



Are women more generous than men? A meta-analysis

David Bilén¹ · Anna Dreber^{2,3} · Magnus Johannesson²

Received: 14 October 2020 / Revised: 11 August 2021 / Accepted: 24 August 2021 /
Published online: 4 September 2021
© The Author(s) 2021

Abstract

We perform a meta analysis of gender differences in the standard windfall gains dictator game (DG) by collecting raw data from 53 studies with 117 conditions, giving us 15,016 unique individual observations. We find that women on average give 4 percentage points more than men (Cohen's $d = 0.16$), and that this difference decreases to 3.1% points (Cohen's $d = 0.13$) if we exclude studies where dictators can only give all or nothing. The gender difference is larger if the recipient in the DG is a charity, compared to the standard DG with an anonymous individual as the recipient (a 10.9 versus a 2.3% points gender difference). These effect sizes imply that many individual studies on gender differences are underpowered; the median power in our sample of standard DG studies is only 9% to detect the meta-analytic gender difference at the 5% significance level. Moving forward on this topic, sample sizes should thus be substantially larger than what has been the norm in the past.

Keywords Dictator game · Altruism · Gender difference · Meta-analysis

JEL codes C91 · J16 · D64

✉ David Bilén
david.bilen@economics.gu.se

Anna Dreber
anna.dreber@hhs.se

Magnus Johannesson
magnus.johannesson@hhs.se

¹ Department of Economics, University of Gothenburg, Gothenburg, Sweden

² Department of Economics, Stockholm School of Economics, Stockholm, Sweden

³ Department of Economics, University of Innsbruck, Innsbruck, Austria

1 Introduction

To what extent are there gender differences in altruism or prosocial behavior? The answer to this question could be important for understanding for instance gender differences on the labor market, voting, volunteer work, and charitable giving. In this paper we do a meta-analysis of gender differences in the dictator game (DG; Kahneman et al. 1986, Forsythe et al. 1994). While giving in this game is not necessarily due to altruistic concerns—for example the results of List (2007), Krupka and Weber (2013), Bardley (2008) and Dana et al. (2006) suggest that DG giving is influenced by the strategy space, reference points and expectations of social norms—this is the most commonly studied game to understand non-strategic prosocial behavior. In the standard DG, one individual—the dictator— anonymously decides how to split a windfall endowment with another individual. In an alternative version of the DG, the dictator decides how much of the endowment to give to a charity (charity DG) (Eckel and Grossman 1996).

We include both of these standard windfall gains versions of the DG and collect raw data from 53 papers, both published work and working papers, with a total of 117 conditions and 15,016 unique individual observations. We only include experiments where monetary endowments are windfalls, the most selfish option is to give nothing, the price of giving is equal to 1, both men and women are represented, and where no reciprocity is involved.

There are several previous papers studying gender differences in DG giving (with early papers including e.g. Eckel and Grossman (1998), Bolton and Katok (1995), Andreoni and Vesterlund (2001), and with more recent reviews including e.g. Croson and Gneezy (2009) and Bertrand et al. (2010)), with individual studies typically finding that women on average give more or that there is no statistically significant gender difference. Most related to our study is the extensive meta-analysis of DG giving by Engel (2011). Engel, using reported coefficients rather than raw data, found that women on average gave 5.8 percentage points more than men.

Analyzing raw data allows us to include studies that collected gender information but either did not include gender in their analysis, or included gender but did not explicitly publish any results related to gender in the paper. Including studies where gender is not the main variable of interest could reduce publication bias; as statistically significant gender differences may be more likely to be published. Compared to Engel (2011), we have a substantially larger sample size for estimating the gender difference and we also explicitly compare DGs where the recipient is a person or a charity.

The paper is organized in the following way. Section 2 describes the inclusion criteria and data. Section 3 describes the meta-analysis methods and Sect. 4 the results. Section 5 concludes.

Table 1 Exclusion criteria used in the meta-analysis

Domain	Inclusion criteria
Source of Endowment	Windfall
Anonymity	Single-blind or Double-blind
Recipient	Anonymous individual or Charity organisation
Price of giving	1
Experimental setting	No restriction (Lab, Field, Online etc)
Reciprocity	No
Endowment	Monetary
Payment scheme	Deterministic or random
Choice set	The dictator can keep (give 0) or give the full endowment
Age of participant	No restriction
Genders represented in experiment	Both

2 Inclusion criteria and data

2.1 Inclusion criteria

Our inclusion criteria are summarized in Table 1. We restrict our analysis to the original windfall version of the dictator game (see Forsythe et al. 1994), where the experimenter unconditionally transfers an endowment to dictators and dictators decide how much of the endowment to give to recipient players. This restriction excludes experiments where participants first earned their endowment from performing a task. We also exclude versions of the Take Game, where the dictators also have the option to take money from recipients. Both double-blind (where neither the recipient nor the experimenter can identify individual dictator decisions) and single-blind studies (where only the recipient is blind to individual dictator decisions) are included. In the initial stages of the project we had planned to only include double-blind studies, but this was revised when we realized that our sample would be too small (around 75% of our sample consists of single-blind studies). We exclude studies without a monetary endowment, but have no further restriction with respect to the size of the monetary endowment and we also allow conditions where only a randomly drawn share of participants are paid. We exclude studies where participants are matched and known to each other (for example spouses).

There is no participant age restriction, but as we require monetary incentives we exclude studies on young children. Only studies where the price of giving is equal to 1 are included (excluding studies with multipliers). We also limit the inclusion to conditions where there is no reciprocity involved, while we allow individuals to play the DG in both roles.

