Gender differences in altruism: a Bayesian hierarchical analysis of dictator games

Draft. Last updated: Sept 2021

Michelle Rao

Abstract

I aggregate evidence on gender differences in dictator game giving from experiments published in all working papers and peer-reviewed journals since 1990. Using a two-stage Bayesian hierarchical model, I find that on average women give around 3 percentage points more than men in studies of dictator games. I show that while this estimate is smaller than that found in previous studies, it is likely to be an upper bound estimate due to publication bias. Using a truncated selectivity model, I estimate the conditional probability of publication as a function of experiment results. My findings suggest that experiments that find positive results (i.e. women contribute more than men) and are statistically different from zero at the 5% level, are around 13 times more likely to be published than statistically significant and negative results.

1 Introduction

Are women more altruistic than men? Given that assumptions on preferences are central to models of individual choice, gender differences in altruism would have far-reaching implications for theoretical and empirical work in economics. For instance, differences in altruism could explain differences in labour market outcomes between men and women, including in wages and occupational choice (Bertrand, 2011; Buser et al., 2014). Recent evidence suggests that these differences also matter at the aggregate level, with gender differences in altruism predicting economic development and gender equality across countries (Falk et al., 2016; Falk and Hermle, 2018). With this motivation in mind, a large body of evidence measures altruism using lab and field experiments. Yet the overall findings from this literature are ambiguous and inconclusive.

In this paper, I study whether women are more altruistic than men by aggregating evidence from first-mover behaviour in dictator games. I collect data on gender differences in dictator game behaviour from all working papers and journals, regardless of whether or not gender was the main topic of interest. My sample covers results on gender differences in giving from 100 dictator games across 35 studies and represents the decisions of 20,265 participants. Considered individually, the conclusions from these experiments are mixed. While Eckel and Grossman (1998), Andreoni and Vesterlund (2001), and Boschini et al. (2018) find that women give more than men when the price of giving is one, Bolton and Katok (1995), Ben-Ner et al. (2004) and Cadsby et al. (2010) find limited evidence for gender differences. Extrapolating a general finding from these studies is difficult since differences in results are likely driven by both sampling variation and genuine variation in experimental design and characteristics.

Using a Bayesian hierarchical model, I quantify the overall giving of women relative to men in dictator games. Compared with classical approaches to meta-analysis, the Bayesian hierarchical model allows me to jointly estimate the overall gender differences in dictator game giving, and the heterogeneity across studies. This allows me to separate between within-study and across-study variation, and consequently,

to estimate the extent to which findings from one study can help us learn about the overall population effect. My approach complements the growing literature in economics that uses Bayesian hierarchical models to aggregate findings across contexts (e.g. Burke et al., 2015; Bandiera et al., 2016; Vivalt, 2016; Meager, 2019).

My findings suggest that women give 3.2 percentage points more of their endowment than men, with 95 percent probability that the true mean is between 1.7 and 4.5 percentage points. Using pooling metrics suggested by Gelman et al. (2006), I find that on average 89 percent of the heterogeneity in effects across studies is explained by sampling variation. Thus, genuine heterogeneity across studies is low and each additional study is likely to be informative of the overall population effect.

I then turn my attention to exploring publication bias. The results from the Bayesian hierarchical model can be interpreted as a best estimate of gender differences in giving, within the context of dictator game experiments that report results on gender and are available in working papers and journals. The extent to which these findings generalise to a broader sample relies on how representative published papers are of dictator game giving in the overall population. If, for instance, papers that find that women give more than men are more likely to be published, then the estimated result from the Bayesian hierarchical model is likely to be an overestimate of the general population effect.

Using a truncated selectivity model, I parametrically estimate the conditional probability of publication, where following Andrews and Kasy (2019), I assume that the publication decisions of researchers and editors are a function of the study results. My results are strongly suggestive of selective publication. Overall, papers that find a significant and positive (i.e. women give more than men) result are over 13 times more likely to be published than papers that find a statistically significant and negative result.

I find evidence that the selection rule is complex, and differs by the topic of the paper and the quality of the journal. Among papers that explicitly study gender, I find evidence for selection based on statistical significance, but not on the sign of the effect. Among high-quality peer-reviewed journals,[1] I find that positive and

[1]I measure high-quality as journals with 5-year average impact factors ranked in the top two

significant results are more likely to be published than positive and insignificant results.

Taken together, the findings from the Bayesian hierarchical model suggest that women give more than men in dictator games in the context of studies where gender results are made available. While this result is smaller than that found in existing studies, it is likely to be an upper bound of the overall population effect since results that are positive and significant are more likely to be published.

My findings relate to the literature on gender differences in social preferences by aggregating findings from dictator games. Existing review articles provide a qualitative assessment of the literature (e.g. Eckel and Grossman, 2008; Croson and Gneezy, 2009). I contribute to these findings by estimating the average differences in giving and quantifying the likely heterogeneity across studies. Relatedly, Engel (2011) uses classical meta-analysis techniques to analyse findings from dictator game experiments. However whereas Engel (2011) separately estimates the average effect and the cross-study heterogeneity and therefore likely underestimates heterogeneity (Rubin, 1981), I am able to jointly estimate these two variables of interest using Bayesian hierarchical methods.

Finally, I highlight a new reason for the gender gap in giving observed in existing papers: publication bias. My findings here are related to the growing economics literature on publication bias (Simonsohn et al., 2014; Brodeur et al., 2016; Andrews and Kasy, 2019). Various other reasons for gender differences in dictator games have been suggested in the literature, including the price of giving (Andreoni and Vesterlund, 2001), gender priming (Boschini et al., 2018), and anonymity of decision-making (Dufwenberg and Muren, 2006). I stress that while these experimental differences are potential sources of gender differences in giving, in the presence of selective publication, findings from the literature are likely to overestimate the differences in the overall population.

quartiles in the Annual Journal Citation Reports.

2 Data and Context

2.1 Selection of studies

To study gender differences in altruism, I focus my attention on behaviour in dictator games. Introduced by Forsythe et al. (1994) and Kahneman et al. (1986), the dictator game is a lab experiment involving two players, often referred to as the proposer and the recipient. The proposer is given a sum of money and decides what proportion of the money to offer to the recipient, versus what proportion to keep for themselves. For rational and purely self-interested agents, the subgame perfect Nash equilibrium of the dictator game is for proposers to keep the entire sum of money for themselves (i.e. to offer zero). Thus, a positive offer in the first stage is often interpreted as evidence for altruistic preferences.

While altruism is sometimes measured using other lab experiments, such as public goods games or ultimatum games, the dictator game is arguably the cleanest experimental measure for altruism because it is simple and involves limited strategic interactions (Camerer and Fehr, 2004; Eckel and Grossman, 2008). Relatedly, first-mover behaviour in the dictator game remains one of the most prominent measures for altruism in the lab, and has been shown to predict individual and aggregate economic outcomes (Becker et al., 2011; Falk et al., 2016; Falk and Hermle, 2018).

I collect data on all relevant dictator games published in working papers and journals published up until the end of 2019 (when this data was collected). I use two data sources and approaches for compiling relevant papers depending on whether the paper was published prior to or post 2010.

For papers published from 2010 onwards, I conduct a keyword search of "dictator game" and the phrases, "altruism", "generosity", "philanthropy", or "intergenerational transfers" on two databases, EconLit and RePEc. This leaves me with a total of 328 unique papers from journals and working papers published from 2010 to 2019.

To narrow down the search to a relevant sample, I focus my study on non-interactive, one-stage dictator games. Common variants of dictator games include giving recipient power, adding multiple stages, or requiring effort to generate the

endowment. As such variants are often designed to measure preferences other than altruism (e.g. risk preferences, reciprocity) I do not include them in my sample.

Of the relevant experimental designs, I include papers in my study if the authors report the average giving in a dictator game of men and women (or the difference between the two) and their associated standard errors; or if the full data is provided, such that these values can be calculated. Note that while most experiments collect data on gender, contributions disaggregated by gender are often not reported. In fact, 6 studies in my search state that there are no gender differences in dictator game giving observed in their experiments, but do not state the average differences in giving, or the standard errors of these differences. Since these papers have to be excluded from my sample, the results from this study are likely to be an overestimate of the true treatment effect. The selection criteria are summarised in Table 1.

Table 1: Selection Criteria

	Selection Criteria	Description
1	Keywords	Dictator Game AND altruism, philanthropy, generosity, or intergenerational transfers
2	Experimental Design	Focus on non-interactive, single stage dictator games. Exclude sequential or multidimensional dictator games; games which give recipient power; games which require effort to generate the endowment
3	Results Reported	Results on either (1) Average contributions of men and women or the gender differences in contributions, and the associated standard errors, or (2) Raw data to calculate.

Of the 328 papers identified in my search, 23 of the papers report average giving in a dictator game of men versus women (with corresponding standard errors); and 4 provide raw data on their experiments, which allows me to calculate the required results. This leaves me with 27 relevant papers for papers published from 2010.

For papers published prior to 2010, I use data from Engel (2011)'s meta-analysis

of dictator games, which includes all working and published papers on dictator games available on EconLit and RePEC up to the end of 2009. Of the 131 papers in his full dataset, 14 studies provide data on average giving of women versus men in dictator games, and their associated standard errors. I review the studies using the selection criteria defined in Table 1, and validate the data against the final reported treatment effects of the papers. Of the 14 papers included in Engel (2011), I exclude the results from 6 papers: 2 papers that have since released new versions or have been published in journals after 2009 (and hence are already in my sample); and 4 papers, since they do not meet my selection criteria.

