



UNIL | Université de Lausanne

Unicentre

CH-1015 Lausanne

<http://serval.unil.ch>

Year : 2021

Influences in decision making: Three essays in behavioral economics

Wettstein Jason

Wettstein Jason, 2021, Influences in decision making: Three essays in behavioral economics

Originally published at : Thesis, University of Lausanne

Posted at the University of Lausanne Open Archive <http://serval.unil.ch>

Document URN : urn:nbn:ch:serval-BIB_8110C5DBC2828

Droits d'auteur

L'Université de Lausanne attire expressément l'attention des utilisateurs sur le fait que tous les documents publiés dans l'Archive SERVAL sont protégés par le droit d'auteur, conformément à la loi fédérale sur le droit d'auteur et les droits voisins (LDA). A ce titre, il est indispensable d'obtenir le consentement préalable de l'auteur et/ou de l'éditeur avant toute utilisation d'une oeuvre ou d'une partie d'une oeuvre ne relevant pas d'une utilisation à des fins personnelles au sens de la LDA (art. 19, al. 1 lettre a). A défaut, tout contrevenant s'expose aux sanctions prévues par cette loi. Nous déclinons toute responsabilité en la matière.

Copyright

The University of Lausanne expressly draws the attention of users to the fact that all documents published in the SERVAL Archive are protected by copyright in accordance with federal law on copyright and similar rights (LDA). Accordingly it is indispensable to obtain prior consent from the author and/or publisher before any use of a work or part of a work for purposes other than personal use within the meaning of LDA (art. 19, para. 1 letter a). Failure to do so will expose offenders to the sanctions laid down by this law. We accept no liability in this respect.



UNIL | Université de Lausanne

Faculté de Droit, des Sciences Criminelles, et d'Administration Publique de
l'Université de Lausanne

IDHEAP

Influences in decision making: Three essays in behavioral economics

Jason Wettstein

Thesis supervisor
Prof. Christian Thöni

Jury

Prof. Laure Athias, University of Lausanne,
Prof. Jordi Brandts, Institute for Economic Analysis (CSIC),
Prof. Adrian Bruhin, University of Lausanne,
Prof. Christian Thöni, University of Lausanne

LAUSANNE, 2021



UNIL | Université de Lausanne

Faculté de Droit, des Sciences Criminelles, et d'Administration Publique de
l'Université de Lausanne

IDHEAP

Influences in decision making: Three essays in behavioral economics

Jason Wettstein

Thesis supervisor
Prof. Christian Thöni

Jury

Prof. Laure Athias, University of Lausanne,
Prof. Jordi Brandts, Institute for Economic Analysis (CSIC),
Prof. Adrian Bruhin, University of Lausanne,
Prof. Christian Thöni, University of Lausanne

LAUSANNE, 2021



UNIL | Université de Lausanne

IDHEAP

Institut de hautes études
en administration publique

IMPRIMATUR

Le Décanat de la Faculté de droit, des sciences criminelles et d'administration publique, sur proposition d'un jury formé du professeur Christian Thoeni, de la professeure Laure Athias, et des professeurs Jordi Brandts et Adrian Bruhin, sans se prononcer sur les opinions du candidat, autorise l'impression de la thèse de Monsieur Jason Wettstein, intitulée :

Influences in decision making: Three essays in behavioral economics

Lausanne, le 30 mars 2021

A handwritten signature in blue ink, appearing to be "N. Soguel", written in a cursive style.

Prof. Nils Soguel
Vice-Doyen de la Faculté de droit,
des sciences criminelles
et d'administration publique

Acknowledgments

First and foremost, I would like to express my gratitude to my supervisor Christian Thöni. Thank you for the great opportunity, all these intellectual discoveries, the challenges, the support, and the amazing working environment.

I would also like to thank the experts of my committee, Laure Athias, Jordi Brandts, and Adrian Bruhin, whose words of advice brought considerable improvements to my thesis.

Next, I thank all of my colleagues and friends for the outstanding environment. Your different insights have provided grist for the mill over the years. But above all, it has been great to have such worthwhile, committed, curious, and chill people to hang with.

Last but not least, I thank all my entourage who have shown the necessary qualities for this journey. Your patience, understanding, and more fundamentally enthusiasm have led me to the end of my thesis.

Bref, many thanks!

Contents

0	Synopsis	8
0.1	Chapter summaries	11
0.1.1	Chapter 1	11
0.1.2	Chapter 2	12
0.1.3	Chapter 3	14
0.2	Perspectives & concluding remarks	15
0.2.1	Chapter 1	15
0.2.2	Chapter 2	16
0.2.3	Chapter 3	18
1	Chapter 1	23
1.1	Introduction	24
1.2	Related Literature	26
1.3	Conceptual framework & experimental procedures	30
1.3.1	Experimental procedures	31
1.3.2	Stereotype elicitation	32
1.4	Hypotheses	35
1.5	Study 1	37
1.5.1	Experimental Design	37
1.5.2	Results	37
1.6	Study 2	40
1.6.1	Experimental design	40
1.6.2	Results	41
1.7	Additional analyses	44
1.7.1	Self-serving bias	44
1.7.2	Distributions of the stereotypes	46
1.7.3	Priming	46
1.8	Discussion	51
1.9	Conclusion	51
1.10	Appendices	57
1.11	Supplementary material	62
2	Chapter 2	66
2.1	Introduction	67
2.2	Experimental Design	68
2.3	Hypotheses	70
2.4	Results	74
2.5	Discussion	88
2.6	Appendices	95

2.7	Supplementary material	99
2.7.1	Randomization checks	99
2.7.2	Social Preferences	102
2.7.3	Income effect	110
3	Chapter 3	113
3.1	Introduction	114
3.2	Methods	115
3.2.1	Four measurements for gender differences	115
3.2.2	GDP and emancipation indexes	120
3.2.3	Models	122
3.3	Results	124
3.3.1	Cross-country analyses - M1	126
3.3.2	Within-country analyses - M2	128
3.3.3	Robustness of paradox to sub-samplings	132
3.3.4	Investigation of omitted variables - M3	135
3.3.5	General remarks	141
3.4	Conclusion	144
3.5	Supplementary material	148
3.5.1	Descriptive statistics	148
3.5.2	Instructions	151
3.5.3	Coding the variables	151
3.5.4	The coefficient from the WVS and the EVS	155
3.5.5	Classification of the items from the WVS.	157
3.5.6	Ordinal analysis of data	157
3.5.7	Indexes	159
3.5.8	Clustering	160
3.5.9	Additional analyses	163
3.5.10	Robustness checks	166
4	General conclusion	182

0 Synopsis

The traditional assumption that humans are perfect rational agents has been challenged and many studies have shown that the cost-benefit analysis in decision making cannot remain the only prerogative of the standard economic model.

Simon (1947) was the pioneer to introduce bounded rationality in a model of decision making. In his model, individuals are not able to evaluate the consequences of every possibility because of limited cognitive resources or a lack of information. Then, Tversky and Kahneman (1974) introduced the concept of judgment heuristics and investigate rationality biases in the process of decision making under uncertainty. Individuals use heuristic methods to make decisions because of bounded rationality. My thesis presents three studies in the field of decision making. Each study has a focus on a different determinant of decision making. I illustrate these determinants with an example.

Let's consider the problem of picking chocolate from a choice of brands. Although there are numerous reasons why someone chooses an option rather than another, I will focus on aspects that relate to my thesis. Let's assume a consumer who has no strict preferences in his chocolate taste. He asks a friend to advise him on this decision. The consumer's friend wears glasses and advises him to pick brand A instead of brand B. His main argument is that brand A is of a higher ethical standard than brand B. Since, the consumer values ethics, he decides to select brand A. When he goes to pick the chocolate, he finds it easily, as the store places chocolate A at eye level while chocolate B is located on a shelf that is not in the consumer's eye-line. For the consumer, all three reasons, i.e., stereotype that people who wear glasses are knowledgeable, how choices are presented, and valuing ethical practices, lead to brand A, thus he picks chocolate A.

This particular example of decision making involves three distinct elements, which I address in my thesis. First stereotypes (Chapter 1), second, choice architecture (Chapter 2), and third values (Chapter 3). All these factors influence decision making.

Chapter 1 investigates stereotypes. Stereotypes are category-based generalization (P. J. Grossman and Lugovsky 2011). People predict and infer behaviors of others based on their attributes. An attribute, in the above example, wearing glasses, is used to categorize an individual into a certain group. The person is part of the "group" that wears glasses as opposed to a "group" that does not. In this particular example, the attribute is associated with knowledge. In short, the consumer stereotypes people who wear glasses as being smarter.

Chapter 2 investigates choice architecture. In 2008, Thaler and Sunstein published a book called “Nudge”. They developed a concept which states that agents are sensitive to the presentation design of choices. A different choice architecture induces different behaviors. These changes nudge (push) individuals towards certain behaviors.

Chapter 3 investigates values. Values are fundamental beliefs, which state preferred ways of living and thinking. Often people have stable values and tend to live according to their values. This concept is rather close to the concept of preferences, which implies that values predict and influence behaviors. This concept encompasses a wide set of preferences, such as priorities, how should the world be, how one should act, but also general beliefs about what is good and what is bad.

Before I elaborate on the contents of my research, I shall mention some elements of the methodology. Besides that all chapters are quantitative analyses based on statistics, Chapters 1 and 2 differ from Chapter 3. The first two chapters are based on experiments conducted using the standard experimental economic methodology in a laboratory and an online setting. In Chapter 1, we use a laboratory experiment to measure stereotypes as this method offers maximum control over the environment. Moreover, since we can incentivize participants based on their answers, we motivate participants to answer truthfully. In Chapter 2, the experiment is used to draw causal inferences of different treatments. Since we assume a random allocation of the participants to treatments, we can infer that any systematic change on average in behaviors can be attributed to the treatment. Thus, experiments remain the gold standard to draw causal inferences. In Chapter 3, I use archival data and econometric tools. Archival data have the advantage to be easily accessible and usually offer large sample sizes. Furthermore, depending on the research question, like in this chapter, an experiment is hardly feasible. Overall, these methodologies are complementary. While in experiments, especially in the lab, we have great control of the environment, we may lack some external validity. On the contrary, when using archival data, the results can often be more generalized, but we have low control over the data generating process.

My thesis chapters do not differ solely in terms of methodology but also in the issue they relate to. Chapters 1 and 3 are both related to gender studies, while Chapter 2 falls within the scope of charitable giving.

Issues related to gender are not only of academic interest but also to policymakers, since targeting gender equality has been on the political agenda of

many countries, including Switzerland.¹ I (i) elaborate on gender stereotypes and how their impact is important for gender equality and (ii) develop on the relevance of studying differences in values for gender equality policies.

The elicitation of gender stereotypes is certainly a first step when addressing gender differences. In line with the social role theory (Eagly and Wood 1999; Wood and Eagly 2012), the reproduction of gender differences in roles is partially sustained through self-categorization and self-stereotyping (Guimond et al. 2006). These processes, although they are assumed to be the consequence of the early division of labor, tend to maintain the actual division of labor. In short, the belief about a gender difference in the labor market enforces discrimination correspondingly.

The influence of stereotypes on behaviors, often termed the activation of stereotypes, has been of scientific interest for many years, i.e., the activation of a stereotype is when stereotypes influence behaviors in the corresponding way (Boschini et al. 2012; Cohn et al. 2015; Krieglmeyer and Sherman 2012). An active stereotype induces different behaviors as opposed to a passive one that does not. Therefore, before we can measure the effect of a stereotype on behaviors, we must be able to measure it with precision. Chapter 1 introduces an experimental tool to measure stereotypes.

Chapter 3 addresses this issue with a different perspective. Since differences in economic outcomes are associated with values (Buser et al. 2014; DeLeire and Levy 2004; Duflo 2012), it is of interest to policymakers to know how these values evolve.

Gender equality is evolving almost everywhere in the world. An interesting question is whether more gender equality in life-circumstances leads to more or less differences in values. If gender differences in values tend to grow with more gender equality in life-circumstances, then differences in economic outcomes might grow. More precisely, this raises the question of whether equal opportunity leads to equal outcomes. If women and men tend to differ more and more in values, equal opportunity might lead to more differences in the outcome. However, on the contrary, if the tendency is towards convergence of values, then equal opportunity should lead to similar outcomes.

Chapter 2 addresses manipulations of choice architecture and its effects on charitable giving. Donations are economically important and concern a substantial fraction of the population. For instance, in the USA, 69% all the population donate or have donated to a charity.² While the implications of

¹see <https://www.eda.admin.ch/deza/en/home/themes-sdc/gender-equality.html>

²see <https://nonprofitssource.com/online-giving-statistics/>

research on this topic is fairly obvious and many studies focus on increasing donations (Andreoni and Payne 2013; A. Gneezy et al. 2010; Karlan and List 2007), this study follows the general research on choice architecture.

While the effects of choice architecture are often used pragmatically to nudge people towards certain behaviors³, studying choice architecture offers insights into the underlying mechanisms of decision making. The process is twofold. While nudges and choice architecture are the consequence of research in behavioral economics, their everyday usability favor more research to understand underlying mechanisms. Chapter 2 goes in that direction.

0.1 Chapter summaries

0.1.1 Chapter 1

This chapter is divided into two studies. While we investigate stereotypes about cooperative behaviors in both studies, Study 1 focuses on gender and Study 2 on political orientation. In both studies, we investigate whether participants believe that being a woman/man or being left-leaning or right-leaning will predict cooperative behavior.

Similar in both studies, we developed a mechanism to elicit stereotypes about cooperative behaviors. The elicitation mechanism is as follows: we gathered, from previous experiments, the average contribution of different subgroups plus the overall average. We use these datasets as references for our experiments. In Study 1, the two subgroups are males and females and in Study 2, left-leaning and right-leaning subjects.

In our experiment, we provide subjects with the ratio in percent of each subgroup in the reference dataset and the overall average contributions. Then, we ask them to give us their guess about the average contribution of either one or the other subgroup, i.e., in Study 1, we asked them to guess the average contribution of males or females with the given information about the overall average contributions. Before ending the elicitation stage, we provide them with the average contribution of both subgroups, i.e., the one for which they provided a guess plus the average contribution of the other subgroup.

³It is nowadays well recognized and used, to the extent that countries, such as France, Uk, and the USA have set in place special units dedicated to the implementation of these nudges in public policies. There is an interesting discussion about the legitimacy of using nudges. The main argument is that the approach is libertarian paternalist (see: Sunstein and Thaler 2003; Thaler and Sunstein 2003). Although the concept seems paradoxical, it combines two ideas, first that the choice is never restricted (libertarian) and second that the influence one tries to make on behaviors is for the good of the choosers (paternalism).

This latter one is conditional on the guess they provide.⁴ In this last stage, they needed to confirm their guess plus the other computed average contribution to complete the elicitation stage. We incentivized subjects, such that their payoff was dependent on how close their guess was to the true value in the reference dataset.

While, in Study 1, the gender attribute is not associated with a particular cooperative behavior, in Study 2, being left-leaning is stereotyped with higher cooperative behaviors. Given the reference dataset and the distribution of the guesses, we are able to elaborate on the accuracy of the stereotypes and to what extent the stereotypes are shared among the participants. In Study 1, we find that gender stereotypes are on average accurate compared to the reference dataset. On the other hand, in Study 2, we find that political orientation stereotypes are inaccurate. While, in the reference dataset, right-leaning contribute on average more, participants, in Study 2, think the opposite. We finally compare the distribution of the stereotypes and find that the shape is close to a normal distribution indicating that participants by and large share the stereotypes.⁵

Nevertheless, the major contribution in this chapter is not particularly the stereotype we elicit rather than the elicitation mechanism we propose, which overcomes some biases of other mechanisms found in the literature. As we incentivized participants for their guess, it is costly to respond in a socially desirable way. Moreover, by providing a baseline, we reduce the overestimation of a guess, i.e., reporting a very high or a very low average contribution if a subject does not know the overall average contribution. And finally, since we ask subjects to confirm the average contribution of both subgroups, we render salient the difference between the two subgroups, which likely reduces the overestimation of this difference.

0.1.2 Chapter 2

This chapter investigates the underlying mechanism in charitable giving. In a previous experiment by Schulz et al. (2018), they provide a list of charities to

⁴The correct estimate of the other subgroup average contribution is dependent on the ratio of each group in the reference dataset. Since this information is not obvious, we compute the other average contribution and asked them to confirm both average contributions.

⁵For instance, a bimodal distribution would indicate that participants have divided stereotypes, i.e., half of the participants thinking that males average contribution is higher than the females one and the other half thinking the opposite.

participants and double the number of donors. The variation in outcomes was substantial but the underlying mechanism that produced this shift remained difficult to explain. This study is a follow-up.

I use the data from a large experiment based on the experimental tool introduced by Kistler et al. (2017). I test two competing theories. One assumes that the shift in the number of donors is caused by lower cognitive cost. Having available options of charities require less cognitive effort than coming up with a charity. The second hypothesis assumes that the list triggers an emotional response that pushed participants towards donating.

The charity stage takes place at the end of an online experiment, where participants go through a series of incentivized tasks. Thus, at the end of the overall experiment, subjects are asked to indicate whether they would be willing to give part of their potential earnings from the previous tasks to a charitable organization. The decision is divided into three steps. First, the willingness to donate, a binary decision (yes/no), second, what percentage they are willing to donate, and third the choice of the charity. In the third step, Kistler et al. (2017) manipulate two dimensions in the provision of a list of charities: the length and whether the list was directly visible or with a drop-down button. They end up with four different list treatments plus a control group where no list was provided.

I find that the number of donors increases with the provision of the list and does not decrease with a longer list. I also find that the drop-down treatments mitigate the number of donors, albeit the effect is only marginally significant. While, the cognitive cost theory assumes a choice overload with the long list, which should dissuade some participants to donate (Gourville and Soman 2005; Iyengar and Lepper 2000), the emotional arousal one does not predict a lower number of donors in the long list (Kogut and Ritov 2005a; Kogut and Ritov 2005b; Small et al. 2007), and likely the opposite. In the long list treatments, there is a higher probability that a participant is presented with his favorite charity. This might induce a stronger emotional arousal. Moreover, if participants want to avoid the emotional arousal, they can strategically not click on the drop-down button (Andreoni, Rao, et al. 2017; Z. Grossman 2014). This mitigates the number of donors. In line with the results, I find support for the emotional arousal theory.