There is also variation between studies in the choice set of dictators. To give an example, endowments may be 4\$ and dictators can give x \$ to the recipient,

where $x \in \{0, 1, 2, 3, 4\}$. We include all studies where dictators are allowed to give or keep the full endowment, making no further restriction on the choice set. Thus, in the extreme case dictators play an all or nothing game and decide whether to either give or keep the endowment, which is the case in one of our included papers (Tinghög et al. 2016). We create a variable we call Partition, to control for the choice set in an experiment, which is defined as the smallest share that can be donated. Finally, we only include studies that have data on gender and where both genders participate (excluding single-gender studies). The search for studies started on the database Econlit with a search for the keyword “Dictator game”. This gave us 513 hits and if the studies fitted our inclusion criteria we sent out an email to request the raw data from the corresponding authors. We did not systematically check for unpublished papers. In May of 2018 we also sent out a request to the experimental email group (ESA) describing our sample of included studies and our inclusion criteria, and we closed the inclusion of studies in September 2018.

Defining inclusion and exclusion criteria involves a certain degree of arbitrariness and there is a tradeoff between including studies that use as similar experimental design as possible, which allows for internal validity, and expanding the inclusion criteria to more heterogeneous designs that may increase the statistical power and the generalizability of results. We potentially lose external validity when for instance not including studies varying the price of giving and studies where the dictator earns the endowment (which excludes experiments such as the all or nothing experiment in Bekkers (2007), where the dictator earns the endowment in the experiment). However, we think the standard DG version is a reasonable starting point for analyzing gender differences in the dictator game. Generosity is unconditional, with no involvement of reciprocity, and involves no efficiency gains by changing the size of the endowment by giving and no party has done more (or less) to earn the endowment. We do include charity organizations as recipients which may increase external validity, as donating directly to anonymous individuals as in the standard DG is rare outside of the laboratory. We did not restrict our inclusion criteria to studies specifically designed to study gender differences, but included all studies meeting our inclusion criteria that had collected data on gender. It could be argued that ideally only studies designed to study gender differences should be included as they may be designed to eliminate confounding experimental designs that could influence the gender gap. However, it is not straightforward to define which studies were *ex ante* designed to study gender differences, and the observable “gender studies” may be published because they found gender differences and not due to their superior designs leading to publication bias. In an attempt to test if observable gender studies differ, we compare results for studies having gender in the title of the paper to the other studies in our data. An additional potential limitation of our data collection is that we did not explicitly search for discussion papers, which may induce publication bias. In Sect. 4.5 we test for publication bias using Egger’s and Begg’s tests to assess the importance of this limitation. The final sample consists of 15,016 unique observations where we also have gender data.

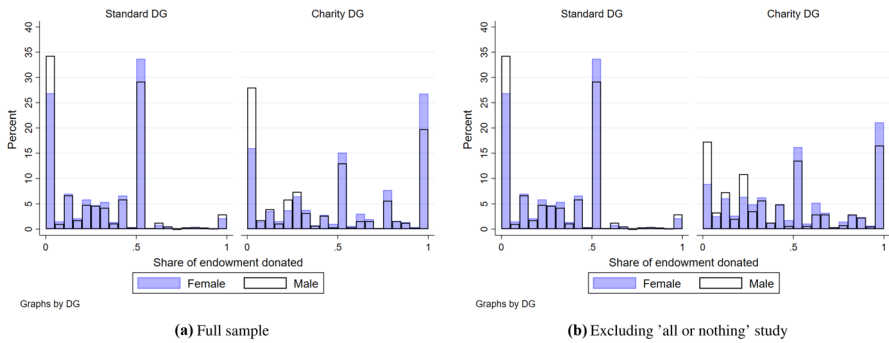


Fig. 1 The full sample contains 15,016 unique individual dictator decisions. There are 11,802 observations in the standard DG and 3214 observations in the charity DG in **a**. Excluding the ‘all or nothing’ study in **b** reduces the sample size in the charity DG to 1812 observations

2.2 Data

In Table 2 we present summary statistics of the included data. The full sample consists of 15,122 observations but for some observations either gender or the dictator’s decision is missing or have been incorrectly coded. Excluding those observations gives us a sample of 15,016 unique individual observations where we have both gender and the donation decision by the subject.¹ In Fig. 1 we plot the distributions of DG giving for each gender and by recipient type. As previously shown by e.g. Engel (2011), the distribution of DG giving in the standard DG is concentrated at giving nothing or half. In the charity DG a substantial fraction of subjects also give the entire endowment. When we exclude the ‘all or nothing’ study (Tinghög et al. 2016) the density on the two extreme points of the distribution decreases somewhat for the charity DG.

3 Meta-analysis methods

By collecting the individual participation data, a meta-analysis can either be done by the traditional approach of pooling effect sizes in a random effects model or with individual regression models. As Burke et al. (2017) note, these methods in general produce similar results and differences largely occur when researchers use different modelling assumptions. We begin by performing traditional random-effects meta-analysis which allows us to estimate the heterogeneity in the gender difference across the conditions included in the meta-analysis (the heterogeneity is captured

¹ If an experiment consists of several independent rounds we calculate the average across all rounds for each participant so that we only have one unique observation per individual in the data set.