2.2 Summary statistics

My final sample comprises results from 35 studies over a total of 100 experiments. A summary of these results is provided in Table 2. On aggregate, the experiments in my sample cover 20,265 distinct allocation decisions in dictator games, of which 53% are decisions by women. On average women contribute 2.7 percentage points more than men, with men contributing 29.7% and women contributing 32.4% of their endowment. The average giving in my sample irrespective of gender is 31.1%, which is broadly consistent with the average giving found in the literature (for instance, Engel (2011) finds an average giving of 28.35%).

Table 2: Average contributions by gender, % stake size

	N	Mean	St. Dev.	Min	Max
Average contribution of men	100	0.297	0.217	0.000	1.052
Average contribution of women	100	0.324	0.208	0.000	1.131
Gender difference in contribution	100	0.027	0.097	−0.292	0.465

Notes: Gender difference in contribution defined as the percentage point difference in contribution of women relative to men. Positive gender difference corresponds to women giving more than men. A contribution of more than 1 corresponds to experiments in which the price of giving is less than 1 (see: Andreoni and Vesterlund, 2001, for an example).

As is common in lab experiments, most studies in my sample have multiple variants of the dictator game within the same paper in an attempt to disentangle how different experimental characteristics may affect average giving (see: Table 3). Common variants of the dictator game include variation in the price of giving (e.g. Andreoni and Vesterlund, 2001; Visser and Roelofs, 2011); the anonymity of the dictator and identity of the recipient (e.g. Cadsby et al., 2010; Dufwenberg and Muren, 2006; Slonim and Garbarino, 2008); gender priming (e.g. Boschini et al., 2012, 2018); and the framing of the game (e.g. Smith, 2015; van Rijn et al., 2019). There is also variation in the characteristics across studies including differences in location and subject population. 20 out of 38 studies in my sample explicitly mention gender (or a related term) as the main topic of their paper. The majority of the experiments are conducted among a population of university students (27 out of 38 studies).

To control for quality, I use a subset of this full sample that are published in *RelevantJournals*[2] for my baseline analysis. In particular, I include papers published in the top 5 economics peer-reviewed journals and the main field journals in behavioural and experimental economics. I also include papers published in the NBER Working Papers series, the IZA Discussion Papers, and the CEPR Discussion Paper series. This leaves me with results from 83 experiments across 29 studies.

In Figure 1, I plot the average contribution of men versus women in dictator games, disaggregated by journal type. Results are closely distributed around the 45 degree line, with marginally more study estimates finding higher contributions by women as compared to men. Results from experiments that are not from my *RelevantJournals* list tend to be noisier and at more extreme values than that found for results published in journals on my list. In Section 6, I explore these relationships more systematically, by estimating how the type of results published may differ by the characteristics of the study.

[2]Full list of Relevant Journals: American Economic Review, Econometrica, Journal of Political Economy, The Quarterly Journal of Economics, Review of Economic studies, Journal of Behavioral and Experimental Economics, Experimental Economics, Journal of Economic Behavior and Organisation, Games and Economic Behavior, Economic Journal, American Economic Journal: Applied Economics, Journal of Economic Psychology, Management Science, NBER Working Paper series, IZA Discussion Papers, CEPR Discussion Paper series

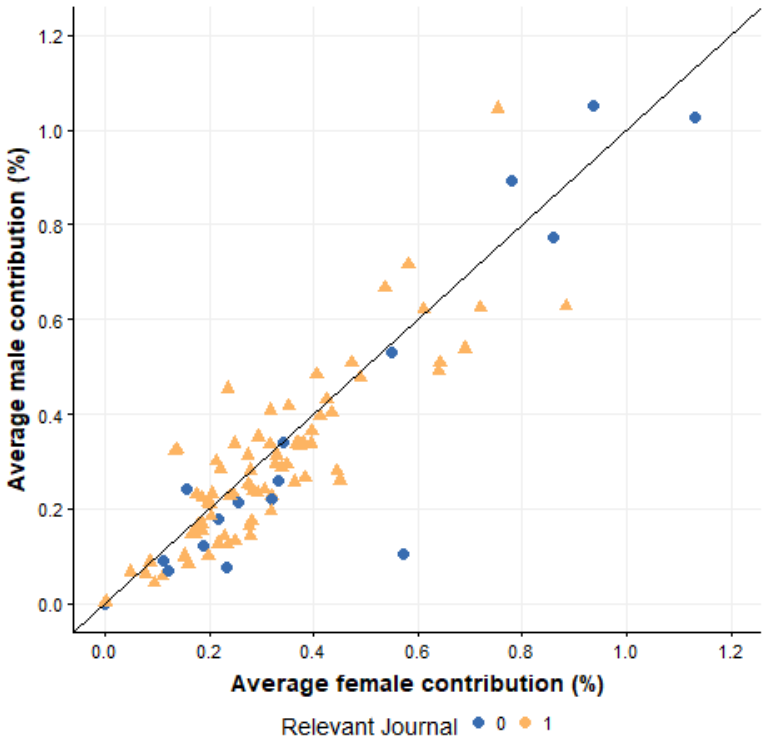

Figure 1: Average contributions of women versus men (% of stake size), by journal type. *RelevantJournals* defined as papers published in the top 5 economics journals, the field journals in behavioral and experimental economics, and the main working paper series (NBER, IZA, and CEPR). See Footnote 2 for full list.

Table 3: Study and experiment characteristics, summary

Study	Relevant Journal[1]	Gender main topic[2]	Number of observations	Share women	Number of relevant experiments	Source of variation in experiments, if multiple
Aguiar et al. (2009)	Yes	Yes	40	0.50	1	
Alevy et al. (2014)	No	Yes	219	0.50	4	Anonymity of dictator, framing
Andreoni and Vesterlund (2001)	Yes	Yes	1136	0.33	8	Price of giving, size of pie
Baltrusch and Wichardt (2018)	No	Yes	1016	0.27	2	Identity of recipient
Ben-Ner et al. (2017)	Yes	No	293	0.67	2	Anonymity of dictator
Berge et al. (2015)	Yes	No	4048	0.60	5	Anonymity of dictator, timing of game
Bezu and Holden (2015)	Yes	Yes	724	0.50	2	Identity of recipient
Boschini et al. (2012)	Yes	Yes	1086	0.64	12	Priming (gender), price of giving
Boschini et al. (2018)	Yes	Yes	889	0.40	4	Priming (gender)
Brandstatter and Guth (2002)	No	No	51	0.61	1	
Brock et al. (2013)	Yes	No	63	0.46	1	
Cadsby et al. (2010)	Yes	Yes	699	0.49	4	Anonymity of dictator
Cason and Mui (1997)	Yes	No	188	NA[3]	1	
Castillo and Cross (2008)	Yes	Yes	107	0.41	4	Price of giving, size of pie
Chaudhry and Saleem (2011)	No	No	238	0.56	1	
Dasgupta (2011)	No	Yes	80	0.50	1	
Dufwenberg and Muren (2006)	Yes	Yes	352	0.48	2	Anonymity of dictator
Eckel and Grossman (1998)	Yes	Yes	120	0.50	1	
Gong et al. (2015)	Yes	Yes	144	0.50	2	Subject population
Grech and Nax (2020) 2019	Yes	No	4120	0.61	1	
Gummerum et al. (2010)	No	No	77	0.55	1	
Halvorsen (2015)	No	No	177	0.40	4	Framing
Heinz et al. (2012)	Yes	Yes	83	0.55	2	Size of pie
Houser and Schunk (2009)	No	Yes	151	0.47	3	Anonymity of dictator
Iida (2015)	Yes	No	168	0.30	2	Subject population
John and Thomsen (2017)	Yes	Yes	985	0.48	1	
Klinowski (2018)	No	Yes	308	0.50	1	
Lazear et al. (2012)	Yes	No	83	0.53	1	
Leibbrandt et al. (2015)	No	No	90	0.33	4	Size of pie, framing
Marlowe (2004)	No	No	43	0.49	1	
Rigdon et al. (2009)	No	No	113	0.55	2	
Saad and Gill (2001)	No	Yes	224	0.48	2	Identity of recipient
Slonim and Garbarino (2008)	Yes	No	580	0.52	2	Identity of recipient
Smith (2015)	Yes	No	144	0.21	3	Framing
Umer (2020)	Yes	No	157	0.50	2	Anonymity of dictator
van Rijn et al. (2017)	Yes	No	166	0.72	1	
van Rijn et al. (2019)	Yes	Yes	573	0.54	4	Priming (guilt)
Visser and Roelofs (2011)	Yes	Yes	530	0.65	5	Price of giving, size of pie
Overall (out of 38 studies)	25	20	20,265	0.53	100	

[1] The *RelevantJournals* are defined as papers published in the top 5 economics journals, the field journals in behavioral and experimental economics, and the main working paper series (NBER, IZA, and CEPR). See Footnote 2 for full list. [2] Gender listed as main topic if a gender related term (e.g. women, men, gender) is used in the title of the paper. [3] Cason and Mui (1997) do not report gender split of participants.

3 Methodology

3.1 Bayesian hierarchical models

In understanding gender differences in giving in the overall population, the main empirical challenge is in how we should be summarising evidence across different experiments and studies. The estimated difference in giving of women versus men ranges from -0.292 to 0.465. As illustrated in Section 2.2 however, this range in estimates could be driven by genuine variation across studies and experiments, due to differences in experimental design and setting. Or alternatively, these differences could be driven by sampling variation in the estimate, specific to either the study or the experiment. Bayesian hierarchical models provide a method to disentangle between these two sources of variation, and have been increasingly used in economics (e.g. Rubin, 1981; Bandiera et al., 2016; Vivalt, 2016; Meager, 2019). By separating between these sources of variation, the methodology allows us to obtain improved estimates of the treatment effect within each study, as well as an estimate of the overall treatment effect for the population.