Nevertheless, this is not the most surprising result. I also find that there is a pitfall of changing the choice architecture in some situations. While the number of donors increases with the provision of the list, the realized level of donation remains unchanged. This result does not replicate the findings from Schulz et al. (2018), as they find that the realized level of donation doubles with the provision of a list. In my study, there are more donors in the list treatments, but they donate less. The treatment variation likely withdraws

their intrinsic motivation to donate summing up to no difference on the realized level of donation. I finally speculate on the elements that decrease intrinsic motivation. I posit that the impression of controls and the lower opportunity cost to donate might explain the shift in intrinsic motivation.

0.1.3 Chapter 3

This chapter tackles a very controversial question in social science. Does emancipation lead to the convergence of values between women and men or, on the contrary, lead to more differences?

I contrast two hypotheses. While the social role theory predicts that gender differences in values should decrease with more gender equality (Croson and U. Gneezy 2009; Eagly and Wood 1999), the resource hypothesis predicts the opposite (Almås et al. 2016; Haushofer and Fehr 2014). The former assumes that gender differences in values reflect gender differences in the labor market. Thus, a more balanced labor market should lead to less differences in values. The latter assumes that gender differences are intrinsic and their manifestation is a matter of opportunity. Gender differences in values are expected to grow since more gender equality should be associated with more equal opportunities.

I gathered longitudinal data from the World Value Survey and the European Value Survey and classify the items of these surveys into two dimensions: life-situations and values. Then, I further subdivided both dimensions into two, objective and subjective life-situations, and self-centered or general values, leaving me with four categories. Afterward, I computed the absolute difference between male and female scores per item, per year, and country of the surveys. Then, I regressed each category independently with indexes, such as the Gender Equality Index, the GDP Index from the United Nations, the Gender Equality Index from the World Economic Forum, ecological stress indicators, and different cultural dimensions, such as religiousness, individualism, and power distance.

While I find that more gender equality and economic growth are unambiguously associated with less gender differences in life-situations, as measured in the surveys, the effect on values is paradoxical. I end up with both divergence and convergence of gender differences in values with emancipation and economic growth depending on whether I run a cross-country analysis or a within-country one, i.e., controlling for country fixed effects. Both evidences are robust and significant. This suggests an endogeneity bias in the model.

While endogeneity bias is inherent in econometrics, it is rather hard to completely overcome (Antonakis et al. 2010). Thus, after testing many sub-

sampling and restriction on the data, I test other sources of gender differences in values. Kaiser (2019) find that ecological stress factors might explain the differences in values. The assumed causal relationship is that societies that had a high prevalence of pathogens tend to be more collectivist and display lower gender differences in values. I find that ecological stress factors are significantly correlated with differences in values. Along the same lines, the different cultural dimensions correlate with gender differences. Nevertheless, the paradox remains robust, even when I control for these additional factors, suggesting that the country fixed effects is able to capture more variability than these control variables and that the model still suffers from endogeneity.

Thus, finding the true underlying relationship between social/economic evolution and values remains unsolved so far. The main contribution is to show the robustness of this paradox, but I conclude that extensive research on other possible causes is necessary to reveal the true relationship between gender equality and gender differences in values.

0.2 Perspectives & concluding remarks

As my thesis chapters are substantially different, I address, in this section, the different perspectives of each chapter for future research as well as concluding remarks.

0.2.1 Chapter 1

Although we present a stereotype elicitation mechanism that partially overcome some biases identified in previous findings, such as social desirability or the overestimation due to unknown baselines, we cannot deduct to what extent participants think the stereotype is informative. This informative component matters when eliciting a stereotype, because the more a stereotype is believed to inform or predict the behaviors of a group the higher is the probability of statistical discrimination.

In our experiment, when participants indicate the average difference between two groups⁶, we do not know whether they have in mind a distribution of contributions. It might well be that they indicate a difference of average contributions between two groups but at the same time, think that this difference is not useful if they have to predict the contribution of the members

⁶The two groups are distinguished by an attribute. Either both groups possess an opposite attribute (e.g. male/female) or one has an attribute, while the other group does not (e.g. wearing glasses).

of a given group.

Let's assume the following regression: $b_{ij} = \alpha + \beta \mathbb{1}_G + \epsilon$, where i represents the belief about j 's contribution; α is the average contribution of the subjects in the baseline group (say, people not wearing glasses), and β is the expected difference between the two groups (say, the glasses effect). Whether a stereotype is informative depends on the R^2 of this regression. If participants estimate a low epsilon, it indicates that they think that the stereotype is informative. Overall, if their estimate about the difference between the two groups is different from 0 ($\beta \neq 0$), and if they think the stereotype is informative, they might well use this attribute to predict behaviors which leads to statistical discrimination.

Therefore, a future experiment should try to combine our stereotype elicitation mechanism with another stage which would elicit how informative subjects think their stereotype is. I suggest the following way: suppose you wish to elicit gender stereotypes about cooperative behaviors. After participants passed the elicitation mechanism, we tell them that we randomly pick one male and one female participant a hundred times from a large population. We, then, ask them to indicate how many times they think that the randomly picked male has a higher contribution than the randomly picked female. For instance, a subject thinks that the female average contribution is higher than the male one in a similar elicitation mechanism as in Chapter 1; how many times, in the random selection, she thinks the contribution of the randomly picked female is above the contribution of the randomly picked male.

The proposed design might overcome some limitations of Chapter 1 but is only one step ahead in the study of stereotypes. I think that the elicitation of stereotypes is still a necessary step to increase the validity of research on stereotype activation and their everyday implications.⁷

0.2.2 Chapter 2

This experiment has an interesting external validity since it reproduces closely the donation decision present on some websites. Nevertheless, it remains difficult to completely disentangle the underlying mechanism in the nudge. Al-

⁷The activation of a stereotype is likely mediated by how informative a stereotype is believed to be and to what extent a group behavior is believed to differ from the average behavior of the other group, i.e., if a person thinks that gender is a very good predictor of performance outcomes in a certain job (informative stereotypes where one gender differs substantially from the other gender), the person will likely discriminate based on gender when hiring someone.

though, I find evidence that the main mechanism that increased the number of donors is the emotional arousal, a possible effect produced by a cognitive cost is not completely discarded. While a body of literature reports the relative importance of external changes in mediating the intrinsic motivation (Deci 1976; Falk and Kosfeld 2006; Frey and Jegen 2001), I assume that the provision of the list undermines the intrinsic motivation. Following this logic, since participants do not need to commit the same cognitive effort to give to a charity, it likely withdraws their intrinsic motivation. Therefore, although the lower cognitive cost does not seem to be the main mechanism to nudge participants into donating, it might still play an important role.

As the decision to donate seems to be multi-dimensional and fairly complex, I identified at least two possibilities for future research. First, investigating the relative importance of extrinsic motivation in the donation decision. In my study, I assumed that the intrinsic motivation is crowded out, which would explain the lower amount participants are willing to donate. I suppose this is a consequence of a change in extrinsic motivation. Since DellaVigna et al. (2012) and Tonin and Vlassopoulos (2013) find that the donation decision is also influenced by the extrinsic motivation to enhance one's self image, future research could replicate the treatment variation of the provision of a list, but mitigate the extrinsic motivation of the participants. I would suggest to add a treatment variation where the donation decision is not anonymous. In sum, the experiment would contain four treatments: a control treatment with no list, a treatment with a list, and two other similar treatments where participants are not anonymous. This experiment might shed light on the possible interaction between intrinsic and extrinsic motivation with nudges in donation decisions.

A second perspective would be to mitigate the intrinsic motivation, which I think is rather complex since we can only suppose that a treatment variation would increase or decrease the intrinsic motivation.⁸ As I assume in Chapter 2, that the lower cognitive cost in the donation decision decreases the intrinsic motivation, I suggest to replicate the treatment variation of the provision of a list but add a treatment variation, where participants have to provide some effort if they are willing to donate. How much extra effort is required is hard to estimate, however, if the effort becomes too consequent, it might completely take over the first effect of the provision of a list. Therefore, a simple treatment would be that participants need to copy the name of the

⁸The intrinsic motivation of participants is a latent variable. Intrinsic motivation is usually inferred from the behaviors of the participants.

charity on another screen to confirm their willingness to donate. This experiment would possibly confirm the withdrawal of intrinsic motivation when participants are nudged to donate.

This study was at first an extension of Schulz et al. (2018) but failed to replicate their main finding. I believe this shows the complexity of donation decisions and suggests the possibility of many interferences in these processes. I finally think that this shows the necessity to replicate studies before claiming a definite conclusion and that even robust findings in one context, such as in face-to-face interaction, might not be applicable in another context, such as in an online setting.

0.2.3 Chapter 3

My investigation ends on a puzzle and therefore asks more questions than it solves. Although many studies find support for divergence (Falk and Hermle 2018; Mac Giolla and Kajonius 2019) and others for convergence (Donnelly and Twenge 2017; Konrad et al. 2000) of gender differences in values with respect to emancipation, I argue, in line with the findings from Connolly et al. (2019), that this relationship likely suffer an endogeneity bias.

The endogeneity bias is almost omnipresent in econometrics and well documented (Antonakis et al. 2014; Hamilton and Nickerson 2003). However, in some cases, inconsistent estimates will not lead to completely different conclusions when they are corrected, i.e., a positive correlation that remains positive even after correction. Even if this is deleterious for scientific research, the possible negative impact on public policies is contained. However, in cases such as in Chapter 3, the correlation changes sign, which might lead to harmful public policies.

Therefore, even if the topic is already well studied, I would still favor more research. Since experiments on this topic are likely inconceivable. I suggest two different perspectives. On one hand, some statistical tools might help uncover the true relationship between gender equality and gender differences in values, such as the regression discontinuity approach. However, the most promising one remains the instrumental variable. For the instrument to be viable, one must find an instrument that satisfies the exclusion restriction, namely one that does not correlate with the error term. Nevertheless, finding such an instrumental variable is very challenging since the dependent variable I use contains a very large set of values. One should probably restrict the number of values to a smaller sub-sample. Moreover, I guess that the often used ones, such as climate change or which country has been colonized are likely to not satisfy the exclusion restriction criterion for the actual set of values.

Although I think the instrument is the most promising tool, one can also try to find other variables that would explain gender differences in values (statistical adjustment). While in Chapter 3, I replicate some of the findings from Kaiser (2019) and confirm that ecological stress and cultural differences could be the cause of gender differences in values, the recent evolution of gender differences remains puzzling. Since there seems to be an important correlation between individualism and gender differences in values, I suggest looking for the determinants of individualism and also tracking the evolution of individualism over time. Kaiser (2019) find that a very important predictor of individualism was the prevalence of pathogens. Therefore, if this causal relationship is robust, it might well be that the countries that were highly affected by the coronavirus would display, with a certain latency, less gender differences in values.

Overall, the puzzle seems to be far from solved yet. As research tends to confirm the relationship between preferences/values and economic outcome, it increases the necessity to know if the difference in values between women and men tends to converge or diverge. Moreover, in terms of policy implication, it begs the question, as mentioned above, of whether reaching equality of opportunity will lead to equality in the outcome. Due to the possible consequences of such findings, I think this topic should remain a research priority in the future years.

References

- Almås, I., Cappelen, A. W., Salvanes, K. G., Sørensen, E. Ø., & Tungodden, B. (2016). Willingness to compete: Family matters. *Management Science*, *62*(8), 2149–2162.
- Andreoni, J., & Payne, A. A. (2013). Charitable giving. *Handbook of public economics*, vol. 5 (pp. 1–50). Elsevier B.V.
- Andreoni, J., Rao, J. M., & Trachtman, H. (2017). Avoiding the ask: A field experiment on altruism, empathy, and charitable giving. *Journal of Political Economy*, *125*(3), 625–653.
- Antonakis, J., Bendahan, S., Jacquart, P., & Lalive, R. (2010). On making causal claims: A review and recommendations. *Leadership Quarterly*, *21*(6), 1086–1120.
- Antonakis, J., Bendahan, S., Jacquart, P., & Lalive, R. (2014). Causality and endogeneity: Problems and solutions. *The Oxford Handbook of Leadership and Organizations*, *1*, 93–117.

- Boschini, A., Muren, A., & Persson, M. (2012). Constructing gender differences in the economics lab. *Journal of Economic Behavior & Organization*, *84*(3), 741–752.
- Buser, T., Niederle, M., & Oosterbeek, H. (2014). Gender, competitiveness and career choices. *Quarterly Journal of Economics*, *129*(3), 1409–1447.
- Cohn, A., Maréchal, M. A., & Noll, T. (2015). Bad boys: How criminal identity salience affects rule violation. *The Review of Economic Studies*, *82*(4), 1289–1308.
- Connolly, F. F., Goossen, M., & Hjerm, M. (2019). Does gender equality cause gender differences in values? Reassessing the gender-equality-personality paradox. *Sex Roles*, 1–13.
- Croson, R., & Gneezy, U. (2009). Gender differences in preferences. *Journal of Economic Literature*, *47*(2), 448–474.
- Deci, E. L. (1976). The hidden costs of rewards. *Organizational Dynamics*, *4*(3), 61–72.
- DeLeire, T., & Levy, H. (2004). Worker sorting and the risk of death on the job. *Journal of Labor Economics*, *22*(4), 925–953.
- DellaVigna, S., List, J. A., & Malmendier, U. (2012). Testing for altruism and social pressure in charitable giving. *The Quarterly Journal of Economics*, *127*(1), 1–56.
- Donnelly, K., & Twenge, J. M. (2017). Masculine and feminine traits on the Bem Sex-Role Inventory, 1993–2012: A cross-temporal meta-analysis. *Sex Roles*, *76*(9), 556–565.
- Duflo, E. (2012). Womens empowerment and economic development. *Journal of Economic Literature*, *50*(4), 1051–79.
- Eagly, A. H., & Wood, W. (1999). The origins of sex differences in human behavior - Evolved dispositions versus social roles. *American Psychologist*, *54*(6), 408–423.
- Falk, A., & Hermle, J. (2018). Relationship of gender differences in preferences to economic development and gender equality. *Science*, *362*(6412).
- Falk, A., & Kosfeld, M. (2006). The hidden costs of control. *American Economic Review*, *96*(5), 1611–1630.
- Frey, B. S., & Jegen, R. (2001). Motivation crowding theory. *Journal of Economic Surveys*, *15*(5), 589–611.
- Gneezy, A., Gneezy, U., Nelson, L. D., & Brown, A. (2010). Shared social responsibility: A field experiment in pay-what-you-want pricing and charitable giving. *Science*, *329*(5989), 325–327.
- Gourville, J. T., & Soman, D. (2005). Overchoice and assortment type: When and why variety backfires. *Marketing Science*, *24*(3), 382–395.

- Grossman, P. J., & Lugovskyy, O. (2011). An experimental test of the persistence of gender-based stereotypes. *Economic Inquiry*, *49*(2), 598–611.
- Grossman, Z. (2014). Strategic ignorance and the robustness of social preferences. *Management Science*, *60*(11), 2659–2665.
- Guimond, S., Chatard, A., Martinot, D., Crisp, R. J., & Redersdorff, S. (2006). Social comparison, self-stereotyping, and gender differences in self-construals. *Journal of Personality and Social Psychology*, *90*(2), 221–242.
- Hamilton, B. H., & Nickerson, J. A. (2003). Correcting for endogeneity in strategic management research. *Strategic Organization*, *1*(1), 51–78.
- Haushofer, J., & Fehr, E. (2014). On the psychology of poverty. *Science*, *344*(6186), 862–867.
- Iyengar, S. S., & Lepper, M. R. (2000). When choice is demotivating: Can one desire too much of a good thing? *Journal of Personality and Social Psychology*, *79*(6), 995–1006.
- Kaiser, T. (2019). Nature and evoked culture: Sex differences in personality are uniquely correlated with ecological stress. *Personality and Individual Differences*, *148*, 67–72.
- Karlan, D., & List, J. A. (2007). Does price matter in charitable giving? Evidence from a large-scale natural field experiment. *American Economic Review*, *97*(5), 1774–1793.
- Kistler, D., Thöni, C., & Welzel, C. (2017). Survey response and observed behavior: Emancipative and secular values predict prosocial behaviors. *Journal of Cross-Cultural Psychology*, *48*(4), 461–489.
- Kogut, T., & Ritov, I. (2005a). The identified victim effect: An identified group, or just a single individual? *Journal of Behavioral Decision Making*, *18*(3), 157–167.
- Kogut, T., & Ritov, I. (2005b). The singularity effect of identified victims in separate and joint evaluations. *Organizational Behavior and Human Decision Processes*, *97*(2), 106–116.
- Konrad, A. M., Ritchie Jr, J. E., Lieb, P., & Corrigan, E. (2000). Sex differences and similarities in job attribute preferences: A meta-analysis. *Psychological Bulletin*, *126*(4), 593–641.
- Krieglmeyer, R., & Sherman, J. W. (2012). Disentangling stereotype activation and stereotype application in the stereotype misperception task. *Journal of Personality and Social Psychology*, *103*(2), 205–224.
- Mac Giolla, E., & Kajonius, P. J. (2019). Sex differences in personality are larger in gender equal countries: Replicating and extending a surprising finding. *International Journal of Psychology*, *54*(6), 705–711.

- Schulz, J. F., Thiemann, P., & Thöni, C. (2018). Nudging generosity: Choice architecture and cognitive factors in charitable giving. *Journal of Behavioral and Experimental Economics*, *74*, 139–145.
- Simon, H. A. (1947). *Administrative behavior*. Macmillan.
- Small, D. A., Loewenstein, G., & Slovic, P. (2007). Sympathy and callousness: The impact of deliberative thought on donations to identifiable and statistical victims. *Organizational Behavior and Human Decision Processes*, *102*(2), 143–153.
- Sunstein, C. R., & Thaler, R. H. (2003). Libertarian paternalism is not an oxymoron. *University of Chicago Law Review*, *70*(4), 1159–1202.
- Thaler, R. H., & Sunstein, C. R. (2003). Libertarian paternalism. *American Economic Review*, *93*(2), 175–179.
- Thaler, R. H., & Sunstein, C. R. (2008). *Nudge: Improving decisions about health, wealth, and happiness*. Yale University Press.
- Tonin, M., & Vlassopoulos, M. (2013). Experimental evidence of self-image concerns as motivation for giving. *Journal of Economic Behavior & Organization*, *90*, 19–27.
- Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*(4157), 1124–1131.
- Wood, W., & Eagly, A. H. (2012). Biosocial construction of sex differences and similarities in behavior. *Advances in experimental social psychology* (pp. 55–123). Elsevier.