Table 2 Descriptive statistics of the data included in the meta-analysis

	<i>N</i>	Mean	SD	Min	Max
Share donated (Full sample)	15,085	0.3245847	0.2927535	0	1
Share donated (Standard DG)	11,829	0.2818374	0.2487024	0	1
Share donated (Charity DG)	3256	0.4798851	0.3763469	0	1
1 if Female	15,050	0.4942193	0.4999832	0	1
Obs with gender and donation data	15,016				
Papers	53				
Conditions	117				
Year of paper		2013.875	3.744847	1998	2018
Partition (the smallest share a dictator can donate)	15,122	0.156878	0.2751402	0.0000667	1
1 if Student sample	15,122	0.4565534	0.4981253	0	1
1 if Econ student	15,122	0.0675837	0.2510384	0	1
1 if Random payment scheme	15,122	0.2526782	0.4345623	0	1
1 if Double-blind	15,122	0.264383	0.4410187	0	1
Experiment setting					
Field	968				
Lab	7890				
Online	5375				
of which Mturk	3211				
Phone	889				
Age					
No info	4521				
< 15	229				
15–30	5911				
31–40	1540				
41–50	1166				
50–60	973				
60+	782				
Region					
Europe	7624				
of which Nordic	4759				
N America	4974				
of which US	4958				
Asia	896				
Africa	933				
S America	153				
Oceania	454				

by the estimate of Tau, which is the standard deviation in the true effect size across the conditions).² We then estimate one stage individual regression models where we

² The estimation is done by the `Ipdmegan`-command by Fischer (2015) in Stata, with the estimation of the between study variance τ^2 by DerSimonian and Laird.

also provide several robustness checks of our results. Several experiments contain multiple conditions. These conditions create a natural clustering of the individual observations in our data. In total we have 117 conditions from 53 papers, which allows us to treat each condition within an experiment as a separate cluster. We estimate a separate effect size for each condition in the random-effects model, and we cluster the standard errors on the condition level in all individual regression models.³ Our definition of a condition follows the definition within each paper. If the same condition within an experiment is conducted in different countries (except if the study is done online (MTurk)), we define these as separate conditions to account for the country level clustering of the experiments.

We estimate Eq. (1) where S_{ij} denotes the share of the endowment donated by participant i in condition j , X is a vector of individual covariates and Z is a vector of treatment condition controls. We also replace Z with a condition fixed effect using dummy variables for each condition.⁴ The gender coefficient β_1 and the interaction between gender and charity recipient β_3 are the coefficients of interest,

$$S_{ij} = \beta_0 + \beta_1 \text{Female} + \beta_2 \text{Charity} + \beta_3 (\text{Female} \star \text{Charity}) + \beta_4 X_{ij} + \beta_5 Z_j + \epsilon_{ij}. \quad (1)$$

4 Results

We use a significance threshold of $p < 0.005$ for “statistically significant evidence” and a threshold of $p < 0.05$ for “suggestive evidence” in our results below in line with the recent recommendation of Benjamin et al. (2018). All our tests are two-sided.

4.1 Random-effects meta-analysis

Figure 2 shows a forest plot of the estimated gender gap for each of the 117 conditions in our sample; and the random effects results are also reported in Online Appendix Table A1. We show the results both separately for the standard DG and the charity DG, and pooled for both DG versions. Women give on average 4 percentage points more than men and the gender gap is statistically significant. The average donation in our data is 32% of the endowment (see Table 2) and women on average give 13% more than men (Cohen’s $d = 0.16$)⁵. The standard deviation in the

³ An alternative strategy would be to cluster on the paper level. However, there are only 12 papers where a charity is the recipient, making this clustering unfeasible for us.

⁴ As the recipient type does not vary within a condition, the binary variable for the charity DG is already captured by the condition fixed effects. However, as gender varies within conditions we can still estimate the interaction effect between gender and charity DG.

⁵ The average donation for men in the whole sample is 30% of the endowment, 27% in the standard DG and 42% in the charity DG. To convert effect sizes to Cohen’s d (the effect size as a fraction of the STD), we calculate the STD for each paper and take the average STD across papers. The STD in the whole sam-

true effect size—the variation between studies over and above sampling variation—is slightly higher than the average effect size at $\hat{\tau} = 4.6$ percentage points. To further assess heterogeneity in the gender gap, we estimate the gender gap in the standard DG and the charity DG respectively.

In the standard DG where the recipient is another participant, women on average donate 2.3 percentage points more than men, with a standard deviation in the true effect size of $\hat{\tau} = 3.3$ percentage points. In the charity DG the gender gap is larger, with women on average giving 10.9 percentage points more than men with a standard deviation in the true effect size of $\hat{\tau} = 6.4$ percentage points. A meta-regression in Table A2 confirms that there is a statistically significant difference in the gender gap between the standard DG and the charity DG. The gender gaps of 2.3 percentage units in the standard DG and 10.9 percentage units in the charity DG imply that women give 9% more than men in the standard DG (Cohen's $d = 0.10$) and 26% more than men in the charity DG (Cohen's $d = 0.35$).

Excluding the 'all or nothing' conditions reduces the gender gap found in the overall sample from 4 to 3.1 percentage points (Cohen's $d = 0.13$), but it remains statistically significant with heterogeneity in the true effect size of $\hat{\tau} = 3.8$ percentage points. In the charity DG the gender gap is reduced from 10.9 to 8 percentage points (Cohen's $d = 0.27$), with heterogeneity $\hat{\tau} = 4.7$ percentage points. The smaller gender gap in the charity DG decreases the meta-regression estimate of the difference in the gender gap between the standard DG and the charity DG from 8.7 to 5.9 percentage points.