Let $\hat{y}_{jk}$ denote the estimated difference in giving of women relative to men in experiment k within study j, such that a positive effect estimate $\hat{y}_{jk}$ is an instance in which the average giving of women is higher than that of men. In my sample, I have a set of $\hat{y}_{jk}$ and their associated standard errors, $\hat{se}_{jk}$, across $j = 1, 2, ..., J$ studies, where each study has $k = 1, 2, ..., K$ experiments.

To set ideas, consider the following hierarchical model for the data:

$$\begin{aligned} &\hat{y}_{jk} \sim N(y_{jk}, \hat{se^2}_{jk}) \;\; j = 1..., J, k = 1, ..., K \\ &y_{jk} \sim N(\mu, \tau^2) \end{aligned} \tag{1}$$

In this model, each experiment k within study j, obtains an estimate of the average treatment effect, $\hat{y}_{jk}$. This estimated treatment effect is normally distributed around the true mean effect of the experiment, y_{jk}, and has known variance, $\hat{se}^2_{jk}$. In turn, each experiment mean, y_{jk}, is drawn from a common distribution that is normally distributed around the population mean, μ, and variance, τ^2.

This simple model provides a structure to understand differences between the overall effect for the population, μ, and the estimated effects for a given experiment, $\hat{y}_{jk}$. It distinguishes between statistical sampling variation, as captured by $\hat{y}_{jk} - y_{jk}$, and genuine variation in the treatment effect, $y_{jk} - \mu$. The model further nests approaches used in classical meta-analysis, as well as contrasting views on how we should be aggregating results across experiments. On the one hand, when $\tau = 0$, the hierarchical model corresponds to the 'full pooling', classical fixed effects model, where we assume that each experiment is estimating a common population effect. In this case, the best estimate of the overall population mean is a weighted average of the estimated treatment effect per experiment, where each estimate is weighted by its precision $(1/\hat{se}_{jk}^2)$. At the other extreme, when $\tau = \infty$ the model corresponds to the 'no pooling' case and returns the original experiment estimates [3]. Under no pooling, we assume that each experiment is estimating its own context specific effect, and hence there is no learning to be done across studies. The hierarchical model is a compromise between these two extremes. The estimated τ^2 gives us a measure of the external validity of study results: intuitively, the smaller is τ, the more each additional experiment estimate, $\hat{y}_{jk}$, tells us about the overall population effect, and hence, the more we should be updating our estimate and beliefs of μ.

The Bayesian hierarchical model builds on the hierarchical model by treating μ and τ as random variables, and assigning distributional assumptions on these variables. This gives us several advantages beyond the hierarchical approach. In treating μ and τ as random variables, we are less likely to underestimate cross-study heterogeneity (Rubin, 1981). The Bayesian approach further allows us to obtain posterior distributions, so that we can obtain probability distributions on our parameter estimates.

The Bayesian hierarchical model is based on four key assumptions: (1) Normality of the estimated experiment effects, $\hat{y}_{jk}$, given parameters y_{jk} and $\hat{se}_{jk}$, where the variance is assumed to be known; (2) Normality of the study-specific mean, y_{jk}, given μ and τ; (3) Exchangeability of the joint distribution of $\{y_{jk}\}_{j=1,k=1}^{J,K}$; and

[3] In practice, when τ is very large (i.e. more than 5 times the standard error) there is also no pooling.

(4) Distributional assumptions on the hyperpriors, μ and τ. I elaborate upon and discuss each of these assumptions in turn.

Assumption (1), the normality of $\hat{y}_j$, follows almost directly from the assumption of internal validity of the inference within each study. Given the sample sizes are all sufficiently large, the central limit theorem justifies the normal distribution. Justification for Assumption (2), is less straight forward, but there are a number of reasons for which normality of the experiment level means is a natural assumption for this analysis. From a frequentist perspective, Efron and Morris (1977) show that under the assumption of normality, shrinkage estimators have smaller mean squared errors than estimators with full pooling. More broadly, McCulloch and Neuhaus (2011) show that inference on μ and τ under the assumption of normality is still generally reliable even when the true underlying distribution is non-normal. From a practical standpoint, the normal-normal hierarchical structure facilitates comparability of estimates with results from classical meta-analyses Gelman et al. (2013), which enables me to compare findings from my analysis to that of Engel (2011).

The third assumption required for the model is that of exchangeability. The data is exchangeable if the joint distribution of $\{y_{jk}\}_{j=1,k=1}^{J,K}$ is invariant to different permutations of the indices. That is, prior to seeing the effect estimates, there is no prior reason to believe that the average contribution of women relative to men would be larger, smaller, or of similar magnitude in any experiment or study versus that of another. In the absence of information to distinguish between the data and effect estimates, Gelman et al. (2013) argue that exchangeability is the best assumption for modelling. When data is available to distinguish between observations, we can structure the model and condition on groups and study characteristics, so that the model instead relies on conditional exchangeability, rather than full exchangeability.

In the context of this paper, there are several potential threats to exchangeability. First and foremost, within each study j, there are k experiments that each provides a distinct estimate of the treatment effect, $\hat{y}_{jk}$. Since experiments within the same study are conducted and designed by the same set of researchers, the effect estimates are likely to be subject to experimenter effects (Rosenthal, 1976). As such, the prior distribution of experiment effects, y_{jk}, within the same study j, are unlikely to be

exchangeable.

To account for this, I adopt a two-stage estimation process outlined in Model 2. In the first stage of the analysis, I obtain study-level effect estimates, $\hat{y}_j$, for all studies with more than one experiment[4]. The first stage Bayesian hierarchical model gives me an effect estimate and associated standard error for each study j. In the second stage of the analysis, I run the full Bayesian hierarchical model on the study-level estimates, $\hat{y}_j$, using either (1) the estimated treatment effect and associated standard error from the first stage if a given study j has more than one relevant experiment, $k > 1$; or (2) the estimated treatment effect and associated standard error reported in a study if the original study has just one relevant experiment ($k = 1$).

First stage: Obtain estimates for $\hat{y}_j$ for $j = 1...J$. For studies where $k = 1$, $\hat{y}_j = \hat{y}_{jk}$. For studies with $k > 1$, $\hat{y}_j$ is the Bayes estimator from:

$$\begin{aligned} \hat{y}_{jk} &\sim N(y_{jk}, \hat{se^2}_{jk}) \;\; k = 1...K \\ y_{jk} &\sim N(y_j, se^2{}_j) \end{aligned} \tag{2}$$

Second stage: Using posterior means of $\hat{y}_j$ and $\hat{se}_j^2$ from the first stage, estimate:

$$\begin{aligned} \hat{y}_j &\sim N(y_j, \hat{se^2}_j) \;\; j = 1...J \\ y_j &\sim N(\mu, \tau^2) \end{aligned}$$

Within the two-stage framework, the assumption of exchangeability now applies to the joint distribution of $\{y_{jk}\}_{k=1}^{K}$ within the same j (in the first stage), and the joint distribution of $\{y_j\}_{j=1}^{J}$ (in the second stage). In a given study, however, there are often variants of lab experiments designed explicitly to tease out gender differences in giving. For instance, in one of their experiments Boschini et al. (2018) remind respondents of their gender prior to playing the dictator game, citing findings from economics and psychology that find evidence of women being more responsive

[4]Note here that if $k = 1$ for a given j, $\hat{y}_j = \hat{y}_{jk}$.

to gender priming and gender stereotypes (e.g. Steele and Aronson, 1995; Benjamin et al., 2010). In these instances, exchangeability of effect estimates is likely to be violated since prior to seeing the data, we would expect the relative giving of women to be higher in experiments with priming, than that of experiments without. Consistent with this reasoning, I exclude all experiments that use priming (e.g. gender, guilt), and 'take' framing (as opposed to 'give')[5] from my main baseline sample.

Finally, to close the model, I specify a prior distribution for the hyperparameters (Assumption (4)). In context of the two-stage estimation, this means that I need to specify prior distributions for $(\hat{y}_j, \hat{se}_j^2)$ in the first stage, and (μ, τ) in the second stage. I use the following prior distributions:

$$\begin{aligned} y_j &\sim N(0, 0.2) \\ se_j &\sim N(0, 0.2) \\ \mu &\sim N(0, 1) \\ \tau &\sim N(0, 1) \end{aligned} \tag{3}$$

Where possible I use weakly informative priors, so that the information in the likelihood dominates and the prior distribution has minimal influence on the posterior distribution. However, as noted by Gelman et al. (2017), the prior distribution will matter for posterior inference when the data is weak. This is particularly relevant in the first stage, where we only have a limited set of experiments per study. I thus adopt a 'tighter' distributional assumption in the first stage; whereas, in the second stage, I can use a relatively weaker prior, in line with the fact that I have stronger data. In Section 5, I show that my results are robust to different prior assumptions.

[5]The 'Take' frame asks the dictator how much money they want to 'take' from the recipient, as opposed to the standard dictator game, which asks how much they want to 'Give'.

4 Baseline results

My baseline sample includes all experiments published in relevant journals, other than those that use priming (e.g. gender, guilt), and 'take' framing (as opposed to 'give'). This leaves me with 69 experiments across 29 studies. I summarise this data in Table 5. The mean gender difference in giving is smaller than that in the full sample (1.3 vs 2.7 percentage points), which follows mechanically from the fact that I have excluded experiments that are designed explicitly to accentuate these gender differences.