1 Chapter 1

Who is cooperative? Stereotypes in gender and political orientation

Christian Thöni & Jason Wettstein

Abstract

We present two studies that elicit explicit stereotypes about cooperative behavior in a laboratory setting. Like in a normal public goods game, subjects are introduced to the procedures and solve control questions. Before they actually play the public goods game, we elicit incentivized estimates about the contributions of subgroups of the population. In Study 1, we elicit gender stereotypes by asking subjects to guess the average contributions of male and female participants from a reference data set. In Study 2, we use the same design to elicit stereotypes in political orientation. We find no systematic gender stereotype, whereas subjects systematically overestimate contributions of the left-leaning subjects relative to the right-leaning subjects. In both studies, the distribution of the guesses are not significantly different from a normal distribution suggesting that participants do not have opposite stereotypes.

1.1 Introduction

Understanding how stereotypes arise, whether they are accurate, and how they influence behavior has been of academic interests for more than 80 years (see e.g. Katz and Braly 1933). Stereotypes can affect choices, when agents or firms deploy statistical discrimination (Arrow 1973; Phelps 1972). While one can debate the fairness implications of statistical discrimination under accurate stereotypes (Fiss 1976; Lippert-Rasmussen 2006; Cass R. Sunstein 1994), it seems obvious that statistical discrimination based on inaccurate stereotypes is undesirable if not harmful. The goal of our research is to measure stereotypes along two dimensions, gender, and political orientation.

While there are countless behaviors for which stereotypes might be found, we focus on stereotypes in cooperative behavior. Cooperative behavior is particularly important, because it is likely one of the strong determinants of whether people are willing to invest in productive teamwork. Strong stereotypes with respect to cooperative behavior might have adverse consequences for those involved. If, e.g., people falsely hold a strong stereotype that women are uncooperative, then they may refrain from entering collaborative relations and fail to realize potential gains from cooperation. Our goal is to elicit beliefs about average behavior of sub-populations in a well-defined strategic environment. Our workhorse to measure cooperation is the tried and tested linear public goods game (henceforth PGG, Ledyard 1995; Zelmer 2003). We use data from previous studies as reference data sets. We ask participants of the current study to guess the average behavior of either male/female participants or left/right-leaning participants in the reference data sets.

Our primary focus was on gender stereotypes in cooperation. In the PGG, individuals' payoffs are highly dependent on the other members of the group. This strategic situation might favor statistical discrimination. We focus on gender stereotypes, as, in principle, gender is an identifiable and static attribute. Therefore, gender stereotypes in cooperation may lead to discrimination. Gender stereotypes would be difficult to counter, because the trigger (gender) is almost perfectly observable. To our surprise, we did not find evidence for a gender stereotype in our data. We ran a second study to investigate whether this lack of stereotype is due to (i) subjects anticipating gender differences correctly, or (ii) because our measurement tool is unable to capture a stereotype. To control for the latter we apply our stereotypes elicitation mechanism to another characteristic. In our choice of characteristics we were restricted by what was available in our data from previous experiments. Among the available variables, such as age, study, nationality, or religious beliefs, stereotypes in political orientation seemed the most promising to us, especially, since these stereotypes are often found to be

exaggerated (Bordalo et al. 2016).⁹

Following Grossman and Lugovskyy (2011) we define stereotyping as “the act of assigning to a member of a particular group a characteristic or trait based solely on the individuals membership in that group”. Research in psychology and sociology offer similar definitions. According to Correll et al. (2010), stereotypes are category-based generalizations that predict or explain the behavior of others. Based on certain attributes, we infer the behaviors of others. This definition also implies that stereotypes do not have an affective nature per se; they could be either positive, neutral or negative.

Krieglmeyer and Sherman (2012) point out that a stereotype is the result of a heuristic process. They explain that due to limited information or cognitive abilities we rely on shortcuts to make sense of the social environment. Hence, we link certain types of behaviors to certain attributes. Bordalo et al. (2016) investigated the formation of stereotypes and highlighted the role of the context and the representativeness of the group. The fabrication of a stereotype lies in the comparison of at least two groups implying that only the difference between the groups matters.¹⁰

Finally, Madon et al. (2006) and Lee et al. (2013) investigate the accuracy of stereotypes. Whenever they are accurate, they are good predictors of average behaviors; when they are biased, the social group can behave in a different way versus the stipulated stereotype or they are simply unrelated. Therefore, belonging to a group might not predict your social behavior. For instance, being an academic might not predict your cooperative behavior. The accuracy of a stereotype can be assessed on two different levels: qualitatively and the intensity. The former refers to the direction of the stereotype, such as whether men or women are more risk averse, and the latter to which extent the stereotype predicts the behaviors, for example, the extent to which women are more risk averse than men. The second level investigates how informative a stereotype can be. To be informative, the stereotype needs to differ sufficiently from the pooled average of all subgroups. If the behavior of the subgroup is too close to the other groups, then the stereotype is uninformative.¹¹ Overall, these definitions stand in favor of an inclusive approach, in contrast to the common understanding of stereotypes, which is closer to

⁹See the discussion on representative types in the result section of Study 2 on page 43.

¹⁰E.g. the gender attribute is often linked with a type of behavior, such as women being more conscientious or more risk averse compared to men. Whether men and women are in the absolute risk seeker or risk averse does not matter, the relative difference does.

¹¹The informative component is a relative notion. One can consider that it is informative as soon as it helps predict better than chance the behaviors.

the concept of prejudice.¹²

1.2 Related Literature

The literature on stereotypes distinguishes two main categories: implicit and explicit stereotypes. Both types are category-based generalizations but they differ in their awareness. Implicit stereotypes affect behavior, but people are not aware of them, while people are aware of their explicit stereotypes and can verbalize them. According to Arendt (2013), implicit stereotypes will lead to explicit stereotypes when the person has enough introspective capacities. In contrast, a dissociation between these two kinds of stereotypes may be induced by coercive social norms, which may provoke a social desirability bias. Nevertheless, this is still open to debate in the literature (Arendt 2013).

While implicit and explicit stereotypes are distinguished by their awareness, the measurement of implicit stereotypes often overlay the measurement of explicit stereotypes, such that any methods measuring implicit stereotypes might also measure explicit stereotypes. A typical example of an implicit method is when participants have to infer the behaviors of other people by their physical appearance. In this example, it is not made salient to participants which stereotypes they should use to infer behaviors. Participants might use hair color, height, or gender to make their predictions. While they might think that gender is a good predictor of a certain behavior, they might also implicitly - without being aware of this stereotype - think that the hair color indicates the same behavior. In this case, this method might target a stereotype participants are aware of: gender, and an implicit one, they are not aware of: hair color. Overall, many methods are neither completely implicit, nor completely explicit, but rather in-between. Therefore, we make the distinction, that explicit methods focus on one stereotype which is made salient to participants and implicit methods do not make salient which stereotypes are elicited.

Eckel and Grossman (2002) used an incentivized implicit mechanism to measure stereotypes in risk preferences. Participants play a lottery game, with two choices, one riskier than the other one. After playing the lottery, subjects had to guess the decisions of other participants. In order to trigger implicit or explicit stereotypes, each participant had to stand up, such as to be seen by all the other participants. The authors found that the gender of a person standing up was significantly correlated with the guesses of the

¹²For the early study of stereotypes, see Katz and Braly (1933).

lottery choices. Participants assumed that women are more risk averse than men. In their experiment, the stereotype is accurate, i.e., it predicts behavior better than chance.

Daruvala (2007) used a similar design in which participants could see each other without standing up. They confirmed that participants predicted lottery choices based on gender (women are again assumed to be more risk averse than men). Ball et al. (2010) confirmed these observations and added the physical prowess with attributes such as height, strength, and attractiveness, as predictors for the guesses. Taller, stronger, and more attractive people were stereotyped as more risk-tolerant. Overall, stereotypes are qualitatively accurate, as they tend to predict risk choices better than chance in these experiments. On the other hand, the differences between the subgroups are often overestimated. For instance, participants assume women to be risk averse, but to a greater extent than what they actually are.

Grossman and Lugovsky (2011) investigated the robustness of gender stereotypes. For this purpose, each participant had to complete a risk attitude elicitation survey before playing a lottery. Later on, participants who had to guess the choices of other subjects received a part of the risk survey, with which one can infer risk preferences, in addition to the visual information. Survey answers are individual information, presumably more accurate than group attributes; however, the authors find that the stereotypes dominate in the prediction process. In particular, the gender stereotype is shown to be robust to the additional information. The robustness was also confirmed by Grossman (2013) who varied the sequence of information. In a random order, participants were either shown the survey answers first, or they saw the person first.

Castillo and Petrie (2010) study social preferences. In their study, participants played 20 rounds of a public goods game and then had to rank other participants from the one they would like to interact with the most to the one they would like to interact with the least for the upcoming rounds. They implemented three different treatments, one with information about the previous contributions, one with a photo of the participants, and one with both. Beliefs about cooperation were elicited with the participants' preferences for their future group members. They found that race was used to predict behavior. However, this did not hold with the information about previous contributions. Furthermore, gender was not used as a predictor even without the contribution information. This suggests that stereotypes take over the individual information only in some situations, possibly depending on the intensity of the stereotype.

The above examples do not allow us to infer whether participants were aware of the stereotypes they use to educate their guess. As it was not made

salient to participants which stereotype to focus on, these methods are rather implicit. In such studies it is difficult to disentangle stereotypes based on general group characteristics from the influence of individual characteristics. If, for example, subjects can infer risk preferences from the physical appearance of other subjects, then differences in average risk-taking might not reflect a general gender stereotype.¹³ These designs might also provide an incentive to overvalue the information¹⁴, artificially increase the effect by using binary choice sets¹⁵, and might be sensitive to intentions¹⁶.

Before we discuss explicit methods, it is worth noting that there exist some approaches in-between explicit and implicit methods. Vyrastekova et al. (2015) investigated gender beliefs in cooperation. Although their topic is highly similar to ours, they investigate the beliefs using a rather implicit method. In their experiment, participants indicated their contribution conditional on the gender composition of their group. In this elicitation mechanism, factors, such as appearance does not play a role. However, differences in conditional contributions cannot be uniquely attributed to stereotypes about cooperation, as there may be other reasons why subjects would want to differentiate their contribution in response to the gender composition in the group.

In our study, we focus on explicit stereotypes in cooperative behavior by using an explicit method. So far, most of the research done on explicit stereotypes comes from psychology and sociology, mainly using surveys.¹⁷

¹³For instance, the gender attribute can superimpose with other attributes such as height or physical strength (Ball et al. 2010). It could be that height is main stereotype people use when predicting risk preferences, but since it correlates with gender, the researcher might misinterpret this as gender stereotype. Consequently, the gender stereotype might vanish if one could control for all other factors.

¹⁴In the mentioned experiments, there are incentives to seek for information. Participants look for clues to make their guesses and this might overvalue the prevalence of the stereotype.

¹⁵In the literature on risk preferences, participants often face a binary choice. They choose between two options, a risky and a comparatively less risky. The same holds for Aguiar et al. (2009), where they investigate gender stereotype and altruism, but only allow binary decisions. The framing forces a decision between two extremes, which is likely to amplify reported stereotypes.

¹⁶According to Fiedler and Bluemke (2005) and Steffens (2004) implicit responses (signals) can be faked, and especially if participants are incentivized to do so. Their studies use implicit association tests (IAT), where they measure the response time after a stimulus. Participants were able to speed up their response time if they received instructions to fake the IAT.

¹⁷Researchers use various questionnaire items: unipolar (e.g. how much do you agree

There is some concern that these research underestimate stereotypes due to social desirability (Schuman 1997), misunderstanding of the context, a lack of precision¹⁸, or a lack of introspective access¹⁹. Intuitively, implicit methods should have a higher risk of false positives²⁰, while explicit methods should have a higher risk of false negative²¹.

Our design for Studies 1 and 2 follows recent work in experimental economics on explicit stereotypes. Brañas-Garza et al. (2018) study explicit gender stereotypes in a dictator game. Subjects, informed about the gender of the dictator, gave guesses about the amount given. Closer to our design, Dieckmann et al. (2016) asked participants to guess the average score of other nationalities in an effort task and in an honesty game. In both experiments, participants earned money depending on how close their guesses were to the true average as observed in previous experiments with the respective population. In our work, we ask participants to guess the average contribution of subgroups from a reference data set in a public goods game, but unlike the previous experiments, we inform subjects about the overall average and of the proportion of each subgroup. We use strong incentives, which should reduce the social desirability bias, as it is costly to express a socially desirable guess when it deviates from the true guess. We also kept the design simple to avoid noise due to complexity and misunderstanding (Dave et al. 2010). We argue that our method is explicit as we make salient which stereotype we aim to elicit. Even if subjects in our experiment have a priori no informative stereotypes, we incentivize them to form an explicit stereotype.

with this trait), bipolar (e.g. rate this group between rude or polite), with percentage points (e.g. how much of this trait do you believe characterizes this group) or similarity ones (e.g. rate these two groups in terms of abilities).

¹⁸E.g. when asked to rate the driving skills of old and young drivers, participants might think of different aspects, such as parking skills, accident rates, or driving skills to different weather conditions.

¹⁹Participants might not be aware of their own perceptions, due to a lack of introspective access (Arendt 2013). Stereotypes exist, they are used to predict the other's behavior but when we ask about it, participants are not able to express them. For instance, discriminating when hiring people even though when asked about any preliminary belief that certain attributes will predict the performance, people think they do not have any.

²⁰Detecting a stereotype even though it does not exist; e.g. a stereotype channeled through an omitted variable.

²¹Not detecting a stereotype, even though it exists; e.g. strong social norms preventing its elicitation.

1.3 Conceptual framework & experimental procedures

We measure explicit stereotypes about cooperative behavior by letting subjects guess the average contributions of sub-populations in previous experiments (the reference data). We focus on the contribution in the first period of a repeated game because that is where we expect sub-populations to differ most clearly. Before we elicit the guess, subjects of the current study run through the exact same procedures (instructions, control questions) as the subjects that produced the reference data. The elicitation of the stereotypes occurs in the same moment, in which subjects in the reference data studies chose their first contribution.

Why should stereotypes matter? Under standard assumptions, the Nash equilibrium of the PGG is zero contributions for all players. In such an environment stereotypes would be irrelevant, as the game is dominance solvable. However, this is not an empirically accurate prediction of human behavior, as demonstrated in countless experimental PGG studies. It has been shown that the predominant behavioral pattern in PGGs is conditional cooperation, i.e., the willingness to cooperate if others cooperate as well (Fischbacher et al. 2001; Thöni and Volk 2018). For a group of conditionally cooperative players, the PGG has multiple equilibria, which means that beliefs importantly influence actions. Consequently, if players have stereotypes about other players' actions, then they adjust their behavior when interacting with a player of the respective group. Fortunately for our purpose, our reference data shows a large variance in first period contributions, such that there is ample room for stereotypes.

How do subjects form beliefs? While we may not understand the exact process, we can posit that beliefs arise from at least three sources of information. Consider a player i who has to form a belief about player j 's contribution. We assume that beliefs are formed based on previously observed behavior in the respective strategic situation (the PGG in our case), individual and group characteristics of player j . We define group characteristics as common identifiable individual characteristics. This implies that individuals have the characteristics of the group they belong to, in addition to individual characteristics that the group does not have.

$$b_{ij}(c_j^{t-1}, I_j, G_j)$$

When individual information regarding j 's past contributions is available (c_j^{t-1}), then a player presumably uses this to predict contributions in the current situation. This is particularly evident in repeated games, where past

contributions strongly affect beliefs in the current period.²² In the absence of information about past behavior in the respective strategic situation, subjects might use other individual information I_j if available, like e.g., physical appearance, observed behavior in other environments etc. A number of studies discussed above study implicit and explicit stereotypes by providing individual information.²³

Our approach is to eliminate the effect of individual information and to focus on the effect of group information (G_j). We provide our subjects with nothing else than a group distinction (male/female, or left/right-leaning) to base their estimates on. Any stereotype we measure is therefore explicit, i.e., it is fully transparent to the subjects that they are asked to use the group affiliation as the determinant for their guesses.

A stereotype might be accurate but uninformative. We can illustrate this in terms of a simple regression analysis. Let $b_{ij} = \alpha + \beta \mathbb{1}_G + \epsilon$ be the function that represents i 's belief about j 's contribution; α is the average of the subjects in the baseline group (say, males), and β measures the expected difference between the two groups (the gender effect). In our experiment, we will elicit $\hat{\beta}_i$, while we will use our reference data set to calculate the true β . We will say that subjects have accurate stereotypes if their $\hat{\beta}_i$ is reasonably close to β . Whether they consider the stereotype to be informative depends on both the slope ($\hat{\beta}_i$) and their estimate about the variance of the error term (ϵ). A very low variance means that a subject expects beliefs to be on average very close to the true value. As a metric for the degree to which subjects think their stereotype is informative we could use their belief about the R^2 of the above-mentioned regression.

1.3.1 Experimental procedures

Our experiment starts like a normal PGG experiment with the subjects reading the instructions of the standard PGG game and answering the control questions. For the control question, they had to calculate their payoff and the payoffs of the other group members for various combinations of contributions.

²²This begs the question of whether past behavior is in fact a reasonable predictor for future behavior in PGGs. The results of Volk et al. (2012) suggest that in fact there is a lot of predictive power in the information about past behavior.

²³This point is related to the green beard hypothesis discussed in evolutionary biology (Dawkins 1976). It asks the question of whether some physical attributes (like a green beard) might evolve to credibly signal a type and help cooperative actors to cooperate only among themselves (see also Fehr and Fischbacher 2005; Frank 2005).