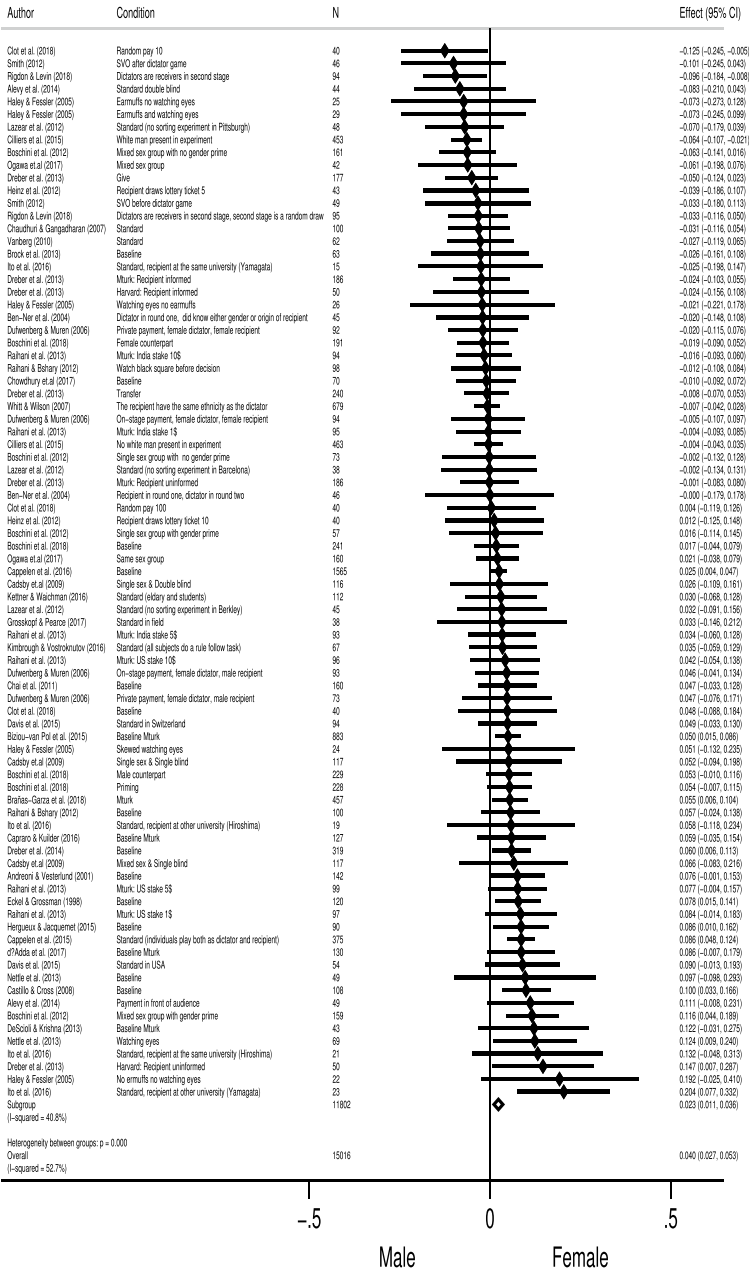
4.2 Individual level regression analysis

In Table 3 we report the results of the individual level regression analysis. In column 1, where we only include a binary variable for the gender of the dictator, women give on average 4.8 percentage points more than men, which is similar to the gender gap of 5.8 percentage points reported in Engel (2011). Controlling for the charity DG in column 2 gives an overall gender gap of 4 percentage points, which is identical to the gender gap found with the random-effects meta-analysis.⁶ In column 3–6 we add an interaction between the female variable and the charity DG. The gender gap is statistically significantly larger in the charity DG compared to the standard DG in all four specifications, with an interaction coefficient of between 9.3 and 9.8 percentage points. Women give on average around 2 percentage points more than men in the standard DG and 11–12 percentage points more in the charity DG. These

Footnote 5 (continued)

ple is 0.250 (0.247 if we exclude 'all or nothing' conditions), while in the standard DG it is 0.233 and in the charity DG it is 0.310 (0.300 if we exclude 'all or nothing' conditions).