Table 4: Average contributions by gender, % stake size - baseline sample

	N	Mean	St. Dev.	Min	Max
Average contribution of men	69	0.289	0.185	0.001	1.045
Average contribution of women	69	0.303	0.175	0.00001	0.883
Gender difference in contribution	69	0.013	0.086	−0.292	0.257

Notes: Gender difference in contribution defined as the percentage point difference in contribution of women relative to men. Positive gender difference corresponds to women giving more than men. A contribution of more than 1 corresponds to experiments in which the price of giving is less than 1 (see: Andreoni and Vesterlund, 2001, for an example). Baseline sample includes all experiments published in $RelevantJournals$, other than those that include priming and framing. $RelevantJournals$ are defined as papers published in the top 5 economics journals, the field journals in behavioral and experimental economics, and the main working paper series (NBER, IZA, and CEPR). See Footnote 2 for full list.

I estimate the two-stage model using my baseline data. Figure 2 summarizes the posterior distribution of the estimated overall effect, μ. On average, women give 3.2 percentage points more than men in dictator games, with 95% probability that the true mean lies between 1.7 and 4.5 percentage points.

To investigate whether my results are driven by sample selection, in Table 5, I estimate the model with two other subsets of my sample: (1) the full dataset, with results from all experiments and studies, irrespective of experimental design or

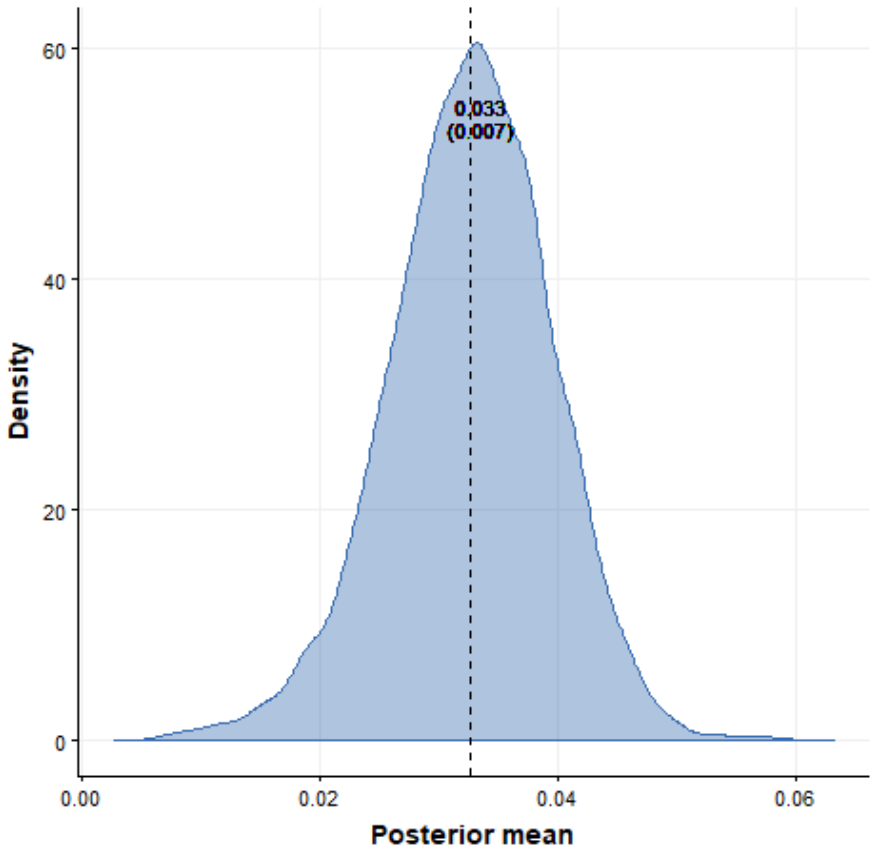

Figure 2: Posterior distribution of effect estimate

journal of publication, and (2) the 'Vanilla' subset, for which I include only standard, one-shot dictator games (i.e. where dictators and recipients are anonymous and the price of giving is equal to one) that are published in my list of relevant journals.

The estimates of the posterior effect remain reasonably stable across all three subsets of the data, and critically, the 95% intervals are positive and do not include zero for any of the subsets. Compared with previous meta-analyses, the estimated difference in contributions is noticeably smaller. For instance, using the random-effects model in a meta-analysis of dictator games, Engel (2011) finds that women give 5.8

Table 5: Posterior estimates of μ, by subsample

					Quantiles				
	N	J	Mean	s.e.	2.5%	25%	50%	75%	97.5%
Baseline	69	29	0.0323	0.0069	0.0171	0.0284	0.0328	0.0370	0.0454
Full sample	100	38	0.0319	0.0046	0.0229	0.0289	0.0319	0.0351	0.0406
Vanilla	31	23	0.0441	0.0120	0.0197	0.0369	0.0443	0.0514	0.0690

Notes: Baseline sample: includes all experiments, other than those that include priming and framing, and that are published in *RelevantJournals*. *RelevantJournals* are defined as papers published in the top 5 economics journals, the field journals in behavioral and experimental economics, and the main working paper series (NBER, IZA, and CEPR). Full sample: includes all experiments and results. Vanilla sample: includes only standard, one-shot dictator games, published in *RelevantJournals*.

percentage points more than men. However, compared to the Bayesian hierarchical model, the random-effects model treats priors, μ and τ, as fixed once estimated and hence likely underestimates cross-study heterogeneity, and overestimates the population effect (Rubin, 1981). Indeed, the estimate of 5.8 percentage points is not included in the 95% probability interval for my Baseline and Full sample datasets [6].

As alluded to in Section 3.1, the Bayesian hierarchical model gives us some indication of external validity by separating between sampling variation within studies and genuine variation across studies. Intuitively, if the genuine heterogeneity across studies is small or at the extreme, if $\tau = 0$ (corresponding to full pooling), then each study is implicitly estimating a common population effect, μ. In this case, pooling together data across studies not only improves our understanding of the common population effect, μ, but also improves our estimate for the study-specific effect, $\hat{y}_j$. In contrast, if heterogeneity across studies is large, or at the extreme if $\tau = \infty$ (corresponding to no pooling), then each study is estimating a separate independent phenomenon and should be considered in isolation. The degree of genuine variation across studies thus provides an indication of the degree to which we can generalise and learn across contexts.

[6]Note here that the Vanilla subset would not be comparable to Engel (2011), since he includes all dictator games in his sample that report gender differences in giving (and not just one-shot, standard dictator games).

As the scale of τ is difficult to interpret and compare across contexts, a common measure of cross-study heterogeneity is instead the pooling metric suggested in Gelman and Pardoe (2006), which measures the genuine variation across studies (τ^2) relative to total variation, ($\tau^2 + \hat{se}_j^2$). More specifically, the degree of pooling λ is given by:

$$\lambda = 1 - \frac{\tau^2}{\tau^2 + E(\hat{se}_j^2)}$$

where $\lambda = 1$ corresponds to the full pooling case, and $\lambda = 0$ corresponds to the no pooling case.

In Table 6, I provide the pooling metrics estimated for each of the studies in the baseline sample. For all but one study, I find that the pooling metric is greater than 0.5, suggesting that study-level estimates are being adjusted towards the population mean. The overall pooling factor across studies suggests that 89% of the heterogeneity in estimated effects is due to sampling variation. Thus, genuine heterogeneity across studies is low and each additional study is informative on the overall population effect.

The intuition of this result can also be seen graphically, in Figure 3. Here, I plot the posterior effect estimates of each study in my baseline sample, and the corresponding 95% probability intervals for a model with full pooling, partial pooling, and with no pooling. Compared to full pooling, the 95% probability intervals of the partial pooling model are larger, capturing the fact that there is some heterogeneity across studies. These bounds are much smaller than that of the original study estimates, however, suggesting that differences in effects across studies are primarily driven by sampling variation, rather than genuine variation.

Table 6: Pooling factors for each study

	Pooling factor
Overall Pooling	0.892
Aguiar et al. (2009)	0.986
Andreoni and Vesterlund (2001)	0.863
Ben-Ner et al. (2017)	0.993
Berge et al. (2015)	0.550
Bezu and Holden (2015)	0.963
Boschini et al. (2012)	0.876
Boschini et al. (2018)	0.924
Brock et al. (2013)	0.963
Cadsby et al. (2010)	0.911
Cason and Mui (1997)	0.916
Castillo and Cross (2008)	0.872
Dufwenberg and Muren (2006)	0.960
Eckel and Grossman (1998)	0.868
Gong et al. (2015)	0.981
Grech and Nax (2020)	0.489
Gummerum et al. (2010)	0.946
Heinz et al. (2012)	0.974
Houser and Schunk (2009)	0.962
Iida (2015)	0.981
John and Thomsen (2017)	0.640
Lazear et al. (2012)	0.919
Leibbrandt et al. (2015)	0.896
Rigdon et al. (2009)	0.982
Slonim and Garbarino (2008)	0.946
Smith (2015)	0.986
Umer (2020)	0.975
van Rijn et al. (2017)	0.871
van Rijn et al. (2019)	0.906
Visser and Roelofs (2011)	0.846

Notes: Pooling factors correspond to the metric suggested by Gelman and Pardoe (2006). The overall pooling factor is an arithmetic mean of the pooling factors across studies.

