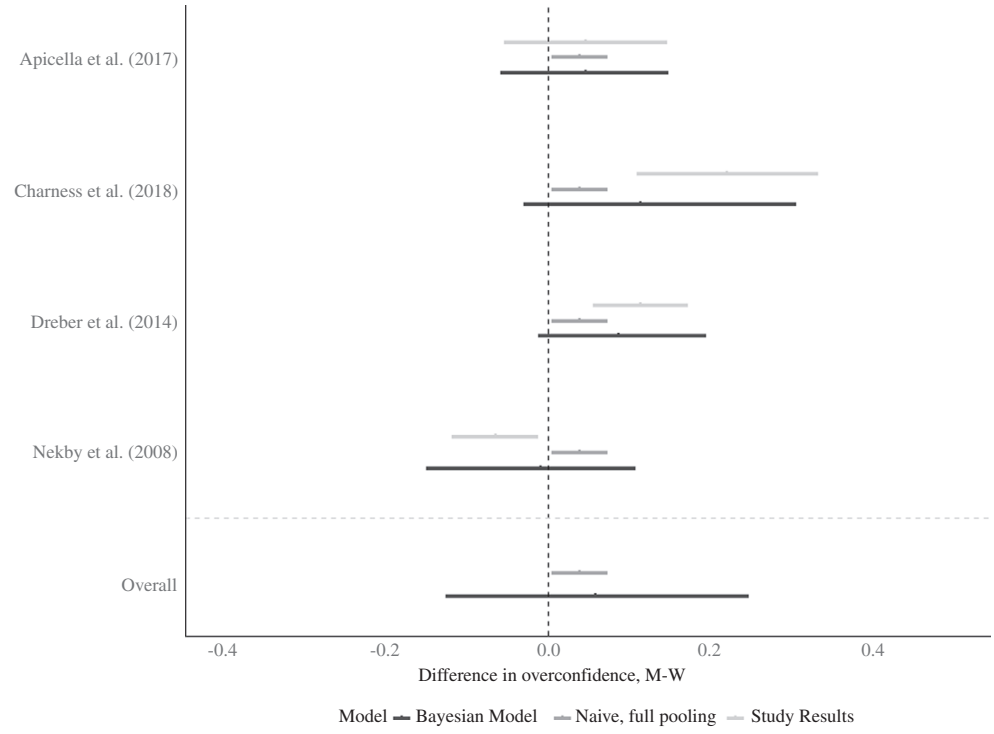

FIGURE A8. Model comparison: gender differences in overconfidence, extensive margin sample.

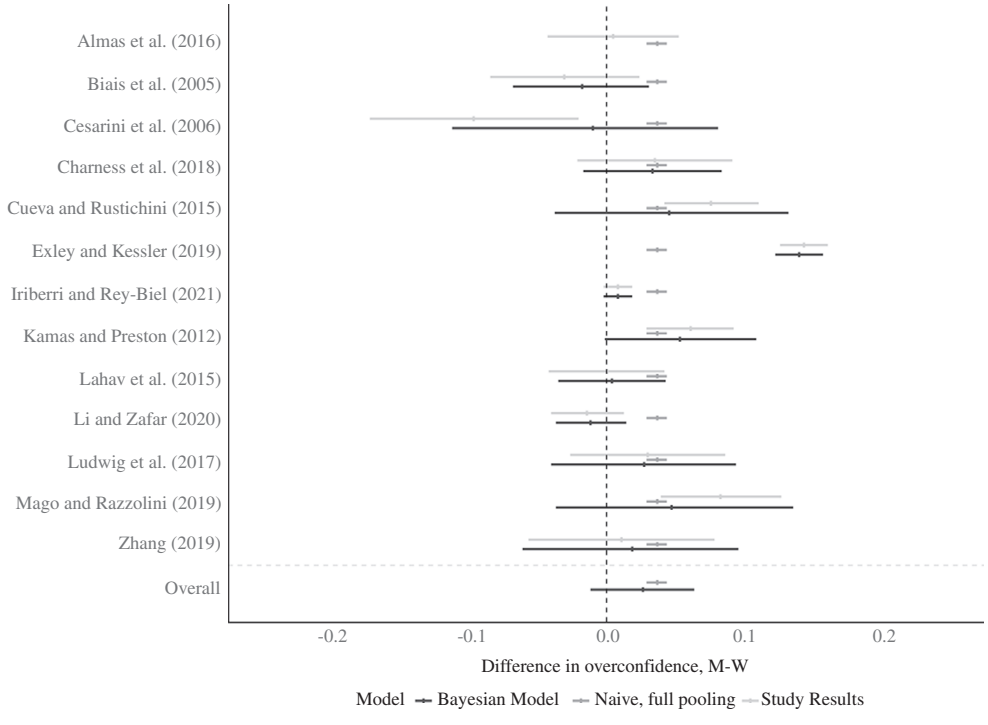

FIGURE A9. Model comparison: gender differences in overconfidence, intensive margin sample.