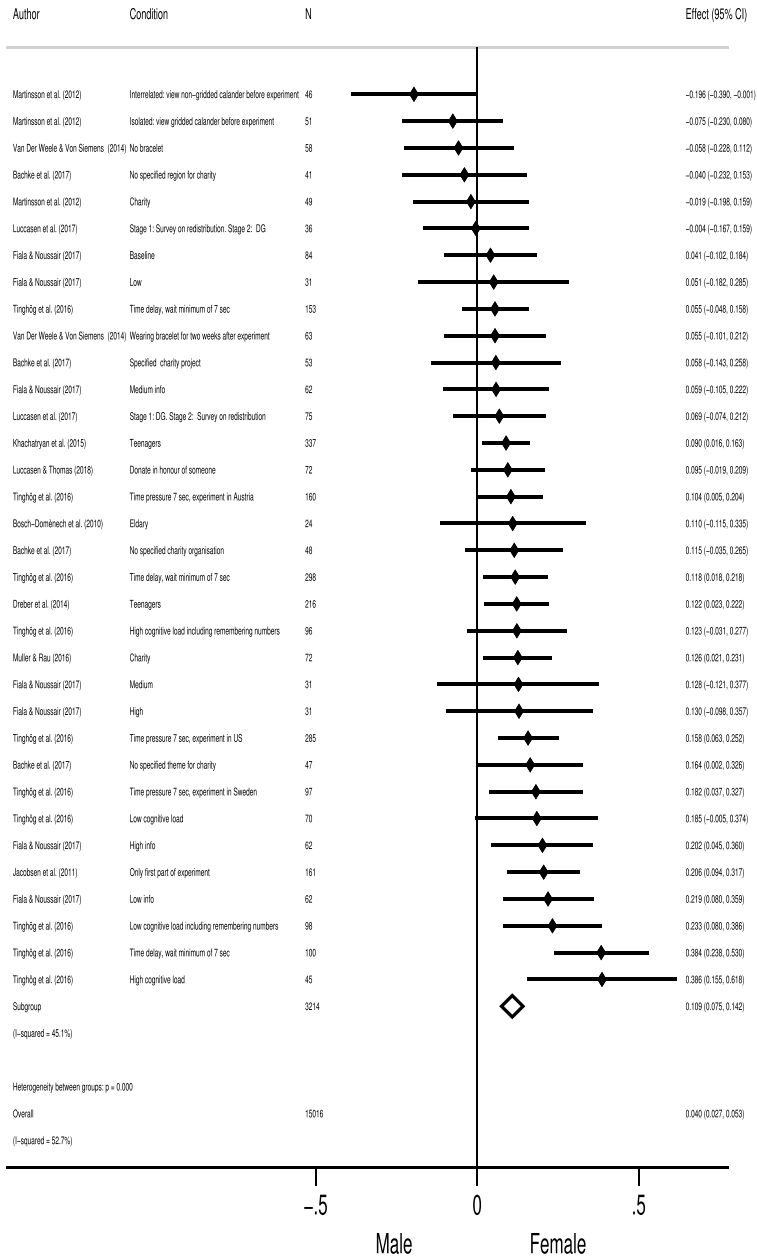
⁶ We also estimate the model in column 3 where we include a dummy for doubleblind studies, instead of the charity DG dummy, that is interacted with the female variable to test if the gender gap is significantly different in doubleblind studies. The interaction coefficient is -1 percentage point but is not statistically significant (interaction coefficient = -0.010 , SE = 0.022 , p value = 0.654). We thank the anonymous reviewers for suggesting this robustness check.



NOTE: Weights are from random-effects model

(a) Standard DG

Fig. 2 Random effects model (estimated with the Ipdmetan command in Stata). **a** Contains experiments with the standard DG and **b** contains experiments with the charity DG. The diamonds indicate the estimated effect size (and the CI) for each sub sample and the pooled (overall) effect size is at the bottom



NOTE: Weights are from random-effects model

(b) Charity DG

Fig. 2 (continued)

gender gaps are statistically significant for both types of DG in all the four regression models, except for the standard DG in column 3 where there is suggestive evidence of a gender difference ($p = 0.0058$). If we exclude the ‘all or nothing’ conditions, the results are similar (see Table 4).

4.2.1 Robustness checks

In a robustness test, we estimate a mixed random effects model where we allow for both the intercept and the gender gap to have random effects on the condition level.⁷ These results are reported in tables A3 and A4. We find very similar results in these estimations, although the pooled gender gap of 4.6 percentage points is slightly higher. The gender gap in the standard DG is around 2 percentage points and the gender gap in the charity DG is 11 percentage points and significantly higher than in the standard DG. The gender gap is statistically significant in both the standard DG and the charity DG, both with and without the ‘all or nothing’ conditions included.

In a second robustness test, we re-estimate our results using a tobit model. The action space is limited to a donation between 0 and 1 even though some subjects may possibly prefer to take from the recipient or give more than the endowment. In the tobit model we allow censoring to occur at both 0 and 1. We report these results in tables A5 and A6, and the coefficients in these tables should be interpreted as the gender gap with respect to the latent (that in theory can take on both negative values and values above 1) dependent variable. The tobit model yields higher estimates of the gender gap, with a gender gap of 6.6 percentage points in the pooled sample. The gender gap is around 4 percentage points in the standard DG and 17 percentage points in the charity DG, and this difference is statistically significant. The gender gap is statistically significant in both the standard DG and the charity DG, both with and without the ‘all or nothing’ conditions included. As can be seen from Fig. 1, men are more likely to donate zero in both the standard DG and the charity DG and women are more likely to donate the full endowment in the charity DG. These differences at the censoring points of 0 and 1 result in a higher estimated gender gap in the tobit model when these observations are interpreted as being censored.

So far we have measured the gender gap as the difference in the share of the endowment donated. As both men and women donate more in the charity DG it is possible that the gender gap is larger in percentage points but not in terms of percentage of the average donation. We test this in an additional robustness check where we instead use the measure from Eq. (2) below, where we divide the individual share donated in the DG (s_{ij}) by individual i that took part in condition j by the average donation in condition j ,

$$\hat{s}_{ij} = \frac{s_{ij}}{s_j}. \quad (2)$$

⁷ We thank the anonymous reviewers for suggesting this robustness check and the robustness check below using the tobit model.

Table 3 OLS results of the estimated gender gap in the DG. Standard errors clustered on the condition level in parentheses

	(1)	(2)	(3)	(4)	(5)	(6)
	Share	Share	Share	Share	Share	Share
Female	0.048*** (0.010)	0.040*** (0.008)	0.020* (0.007)	0.020*** (0.006)	0.023*** (0.006)	0.021*** (0.006)
Charity DG		0.194*** (0.029)	0.144*** (0.029)	0.141*** (0.035)		
Charity DG * Female			0.094*** (0.018)	0.093*** (0.017)	0.097*** (0.016)	0.098*** (0.016)
Constant	0.300*** (0.012)	0.263*** (0.012)	0.272*** (0.011)	0.463*** (0.050)	0.302*** (0.007)	0.362*** (0.025)
Condition fixed effects	No	No	No	No	Yes	Yes
Individual controls ^a	No	No	No	Yes	No	Yes
Treatment controls ^b	No	No	No	Yes	No	No
Female +(Charity DG * Female)			0.113*** (0.016)	0.114*** (0.015)	0.120*** (0.015)	0.119*** (0.014)
Observations	15,016	15,016	15,016	15,016	15,016	15,016
Number of conditions	117	117	117	117	117	117

^aIndividual controls: Student characteristics, age and region

^bTreatment controls: Double-blind, setting characteristics, random payment and partitioning of endowment

* $p < 0.05$, ** $p < 0.005$, *** $p < 0.001$

Multiplying this measure with 100 allows us to interpret each observation in terms of percent of the average donation within the condition that the participant took part.