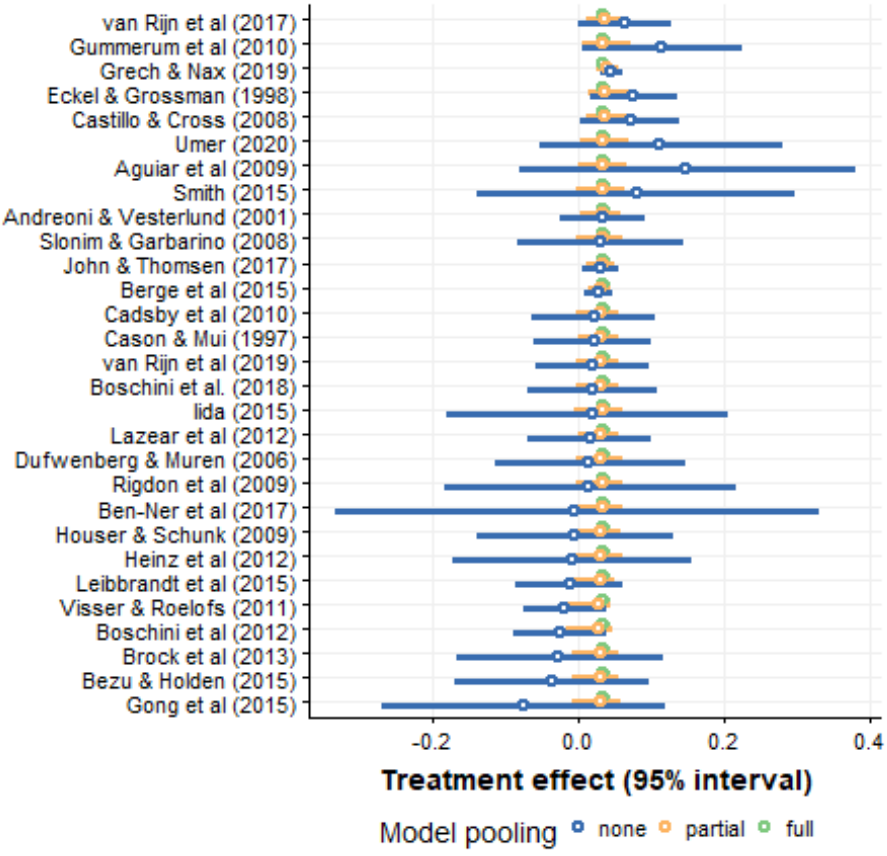

Figure 3: Model comparison - posterior effect estimates of μ by study

5 Robustness checks

The validity of posterior inference is critically dependent on the set of assumptions on the probability model, as laid out in Section 3.1. It is thus necessary to assess the fit and sensitivity of the model to these assumptions. In this section, I conduct a series of posterior predictive checks and explore the sensitivity of the analysis to different distributional assumptions on the priors.

5.1 Posterior predictive checks

If the model is suited to the setting, simulations under the posterior predictive distribution should look similar to the distribution of the true data. That is, after estimation, it should seem plausible that the data was generated with the chosen model (Gelman et al., 2013). While the use of posterior predictive checks violates the likelihood principle, in that the data is being used twice (for estimation and for model checking), Meng et al. (1994) and Gelman et al. (2013) argue that, at the very least, we should look for systematic differences between the data and simulations from the posterior predictive distribution to understand the limitations of the model.

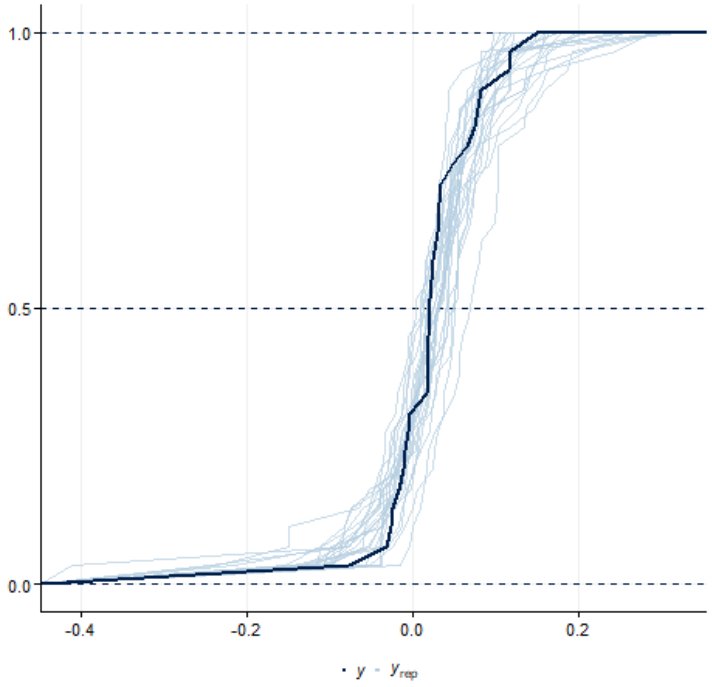

Figure 4: Cumulative density of data, y, overlaid with cumulative density of 25 simulations from the posterior predictive distribution, y^{rep}

In Figure 4, I overlay the cumulative density of the data with that of simulations

from the posterior predictive distribution. For simplicity, I suppress the subscripts and let y denote observations from my data, and y^{rep} denote simulations of the data from the posterior predictive distribution. The cumulative density of the simulated data closely resembles that of the true data, suggesting that it is plausible that the data could be generated by the model.

I further construct measures of the fit by considering a series of relevant test statistics for the posterior predictive distribution. For each test quantity, $T(\hat{y})$, I calculate the corresponding Bayesian p-value, p_b as follows:

$$p_b = Pr(T(\hat{y}^{rep}, \theta) \geq T(\hat{y}, \theta) \mid \hat{y})$$

In practice, the Bayesian p-value is calculated as the proportion of simulations from the posterior predictive distribution, for which the simulated value of the test statistic is greater than the test quantity calculated from the data. The closer is the p-value to 0 or to 1, the less likely it is that the data would be generated under the posterior predictive distribution implied by the model.

In Figure 5 I consider four test-statistics of interest: the maximum, minimum, median, and mean of study effects. I plot the posterior predictive distributions for each of these test statistics, using the value of the test statistic for 1000 simulations of the predictive data. For each of these, the Bayesian p-value is sufficiently far away from 0 and 1, which suggests that the model generates predicted values that are close to the sample data.

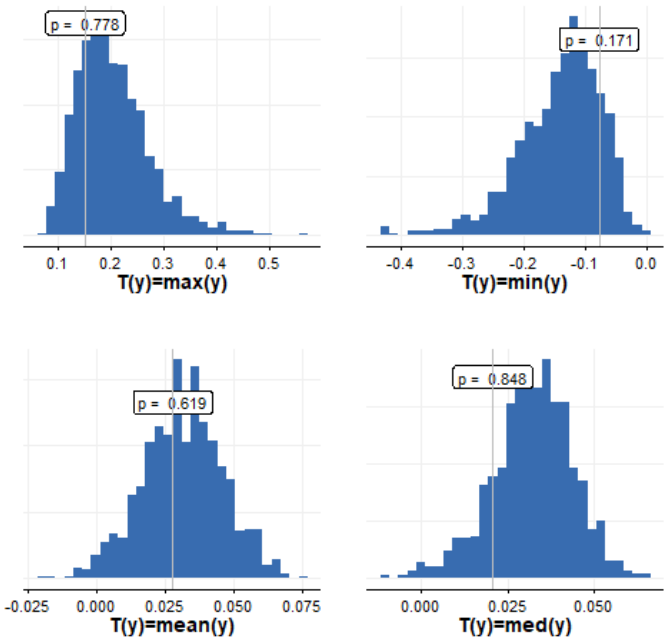

Figure 5: Posterior predictive distribution and associated p-value for four test statistics. Vertical lines denote the value of the test statistic for the data.

5.2 Prior checks

A second concern on inference is the sensitivity of results to the choice of the prior distribution. In Table 7 I explore the sensitivity of my estimates to 12 alternative choices of the prior distribution. For each of these specifications, I center the prior distribution around a zero mean, consistent with the assumption of a null effect unless proven otherwise by the data (as is the approach with hypothesis testing). The posterior mean and 95% interval for μ remain stable for the range of different distributional assumptions.

Table 7: Prior checks - estimates of posterior mean

Model Priors	Mean	s.e.	2.5%	50%	97.5%
$\mu \sim$normal(0,1); $\tau \sim$normal(0,1)	0.0324	0.0072	0.0171	0.0328	0.0454
$\mu \sim$cauchy(0,1); $\tau \sim$normal(0,1)	0.0328	0.0069	0.0184	0.0329	0.0458
$\mu \sim$normal(0,10); $\tau \sim$normal(0,1)	0.0324	0.0072	0.0172	0.0327	0.0457
$\mu \sim$cauchy(0,10); $\tau \sim$normal(0,1)	0.0324	0.0072	0.0173	0.0328	0.0455
$\mu \sim$normal(0,1); $\tau \sim$normal(0,10)	0.0326	0.0070	0.0175	0.0328	0.0455
$\mu \sim$cauchy(0,1); $\tau \sim$normal(0,10)	0.0323	0.0072	0.0171	0.0326	0.0453
$\mu \sim$normal(0,10); $\tau \sim$normal(0,10)	0.0320	0.0073	0.0159	0.0324	0.0449
$\mu \sim$cauchy(0,10); $\tau \sim$normal(0,10)	0.0325	0.0070	0.0174	0.0330	0.0456
$\mu \sim$normal(0,1); $\tau \sim$uniform(0,1)	0.0326	0.0067	0.0188	0.0329	0.0450
$\mu \sim$cauchy(0,1); $\tau \sim$uniform(0,1)	0.0323	0.0077	0.0170	0.0329	0.0451
$\mu \sim$normal(0,10); $\tau \sim$uniform(0,10)	0.0327	0.0070	0.0178	0.0329	0.0457
$\mu \sim$cauchy(0,10); $\tau \sim$uniform(0,10)	0.0325	0.0071	0.0172	0.0329	0.0459

6 Publication Bias

The results from the Bayesian hierarchical model can be interpreted as the overall gender differences in giving for dictator games, within settings for which researchers have conducted dictator games and critically, within the population of working papers and journals that publish results on gender differences in dictator game giving. The extent to which the result can be applied to our broader understanding of gender differences in altruism thus depends on the external validity of behaviour in dictator games (and lab experiments more generally), and the degree of publication bias. While there is extensive literature on the former issue (e.g. List, 2007; Levitt and List, 2007; Benz and Meier, 2008; Franzen and Pointner, 2013), I now turn my focus to exploring the extent of publication bias.