We checked their answers and their understanding before we started the experiment. Up to the point where the experiment starts, we do not inform them about the stereotype elicitation. This procedure was chosen to ensure that subjects had the same information about the game as the experimental subjects that produced the reference data.

The PGG was played following the protocol of Herrmann et al. (2008): participants are matched in groups of four. Group composition remained the same during the whole experiment (partner matching). In each period, participants received an endowment of 20 ECU (Experimental Currency Units) and had to decide how many ECU to allocate to a public good. The monetary payoff of the stage game is:

$$\pi_i(g_i, g_{-i}) = 20 - g_i + \frac{2}{5} \sum_{j=1}^4 g_j, \quad (1)$$

where g_i is player i 's contribution and $\frac{2}{5}$ is the marginal per capita return. Under standard assumptions, this game has a unique Nash equilibrium, in which nobody contributes. Every ECU contributed pays back 0.4 to the contributor, and 1.6 to the group, thus constituting a social dilemma where contributing is costly to the individual but beneficial for the group.

1.3.2 Stereotype elicitation

After instructions and control questions subjects proceeded to the stage where we elicited stereotypes. We asked participants to guess the average contributions of sub-populations 0 and 1 (men/women or left/right-leaning subjects) in the first period of a standard public goods game from a reference data. All instructions for this stage were presented on the screen. Participants went through three steps to enter their guess of the average contribution by sub-population. The first screen explained that they would have to guess the average contribution of the two groups in the first period of the public goods game from the reference data (Element 1 in Figure 1.1, henceforth: E1.1.1).²⁴ The monetary profit function of the stereotype elicitation stage follows $\pi_i(\hat{\beta}_i) = \max\{0, 100 - 500|\hat{\beta}_i - \beta|\}$, rounded up to multiples

²⁴We asked to guess the average contribution in the first period only because this is arguably the moment where differences between subgroups of the subject pool are most pronounced. Over the course of a repeated interaction, things get much more complicated, as the subgroups might display differences in their strategic reaction to the behavior of other subjects.

of 10. We communicated the incentives in the instructions using a table, indicating that an absolute deviation of .02 or less would pay 100, and for any increase in the deviation of .02 the payoff would be reduced by 10 units, such that deviations larger than 0.2 would not be rewarded (E1.1.2). The first screen also contained information about the reference data we used to calculate β . We emphasized that the data was generated under highly similar circumstances, with a similar subject pool, i.e., students, the same age, and under similar experimental conditions (E1.1.3). Participants had to press a button to move to the next screen (E1.1.4). No time limit was imposed at this stage. Figure 1.1 is a translated screenshot of the first step from Study 1, which investigates gender stereotypes in contribution. From the top to the bottom of the screen, participants had access to the general matter of these stages, i.e, what question participants will have to answer (E1.1.1), the profit function (E1.1.2), information regarding the reference data (E1.1.3), and the button to get to next step (E1.1.4, see Figure A1.1 in the Appendix for the original screenshot and its translation).

Stereotype elicitation: Step 1 - Gender stereotype

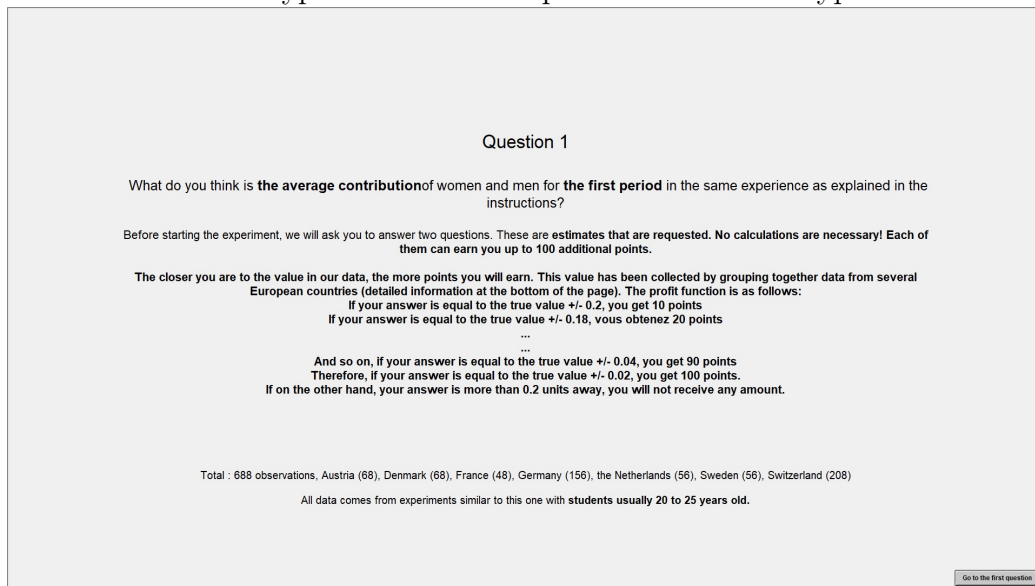


Figure 1.1: Translated screenshot - Study 1 - Screen 1

In the second screen, participants were provided with the information about the overall average contribution in the first period of the reference data set and were given the percentages of subjects in group 0 and 1, f_0 and f_1 , respectively (Element 1 in Figure 1.2, henceforth: E1.2.1). Formally

the overall average corresponds to $f_0\alpha + f_1(\alpha + \beta)$. We then asked them to indicate their guess. We randomized the order of group 0 and 1 to control for order effects, i.e., subjects either enter $\hat{\alpha}_i$, or $\hat{\alpha}_i + \hat{\beta}_i$. Subjects were randomly assigned to one of two questions. At the bottom of the screen in the second step we repeated the information about the reference data set (E1.2.2). When subject entered a number they had to press a “calculate” button (E1.2.3). Figure 1.2 is a translated screenshot of Study 1, from the second step in the elicitation mechanism of gender stereotypes in cooperation. On the left part of the screen, participants had information about the overall average contribution and the percentage of females and males in the reference data set (E1.2.1). On the right side, they had to indicate their guess about the females or males average contribution. Here, the question is the following: “What do you think is the mean contribution of female participants in this kind of experiment?”. Underneath this question, they have the calculate button (E1.2.3), which would make appear the confirmation step (Figure 1.3). On the bottom of the screen, participants still see the information about the reference data set (E1.2.2, see Figure A1.2 in the Appendix for the original screenshot and its translation).

Stereotype elicitation: Step 2 - Gender stereotype

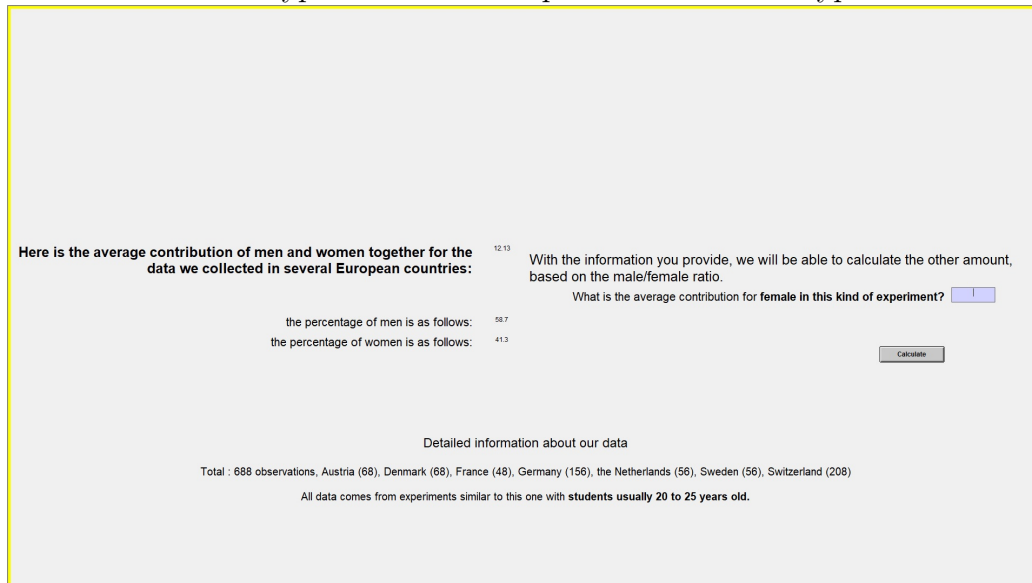


Figure 1.2: Translated screenshot - Study 1 - Screen 2

As we wanted to minimize confusion in the elicitation of the stereotypes we added a third step, in which subjects were asked to confirm their entry.

We used their guess about the average contribution of group 0 (1) to calculate the respective average of the other group, such that the values are consistent with the overall weighted average. Figure 1.3 shows an example for the gender stereotype elicitation. Given this information, subjects could either confirm the values or press the “change” button to return to the previous screen and enter a new value. Upon pressing calculate they would again be presented the screen in Figure 1.3. Subjects could change their entries as often as they wanted.

Confirmation of the guesses - Step 3

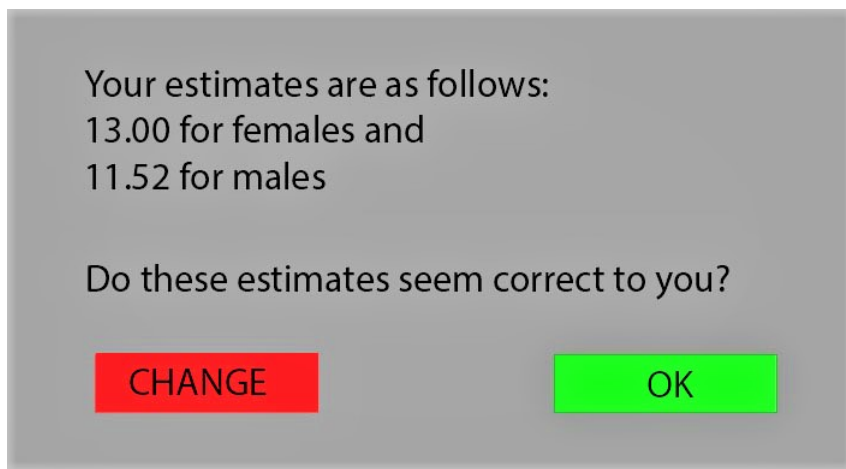


Figure 1.3: Screenshot Screen 3

After all subjects had finished the stereotype elicitation, the experiment proceeded with the 10-period public goods game, followed by a post-experimental questionnaire.

1.4 Hypotheses

The null hypothesis for both stereotype elicitation experiments is that subjects' estimations about the average differences between the two genders and political leanings are not systematically different from zero ($\hat{\beta}_i \sim 0$)

While the literature on gender effects suggests no gender differences in cooperation (Balliet et al. 2011), we expect to observe systematic gender stereotypes in cooperation. We expect our subjects to overestimate female cooperation relative to male cooperation.

We think it is safe to assume that our subjects are not familiar with the meta-analytic literature on gender differences in cooperation. Therefore,

when trying to estimate the average contribution of a sub-group, participants likely use other sources than the scientific literature to make their guess. An obvious one is through personality traits, especially agreeableness. Agreeableness is one of the five major personality traits and refers to behaviors that are generally perceived as kind, sympathetic, and cooperative (Graziano and Eisenberg 1997; Thompson 2008). Gender differences in agreeableness are suggested by social role theory. The social role theory posits that sex differences have their origin in different social experiences (Eagly and Wood 1999; Wood and Eagly 2012). Due to biological differences, such as pregnancy, or average strength and height, women and men have historically taken on different roles (Balliet et al. 2011). These differences have raised different expectations by gender. While men are expected to be more agentic, that is assertive and competitive, women are expected to be more communal in orientation, less selfish, and more friendly (Eagly 2009). In other words, women are expected to be more agreeable. In the same vein, in a large study in 25 countries, Williams et al. (1999) find that women are expected to be more agreeable than men. This suggests that these historical different expectations still prevails. While the reproduction of these differences happens through different channels, such as self-categorization and self-stereotyping (Guimond et al. 2006), many research bring evidence that women display a higher level of agreeableness (Costa Jr et al. 2001; McCrae and Terracciano 2005). Finally, as agreeableness is associated with higher contributions in public goods game (Volk et al. 2011), we expect participants to stereotype women as more agreeable and thus to contribute more.

Hypothesis 1 *Subjects expect female average contribution to be higher than male average contribution, i.e., $\hat{\beta}_i > 0$*

For the political stereotype in cooperation, the scientific literature reports some correlations between political orientation and trust or cooperative behaviors (Fehr, Fischbacher, et al. 2002). However, others find that these correlations are not robust (Anderson et al. 2005). But again, how people stereotype parties might not be solely derived, if at all, from the scientific literature.

The left-right continuum is indeed a construct that has its limitations when trying to capture political attitudes. Nevertheless, some general tendencies are fairly uncontroversial. For instance, left-leaning people tend to favor more involvement of the government and more redistribution than the right-leaning people.

According to Wilson et al. (2013), the left-wing is associated with higher spending on social programs and public spending. As the public goods game

mimics some features of governmental redistribution schemes, such as the presence of a public good and the equal redistribution of this public good, participants likely stereotype left-leaning individuals to be more cooperative. Therefore, we expect participants to associate left-leaning individuals with a higher level of contribution.

Hypothesis 2 *Subjects expect left-leaning individuals' average contribution to be higher than right-leaning individuals' average contribution, i.e., $\hat{\beta}_i > 0$*

1.5 Study 1

1.5.1 Experimental Design

For Study 1 we conducted two sessions with a total of 48 subjects. Most of the participants were first-year bachelor students from various fields of study. The sessions lasted for 90 minutes. The mean age was 20 years. Participants received 10 CHF (Swiss francs) as a show-up fee in addition to the earnings from the experimental tasks. We conducted our laboratory experiment at the LABEX (HEC Lausanne). We recruited participants via ORSEE (Greiner 2015) and ran the experiment using zTree (Fischbacher 2007).

As reference data set for the “true” gender effect we use a subset of the cross-cultural PGG data reported in Herrmann et al. (2008). We use only the observations from subject pools that are culturally close to the subjects in our laboratory.²⁵ The observations in the reference data set stem from subject pools with very similar socio-economic characteristics as the subject pool in Lausanne. We will refer to this data set as RD1.

1.5.2 Results

In RD1, the average contribution of female participants in the first period of the PGG was 11.99 (SD=6.27), and 12.23 (SD=7.07) for male participants. Thus, the “true” gender effect for our experiment is $\beta_1 = -0.24$, indicating that males contribute slightly more than females. Relative to the possible range of $[-20, 20]$ the difference is small and, despite the large sample, does not reach significance ($p = 0.379$, Wilcoxon rank-sum). This is in line with the literature on gender effects in cooperation (Balliet et al. 2011; Thöni, Volk, and Cortina 2020). Moreover, the predictive power of gender on the

²⁵Overall, we have 688 observations: from Austria (68), Denmark (68), France (48), Germany (156), Netherlands (84), Sweden (56) and Switzerland (208).

contribution in the first period of the PGG is virtually zero ($R^2 = 0.0003$) in RD1. This means that an accurate stereotype is uninformative.

In Study 1, participants reported on average 11.64 for the female contribution ($SD = 2.19$) and 12.48 for the male contribution ($SD = 1.54$).²⁶ To our surprise, we find no support for Hypothesis 1: Our point estimate for the average stereotype between female and male contributions is close to zero (-0.84). Qualitatively we find that 44.7 percent of the subjects think that the female average contribution is higher than the male average, while 51.1 percent think the opposite, and 3.1 percent guess that the two are exactly equal, which is very close to the true value of -0.24 . We use a one sample student t-test to test whether the distribution of the difference between the female guess and the male guess ($\hat{\beta}_i$) is systematically different from zero. The result indicates that the difference of guesses is not statistically different from zero ($t(46) = -0.27$, $p = 0.129$).

Finding similar averages does, however, not necessarily mean that subjects have accurate stereotypes. It could be that half of the subjects believe that women are far more cooperative than men and half of the subjects think the opposite is true.

Figure 1.4 shows a kernel density plot of the guesses and the true β_1 as observed in RD1 (vertical line).²⁷ The density ranges from the lowest guess (-7.04) to the highest guess observed in the data (6.37). The plot shows that the guesses about the gender differences are approximately normally distributed around the true value.

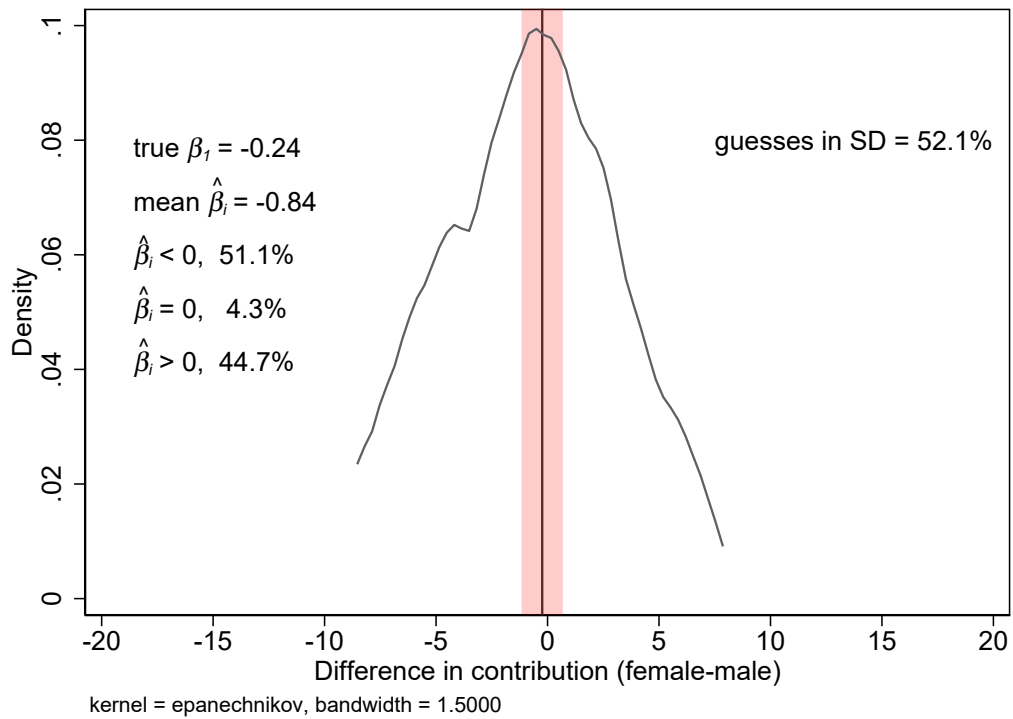
Given that the true β_1 for gender differences is close to zero it is not surprising that we do also not find significant differences between the guesses and the true gender effect in cooperation. A t-test comparing the guesses to the true gender effect results in $t(46) = -1.104$, $p = 0.275$.