Our results are confirmed by this standardization when we include all the DG studies in Table A7. Women on average give around 9% more than men in the standard DG and 25% more in the charity DG; and this difference is statistically significant. In Table A8 we carry out this analysis excluding the “all or nothing” DG study. This reduces the gender gap in the charity DG by one fifth to around 20%, and the gender gap is not statistically significantly larger in the charity DG any more (but there is suggestive evidence for a larger gender gap in the charity DG in all models). When we measure donations in relative terms the evidence of a larger gender difference in the charity DG is thus less strong, as the donations are larger on average in the charity DG compared to the standard DG.

4.3 Gender in the title of the studies

Of the 53 (117) papers (conditions) included in our study 16 (31) have gender in the title of the paper, and we test if the gender difference differs between papers with

Table 4 OLS results of the gender difference in the DG, excluding the “all or nothing” DG study. Standard errors clustered on the condition level in parentheses

	(1)	(2)	(3)	(4)	(5)	(6)
	Share	Share	Share	Share	Share	Share
Female	0.035*** (0.009)	0.028*** (0.007)	0.020* (0.007)	0.021*** (0.006)	0.023*** (0.006)	0.021*** (0.006)
Charity DG		0.184*** (0.036)	0.147*** (0.034)	0.137*** (0.035)		
Charity DG * Female			0.066** (0.022)	0.067** (0.020)	0.065*** (0.018)	0.066*** (0.017)
Constant	0.289*** (0.012)	0.268*** (0.011)	0.272*** (0.011)	0.496*** (0.046)	0.317*** (0.007)	0.373*** (0.026)
Condition fixed effects	No	No	No	No	Yes	Yes
Individual controls ^a	No	No	No	Yes	No	Yes
Treatment controls ^b	No	No	No	Yes	No	No
Female + (Charity DG * Female)			0.086*** (0.021)	0.087*** (0.019)	0.088*** (0.017)	0.087*** (0.016)
Observations	13,614	13,614	13,614	13,614	13,614	13,614
Number of conditions	107	107	107	107	107	107

^aIndividual controls: Student characteristics, age and region

^bTreatment controls: Double-blind, setting characteristics, random payment and partitioning of endowment

* $p < 0.05$, ** $p < 0.005$, *** $p < 0.001$

and without gender in the title.⁸ If we observe such a difference we cannot tell if this is due to that studies explicitly designed to study gender differences lead to different results, or if the difference is due to that studies with gender in the title were framed as studies of gender differences and published because they found a significant gender difference. We return to the issue of publication bias in Sect. 4.5.

A meta-regression in Table A9 provides no evidence that the gender gap differs between papers with gender in the title and the other DG studies. We also estimate Eq. (1) in an OLS model in Table A10, where we include a dummy for “gender in the title”, that is interacted with the female variable to test if the gender gap is significantly larger in papers with gender in the title. The null results from the meta-regression are confirmed in the OLS model. In Table A10 we also report the estimated gender difference among the subset of papers with gender in the title (it is the sum of the gender coefficient and the interaction coefficient) and it is between 3.8 and 3.9 percentage points when we include both types of DGs, and between 2.6 and 2.9 percentage points for the standard DG and between 10 and 10.8 percentage

⁸ With “gender in the title”, we do not explicitly mean the word “gender”, so we include Eckel and Grossman (1998) that mention “women” and “men” in the title.

points for the charity DG. The gender difference is statistically significant in all models except in model 3 where there is suggestive evidence.

4.4 Statistical power

In Table A11 we summarize the statistical power to detect the gender gap found in this meta-analysis. The power estimates are based on estimating the mean difference between males and females using two-sided hypothesis testing, and they are based on the average standard deviation of the standard DG papers ($STD = 0.233$) and the charity DG papers ($STD = 0.310$). We use the random effects results of a gender difference of 2.3 percentage units in the standard DG and 10.9 percentage units in the charity DG to estimate power in each DG type; but we also report power for the overall gender difference of 4 percentage units for both types of DGs. We do the power calculations for tests at the 5% level, as that is most commonly used in the literature. But we also report results for the more stringent 0.5% threshold used in this paper.

We calculate the sample size as the total number of observations in a paper, which means that we sum over all conditions within the paper. The median sample size in the standard DG papers is 130 observations, which yields a statistical power of only 9% (16%) to detect an effect size of 2.3 (4) percentage points. To reach 80% power a paper would need around 3224 (1068) observations to detect an effect size of 2.3 (4) percentage points in the standard DG. The median sample size for the charity DG papers is 192 observations, yielding 68% (14%) power to detect an effect size of 10.9 (4) percentage points. All the above estimates are based on tests at the 5% level, and using the more stringent 0.5% threshold leads to even lower power (see Table A11).

Some of the included DG studies were not designed to study gender differences, which may explain the inadequate power. However, the power for studies that have gender in the title of the paper are only slightly higher. There are 13 papers in the standard DG and three papers in the charity DG that have gender in the title of the paper. The median sample size of the 13 standard DG studies is 191, which gives 10% (22%) statistical power to detect a 2.3 (4) percentage units gender difference. The median sample size of the three charity DG studies is 216, which gives 73% (16%) statistical power to detect a 10.9 (4) percentage units gender difference.