In particular, the findings from the Bayesian hierarchical model would potentially be biased in the presence of publication bias, that is, if certain types of results are systematically more likely to be published. Importantly, in exploring 'publication bias' I am unable to distinguish between the decisions of the journal and the decisions of the researcher, otherwise known as the 'file drawer' problem (Rosenthal, 1979). The issue of the 'file drawer' problem is particularly relevant to this setting, since

almost all studies of the dictator game collect data on gender, but only a select subset report the average giving of women versus men. In fact, authors of 6 studies surveyed in my data collection state explicitly that they do not find statistically significant gender differences in giving, and hence do not report the results. Are researchers more likely to report gender results if they find large differences in giving? Similarly, are editors more likely to publish results that find a large effect? In this section, I explore the extent to which this may be true.

I start the section by documenting the patterns in the distribution and variation in estimated treatment effects and standard errors across studies. Then, I follow Andrews and Kasy (2019) in estimating the conditional probability of publication using a truncated selectivity model. Under the assumption that the latent variables are independently and identically distributed, the model allows me to parametrically estimate how the probability of publication varies with the study results. Finally, I present the results and discuss the implications of this analysis.

6.1 Distribution of estimates

As a first pass for exploring the degree of publication bias, it is useful to consider the distribution of test statistics, point estimates, and standard errors of the full set of experiments. I follow Andrews and Kasy (2019) and Brodeur et al. (2016) in considering the distribution of the z-statistics (the ratio of the effect size to the standard error) above and below the 5% significance level threshold. Intuitively, absent publication bias, there should not be any bunching or jumps in the test statistics on either side of the significance thresholds.

I focus here on three subsets of the data that may be of interest. First, the $FullData$, comprising of all 100 experiments in my sample. Second, the $GenderTopic$ subset, comprising of the 65 experiments from the 20 studies that explicitly refer to a gender-related term in the title of the paper. Third, the $TopJIF$ subset, comprising of the 57 experiments in my full sample from the 19 studies published in top peer-reviewed journals. As a proxy measure for journal quality, I use the Journal Impact Factors (JIF) published in the 2019 Journal Citation Reports, which give a measure

of the impact and influence of an academic journal. I include in my $TopJIF$ the subset of results from papers published in peer-reviewed journals that are ranked in the top two quartiles of the 5 year average JIF indicators [7] .

In Figure 6, I construct a binned-density plot of the z-statistic for the full dataset, the *GenderTopic* subset, and the $TopJIF$ subset. Similar to Brodeur et al. (2016) I observe jumps in the distribution around the cutoffs for -1.96, 0, and 1.96 for the full dataset. This pattern is broadly similar for the subsets with slight differences: while for the *GenderTopic* subset, there does not appear to be a jump in the data around 1.96; for the $TopJIF$ subset, the jump in the density is noticeably smaller around zero.

Next, I construct funnel plots of the effect estimate against the standard errors in Figure 7, as suggested by Andrews and Kasy (2019). Absent publication bias, as the standard error of a study increases, the effect estimates should get noisier and be symmetrically split to the right and left of the true effect. As with the density plots, any bunching around the significance thresholds (as illustrated by the dotted lines) would again be suggestive of some degree of selective publication. As seen in Figure 7 there is a mass of effect estimates asymmetrically bunched around positive effect sizes that are statistically distinguishable from zero at the 5% level. This is seen for all three subsets, but particularly evident for the full sample, as seen in panel A.

[7] Sutter and Kocher (2001) find that the JIF rankings in economics remain stable over time: 95% of economics journals remain in the same or neighbouring quartile over a 10-year period; and there is even less variation in JIF for the Top 15 journals. Thus, although papers in my sample are published at different times, the 5-year average JIF, is likely to be a good proxy for journal quality at the time of publication.

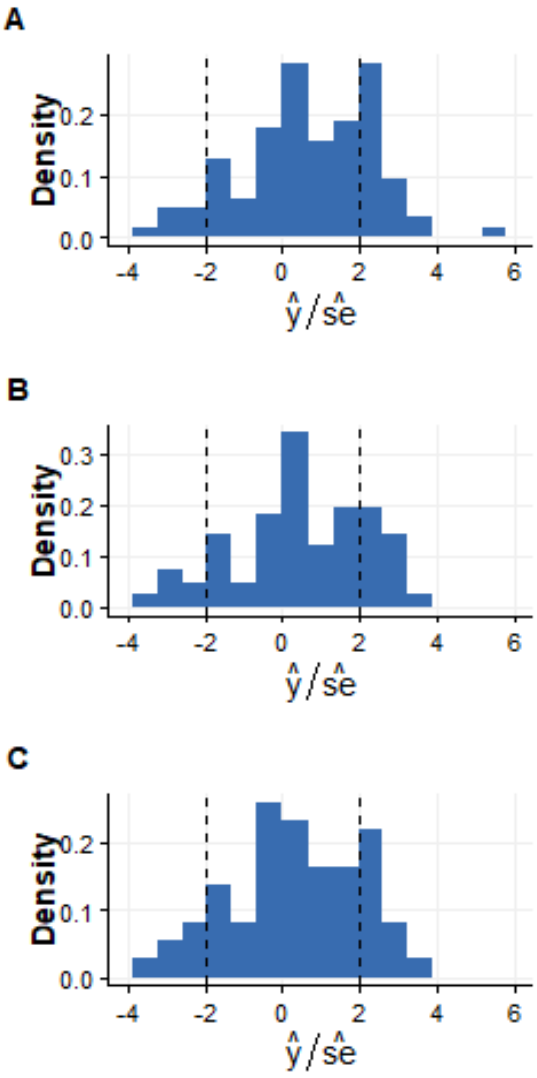

Figure 6: Binned density plot for three subsets of data. The dotted lines mark where the z-statistic is equal to 1.96 and -1.96. Panel A: full dataset; Panel B: Gender topic subset, including only observations with 'gender' in the title; Panel C: Top JIF subset, including only observations published in peer-reviewed journals placed in Q1 & Q2 of 5 year Impact Factor rankings

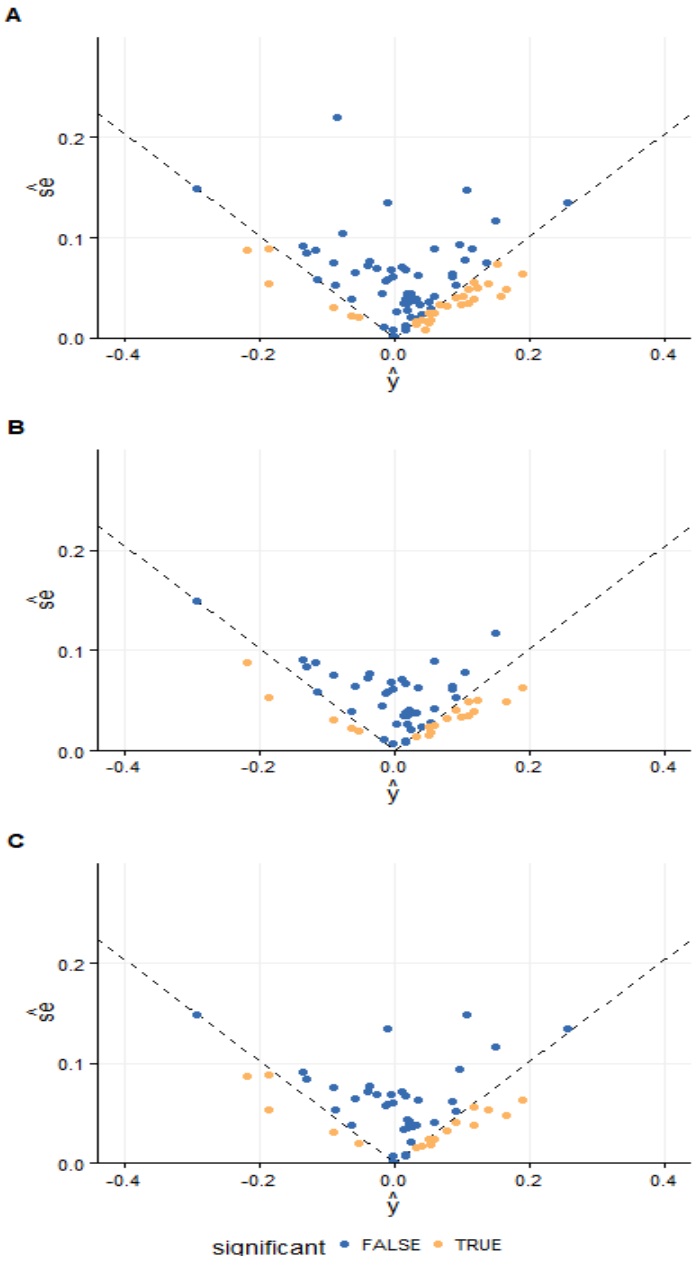

Figure 7: Scatter plot of effect estimate and standard error, by whether or not observation is statistically distinguishable from zero, at the 5% level. The dotted lines mark where $|\hat{y}/\hat{se}| = 1.96$. Panel A: full dataset; Panel B: *GenderTopic* subset, including only observations with 'gender' (or gender related term) in the title; Panel C: $TopJIF$ subset, including only observations published in peer-reviewed journals placed in Q1 & Q2 of 5 year JIF ranking

6.2 Methodology

I follow Andrews and Kasy (2019) in modelling publication bias as a truncated sampling process, in which studies are selected for publication only on the basis of the results. Let us distinguish between latent (unobserved) variables, denoted by an asterisk (e.g. $\hat{y}^*_{jk}, y^*_{jk}$), which capture the full set of experimental results; and observed variables (e.g. $\hat{y}_{jk}, y_{jk}$), which capture the subset of the latent results that are published in journals or in working papers. In particular, we observe $\hat{y}^*_{jk}$ only if $D_{jk} = 1$, that is, if the result is published.