While there is no systematic effect in gender stereotypes, there are still stereotypes that are far off the true values to be found in the data. In order to interpret the accuracy of the stereotypes, we compute the percentage of participants within one standard deviation of the gender effect in the

²⁶One subject entered a guess for the averages which is led to values of the opposite gender outside of the admissible range ($0 - 20$). We dropped this outlier from the main analysis. None of the qualitative results change if we include the outlier.

²⁷All kernel density figures have the same bandwidth to facilitate comparisons between the distributions. We choose a slightly smaller bandwidth than the one given by the Silverman's rule of thumb (Silverman 1986) in Study 1 to equal the one in Study 2. Therefore, our figures tend to be slightly under-smoothed.

Distribution of gender stereotypes in cooperation



Gender stereotype elicitation: guesses about the difference in contribution between female and male subjects, the true difference (female contribution minus male one) from the reference data (vertical line). Standard deviation from the reference data (vertical lighter line). True β_1 is the value from RD1. Mean $\hat{\beta}_i$ and the percentages, above, equal, and below 0, are statistics from Study 1. On the right side, the percentage of guesses within the standard deviation of the reference data set.

Figure 1.4: Kernel Density - $\hat{\beta}_i$

reference data set RD1 (shaded area in Figure 1.4).²⁸ We find that a small majority, 52%, of the subjects are within one standard deviation of the true value.

To conclude, we find no evidence for our hypothesis that females are expected to be more cooperative than males. Subjects' guesses are on average highly accurate, and a small majority of the subjects indicates a guess which is very close to the true values (± 1 SD).

We expected a gender stereotype in cooperation based on two premises: participants associate women with "agreeableness", and they associate agreeableness with higher cooperation. As we did not find support for a systematic gender stereotype, we suppose that participants did not make one or both of these associations. More precisely, they may not associate cooperative behaviors with personality traits, or they think that personality traits predict better economic behaviors than gender and do not make assumptions regarding the distribution of these personality traits among genders, i.e., they do not think that females are more agreeable than males. This latter explanation would be in line with Heckman et al. (2019) and Jagelka (2020), who show that personality traits dominate demographic factors in predicting economic behaviors.

1.6 Study 2

1.6.1 Experimental design

Similar to Study 1, we recruited 48 student participants, all of which were students of the University of Lausanne, mainly in the first year of a bachelor's degree. We ran two sessions of 90 minutes. The mean age was 19.6 years, they received 10 CHF each as a show-up fee, and an extra payment depending on their performance.

In Study 2, we (i) replicate the elicitation of gender stereotypes from Study 1 and (ii) we investigate cooperation stereotypes in political orientation. For the latter, we use a new reference data set, which was generated with the exact same subject pool at the University of Lausanne. The data stems from the study of Kistler et al. (2016), who conducted a series of repeated public goods games. The reason for using a different reference data

²⁸As the sample in reference data set is not balanced for gender, we compute the standard deviation by bootstrapping one male and one female and matching them with each other to obtain the standard deviation of their contribution for the first period of the public goods game. We repeated the operation 100,000 times.

set as for the gender stereotypes is twofold. First, the reference data set RD1 does not contain the variables for political leaning in all locations. Second, as opposed to gender, political leaning is a much more elusive concept. The notion of what constitutes left- and right-leaning depends importantly on the political landscape of the country the subjects live in.²⁹ They ran the experiment in 2016 in the same location under very similar circumstances, with a similar subject pool (we will refer to this as the reference data set RD2, see Table A1.2 in the Appendix for summary statistics). As for the experimental design, we kept the exact same setting as in Study 1 and added extra stages to elicit the political stereotype. We elicit the political stereotype as follows: we ask subjects to indicate the average total contribution of left-leaning or right-leaning individuals given the total average contribution. In our reference data, political orientation was measured on a 10-point scale. Participants, in the reference experiment, were asked to indicate their position on this 10 point left-right scale. Subsequently, we explained that we classified subjects who entered a value between 1 and 5 as left-leaning, and those who were between 6 and 10 as right-leaning. Similar to Study 1 the elicitation of political orientation stereotypes takes place in three steps. In addition to the information about the reference data set, we provided a screenshot of the exact political orientation question that participants of the experiment for the reference data answered.

Furthermore, in Study 2 we asked participants to indicate what they believed was the morally fair contribution for the first period before starting to play the actual public goods game. We mentioned that this question would not earn them any ECU. We wanted to elicit the moral contribution to be able to contrast the stereotype to a moral benchmark afterwards. The rest is *ceteris paribus* as in Study 1.

1.6.2 Results

In Study 2, participants reported on average 11.81 for the female contribution ($SD = 2.64$) and 12.36 for the male contribution ($SD = 1.86$). Our point estimate for the average stereotype between female and male contributions is -0.55 . The difference between the guess for female and the guess for male is not different from zero ($t(47) = -0.84$, $p = 0.400$) nor to the true

²⁹In Lausanne at the time of the experiment, the executive was more left-wing, and the same holds for the executive part at the cantonal level. For the legislative part, the communal council was left-leaning whereas there was a right-leaning majority at the cantonal level.

value from RD1 ($t(47) = -0.48, p = 0.631$). We confirm the results from Study 1. Qualitatively, we find that 39.6 percent think that female average contribution is higher than the male average contribution, while 58.3 percent think the opposite and 2.1 percent think that both genders contribute equally (see Figure A1.5 in the Appendix for the distribution of the gender stereotype in Study 2).

While we confirm the lack of systematic gender stereotypes, we find differences with respect to political orientation. In RD2, left-leaning subjects contributed on average 10.69 ($SD = 6.78$) in the first period of the PGG whereas right-leaning subjects contributed 12.32 ($SD = 7.15$). Thus, the “true” political effect for our experiment is $\beta_2 = -1.64$. The difference between the two subgroups is significant ($p = 0.023$, Wilcoxon rank-sum test). In the same vein, the predicting power - R^2 - of political orientation on the contribution in the first period of the PGG is 0.0133 in RD2. Therefore, assuming the information about the members’ political orientation in a group is available, an accurate stereotype will predict contributions somewhat better than chance.

Participants reported on average 12.07 for left-leaning ($SD = 1.98$) and 10.25 for right-leaning subjects ($SD = 3.62$). We find support for Hypothesis 2: our point estimate for the average political stereotypes is 1.82. Qualitatively, we find that 68.8% of the subjects think that the left-leaning average is higher than the right-leaning average, 27.1% think the opposite, and 4.2% think that right and left-leaning subjects contribute on average the same amount.

We use a one sample student t-test to compare the distribution of the difference between the left- and right-leaning subjects to zero.³⁰ The result indicate that the difference of guesses is statistically different from 0 ($t(47) = 2.25, p = 0.029$). As the guesses systematically deviate from zero, we can infer that on average, the left-leaning individuals are perceived as more cooperative. Figure 1.5 shows a kernel density plot of the guesses and the true β_2 as observed in RD2 (vertical line). The guesses range from -13.26 to 12.90 in the data.

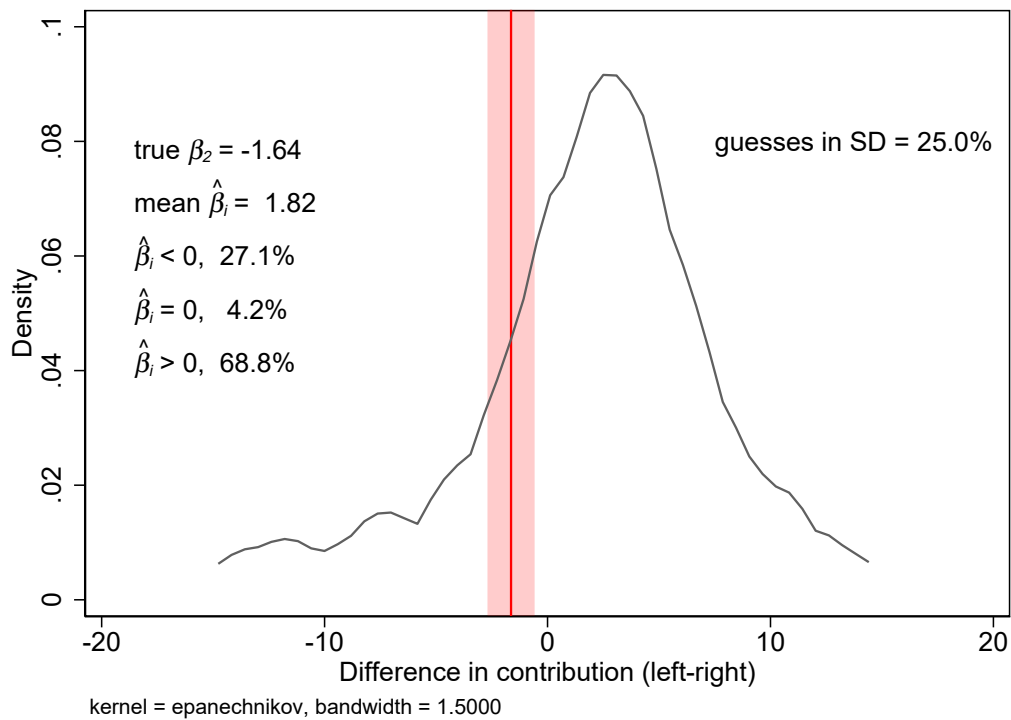
As in Study 1, we estimate the accuracy of the stereotypes. We find that only 25.0% of the subjects reported a value close to the true value.³¹

³⁰ $\hat{\beta}_2$ is the parameter for the difference between the guess for the average left-leaning contribution minus the guess for the average right-leaning contribution.

³¹The sample in the reference data set is politically unbalanced, therefore, we use the same procedure as in Study 1 by bootstrapping one left-leaning subject and one right-leaning subject. We repeated the operation 100,000 times.

The cooperation stereotype about political orientation is fairly inaccurate compared to our RD2 ($t(47) = 3.71, p = 0.000$). Bordalo et al. (2016) investigated political stereotypes and found that political stereotypes are exaggerated compared to true means and that they are mostly the consequence of representative types. In politics, there is often an opposition of ideas, where one group defends one idea and another one defends the opposite idea. In a debate of ideas, the focus is on contradictions, which might exacerbate the opposition. This could explain the systematic deviation in our data about political stereotypes.

Distribution of political orientation stereotypes in cooperation



Political stereotype elicitation: guesses about the difference (left-leaning contribution minus right-leaning contribution) in contribution between left-leaning subjects and right-leaning ones, the true difference from the reference data (vertical line). Standard deviation from the reference data (vertical lighter line). True β_2 is the value from RD2. Mean $\hat{\beta}_i$ and the percentages, above, equal, and below 0, are statistics from Study 2. On the right side, the percentage of guesses within the standard deviation of the reference data set.

Figure 1.5: Kernel Density - $\hat{\beta}_i$

1.7 Additional analyses

In this section, we (i) explore the self-serving bias, (ii) elaborate on the distribution of the stereotypes, and (iii) investigate priming effects of the elicitation stages. For the following analysis we combine the data from Study 1 and 2.

1.7.1 Self-serving bias

We use our subject’s gender and political orientation as explanatory variables for the stereotypes. A self-serving bias would occur if more participants rate their in-group as more cooperative than their out-group, i.e., being a male and thinking that males contribute more than females or being a female and thinking that females contribute more than males. In Table 1.1 we split the samples by gender and by which gender is believed to contribute more. We find that gender and stereotypes about the gender contribution are independent ($p = 0.677$, Fisher’s exact), suggesting that self-serving bias is not an important determinant of gender stereotypes.

In group self-serving bias - Gender stereotypes		
Self-serving bias	Male contribute more	Female contribute more
Male participants	24	20
Female participants	30	20

Pearson $\chi^2(1) = 4.69$ $Pr = 0.594$
Fisher’s exact = 0.677.

Table 1.1: Contingency table

While neither male nor female participants rate their in-group systematically as more cooperative than the other group, we observe a self-serving bias for the political orientation stereotypes ($p = 0.044$, Fisher’s exact). Table 1.2 reports the frequency distribution of the self-assessed political orientation of the participants and the stereotypes about which group, between the left- and the right-leaning, contributes more. While right-leaning participants are balanced in who they think contributes more, left-leaning participants tend to think that left-leaning people contribute on average more than the right-leaning ones.

Next we investigate subjects’ views about the morally right contribution. To our surprise, participants reported on average only 12.16 (SD= 6.30) as

the moral contribution. This challenges the general assumption that cooperating is positively valued and thus also that rating your in-group as more cooperative enhances your self-esteem. The moral reported contribution is also a very significant predictor of the first period's contribution ($p = 0.000$), suggesting that participants play according to what they believe is the moral contribution. The coefficient of correlation - R - between the moral benchmark and the contribution in the first period is 0.54. While these R coefficients differ slightly between left-leaning subjects and right-leaning ones with 0.49 and 0.59 respectively, they differ much more across gender. While the R coefficient for males reaches 0.70, it reaches only 0.32 for females. Females/males or left/right-leanings moral benchmarks are not statistically different. While we find that 41.67% of the participants contribute less than their moral benchmark, 43.75% contribute the same amount, and 14.58% contribute more than their moral benchmark. Interestingly, females and left-leanings have the most important decrease between their moral benchmark and their contribution in Period 1 with on average a decrease of 2.62 ECU for females and 3.53 for left-leanings. This result is not explained by females or left-leanings reporting a systematic higher number than males or right-leanings in the moral benchmark. For instance, left-leaning subjects report a higher number for the moral benchmark and contribute less compared to right-leanings, but females report a lower number and contribute less compared to males. Finally, there is conflicting evidence regarding how contribution is morally perceived. On one hand, the moral benchmark is on average for each group higher than the contribution in Period 1. This suggests that participants think that they should morally contribute more than they do. However, on the other hand, the overall average moral benchmark is far from the maximum possible contribution.

In group self-serving bias - Political orientation stereotypes		
Self-serving bias	Left leaning contribute more	Right leaning contribute more
Right-leaning participants	9	8
Left-leaning participants	24	5

Pearson $\chi^2(1) = 4.69$ $Pr = 0.030$
Fisher's exact = 0.044.

Table 1.2: Contingency table

1.7.2 Distributions of the stereotypes

The elicitation mechanism allows us to investigate the distribution of a stereotype on a group level. Aside from the graphical interpretation, the measures from the normalized third, the skewness, and fourth, the kurtosis, moments provide tools to study the group stereotype. Table 1.3 reports the kurtosis and the skewness coefficients. We find that both guesses have a very similar shape to a normal distribution and that they are relatively symmetric. The skewness and the kurtosis coefficients for the normal distribution are 0 and 3, respectively. The first is a measure of symmetry and the second of the distribution of probability mass around the center. Skewness and kurtosis test for normality returns a p -value of 0.480 for the gender guesses and 0.067 for the political guesses, both above the 5% level that would allow us to reject that they are normally distributed. On the other hand, we obtain a p -value of 0.03 for moral reported contributions, thus rejecting the normal distribution. In other words, participants reported guesses that are close to each other possibly indicating that individual stereotypes might derive from group stereotypes (Gilmour 2015; Le Pelley et al. 2010).

As observable in the density figures, the narrow shape and the symmetry lead us to believe that these stereotypes do not have contradictory components. An example of this would be that half of the participants highly overestimate the level of cooperation of females, while the other half highly underestimate it. However, we observe, that for the gender stereotype, most of the participants indicated a very small either positive or negative stereotype. Participants assumed that the contribution rate of men and women were almost equal. The distribution for the political stereotype also indicates a shared stereotype even if this latter stereotype is mostly exaggerated positively compared to the reference data value. As we did not provide any baseline in the moral stage, the comparison with the stereotypes is rather spurious. Nevertheless, we observe that what constitutes a moral contribution is not shared among the participants.

1.7.3 Priming

Because participants played the public goods game after either only the gender stereotype elicitation stage or the gender stereotype elicitation stage, the political one, and the moral one, we investigate the effect of these stages on the contributions in Period 1. These stages are unusual because we provide participants with more information, i.e., we give them the average contribution amounts from previous experiments.

There is a vast literature related to the social identity theory, that in-

Dispersion of	N	Skewness	Kurtosis
Gender guesses	95	0.10	3.43
Political guesses	48	-0.66	3.70
Moral contribution	48	-0.11	1.89

We aggregated the data from both Study 1 and 2 for the gender kurtosis and skewness coefficient.³²

Table 1.3: Descriptive statistics

investigates the salience of stereotypes and their influence on actual behavior (Akerlof and Kranton 2000). Guimond et al. (2006) show that the identification to a social group leads to stereotyped behaviors attributed to this group. Boschini et al. (2012) show that rendering a stereotype salient activates behaviors accordingly, i.e., subjects, who identify themselves with a social group comply with the stereotyped identity of the group. In a dictator game experiment, they render salient the stereotype that females are more altruistic than males and find that it increases gender differences in generosity in the direction of the stereotype in a gender-mixed environment. More recently, Cohn et al. (2015) show that the reinforcement of social identity leads to stereotyped behavior. In their experiment, they show that rendering salient the criminal identity triggers more stealing. They distinguish between priming effects and social identity as participants that did not have the “criminal identity”, i.e., who were not convicted, did not steal more after being primed by stereotypes related to crime. This literature shows the interest in investigating the activation of stereotypes in social interaction.