4.5 Publication bias

Figure A1 shows funnel plots for the full sample, the standard DG sample, and the charity DG sample with the estimated effect sizes on the x-axis and the corresponding standard errors on the y-axis. An asymmetric plot could be evidence of publication bias, where only significant studies are published. The outliers at the far right in Figure A1 (a) and A1 (c) are two of the ‘all or nothing’ conditions.

The funnel plots in Figure A1 do not provide any clear visual evidence of publication bias. In Figure A2 we restrict the funnel plots to papers with gender in the paper's title. There is no clear visual evidence of any asymmetry in these plots either.⁹

To statistically test if there is evidence of publication bias we have employed Egger's and Begg's tests of publication bias in Table A12. We find no evidence of publication bias in either the pooled sample or when looking at each dictator game separately. We furthermore carry out these tests including only papers with gender in the title, but we do not find a statistically significant publication bias in these tests either.

5 Discussion

Our results suggest that women give more than men on average in both the standard and the charity DG, but the gender gap is modest in size (4 percentage points in the pooled data and a Cohen's d of 0.16). This is similar to the gender gap in the meta-analysis by Engel (2011). Looking at the standard DG and the charity DG separately, we find that the gender gap is 2.3 percentage points (Cohen's $d = 0.10$) in the former and 10.9 percentage points (Cohen's $d = 0.35$) in the latter (and if we exclude the "all or nothing study" this gender gap decreases to 8 percentage points; Cohen's $d = 0.27$). It is interesting to compare these results to the recent study by Falk et al. (2018), measuring economic preferences in a global preference survey with a sample size of about 80,000 individuals. They measured altruism by combining the answers to two survey questions. One of these was a hypothetical donation question similar to the charity DG and the other question measured the willingness to give to good causes on an 11-point scale. They found significantly higher altruism for women than for men, with an estimated gender difference of 0.10 Cohen's d units. We find the same effect size as Falk et al. (2018) in the standard DG. For the charity DG we find a larger effect size than in Falk et al. (2018), but this effect size is also less precisely estimated in our study.¹⁰

The estimated gender differences in our study implies that the typical DG study in the literature is underpowered to test for gender differences. Power problems have previously been reported in economics in general (Ioannidis et al. 2017) as well as for the DG (Ortmann and Zhang 2015). However, for some of the datasets included in the meta-analysis, the researchers may never have had the intention to study gender differences and may have been well powered to study their main research question. Our power results should thus mainly guide future

⁹ For the charity DG there are only three papers with gender in the title making it impossible to test for publication bias, but for completeness we show a funnel plot of these three observations as well in Figure A2.

¹⁰ Falk et al. (2018) do not report gender differences separately for the hypothetical donation question that is similar to the charity DG and the 11-point scale donation question. We therefore cannot compare the results of the real and hypothetical charity DG questions across our studies (but only compare our results to their combined altruism measure).

research that aims to explore gender differences in DG giving and moving forward researchers may need substantially larger sample sizes than what has previously been the norm.

There are also several additional caveats to our conclusions. First, it is not clear whether important datasets are missing from our analysis, and whether the inclusion of these would change any of our conclusions. Publication bias may lead to inflated effect sizes in meta-analysis, which was also observed in a recent study by Kvarven et al. (2020) comparing meta-analyses to pre-registered multiple-laboratory replication projects. We tested for publication bias and found no evidence in this direction; this result, however, could be biased by the exclusion of unpublished papers from the sample. We find substantial heterogeneity in the gender gap between conditions. In our paper we explore heterogeneity with respect to if the recipient is a charity organization or another participant, but there is much more work that can be done in this regard. One should also be careful to generalize our findings to also hold in other designs such as when the price of giving varies or when the dictator earns the endowment. For example, Andreoni and Vesterlund (2001) report gender differences to be conditional on the price of giving.

Our results suggest a larger gender difference in the charity DG compared to the standard DG, although the strength of this evidence depends on if the “all or nothing” charity DG study is included or not and if the difference is measured in absolute or relative terms (as the average donations are higher in the charity DG). To draw strong conclusions about whether the gender difference is larger in the charity DG than the standard DG, it would be interesting to conduct a well-powered study to directly compare the gender difference in these two versions of the DG. A possible explanation for a larger gender difference in the charity DG could be that the charity DG is more closely related to empathy and altruism, whereas the standard DG is more related to fairness preferences (deviating from the 50/50 norm). In the standard DG it is unusual to observe donations over 50% of the endowment, whereas donating 100% of the endowment is relatively common in the charity DG. Altruism as a motivation for donations is consistent with a stronger tendency for such corner solutions of donating all or nothing. Further work is needed to better understand if the two types of DGs measure different forms of social preferences.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s40881-021-00105-9>.

Acknowledgements We thank Eva Ranehill, Valerio Capraro, Lenka Fiala and two anonymous reviewers for comments on an earlier draft. In a previously circulated version of the paper we had failed to include some treatments from Clot et al. (2018) and Fiala and Noussair (2017), which Lenka Fiala kindly brought to our attention. This explains why the reported numbers differ slightly for some of the results in this version compared to the previous version. We are very grateful to the original authors of the various DG studies for generously sharing their data; without this data this project obviously would not have been possible. We are also grateful for generous financial support from the Jan Wallander and Tom Hedelius Foundation (Svenska Handelsbankens Forskningsstiftelser), the Knut and Alice Wallenberg Foundation and the Marcus and Marianne Wallenberg Foundation (AD is a Wallenberg Scholar), Riksbankens Jubileumsfond and the Swedish Foundation for Humanities and Social Sciences.