ROBUSTNESS ANALYSIS

TABLE A3
ALTERNATIVE FUNCTIONAL FORMS ON PRIORS: OVERCONFIDENCE, MEN

Model priors		$\hat{\beta}^p$	$\hat{\sigma}^p$	2.5%	50%	97.5%
$\beta \sim$ normal(0,1)	$\sigma \sim$ normal(0,1)	0.3915	0.1503	0.0855	0.3935	0.6868
$\beta \sim$ cauchy(0,1)	$\sigma \sim$ normal(0,1)	0.3853	0.1492	0.0810	0.3869	0.6792
$\beta \sim$ normal(0,10)	$\sigma \sim$ normal(0,1)	0.3999	0.1538	0.0890	0.4007	0.7065
$\beta \sim$ cauchy(0,10)	$\sigma \sim$ normal(0,1)	0.3999	0.1537	0.0885	0.4006	0.7074
$\beta \sim$ normal(0,1)	$\sigma \sim$ normal(0,10)	0.3906	0.1551	0.0769	0.3923	0.6954
$\beta \sim$ cauchy(0,1)	$\sigma \sim$ normal(0,10)	0.3837	0.1540	0.0680	0.3864	0.6850
$\beta \sim$ normal(0,10)	$\sigma \sim$ normal(0,10)	0.3952	0.1681	0.0613	0.3993	0.7107
$\beta \sim$ cauchy(0,10)	$\sigma \sim$ normal(0,10)	0.4015	0.1568	0.0884	0.4017	0.7176
$\beta \sim$ normal(0,1)	$\sigma \sim$ uniform(0,1)	0.3907	0.1524	0.0793	0.3924	0.6932
$\beta \sim$ cauchy(0,1)	$\sigma \sim$ uniform(0,1)	0.3854	0.1518	0.0746	0.3874	0.6843
$\beta \sim$ normal(0,10)	$\sigma \sim$ uniform(0,10)	0.3993	0.1575	0.0832	0.3996	0.7147
$\beta \sim$ cauchy(0,10)	$\sigma \sim$ uniform(0,10)	0.4000	0.1563	0.0854	0.4004	0.7123

TABLE A4
ALTERNATIVE FUNCTIONAL FORMS ON PRIORS: OVERCONFIDENCE, WOMEN

Model priors		$\hat{\beta}^p$	$\hat{\sigma}^p$	2.5%	50%	97.5%
$\beta \sim$ normal(0,1)	$\sigma \sim$ normal(0,1)	0.3524	0.1541	0.0382	0.3538	0.6570
$\beta \sim$ cauchy(0,1)	$\sigma \sim$ normal(0,1)	0.3472	0.1526	0.0357	0.3490	0.6474
$\beta \sim$ normal(0,10)	$\sigma \sim$ normal(0,1)	0.3601	0.1562	0.0444	0.3604	0.6712
$\beta \sim$ cauchy(0,10)	$\sigma \sim$ normal(0,1)	0.3617	0.1557	0.0499	0.3621	0.6727
$\beta \sim$ normal(0,1)	$\sigma \sim$ normal(0,10)	0.3906	0.1551	0.0769	0.3923	0.6954
$\beta \sim$ cauchy(0,1)	$\sigma \sim$ normal(0,10)	0.3837	0.1540	0.0680	0.3864	0.6850
$\beta \sim$ normal(0,10)	$\sigma \sim$ normal(0,10)	0.3952	0.1681	0.0613	0.3993	0.7107
$\beta \sim$ cauchy(0,10)	$\sigma \sim$ normal(0,10)	0.4015	0.1568	0.0884	0.4017	0.7176
$\beta \sim$ normal(0,1)	$\sigma \sim$ uniform(0,1)	0.3907	0.1524	0.0793	0.3924	0.6932
$\beta \sim$ cauchy(0,1)	$\sigma \sim$ uniform(0,1)	0.3854	0.1518	0.0746	0.3874	0.6843
$\beta \sim$ normal(0,10)	$\sigma \sim$ uniform(0,10)	0.3993	0.1575	0.0832	0.3996	0.7147
$\beta \sim$ cauchy(0,10)	$\sigma \sim$ uniform(0,10)	0.3341	0.1394	0.0472	0.3371	0.6044