We test if women/men or left/right-leaning participants would contribute in line with their guesses in Table 1.4. We use their reported guesses, their gender, and their reported political orientation as explanatory variables for their contribution in the first period. We classify participants according to their belief about their average group contribution. If they think that their group is on average a higher contributor than the opposite group, i.e., males think males contribute more, we identify them as better contributors (BC). On the contrary, if they think that their group contributes on average a lower amount, we classify them as lower contributors (LC). In the current experiment, participants had to indicate their gender and if they identify themselves as left- or right-leaning, such that no option in between was possible. Therefore, BC and LC are mutually exclusive categories, but they are also close to exhaustive unless the participant provided an equal value either

for both gender or both political orientations.

If the activation of the stereotype is effective, we should observe that the ones that believe they are better contributors to have a positive coefficient and the ones that believe their in-group is a lower contributor to have a negative coefficient. Therefore, the values of interests here are the signs of the coefficients and their significance rather than the effect sizes. We observe a significant coefficient for the political orientation but the sign does not go in the expected direction for the BC. Therefore, unless the activation of the stereotype is effective only for lower contributors, which is not supported by other empirical findings as cited above, we do not find any evidence for the activation of the stereotypes.

The lack of stereotype activation might be attributed to a low identification with the group. In our experiment, participants knew neither the gender nor the political orientation of the other members of their group. This has two implications: the utility derived from identifying yourself to a group might be lower and you might identify yourself less easily since you do not know which attributes are present in your group. Furthermore, participants might have imperfect self-knowledge regarding their political orientation, i.e., participants face uncertainty regarding either their preferences or how these preferences are distributed on the left-right political continuum (Bénabou and Tirole 2003). Since participants do not strongly identify with a political group, they might comply less with the stereotyped behavior of this group. We find support for this hypothesis as we observe that a substantial part of the participants think they are politically close to the center, i.e., we find that 39.6% of the participants chose either 5 or 6 on the 10-points Likert scale left-right continuum. This suggests that some participants would not fully self-identify as being left-leaning or right-leaning.

Still in line with the social identity theory, individuals gain utility from behaving in a way to fit in or be accepted by other individuals. This often results in behaviors close to the norms (Akerlof and Kranton 2000). We find a significant lower mean contribution in Study 2 compared to RD1, from 12.1 on average to 9.56 ($p = 0.011$, Student's t-test) and a marginally significant lower contribution compared to RD2 from 11.3 to 9.56 ($p = 0.095$, Student's t-test).

As the subject pool is virtually the same between the current studies and RD2, we compare the occurrence of high contributions (18 or more, 19 or more, and 20) and low contributions (1 or lower, and 0) in the first period to RD2 in Table 1.5. We find a decrease in the number of high contributions, but no decrease in the number of low contributions. The coefficients are already negative in Study 1, but become significant ($p < 0.05$) in Study 2. In line with Fischbacher et al. (2001), most of the participants are conditional

Investigation of priming effects on contribution amount in Period 1 from the elicitation stages

	Dependent variable: Contribution in Period 1			
	Study 1	Pooled	Study 2	Pooled
In Group BC - Gender	3.690 (2.387)	-1.485 (3.793)		
In Group LC - Gender	4.180 ⁺ (2.267)	-1.095 (3.781)		
In Group BC - Political			-10.469** (1.403)	
In Group LC - Political			-8.429** (1.612)	
In Group BC - Both				-0.761 (1.603)
In Group LC - Both				1.517 (1.622)
Constant	7.500** (1.826)	11.667** (3.658)	19.000** (0.730)	10.310** (1.111)
<i>F</i> -test	1.8	0.1	34.2	1.0
Prob > <i>F</i>	0.171	0.906	0.000	0.378
<i>R</i> ²	0.016	0.002	0.114	0.017
<i>N</i>	48	96	48	96

Notes: OLS estimates. In-group BC refers always to participants that are members of the group, either gender or political orientation and rated their in-group as better contributors. In-group LC refers to participants that are members of the group and rated their in-group as lower contributors. Both BC refers to participants who are in both groups: political and gender and who rated their in-group as better contributors. Both LC refers to participants who are in both groups: political and gender and who rated their in-group as lower contributors. All LC and BC categories are dummy variables. Pooled indicates that data from Study 1 and Study 2 are combined. Robust standard errors in parentheses. ⁺ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 1.4: Priming for the 1st period

cooperators, i.e., their contributions are based on others' contributions. In general, in the first period, there is normally no baseline. However, in our experiment, we inform subjects about the average contribution in RD1 or RD2. This likely serves as a reference point for the first period and may trigger conditional behaviors. It also seems that displaying one baseline does not shift significantly behaviors but two baselines do.

Investigation occurrences of extreme contributions

Contribution amount	20	19-20	18-20	0-1	0
Study 1	-0.018 (0.069)	-0.021 (0.069)	-0.043 (0.069)	-0.006 (0.004)	-0.022 (0.043)
Study 2	-0.122* (0.059)	-0.125* (0.059)	-0.126* (0.062)	0.015 (0.021)	-0.001 (0.047)
Constant	0.289** (0.024)	0.292** (0.024)	0.314** (0.025)	0.006 (0.004)	0.106** (0.016)
<i>F</i> -test	2.1	2.2	2.2	.	0.1
Prob > <i>F</i>	0.119	0.108	0.116	.	0.875
<i>R</i> ²	0.007	0.007	0.007	0.004	0.000
<i>N</i>	456	456	456	456	456

Notes: OLS estimates. The dependent variables are dummies, 20 represent the occurrence of a contribution of 20 tokens in the first period, 19-20 represent either 19 or 20, etc. Study 1 and Study 2 are also dummy variables, they are compared to the dataset from Kistler et al. (2016). Robust standard errors in parentheses. + $p < 0.1$, * $p < 0.05$, ** $p < 0.01$.

Table 1.5: Contribution in the 1st period

Other possible channels can, as well, explain this shift, such as compliance to norms³³ or experimenter demand effect³⁴ (see Zizzo 2010). Overall, our experiment shows the implication of displaying a baseline to participants on the cooperation level.

The presence of the moral elicitation stage in between might possibly bias the contribution in the first period or at least bring noise. Therefore, we cannot make strong inferences with respect to the shift we observe and the provision of baselines.

³³In the case study of organ donation, displaying that an important number of people are already registered to donate their organs triggers more compliance. This effect is attributed to the desirability to comply with the majority (see Thaler and Cass R. Sunstein (2008))

³⁴By displaying the norm, we emphasize the experimenter expected behavior and participants comply to it

1.8 Discussion

To the best of our knowledge, our design differs from previous studies as we provide participants with the combined average of both subgroups. Subjects have information about the exact average and indicate a difference between two subgroups rather than an absolute value. This information reduces the possible overestimation bias that may occur when participants are not aware of average contributions. As the kurtosis and skewness show, most of the guesses are close to each other supporting this assumption. Nevertheless, the presence of the confirmation stage, which highlights the guessed difference of contribution between the groups, might reduce stereotypes, thus leading to an underestimation of the stereotypes. Firstly, whenever we ask participants to quantify differences as opposed to asking them merely which group contributes more, we complicate their task and this might affect their cognitive process. Secondly, it is probably rare to be presented with both groups estimated mean contributions at the same time, such that participants can directly compare the two numbers.

Finally, our design allows us to compute the difference between the guesses to quantify the size of the stereotype. As mentioned above, these numbers are more likely to be underestimated. We observe that left-leaning individuals are expected to contribute on average 9.10% more than right-leaning ones and men are expected to contribute only 3.47% more than women.³⁵ As mentioned in the introduction, the accuracy of a stereotype is composed of two dimensions: the direction, qualitatively, and the intensity, quantitatively. In the literature, stereotypes are often accurate for the direction but overestimated in the value, however, in Study 2 we observe both inaccuracies.

1.9 Conclusion

Investigating stereotypes is a useful proxy to understand our environment and behaviors. For policymakers, the interest is twofold. Firstly, eliciting stereotypes might help design more adequate public policies. Secondly, since some stereotypes are socially undesirable, knowing their extent might help policymakers to choose, which stereotype to focus on. In our experiment, we introduce a novel design to elicit stereotypes. Our elicitation mechanism

³⁵As shown by our analysis, this difference is significantly different from zero, which is not the case for the gender stereotype. These percentages are the results of the differences between the subgroups divided by the maximum contribution amount (20 in this public goods game).

allows us to infer the accuracy of the stereotype, qualitatively and quantitatively, but also to what extent do participants share this stereotype.

In this paper, we elicited different cooperation stereotypes and find that gender is on average not associated with a higher or lower level of cooperation, however, we find a perception gap between the political orientation and the cooperation rate. We find that participants overestimate the contribution of left-leaning individuals compared to the right-leaning ones. On top of that, we find that gender and political stereotype in cooperation are on average shared, as suggested by the distribution of the stereotypes.

In practice, individuals might choose their relationships based on these stereotypes. The selection of the other members of your group is a relevant factor in the public goods game since individual outcomes will depend highly on other's contributions. Therefore, if someone wants to maximize her profit from cooperation, she will try to target the best cooperators. The observable attributes of a group can have implications, such as leading to statistical discrimination (Arrow 1973; Phelps 1972).

References

- Aguiar, F., Brañas-Garza, P., Cobo-Reyes, R., Jimenez, N., & M. Miller, L. (2009). Are women expected to be more generous? *Experimental Economics*, 12(1), 93–98.
- Akerlof, G. A., & Kranton, R. E. (2000). Economics and identity. *The Quarterly Journal of Economics*, 115(3), 715–753.
- Anderson, L. R., Mellor, J. M., & Milyo, J. (2005). Do liberals play nice? The effects of party and political ideology in public goods and trust games. *Advances in applied microeconomics*. Emerald Group Publishing Limited.
- Arendt, F. (2013). Dose-dependent media priming effects of stereotypic newspaper articles on implicit and explicit stereotypes. *Journal of Communication*, 63(5), 830–851.
- Arrow, K. J. (1973). The theory of discrimination. In O. Ashenfelter & A. Rees (Eds.), *Discrimination in labor markets* (pp. 3–33). Princeton University Press.
- Ball, S., Eckel, C. C., & Heracleous, M. (2010). Risk aversion and physical prowess: Prediction, choice and bias. *Journal of Risk and Uncertainty*, 41(3), 167–193.
- Balliet, D., Li, N. P., Macfarlan, S. J., & Van Vugt, M. (2011). Sex differences in cooperation: A meta-analytic review of social dilemmas. *Psychological Bulletin*, 137(6), 881–909.

- Bénabou, R., & Tirole, J. (2003). Self-knowledge and self-regulation: An economic approach. *The Psychology of Economic Decisions*, 1, 137–167.
- Bordalo, P., Coffman, K., Gennaioli, N., & Shleifer, A. (2016). Stereotypes. *The Quarterly Journal of Economics*, 131(4), 1753–1794.
- Boschini, A., Muren, A., & Persson, M. (2012). Constructing gender differences in the economics lab. *Journal of Economic Behavior & Organization*, 84(3), 741–752.
- Brañas-Garza, P., Capraro, V., & Rascón-Ramírez, E. (2018). Gender differences in altruism on mechanical turk: Expectations and actual behaviour. *Economics Letters*, 170, 19–23.
- Castillo, M., & Petrie, R. (2010). Discrimination in the lab: Does information trump appearance? *Games and Economic Behavior*, 68(1), 50–59.
- Cohn, A., Maréchal, M. A., & Noll, T. (2015). Bad boys: How criminal identity salience affects rule violation. *The Review of Economic Studies*, 82(4), 1289–1308.
- Correll, J., Judd, C. M., Park, B., & Wittenbrink, B. (2010). Theories of fairness and reciprocity: Evidence and economic applications. In M. Dewatripont, L. P. Hansen, & S. J. Turnovsky (Eds.), *Measuring prejudice, stereotypes and discrimination* (pp. 45–62). Sage Thousand Oaks, CA.
- Costa Jr, P. T., Terracciano, A., & McCrae, R. R. (2001). Gender differences in personality traits across cultures: Robust and surprising findings. *Journal of Personality and Social Psychology*, 81(2), 322.
- Daruvala, D. (2007). Gender, risk and stereotypes. *Journal of Risk and Uncertainty*, 35(3), 265–283.
- Dave, C., Eckel, C. C., Johnson, C. A., & Rojas, C. (2010). Eliciting risk preferences: When is simple better? *Journal of Risk and Uncertainty*, 41(3), 219–243.
- Dawkins, R. (1976). *The selfish gene* (Vol. 214). Oxford University Press.
- Dieckmann, A., Grimm, V., Unfried, M., Utikal, V., & Valmasoni, L. (2016). On trust in honesty and volunteering among Europeans: Cross-country evidence on perceptions and behavior. *European Economic Review*, 90, 225–253.
- Eagly, A. H. (2009). The his and hers of prosocial behavior: An examination of the social psychology of gender. *American Psychologist*, 64(8), 644–658.
- Eagly, A. H., & Wood, W. (1999). The origins of sex differences in human behavior - Evolved dispositions versus social roles. *American Psychologist*, 54(6), 408–423.

- Eckel, C. C., & Grossman, P. J. (2002). Sex differences and statistical stereotyping in attitudes toward financial risk. *Evolution and Human Behavior*, *23*(4), 281–295.
- Fehr, E., & Fischbacher, U. (2005). Altruists with green beards. *Analyse & Kritik*, *27*(1), 73–84.
- Fehr, E., Fischbacher, U., Von Rosenblatt, B., Schupp, J., & Wagner, G. G. (2002). A nation-wide laboratory: Examining trust and trustworthiness by integrating behavioral experiments into representative survey. *Schmollers Jahrbuch*, *4*(122), 519–542.
- Fiedler, K., & Bluemke, M. (2005). Faking the IAT: Aided and unaided response control on the implicit association tests. *Basic and Applied Social Psychology*, *27*(4), 307–316.
- Fischbacher, U. (2007). Z-Tree: Zurich toolbox for ready-made economic experiments. *Experimental Economics*, *10*(2), 171–178.
- Fischbacher, U., Gächter, S., & Fehr, E. (2001). Are people conditionally cooperative? Evidence from a public goods experiment. *Economics Letters*, *71*(3), 397–404.
- Fiss, O. M. (1976). Groups and the equal protection clause. *Philosophy & Public Affairs*, 107–177.
- Frank, R. H. (2005). Altruists with green beards: Still kicking? *Analyse & Kritik*, *27*(1), 85–96.
- Gilmour, J. (2015). Formation of stereotypes. *Behavioural Sciences Undergraduate Journal*, *2*(1), 67–73.
- Graziano, W. G., & Eisenberg, N. (1997). Agreeableness: A dimension of personality. *Handbook of personality psychology* (pp. 795–824). Elsevier.
- Greiner, B. (2015). Subject pool recruitment procedures: Organizing experiments with orsee. *Journal of the Economic Science Association*, *1*(1), 114–125.
- Grossman, P. J. (2013). Holding fast: The persistence and dominance of gender stereotypes. *Economic Inquiry*, *51*(1), 747–763.
- Grossman, P. J., & Lugovskyy, O. (2011). An experimental test of the persistence of gender-based stereotypes. *Economic Inquiry*, *49*(2), 598–611.
- Guimond, S., Chatard, A., Martinot, D., Crisp, R. J., & Redersdorff, S. (2006). Social comparison, self-stereotyping, and gender differences in self-construals. *Journal of Personality and Social Psychology*, *90*(2), 221–242.
- Heckman, J. J., Jagelka, T., & Kautz, T. D. (2019). *Some contributions of economics to the study of personality* (tech. rep.). National Bureau of Economic Research.

- Herrmann, B., Thöni, C., & Gächter, S. (2008). Antisocial punishment across societies. *Science*, *319*(5868), 1362–1367.
- Jagelka, T. (2020). *Are economists' preferences psychologists' personality traits? A structural approach*.
- Katz, D., & Braly, K. (1933). Racial stereotypes of one hundred college students. *The Journal of Abnormal and Social Psychology*, *28*(3), 280–290.
- Kistler, D., Su, N., & Thöni, C. (2016). Salience in public goods games. *Working paper*.
- Krieglmeyer, R., & Sherman, J. W. (2012). Disentangling stereotype activation and stereotype application in the stereotype misperception task. *Journal of Personality and Social Psychology*, *103*(2), 205–224.
- Le Pelley, M. E., Reimers, S. J., Calvini, G., Spears, R., Beesley, T., & Murphy, R. A. (2010). Stereotype formation: Biased by association. *Journal of Experimental Psychology: General*, *139*(1), 138–161.
- Ledyard, J. O. (1995). Public goods: A survey of experimental research. In J. H. Kagel & A. E. Roth (Eds.), *The handbook of experimental economics* (pp. 111–194). Princeton University Press.
- Lee, Y.-T., McCauley, C., & Jussim, L. (2013). Stereotypes as valid categories of knowledge and human perceptions of group differences. *Social and Personality Psychology Compass*, *7*(7), 470–486.
- Lippert-Rasmussen, K. (2006). The badness of discrimination. *Ethical Theory and Moral Practice*, *9*(2), 167–185.
- Madon, S., Guyll, M., Hilbert, S. J., Kyriakatos, E., & Vogel, D. L. (2006). Stereotyping the stereotypic: When individuals match social stereotypes. *Journal of Applied Social Psychology*, *36*(1), 178–205.
- McCrae, R. R., & Terracciano, A. (2005). Universal features of personality traits from the observer's perspective: Data from 50 cultures. *Journal of Personality and Social Psychology*, *88*(3), 547–561.
- Phelps, E. S. (1972). The statistical theory of racism and sexism. *The American Economic Review*, *62*(4), 659–661.
- Schuman, H. (1997). *Racial attitudes in America: Trends and interpretations*. Harvard University Press.
- Silverman, B. W. (1986). *Density estimation for statistics and data analysis* (Vol. 26). Chapman; Hall.
- Steffens, M. C. (2004). Is the implicit association test immune to faking? *Experimental Psychology*, *51*(3), 165–179.
- Sunstein, C. R. [Cass R]. (1994). The anticaste principle. *Michigan Law Review*, *92*(8), 2410–2455.
- Surowiecki, J. (2005). *The wisdom of crowds*. Anchor.