Funding Open access funding provided by University of Gothenburg.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Andreoni, J., & Vesterlund, L. (2001). Which is the fair sex? Gender differences in altruism. *The Quarterly Journal of Economics*, *116*(1), 293–312.
- Bardsley, N. (2008). Dictator game giving: altruism or artefact? *Experimental Economics*, *11*(2), 122–133.
- Benjamin, D.J., Berger, J.O., Johannesson, M. et al. (2018) Redefine statistical significance. *Nature Human Behaviour* *2*, 6–10 . <https://doi.org/10.1038/s41562-017-0189-z>.
- Bekkers, R. (2007). Measuring altruistic behavior in surveys: The all-or-nothing dictator game. *Survey Research Methods* *1*(3):139–144, <https://doi.org/10.18148/srm/2007.v1i3.54>, <https://ojs.ub.uni-konstanz.de/srm/article/view/54>
- Bertrand, M., Goldin, C., & Katz, L. F. (2010). Dynamics of the gender gap for young professionals in the financial and corporate sectors. *American Economic Journal: Applied Economics*, *2*(3), 228–55. <https://doi.org/10.1257/app.2.3.228>, <https://www.aeaweb.org/articles?id=10.1257/app.2.3.228>
- Bolton, G. E., & Katok, E. (1995). An experimental test for gender differences in beneficent behavior. *Economics Letters*, *48*(3), 287–292.
- Burke, D. L., Ensor, J., & Riley, R. D. (2017). Meta-analysis using individual participant data: one-stage and two-stage approaches, and why they may differ. *Statistics in Medicine*, *36*(5), 855–875. <https://doi.org/10.1002/sim.7141>, <https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.7141>
- Clot, S., Grolleau, G., & Ibanez, L. (2018). Shall we pay all? an experimental test of random incentivized systems. *Journal of Behavioral and Experimental Economics*, *73*, 93–98. <https://doi.org/10.1016/j.socec.2018.01.004>, <http://www.sciencedirect.com/science/article/pii/S2214804318300363>
- Crosan, R., & Gneezy, U. (2009). Gender differences in preferences. *Journal of Economic literature*, *47*(2), 448–474.
- Dana, J., Cain, D. M., & Dawes, R. M. (2006). What you don't know won't hurt me: Costly (but quiet) exit in dictator games. *Organizational Behavior and Human Decision Processes*, *100*(2), 193–201.
- Eckel, C. C., & Grossman, P. J. (1996). Altruism in anonymous dictator games. *Games and Economic Behavior*, *16*(2), 181–191. <https://doi.org/10.1006/game.1996.0081>, <http://www.sciencedirect.com/science/article/pii/S0899825696900810>
- Eckel, C. C., & Grossman, P. J. (1998). Are women less selfish than men?: Evidence from dictator experiments. *The Economic Journal*, *108*(448), 726–735.
- Engel, C. (2011). Dictator games: a meta study. *Experimental Economics*, *14*(4), 583–610.
- Falk, A., Becker, A., Dohmen, T., Enke, B., Huffman, D., & Sunde, U. (2018). Global evidence on economic preferences. *The Quarterly Journal of Economics*, *133*(4), 1645–1692.
- Fiala, L., & Noussair, C. N. (2017). Charitable giving, emotions, and the default effect. *Economic Inquiry*, *55*(4), 1792–1812. <https://doi.org/10.1111/ecin.12459>
- Fisher, D. J. (2015). Two-stage individual participant data meta-analysis and generalized forest plots. *The Stata Journal*, *15*(2), 369–396. <https://doi.org/10.1177/1536867X1501500203>.
- Forsythe, R., Horowitz, J. L., Savin, N., & Sefton, M. (1994). Fairness in simple bargaining experiments. *Games and Economic Behavior*, *6*(3), 347–369.
- Ioannidis, J. P. A., Stanley, T. D., & Doucouliagos, H. (2017). The power of bias in economics research. *The Economic Journal*, *127*(605), F236–F265. <https://doi.org/10.1111/eoj.12461>, <https://onlinelibrary.wiley.com/doi/abs/10.1111/eoj.12461>
- Kahneman, D., Knetsch, J.L., & Thaler, R.H. (1986). Fairness and the assumptions of economics. *Journal of Business*, *59*, S285–S300

- Krupka, E. L., & Weber, R. A. (2013). Identifying social norms using coordination games: why does dictator game sharing vary? *Journal of the European Economic Association*, *11*(3), 495–524.
- Kvarven, A., Strømmand, E., & Johannesson, M. (2020). Comparing meta-analyses and preregistered multiple-laboratory replication projects. *Nature Human Behaviour*, *4*(4), 423–434. <https://doi.org/10.1038/s41562-019-0787-z>, <http://www.nature.com/articles/s41562-019-0787-z>
- List, J. A. (2007). On the interpretation of giving in dictator games. *Journal of Political Economy*, *115*(3), 482–493.
- Ortmann, A., Zhang, L. (2015). Sample size null hypothesis significance testing. Greenwood Press/ABC-CLIO, pp 373–374
- Tinghög, G., Andersson, D., Bonn, C., Johannesson, M., Kirchler, M., Koppel, L., & Västfjäll, D. (2016). Intuition and moral decision-making—the effect of time pressure and cognitive load on moral judgment and altruistic behavior. *PLoS ONE*, *11*(10), 1–19.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.