Assume $(\hat{y}^*_{jk}, y^*_{jk}, \hat{se}^{2^*}_{jk}, D_{jk})$ are jointly iid across j and k with

$$\begin{aligned}
\hat{y}^*_{jk} &\sim N(y^*_{jk}, \hat{se}^{2^*}_{jk}) \\
y^*_{jk} &\sim N(\mu^*, \tau^{2^*}) \\
D_{jk} \mid \hat{y}^*_{jk}, y^*_{jk}, \mu^* &\sim Ber(p(Z^*)) \\
\text{where } \hat{y}_{jk} &= \begin{cases} \hat{y}^*_{jk} & \text{if } D_{jk} = 1 \\ unobserved & \text{if } D_{jk} = 0 \end{cases} \\
\text{and } p(\hat{y}_{jk}/\hat{se}_{jk}) &\propto \begin{cases} \beta_{p,1} & \hat{y}_{jk}/\hat{se}_{jk} < -1.96 \\ \beta_{p,2} & -1.96 \leq \hat{y}_{jk}/\hat{se}_{jk} < 0 \\ \beta_{p,3} & 0 \leq \hat{y}_{jk}/\hat{se}_{jk} < 1.96 \\ 1 & \hat{y}_{jk}/\hat{se}_{jk} \geq 1.96 \end{cases}
\end{aligned}$$

In the above model, an experiment is published with probability $p(Z)$, where Z is the z-statistic, calculated as the ratio of the estimated treatment effect and corresponding standard error. I assume that the probability of publication differs by the intervals of the test statistic around the 5% significant level (where the null hypothesis is a zero effect size), and allow for asymmetric selection depending on the sign of the estimated result. Relative to experiments that find a positive result that is significant and distinguishable from zero at the 5% level, positive and insignificant results are $\beta_{p,3}$ as likely to be published, negative and insignificant results are $\beta_{p,2}$ as

likely to be published, and finally, negative and significant results are $\beta_{p,1}$ as likely to be published.

Under the assumption that the latent variables are independently and identically distributed, Andrews and Kasy (2019) show that we can parametrically identify and estimate $p(z)$ up to scale. Note here that while the independence of latent variables cannot be tested by construction (since we do not observe studies that are not published), a clear violation in this setting is the fact that I observe results from multiple experiments within the same study. To account for this, I assume conditional independence and cluster standard errors by study j. In the proceeding section, I estimate the conditional probability of publication, $p(z)$, using the maximum likelihood estimation set out in Andrews and Kasy (2019).

6.3 Results

I estimate the conditional probability of publication across the three sub-samples of my data: the Full dataset, the *GenderTopic* subset; and the $TopJIF$ subset. The results are presented in Table 8.

Using the full sample of experiments, I find strong evidence of selection based on statistical significance. As seen in columns 4-6 of Table 8, positive results (where women give more than men) that are statistically distinguishable from zero at the 5% level are over 13 times more likely to be published than statistically significant negative results that find that men give more than women; and over 3 times more likely to be published than results that are negative and statistically insignificant. While the magnitude of $\beta_{p,3}$ suggests that results that are positive and statistically significant are more likely to be published than those that are positive and statistically insignificant, this difference is not significant at conventional levels.

Restricting the full sample of experiments now to papers that explicitly study gender, the *GenderTopic* subset, I find evidence for selection based on statistical significance, but not on the sign. Among dictator games that study gender differences, experiments that find a statistically significant and positive result are over 4 times more likely to be published than a negative and insignificant result, and over

Table 8: Estimates of $p(z)$, by subset

			$p(z)$			Interpretation[1]		
			(1)	(2)	(3)	(4)	(5)	(6)
	Studies	N	$\beta_{p,1}$	$\beta_{p,2}$	$\beta_{p,3}$	$1/\beta_{p,1}$	$1/\beta_{p,2}$	$1/\beta_{p,3}$
Full sample	38	100	0.074	0.312	0.705	13.514	3.205	1.418
			(0.131)	(0.298)	(0.529)			
GenderTopic	20	65	0.245	0.224	0.403	4.082	4.464	2.481
			(0.456)	(0.193)	(0.232)			
TopJIF	19	57	1.543	0.423	0.333	0.648	2.364	3.003
			(1.077)	(0.334)	(0.241)			

[1] A positive and significant result is $1/\beta_{p,1}$ more likely to be published than a negative and significant result; $1/\beta_{p,2}$ more likely to be published than a negative and insignificant result; and $1/\beta_{p,2}$ more likely to be published than a positive and insignificant result.
[2] *GenderTopic* subset, includes only observations with 'gender' (or a gender related term) in the title; $TopJIF$ subset, includes only observations published in peer-reviewed journals placed in Q1 & Q2 of the 5 year Journal Impact Factor rankings in the 2019 Journal Citation Reports.

2 times more likely to be published than a positive and insignificant result.

Selection based on statistical significance is less severe in the $TopJIF$ subset, the sample of results from peer-reviewed journals with the highest journal impact factors. Positive and significant results are around three times more likely to be published than positive and non-significant results. Compared with the two other subsets, the relative probability of publishing negative results (significant or insignificant) is higher compared with the two other subsets. In fact, the magnitude of $\beta_{p,1}$ suggests that negative and significant results are more likely to be published than positive and significant results, although this difference is not statistically significant at conventional frequentist levels.

6.4 Implications for Bayesian inference

What do these results mean for Bayesian inference? The implications for posterior inference depend on the distributional assumptions on the hyperparameters, μ and τ. Andrews and Kasy (2019) distinguish between two extreme classes of priors: un-

related parameters and common parameters[8]. Whereas under unrelated parameter priors, posterior inference is unaffected by publication bias, under common parameters priors, inference is affected and the posterior distribution would need to be adjusted using the truncated likelihood. Similarly, Yekutieli (2012) show that under 'fixed', non-informative priors, Bayesian inference needs to be adjusted for selection.

In the context of this study the hyperpriors, μ and τ, likely lie between the two extremes of unrelated and common parameters. Hence posterior inference from the two-stage model is likely to be affected by selection.

Ideally, I would quantitatively adjust the posterior effect estimates to account for selective publication. As seen in Section 6 however, the form of selection bias appears to operate in a complex way, and the conditional probability of publication differs by both the topic of the study and by the quality of the journal. Hence a blanket uniform adjustment of the posterior treatment effect is unlikely to be forthcoming.

Taken together, these results suggest that the estimate from the Bayesian hierarchical model is likely to provide an upper bound estimate of the overall effect for the wider population.

7 Conclusion

By aggregating results from dictator game experiments, I make two key contributions. First, I estimate the average gender difference in dictator game giving using a Bayesian hierarchical model that allows me to separate between sampling variation and genuine heterogeneity across studies. Second, I contribute to the interpretation of these studies, by exploring how the prevalence of publication bias affects the results available in published and working papers.

I find that given the available evidence, women give 3 percentage points more than men in dictator games. This effect is smaller than that found in the most frequently cited studies, and the estimated 95% probability interval of 1.7 to 4.5 percentage

[8]Andrews and Kasy (2019) define unrelated priors as the case in which the prior distribution is a point mass around a value; whereas common parameters priors, are such that the prior distribution assigns positive probability to point-measures of the prior.

points rules out existing estimates of the aggregate gender effect (e.g. Engel, 2011). I show that the observed gender differences are likely driven by publication bias, whereby papers are selected based on statistical significance. Thus, while the average giving of women relative to men is 3 percentage points among published results, the true effect for the wider population is likely to be smaller.

Given that lab experiments routinely collect data on gender (but may or may not report the findings), my results also highlight the importance of data transparency to facilitate comparability across studies.

While previous research argues that gender differences in dictator game giving are driven by experimental design, I show that even in the presence of contextual differences, estimates of gender differences in altruism are likely to overestimate the effect due to selective publication. Although I do not explicitly study the role of experimental characteristics in this paper, understanding the relative importance of publication bias versus experimental design would be an interesting direction for future research.

References

F. Aguiar, P. Brañas-Garza, R. Cobo-Reyes, N. Jimenez, and L. M. Miller. Are women expected to be more generous? *Experimental Economics*, 12(1):93–98, 2009.

J. E. Alevy, F. L. Jeffries, and Y. Lu. Gender-and frame-specific audience effects in dictator games. *Economics Letters*, 122(1):50–54, 2014.

J. Andreoni and L. Vesterlund. Which is the fair sex? Gender differences in altruism. *The Quarterly Journal of Economics*, 116(1):293–312, 2001.

I. Andrews and M. Kasy. Identification of and correction for publication bias. *American Economic Review*, 109(8):2766–94, 2019.

M. Baltrusch and P. C. Wichardt. Gender effects in dictator game giving: Women favour female recipients. 2018.

O. Bandiera, G. Fischer, A. Prat, and E. Ytsma. Do women respond less to performance pay? building evidence from multiple experiments. 2016.

A. Becker, T. Deckers, T. Dohmen, A. Falk, and F. Kosse. The relationship between economic preferences and psychological personality measures. 2011.

A. Ben-Ner, F. Kong, and L. Putterman. Share and share alike? gender-pairing, personality, and cognitive ability as determinants of giving. *Journal of Economic Psychology*, 25(5):581–589, 2004.

A. Ben-Ner, J. A. List, L. Putterman, and A. Samek. Learned generosity? an artefactual field experiment with parents and their children. *Journal of Economic Behavior & Organization*, 143:28–44, 2017.