- Thaler, R. H., & Sunstein, C. R. [Cass R.]. (2008). *Nudge: Improving decisions about health, wealth, and happiness*. Yale University Press.
- Thompson, E. R. (2008). Development and validation of an international english big-five mini-markers. *Personality and Individual Differences*, *45*(6), 542–548.
- Thöni, C., & Volk, S. (2018). Conditional cooperation: Review and refinement. *Economics Letters*, *171*, 37–40.
- Thöni, C., Volk, S., & Cortina, J. M. (2020). Greater male variability in cooperation: Meta-analytic evidence for an evolutionary perspective. *Psychological Science*, *32*(1), 50–63.
- Volk, S., Thöni, C., & Ruigrok, W. (2011). Personality, personal values and cooperation preferences in public goods games: A longitudinal study. *Personality and Individual Differences*, *50*(6), 810–815.
- Volk, S., Thöni, C., & Ruigrok, W. (2012). Temporal stability and psychological foundations of cooperation preferences. *Journal of Economic Behavior & Organization*, *81*(2), 664–676.
- Vyrastekova, J., Sent, E.-M., & van Staveren, I. P. (2015). Gender beliefs and cooperation in a public goods game. *Economic Bulletin*, A117–1153.
- Williams, J. E., Satterwhite, R. C., & Best, D. L. (1999). Pancultural gender stereotypes revisited: The five factor model. *Sex Roles*, *40*(7-8), 513–525.
- Wilson, J. Q., DiIulio Jr, J. J., & Bose, M. (2013). *American government: Brief version*. Cengage Learning.
- Wood, W., & Eagly, A. H. (2012). Biosocial construction of sex differences and similarities in behavior. *Advances in experimental social psychology* (pp. 55–123). Elsevier.
- Zelmer, J. (2003). Linear public goods experiments: A meta-analysis. *Experimental Economics*, *6*(3), 299–310.
- Zizzo, D. J. (2010). Experimenter demand effects in economic experiments. *Experimental Economics*, *13*(1), 75–98.

1.10 Appendices

Summary statistics - The European data set - RD1

Country	N	% female	Mean age	Average contribution
Austria	68	45.6	24.4	13.5
Denmark	68	27.9	24.4	14.1
France	48	39.6	21.2	9.3
Germany	156	53.8	21.8	12.2
Netherlands	84	48.8	21.8	11.3
Sweden	56	26.8	24.3	14.9
Switzerland	208	36.1	21.1	11.2

Table A1.1: Descriptive statistics

Summary statistics - The Lausanne data set - RD2

Location	Lausanne
N	360
% female	43.1
Mean age	21.0
Average contribution	11.3
% Right-leaning	35.4
Median split % right-leaning	60.4

Table A1.2: Descriptive statistics

Stereotype elicitation: Stage 1 (gender stereotype)

Question 1

Quelles est selon vous la **contribution moyenne** des femmes et des hommes pour la **première période** dans la même expérience que celle expliquée dans les instructions?

Avant de commencer l'expérience, nous allons vous demander de répondre à des questions. Ce sont des **estimations** qui sont demandées. **Aucun calcul** n'est nécessaire! Chacune d'entre elles peut vous faire gagner jusqu'à 100 points additionnels.
Plus vous serez proche de la valeur de nos données, plus vous engrangerez de points. Cette valeur a été collectée en regroupant les données de plusieurs pays d'Europe (informations détaillées en bas de page)

La fonction de gain est la suivante :
Si votre réponse est égale à la valeur réelle +/- 0.2, vous obtenez 10 points
Si votre réponse est égale à la valeur réelle +/- 0.18, vous obtenez 20 points
...
Et ainsi de suite, si votre réponse est égale à la valeur réelle +/- 0.04, vous obtenez 90 points
De fait si votre réponse est égale à la valeur réelle +/- 0.02 vous obtenez 100 points.
Si par contre, votre réponse est éloignée de plus de 0.2 unités, vous ne recevrez aucun montant

Informations détaillées sur nos données

Total : 688 observations, Autriche (68), Danemark (68), France (48), Allemagne (156), Pays-Bas (56), Suède (56), Suisse (208)
Toutes les données proviennent d'expérience similaire à celle-ci avec des **étudiants** habituellement de 20 à 25 ans

[Passer à la première question](#)

[Translation from top to bottom :]

What do you think is the **mean contribution** of men and women in the **first period** in the same experiment as the one explained in the instructions?

Before starting the experiment, we will ask you to answer questions. These are estimations, no calculation is needed. In each one, you can gain up to 100 additional points.

The closer you are to our data value, the more points you get. This value has been collected by grouping the data of several European countries (detailed information are present at the bottom of this page)

The payoff function is as follows:

If your answer equals the true value +/- 0.2, you get 10 points

If your answer equals the true value +/- 0.18, you get 20 points

...

...

And so on, if your answer equals the true value +/- 0.04, you get 90 points

Therefore if your answer equals the true value +/- 0.02, you get 100 points

If however, your answer is further than 0.2 units you will not gain any gain

Detailed information about our data

Total: 688 observations, Austria (68), Denmark (68), France(48), Germany(156),

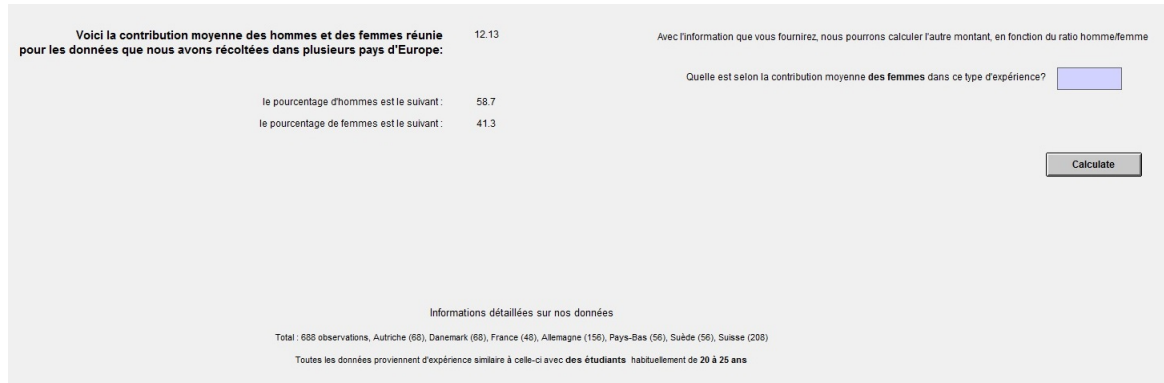
Netherlands(56), Sweden(56), Switzerland (208)

All the data come from similar experiments to this one with students usually between 20 and 25 years.

[go to the first question](#)

Figure A1.1: Screenshot - Study 1 - Stage 1

Stereotype elicitation: Stage 2 (gender stereotype)



Voici la contribution moyenne des hommes et des femmes réunie pour les données que nous avons récoltées dans plusieurs pays d'Europe: 12.13

Avec l'information que vous fournirez, nous pourrions calculer l'autre montant, en fonction du ratio homme/femme

le pourcentage d'hommes est le suivant: 58.7

le pourcentage de femmes est le suivant: 41.3

Quelle est selon la contribution moyenne des femmes dans ce type d'expérience?

Calculate

Informations détaillées sur nos données

Total : 688 observations, Autriche (66), Danemark (68), France (48), Allemagne (156), Pays-Bas (56), Suède (56), Suisse (208)

Toutes les données proviennent d'expérience similaire à celle-ci avec des étudiants habituellement de 20 à 25 ans

[Translation from top to bottom]

Here is the mean contribution of women and men together from the data we collected in several European countries:

The percent of men is the following:

The percent of women is the following:

The right side : with the information you will provide, we will be able to calculate the other number according to the ration male/female

What do you think is the mean contribution of **women** in this kind of experiment?

the bottom: same as the page before with the detailed information about our data.

calculate

Figure A1.2: Screenshot - Study 1 - Stage 2

Third step: confirmation to elicit the gender stereotype

Voici la contribution pour les données que

Vos estimations sont les suivantes :
13.00 pour les femmes et
11.52 pour les hommes.

Ces valeurs vous semblent-elles correctes?

CHANGER **OK**

culer l'autre montant, en fonction du ratio homme/femme

mes dans ce type d'expérience? 13

te contribution pour les hommes: 11.52

Calculate

Informations détaillées sur nos données

Total : 688 observations, Autriche (68), Danemark (68), France (48), Allemagne (156), Pays-Bas (56), Suède (56), Suisse (208)

Toutes les données proviennent d'expérience similaire à celle-ci avec des étudiants habituellement de 20 à 25 ans

Translation from top to bottom (only the dark grey part): Your estimations are as follows:

13.00 for the women and

11.52 for the men. Do these values seem correct for you?

Red button: Change

Green button: OK

Figure A1.3: Screenshot - Study 1 - Step 3

First step to elicit the political stereotype

Question 2

Quelles est selon vous la **contribution moyenne** des individus plus à gauche politiquement et des individus plus à droite politiquement pour la **première période** dans la même expérience que celle expliquée dans les instructions?

Le étudiants ayant passé l'expérience devait répondre à la question que vous voyez à la suite. De fait si la personne était entre 1 et 5, nous la considérons comme plus à gauche, et si elle était entre 6 et 10 comme étant plus à droite.

La question exposée aux étudiants :

A propos de politique, les gens parlent de gauche et de droite. Vous-même, où vous situez-vous sur cette échelle d'une façon générale?

gauche									droite
1	2	3	4	5	6	7	8	9	10
<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Comme auparavant, cette question peut vous faire gagner jusqu'à 100 points additionnels.

Plus vous serez proche de la valeur de nos données, plus vous engrangerez de points. Cette valeur a été collectée en regroupant les données de plusieurs expérience à l'université de Lausanne (informations détaillées en bas de page)

La fonction de gain est la suivante:
Si votre réponse est égale à la valeur réelle +/- 0.2, vous obtenez 10 points
Si votre réponse est égale à la valeur réelle +/- 0.18, vous obtenez 20 points
...

Et ainsi de suite, si votre réponse est égale à la valeur réelle +/- 0.04, vous obtenez 90 points
De fait si votre réponse est égale à la valeur réelle +/- 0.02 vous obtenez 100 points
Si par contre, votre réponse est éloignée de plus de 0.2 unités, vous ne recevrez aucun montant.

Informations détaillées sur nos données

Total : 336 observations à l'université de Lausanne (étudiants de l'Unil et de l'EPFL) Toutes les données proviennent d'expérience similaire à celle-ci avec des étudiants habituellement de 20 à 25 ans dans ce LABEX

[Passer à la deuxième question](#)

Translation from top to bottom: what do you think is the mean contribution of individuals leaning more towards the left and individuals leaning more towards the right for the first period in the same experiment as the one explained in the instructions.

The students had to answer the following question that you can see below. Therefore if the person was between 1 and 5 we considered her left-leaning and if she answered between 6 and 10 as right-leaning.

Like before you can gain up to 100 additional points for this question.

The closer you are to our data value, the more points you get. This value has been collected by grouping the data of several experiments at the University of Lausanne (detailed information are present at the bottom of this page)

The gain function is the following:

If your answer equals the true value ± 0.2 , you get 10 points

If your answer equals the true value ± 0.18 , you get 20 points

...

...

And so on, if your answer equals the true value ± 0.04 , you get 90 points

Therefore if your answer equals the true value ± 0.02 , you get 100 points

If however, your answer is further than 0.2 units you will not gain any gain

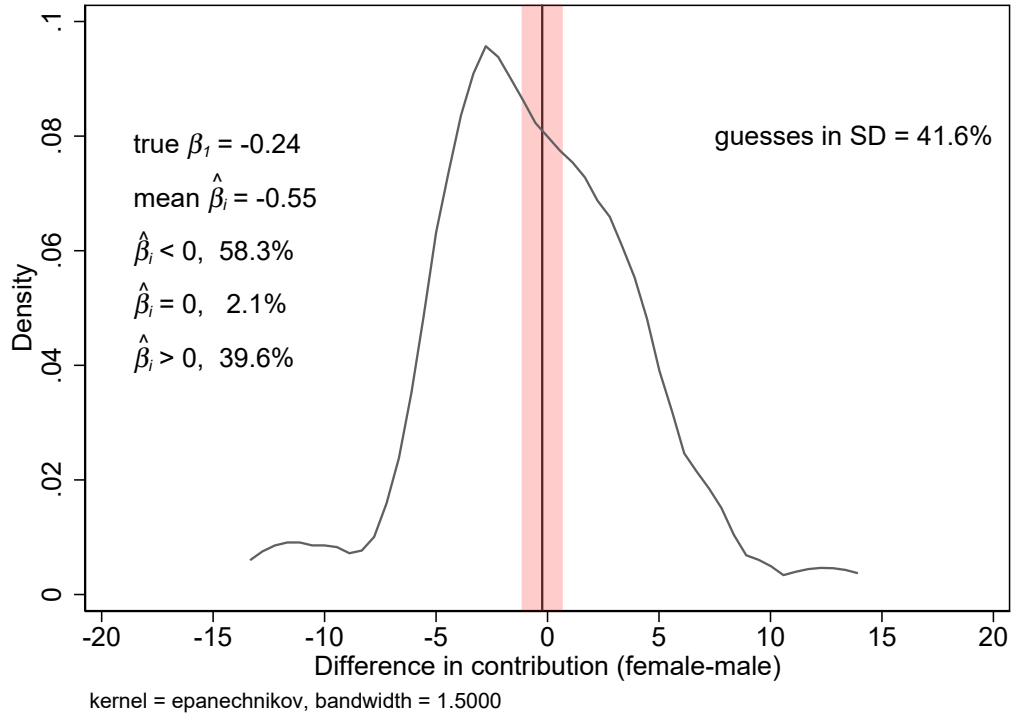
Detailed information about our data

Total: 336 observations from the University of Lausanne (students from the UNIL and EPFL). All the data comes from similar experiments to this one with students usually between 20 and 25 years in this LABEX.

The button: go to the second question

Figure A1.4: Screenshot - Study 2 - Step 1

Distribution of gender stereotypes in cooperation - Study 2



Gender stereotype elicitation: guesses about the difference in contribution between female and male subjects, the true difference (female contribution minus male one) from the reference data (vertical line). Standard deviation from the reference data (vertical lighter line). True β_1 is the value from the European data set (RD1). Mean $\hat{\beta}_i$ and the percentages, above, equal, and below 0, are statistics from Study 2. On the right side, the percentage of guesses within the standard deviation of the reference data set.

Figure A1.5: Kernel Density - $\hat{\beta}_i$

1.11 Supplementary material

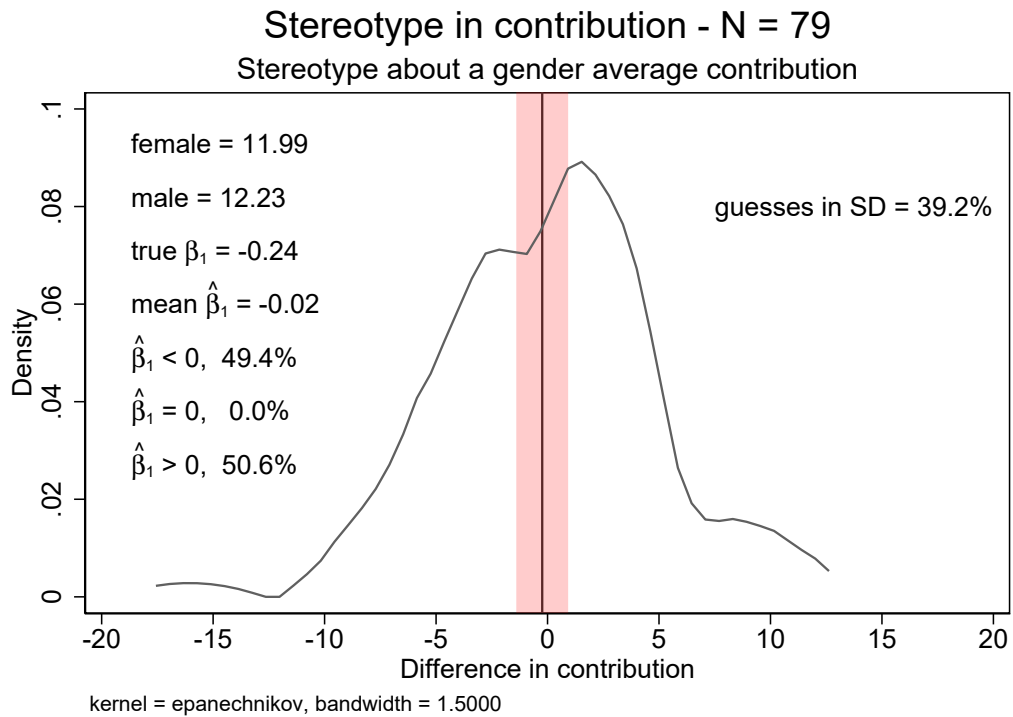
The supplementary material includes additional analyses on data from non-controlled environments.

We ran part of the experiment in classrooms once with Master and Ph.D. students and once with Bachelor students. Therefore, the data could not be used in the main paper but we still use it as a robustness check for our stereotype elicitation mechanism. Both Figures S1.1 and S1.2 report data taken in classrooms. We did not control for the understanding of the public goods game but still explained the game and provided an incentive for the answer. We paid 10 CHF (approximately 10 US dollars) to the person who

provided the closest guess to the true average from our European reference data set (RD1).

We find similar results as in Study 1. In both groups, the guesses are not statistically different from 0 ($t(48) = -0.48$, $p = 0.631$ for Master and PhD students, $t(78) = -0.03$, $p = 0.975$ for Bachelor students) nor to the true value from RD1 ($t(48) = -0.12$, $p = 0.904$ for Master and PhD students, $t(78) = -0.41$, $p = 0.679$ for Bachelor students). As we can observe, there are some individual gender stereotypes. For instance, a difference in gender contribution of more than 10 ECU denotes a belief that gender is rather a good predictor of contribution. Nonetheless, we still end up, with an average gender stereotype close to the true mean from our reference data, perhaps, a brief illustration of the wisdom of the crowds (Surowiecki 2005).