TABLE A5
ALTERNATIVE FUNCTIONAL FORMS ON PRIORS: GENDER DIFFERENCES IN OVERCONFIDENCE

Model priors		$\hat{\beta}^p$	$\hat{\sigma}^p$	2.5%	50%	97.5%
$\beta \sim$ normal(0,1)	$\sigma \sim$ normal(0,1)	0.0936	0.0542	−0.0155	0.0942	0.1995
$\beta \sim$ cauchy(0,1)	$\sigma \sim$ normal(0,1)	0.0932	0.0542	−0.0156	0.0939	0.1988
$\beta \sim$ normal(0,10)	$\sigma \sim$ normal(0,1)	0.0938	0.0543	−0.0153	0.0944	0.1997
$\beta \sim$ cauchy(0,10)	$\sigma \sim$ normal(0,1)	0.0937	0.0543	−0.0157	0.0944	0.1997
$\beta \sim$ normal(0,1)	$\sigma \sim$ normal(0,10)	0.0941	0.0543	−0.0149	0.0946	0.2003
$\beta \sim$ cauchy(0,1)	$\sigma \sim$ normal(0,10)	0.0928	0.0541	−0.0159	0.0933	0.1981
$\beta \sim$ normal(0,10)	$\sigma \sim$ normal(0,10)	0.0939	0.0545	−0.0157	0.0945	0.2001
$\beta \sim$ cauchy(0,10)	$\sigma \sim$ normal(0,10)	0.0936	0.0543	−0.0154	0.0941	0.1996
$\beta \sim$ normal(0,1)	$\sigma \sim$ uniform(0,1)	0.0937	0.0544	−0.0159	0.0942	0.2001
$\beta \sim$ cauchy(0,1)	$\sigma \sim$ uniform(0,1)	0.0931	0.0539	−0.0152	0.0937	0.1988
$\beta \sim$ normal(0,10)	$\sigma \sim$ uniform(0,10)	0.0938	0.0542	−0.0151	0.0944	0.1996
$\beta \sim$ cauchy(0,10)	$\sigma \sim$ uniform(0,10)	0.0938	0.0544	−0.0154	0.0944	0.2001

TABLE A6
RUBIN MODEL: POSTERIOR MEAN OF OVERCONFIDENCE OF MEN AND WOMEN ACROSS SUBSAMPLES

	N	J	$\hat{\beta}^p$	$\hat{\sigma}^p$	Percentiles				
Sample					2.5	25	50	75	97.5
Extensive margin sample, men	7	2	0.512	0.098	0.313	0.455	0.512	0.569	0.707
Extensive margin sample, women	7	2	0.442	0.108	0.223	0.379	0.442	0.506	0.662
Intensive margin sample, men	17	7	0.037	0.019	-0.000	0.025	0.037	0.050	0.075
Intensive margin sample, women	17	7	0.027	0.020	-0.012	0.014	0.027	0.040	0.066
Full sample, men	24	9	0.409	0.091	0.229	0.349	0.409	0.469	0.591
Full sample, women	24	9	0.315	0.085	0.147	0.259	0.315	0.371	0.484

TABLE A7
RUBIN MODEL: POSTERIOR MEAN OF GENDER DIFFERENCES IN OVERCONFIDENCE ACROSS SUBSAMPLES

	N	J	$\hat{\beta}^p$	$\hat{\sigma}^p$	Percentiles				
Sample					2.5	25	50	75	97.5
Extensive margin sample	13	4	0.074	0.039	−0.002	0.048	0.073	0.098	0.154
Intensive margin sample	26	13	0.034	0.013	0.007	0.026	0.034	0.043	0.060
Full sample	39	16	0.127	0.037	0.054	0.103	0.127	0.151	0.202

TABLE A8
RUBIN MODEL: POSTERIOR MEAN OF GENDER DIFFERENCES IN OVERCONFIDENCE WITH CITATION-WEIGHTED S.E

Sample	Data	N	$\hat{\beta}^p$	$\hat{\sigma}^p$	Percentiles				
					2.5	25	50	75	97.5
Extensive margin sample	Original s.e.	13	0.074	0.039	−0.002	0.048	0.073	0.098	0.154
Extensive margin sample	Citation-adjusted s.e.	13	0.080	0.041	−0.001	0.054	0.079	0.105	0.162
Intensive margin sample	Original s.e.	26	0.034	0.013	0.007	0.026	0.034	0.043	0.060
Intensive margin sample	Citation-adjusted s.e.	26	0.029	0.018	−0.006	0.017	0.029	0.040	0.064

TABLE A9
RUBIN MODEL: POSTERIOR MEAN OF GENDER DIFFERENCES IN OVERCONFIDENCE WITH ALTERNATIVE PRIORS

Prior on β	N	$\hat{\beta}^p$	$\hat{\sigma}^p$	Percentiles				
				2.5	25	50	75	97.5
$\beta \sim N(0,\ 1)$	39	0.127	0.037	0.054	0.103	0.127	0.151	0.202
$\beta \sim N(0,\ 0.078^2)$	39	0.003	0.006	0.061	0.106	0.130	0.153	0.203
$\beta \sim N(1.46,\ 0.078^2)$	39	1.463	0.006	1.451	1.459	1.463	1.467	1.475
$\beta \sim N(1.46,\ 1)$	39	0.129	0.035	0.058	0.106	0.130	0.152	0.201