D. J. Benjamin, J. J. Choi, and A. J. Strickland. Social identity and preferences. *American Economic Review*, 100(4):1913–28, 2010.

M. Benz and S. Meier. Do people behave in experiments as in the field?—evidence from donations. *Experimental economics*, 11(3):268–281, 2008.

L. I. O. Berge, K. Bjorvatn, S. Galle, E. Miguel, D. N. Posner, B. Tungodden, and K. Zhang. How strong are ethnic preferences? Technical report, National Bureau of Economic Research, 2015.

M. Bertrand. New perspectives on gender. In *Handbook of labor economics*, volume 4, pages 1543–1590. Elsevier, 2011.

S. Bezu and S. T. Holden. Generosity and sharing among villagers: Do women give more? *Journal of Behavioral and Experimental Economics*, 57:103–111, 2015.

G. E. Bolton and E. Katok. An experimental test for gender differences in beneficent behavior. *Economics Letters*, 48(3-4):287–292, 1995.

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

A. Boschini, A. Dreber, E. von Essen, A. Muren, and E. Ranehill. Gender and altruism in a random sample. *Journal of behavioral and experimental economics*, 77:72–77, 2018.

H. Brandstatter and W. Guth. Personality in dictator and ultimatum games. *Central European Journal of Operations Research*, 10(3), 2002.

J. M. Brock, A. Lange, and E. Y. Ozbay. Dictating the risk: Experimental evidence on giving in risky environments. *American Economic Review*, 103(1):415–37, 2013.

A. Brodeur, M. Lé, M. Sangnier, and Y. Zylberberg. Star wars: The empirics strike back. *American Economic Journal: Applied Economics*, 8(1):1–32, 2016.

M. Burke, S. M. Hsiang, and E. Miguel. Climate and conflict. *Annu. Rev. Econ.*, 7 (1):577–617, 2015.

T. Buser, M. Niederle, and H. Oosterbeek. Gender, competitiveness, and career choices. *The Quarterly Journal of Economics*, 129(3):1409–1447, 2014.

C. B. Cadsby, M. Servátka, and F. Song. Gender and generosity: does degree of anonymity or group gender composition matter? *Experimental economics*, 13(3): 299–308, 2010.

C. F. Camerer and E. Fehr. Measuring social norms and preferences using experimental games: A guide for social scientists. *Foundations of human sociality: Economic experiments and ethnographic evidence from fifteen small-scale societies*, 97:55–95, 2004.

T. N. Cason and V.-L. Mui. A laboratory study of group polarisation in the team dictator game. *The Economic Journal*, 107(444):1465–1483, 1997.

M. E. Castillo and P. J. Cross. Of mice and men: Within gender variation in strategic behavior. *Games and Economic Behavior*, 64(2):421–432, 2008.

T. T. Chaudhry and M. Saleem. Norms of cooperation, trust, altruism, and fairness: Evidence from lab experiments on pakistani students. *Lahore Journal of Economics*, 16, 2011.

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

U. Dasgupta. Do procedures matter in fairness allocations? experimental evidence in mixed gender pairings. *Economics Bulletin*, 31(1):820–829, 2011.

M. Dufwenberg and A. Muren. Generosity, anonymity, gender. *Journal of Economic Behavior & Organization*, 61(1):42–49, 2006.

C. C. Eckel and P. J. Grossman. Are women less selfish than men?: Evidence from dictator experiments. *The economic journal*, 108(448):726–735, 1998.

C. C. Eckel and P. J. Grossman. Differences in the economic decisions of men and women: Experimental evidence. *Handbook of experimental economics results*, 1: 509–519, 2008.

B. Efron and C. Morris. Stein's paradox in statistics. *Scientific American*, 236(5): 119–127, 1977.

C. Engel. Dictator games: A meta study. *Experimental Economics*, 14(4):583–610, 2011.

A. Falk and J. Hermle. Relationship of gender differences in preferences to economic development and gender equality. *Science*, 362(6412), 2018.

A. Falk, A. Becker, T. J. Dohmen, D. Huffman, and U. Sunde. The preference survey module: A validated instrument for measuring risk, time, and social preferences. 2016.

R. Forsythe, J. L. Horowitz, N. E. Savin, and M. Sefton. Fairness in simple bargaining experiments. *Games and Economic behavior*, 6(3):347–369, 1994.

A. Franzen and S. Pointner. The external validity of giving in the dictator game. *Experimental Economics*, 16(2):155–169, 2013.

A. Gelman and I. Pardoe. Bayesian measures of explained variance and pooling in multilevel (hierarchical) models. *Technometrics*, 48(2):241–251, 2006.

A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin. *Bayesian data analysis*. CRC press, 2013.

A. Gelman, D. Simpson, and M. Betancourt. The prior can often only be understood in the context of the likelihood. *Entropy*, 19(10):555, 2017.

A. Gelman et al. Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian analysis*, 1(3):515–534, 2006.

B. Gong, H. Yan, and C.-L. Yang. Gender differences in the dictator experiment: evidence from the matrilineal mosuo and the patriarchal yi. *Experimental economics*, 18(2):302–313, 2015.

P. D. Grech and H. H. Nax. Rational altruism? on preference estimation and dictator game experiments. *Games and Economic Behavior*, 119:309–338, 2020.

M. Gummerum, Y. Hanoch, M. Keller, K. Parsons, and A. Hummel. Preschoolers' allocations in the dictator game: The role of moral emotions. *Journal of Economic Psychology*, 31(1):25–34, 2010.

T. U. Halvorsen. Are dictators loss averse? *Rationality and Society*, 27(4):469–491, 2015.

M. Heinz, S. Juranek, and H. A. Rau. Do women behave more reciprocally than men? gender differences in real effort dictator games. *Journal of Economic Behavior & Organization*, 83(1):105–110, 2012.

D. Houser and D. Schunk. Social environments with competitive pressure: Gender effects in the decisions of german schoolchildren. *Journal of Economic Psychology*, 30(4):634–641, 2009.

Y. Iida. Task-based income inequalities and redistribution preferences: A comparison of china and japan. *Journal of Behavioral and Experimental Economics*, 55:91–102, 2015.

K. John and S. L. Thomsen. Gender differences in the development of other-regarding preferences. 2017.

D. Kahneman, J. L. Knetsch, and R. H. Thaler. Fairness and the assumptions of economics. *Journal of business*, pages S285–S300, 1986.

D. Klinowski. Gender differences in giving in the dictator game: the role of reluctant altruism. *Journal of the Economic Science Association*, 4(2):110–122, 2018.

E. P. Lazear, U. Malmendier, and R. A. Weber. Sorting in experiments with application to social preferences. *American Economic Journal: Applied Economics*, 4 (1):136–63, 2012.

A. Leibbrandt, P. Maitra, and A. Neelim. On the redistribution of wealth in a developing country: Experimental evidence on stake and framing effects. *Journal of Economic Behavior & Organization*, 118:360–371, 2015.

S. D. Levitt and J. A. List. What do laboratory experiments measuring social preferences reveal about the real world? *Journal of Economic perspectives*, 21(2): 153–174, 2007.

J. A. List. On the interpretation of giving in dictator games. *Journal of Political economy*, 115(3):482–493, 2007.

F. W. Marlowe. What explains hadza food sharing. *Research in economic Anthropology*, 23(4):69–88, 2004.

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

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

X.-L. Meng et al. Posterior predictive p-values. *The annals of statistics*, 22(3): 1142–1160, 1994.

M. Rigdon, K. Ishii, M. Watabe, and S. Kitayama. Minimal social cues in the dictator game. *Journal of Economic Psychology*, 30(3):358–367, 2009.

R. Rosenthal. Experimenter effects in behavioral research. 1976.

R. Rosenthal. The file drawer problem and tolerance for null results. *Psychological bulletin*, 86(3):638, 1979.

D. B. Rubin. Estimation in parallel randomized experiments. *Journal of Educational Statistics*, 6(4):377–401, 1981.

G. Saad and T. Gill. The effects of a recipient's gender in a modified dictator game. *Applied Economics Letters*, 8(7):463–466, 2001.

U. Simonsohn, L. D. Nelson, and J. P. Simmons. P-curve: a key to the file-drawer. *Journal of experimental psychology: General*, 143(2):534, 2014.

R. Slonim and E. Garbarino. Increases in trust and altruism from partner selection: Experimental evidence. *Experimental Economics*, 11(2):134–153, 2008.

A. Smith. On the nature of pessimism in taking and giving games. *Journal of Behavioral and Experimental Economics*, 54:50–57, 2015.

C. M. Steele and J. Aronson. Stereotype threat and the intellectual test performance of African Americans. *Journal of personality and social psychology*, 69(5):797, 1995.

M. Sutter and M. G. Kocher. Tools for evaluating research output: Are citation-based rankings of economics journals stable? *Evaluation review*, 25(5):555–566, 2001.

H. Umer. Revisiting generosity in the dictator game: Experimental evidence from pakistan. *Journal of Behavioral and Experimental Economics*, 84:101503, 2020.

J. van Rijn, B. Barham, and R. Sundaram-Stukel. An experimental approach to comparing similarity-and guilt-based charitable appeals. *Journal of behavioral and experimental economics*, 68:25–40, 2017.

J. van Rijn, E. J. Quiñones, and B. L. Barham. Empathic concern for children and the gender-donations gap. *Journal of Behavioral and Experimental Economics*, 82: 101462, 2019.

M. S. Visser and M. R. Roelofs. Heterogeneous preferences for altruism: Gender and personality, social status, giving and taking. *Experimental Economics*, 14(4): 490–506, 2011.

E. Vivalt. How much can we generalize from impact evaluations? 2016.

D. Yekutieli. Adjusted bayesian inference for selected parameters. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(3):515–541, 2012.