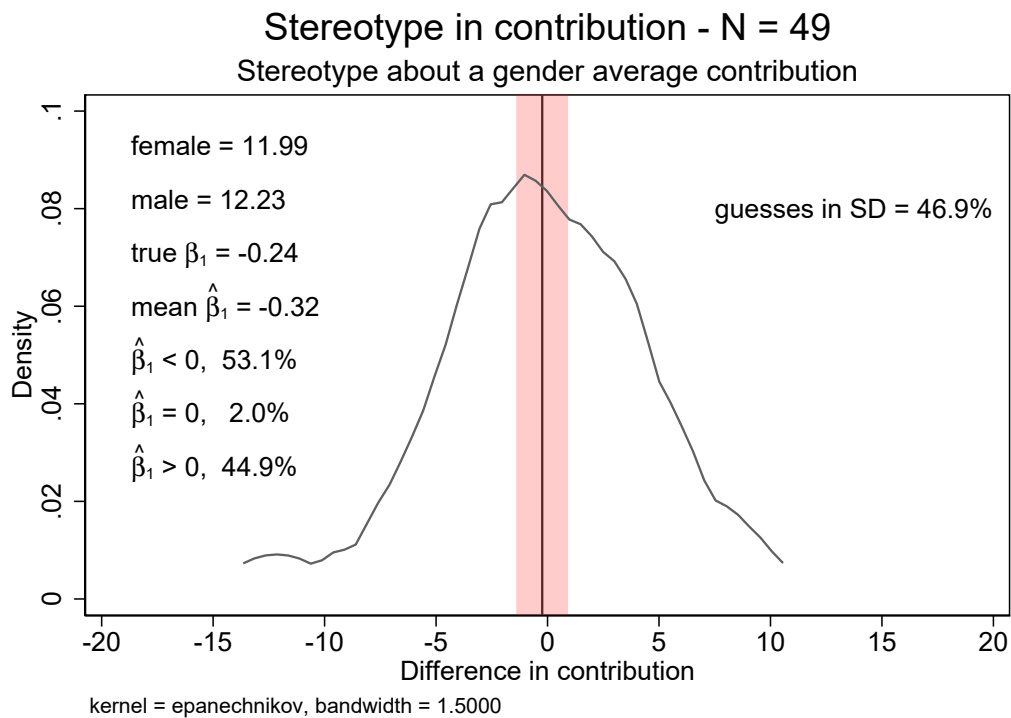
Kernel distribution of $\hat{\beta}_1$ - Bachelor students, 1st year)



Gender stereotype elicitation: guesses about the difference in contribution between female and male subjects, the true difference (female contribution minus male one) from the reference data (vertical line). Standard deviation from the reference data (vertical lighter line). Female, male, and true β_1 are the values from RD1. Mean $\hat{\beta}_1$ and the percentages, above, equal, and below 0, are statistics from Bachelor students in Lausanne. On the right side, the percentage of guesses within the standard deviation of the reference data set.

Figure S1.1: Kernel Density

Kernel distribution of $\hat{\beta}_1$ - Master and PhD students



Gender stereotype elicitation: guesses about the difference in contribution between female and male subjects, the true difference (female contribution minus male one) from the reference data (vertical line). Standard deviation from the reference data (vertical lighter line). Female, male, and true β_1 are the values from RD1. Mean $\hat{\beta}_1$ and the percentages, above, equal, and below 0, are statistics from Master and PhD students in Lausanne. On the right side, the percentage of guesses within the standard deviation of the reference data set.

Figure S1.2: Kernel density

2 Chapter 2

Choice architecture in charitable giving: An experiment on nudging

Jason Wettstein

Abstract

I study the effect of changes in the choice architecture in charitable giving. In an online experiment with 1,338 participants, respondents decide if they want to donate a part of their experimental earnings to a charity. Subjects are presented with a list of charitable organizations to choose from. I investigate the effect of providing a short or long list of charities to participants, either directly visible or with a drop-down button, and compare donations to a control group where no list was provided. I find that providing a list increases the proportion of donors, but in contrast to previous research, attracts only small donations. The list works as a nudge to increase donations at the extensive margin, but crowds out the intrinsic motivation at the intensive margin, resulting in no systematic change on the realized level of donation. The comparison between the different list treatments enables me to investigate the underlying process in the donation decision and I find that the shift in the propensity to donate is channeled through emotions rather than by cognitive costs.

2.1 Introduction

Charitable giving is a widespread phenomenon. In the USA, approximately 69% of the population donate or have donated to a charity at some point in time.³⁶ While this number reflects the popularity of this behavior, the question of why people donate remains a subject of an academic debate. Particularly puzzling are the underlying mechanisms that govern the decision to donate.

This study investigates the influence of the choice architecture on the donation decision. Choice architecture is a term introduced by Thaler and Sunstein (2008) and refers to the different possible presentations of choices to individuals and their influences on decision-making. First, I investigate whether the provision of a list of charities increases donations. Second, if the list increases donations, is the increase due to a lower cognitive cost or by emotional arousal. I investigate this change in the choice architecture, because previous results suggest that this change, while being close to a free lunch for charities, may have substantial effects on donations. The previous literature (Schulz et al. 2018) is, however, unable of identifying the causes of the treatment effect.

In the experiment, participants decide whether they are willing to donate part of their potential earnings from a set of previous tasks to a charity. In the donation decision, they either have no charity proposed, a short or a long list of charities, or a short or long list available using a drop-down button. In this study, I investigate the effect of the treatment variations on three dependent variables: (i) the propensity to donate (the extensive margin); (ii) the amount donated (the intensive margin); (iii) and the realized level of donation (the propensity times the amount).

According to the choice overload literature, a longer list, of 20 charities, in contrast to a shorter list, of 4 charities, should lead to less donations. According to the strategic ignorance literature, the drop-down treatment should mitigate the donation decision, as participants who do not want to donate do not press the drop-down, such that they avoid the emotional arousal.

I find that the number of donors increases with the provision of a list. However, the long list does not decrease the number of donors compared to the short list of charities. Furthermore, the drop-down is only marginally significant in mitigating the donation decision. Surprisingly the increase in the number of donors with the provision of the list does not increase the re-

³⁶See <https://nonprofitssource.com/online-giving-statistics/>

alized level of donation. The explanation lies in the donation amount, which decreases with the provision of the list. The results suggest that emotional arousal plays a bigger role than the cognitive cost in explaining the treatment differences on the propensity to donate. I also discuss the possibility that the provision of the list may crowd out the intrinsic motivation to donate by rendering the donation cognitively less demanding.

This research complements a growing literature with a focus on increasing donations from individuals (A. Gneezy et al. 2010; Karlan and List 2007; Meier 2007; Morgan and Sefton 2000). It also complements research linked to charitable giving in online settings, either in connection with purchasing behavior (McManus and Bennet 2011) or when effort provision was associated with donation (Tonin and Vlassopoulos 2015). However, this study contrasts a study by Schulz et al. (2018), who document a large increase in the number of donors with the provision of a list of five charities. Unlike their study, the list negatively affects the donation amount conditional on donating in the present study. In short, while they show a large increase in the realized level of donation, this study does not.

Finally, I bring evidence in favor of emotions as a trigger for the decision to donate, but at the same time shed light on possible contextual differences when changing the “choice architecture”. I find that facilitating the donation decision, at least in the online context, likely crowds out the intrinsic motivation to donate.

2.2 Experimental Design

The data for this study comes from a large online study, which consisted of several different experimental tasks.³⁷ The design of the main experimental tasks are reported in Kistler et al. (2017). They ran the experiment at three distinct points in time (henceforth: waves) in August 2013, October 2017 and December 2017. They invited 4000 students to participate in the online experiment from two different locations: the University of Hamburg and the University of Magdeburg. A total of 1,338 students completed the online experiment; 842 students from the University of Magdeburg and 496 from the University of Hamburg. 52% of the participants were females and the mean age was 24 years ($SD = 4.00$).

In this experiment, all the previous tasks mentioned above were incen-

³⁷The participants played, in order, a public goods game, a property rights game, an honesty elicitation game, a risk elicitation game, and finally the charity stage.

tivized. For each task either a randomly picked group of participants or a subject, depending on the task, received their/her earnings. Participants knew that they could potentially earn a considerable amount, i.e., on average 400€ among those who received money. At the end of the online study, participants passed through the charitable giving stage, which is split into three distinct steps. This design intends to mimic reality. For instance, in many online donation platforms or when making bank transfers, the decision is split in different steps. In the first step, participants had to indicate whether they would be willing to donate parts of their potential earnings to a charity. The first decision is binary (yes/no). In the second step, they had to indicate the amount, conditional on having said “yes” in the previous step. The two first steps are the same across treatments. In the third step the experimenters implemented the following treatments:

- Control group
- Short list (List Short)
- Long list (List Long)
- Drop-down short list (DD Short, only in Wave 1 and 3)
- Drop-down long list (DD Long, only in Wave 2 and 3)

In this third step, only the presentation of the possible charities, to which participants could donate, varies. In all treatments, no choice was enforced. Participants could always specify a charity that was not listed. In the drop-down treatments, the blank field to specify a charity of their choice was only visible after clicking on the drop-down button. Moreover, participants did not know whether they would have the short or the long of charities before clicking on the drop-down button.

In the short list treatments, the participants could choose between four major charities: WWF, Amnesty International, The Red Cross, Doctors without borders (treatments List Short and DD Short). In the long list treatments, they could choose among 20 charities. In addition to the charities from the short list, the list contained a number of minor charities: UNICEF, DKMS, Deutscher Tierschutzbund, Kinderhospiz, SOS Kinderdorf, Aktion Kleiner Prinz, Terre des hommes, World Vision, Save the Children, Plan international, Welthungerhilfe, Kindernothilfe, adventiat, DGzRS, Transparency International, Weisser Ring (treatments Long list and DD Long).³⁸ The difference between the already displayed lists (List Short and List Long) and

³⁸Some of the charities mentioned here operate only in Germany and therefore have only a German name. The distinction between major and minor charities is artificial and accounts for the amount of donation a charity receives each year.

the drop-down treatments (DD Short and DD Long) was that participants had to click to see the list. The participants saw the entire list, without scrolling down in all the list treatments.³⁹

After the donation decision, the experimenters implemented three extra stages: trust in the charities, a cognitive reflection test, and the possibility to opt-out.

At the time of the experiment, it was made sure that none of the charities listed were involved in a scandal. Furthermore, trust questions to control that the charities listed did not have an adverse effect were implemented. On a 4-points Likert scale, participants rated their trust in all the charities present in the short list plus two extra ones from the long list: Terres des Hommes and Deutscher Tierschutzbund.⁴⁰

In addition to the charity question, participants also had to take the cognitive reflection test (CRT, Frederick 2005). The CRT evaluates an individual's ability to revise an intuitive but wrong first answer with a correct reflexive one. Intuition, which is also called System 1, is defined as a quickly executed deliberation. A typical CRT question is: "A baseball bat and a ball cost together 1.10 euros. The baseball bat costs 1 euro more than the ball. How much does the ball cost (in cents)". A subject who answers 10 cents presumably relies more on fast heuristics, than a subject who comes up with the correct answer (5 cents). After answering a set of three CRT questions, subjects are classified as intuitive types if they provide a wrong answer 2 out of 3 times (intuitive types).

At the very end of the experiment, subjects entered a stage where they could reconsider their donation decision. The opt-out stage allowed participants that agreed to donate to a charity to modify the amount they intended to donate. Screenshots of the experiment are shown in the appendices.

2.3 Hypotheses

A strict reading of the standard economic theory predicts no donation in an anonymous environment like in this experiment. Thus in order to explain donation from citizens, I need to assume some sort of social preferences.

A potential explanation for donations comes from the concept of pure altruism. Pure altruists value the utility of other people positively. However,

³⁹The experimenters specifically told participants to not use a smartphone, but, even if they did, the exact same screen, as on computer, appeared.

⁴⁰The experimenters asked the trust questions in Waves 2 and 3 in all treatments.

under the neutrality hypothesis, pure altruists gain no additional utility from transferring money themselves. For pure altruists government transfers to others crowd out donations. In this theory, public funding, through lump-sum taxes, is Pareto optimal and is the best possible choice for pure altruists. According to this theory, donating to a private charity is puzzling.

Due to the lack of explanatory power of the theory of pure altruism, Andreoni (1989; 1990) introduces the concept of the warm-glow of giving. The warm-glow hypothesis posits that utility is generated by the *act* of donating, rather than the effect it has on the receiver's utility. In short, giving brings to the giver a sense of joy and satisfaction. These impure altruists, who may be motivated both by pure altruism and by the warm-glow of giving, can choose to donate to a private charity. This theoretical framework can also explain the donation to inefficient charities. For instance, people might favor an identifiable victim (Jenni and Loewenstein 1997) rather than giving to a charity which would maximize the marginal return (MacAskill 2015). Overall, the warm-glow theory explains donations to private charity as well as donations from impure altruists.

The warm-glow is mostly based on an intrinsic motivation to donate. Since donating does not rely on any return from the receivers, it is not a strategic decision. The utility of an agent comes solely from giving.

While the warm-glow theory focuses on the intrinsic motivation to donate, there are also extrinsic motivations to donate. For instance, DellaVigna, List, and Malmendier (2012) show that peer pressure increases donation and that participants are concerned about their self-image, but care about the cost of signaling (Tonin and Vlassopoulos 2013). Apart from norms and reputation, people may also donate for purely financial reasons such as a tax reduction.

The donation decision is a function of both intrinsic and extrinsic motivation. More precisely, a participant will donate if the utility she gets from giving surpasses the cost, which could be cognitive and monetary. The utility is derived from different sources - intrinsic and extrinsic - such as the 'warm-glow', signaling, peer pressure, and financial incentives.

The hypotheses I will forward all concern the extensive margin as empirical findings from developmental psychology (Blake and Rand 2010; Liu et al. 2016) and from a previous experiment on charitable giving (Schulz et al. 2018) suggest that the process governing the donation decision is different for the decision to donate, "yes" or "no", and for the amount.

The most extreme assumption of the warm-glow theory, where the utility is generated merely by donating, predicts no difference between the treatments. However, any relaxation of this assumption, such as impure altruists with mixed motives - warm-glow and pure altruistic - predicts that the presence of the list of charities increases charitable giving as the agent has a lower

cost - the transaction cost - to give to a charity and can still feel the joy of giving. The lower cost increases the extrinsic motivation and the intrinsic motivation is assumed to be unchanged.

The main hypothesis (H1) is essentially a replication of the surprisingly large treatment effect documented by Schulz et al. (2018), who provide a list of five charities and this list almost doubles the number of donors without changing the mean amount conditional on donating. The list is a nudge⁴¹ that pushes participants to donate.

Hypothesis 1 *Providing participants with a list increases charitable giving. The list increases the number of donors and does not affect the donation amount conditional on donating.*

The design of the experiment manipulates two dimensions, the length of the list and the visibility of the list with a drop-down button. The four different treatments enable me to investigate the underlying mechanism. I challenge two hypotheses giving different predictions regarding the length of the list: cognitive cost (H2) and emotional arousal (H3). In other words, do participants increase donations because it is easier or because the list triggers an emotional response? The cognitive cost hypothesis predicts that the short list will trigger the highest level of donation and that the long list will trigger a choice overload for the participants, thus reducing the level of donations. The emotional arousal hypothesis, on the other hand, predicts an increase in the level of donations with the list visible vs. non-visible and, eventually, a slight increase with a longer list as participants have a higher probability of finding a charity that matches their preferences in the long list than in the short list.

The cognitive cost hypothesis is based on the assumption that the provision of the list renders the donation easier. Put differently, it is less demanding, cognitively, to choose a charity than to think about one. However, when participants have too many choices they need to engage in a reflection to decide which charity to pick. The necessity to reflect discourages some participants from donating. The short list reduces the opportunity costs, which increases the extrinsic motivation to donate. This hypothesis is supported by empirical findings in consumer behavior. Gourville and Soman (2005) and Iyengar and Lepper (2000) argue that having too many choices is cognitively more demanding than having only a few. For instance, Iyengar and Lepper (2000) find a strong decrease in the consumption of ice cream,

⁴¹see Thaler and Sunstein (2008)

from 30% to 3%, when shifting from 6 flavors to 24. This burden of choice reduces consumption and satisfaction. Chernev (2003) finds that a large set of choices induces even less consumption when the subjects have no strict preferences before choosing. Furthermore, Schwartz (2004) points out that dissatisfaction is prevalent when too many choices are available, which, in charitable giving, could be interpreted as a decrease in the impression of effectiveness. Intuitively, the burden of choice might act through two channels: the impression of effectiveness and the difficulty to choose. One might have the impression, when faced with too many charities, that picking one involves not picking many others, but also, being rather ineffective, because there are too many needs.

Hypothesis 2 *Providing a short list will bring the highest increase in the number of donations (extensive margin) and the long list leads to a lower increase than the short list (List Short vs List Long and DD Short vs DD Long). The relation between the number of charities (0, short, long) and the density of donations has an inverted U-shape.*

The emotional arousal hypothesis predicts that seeing the list triggers donation, but does not predict, to the best of my knowledge, a decrease in donations between the short and the long list. If anything, the longer the list, the higher the probability of matching someone's personal preferences. Broadly speaking, the list is a reminder of the good you can do. Subjects may associate the names of charities with concrete needs. The list is one step down from abstract to concrete. Avoiding thinking about real needs might be easier without a list than with a list, as it would probably even be harder with pictures. The giver might have more warm-glow when he sees the list, perhaps because the victims are easier to identify (Jenni and Loewenstein 1997). The list is assumed to increase the intrinsic motivation to donate. In line with Kogut and Ritov (2005a), Kogut and Ritov (2005b), and Small et al. (2007), I expect that when individuals experience an emotional response they are more willing to donate or to be altruistic. The list might also nudge participants to donate as they may want to reduce the cognitive dissonance between how altruistic they perceive themselves to be and their actions.

In line with this hypothesis, the drop-down button is a means to avoid the emotion triggered. If participants want to preclude themselves from the thought of concrete needs, they can simply not click on the drop-down button. Andreoni et al. (2017), Grossman (2014), and Grossman and van der Weele (2017) show that participants who expect to be asked to make decision, can strategically avoid the information or even avoid taking the decision. In the drop-down treatments, I expect participants who do not want to donate to

strategically avoid seeing the list. Therefore, these drop-down treatments should mitigate the treatment effect of the list.

Hypothesis 3 *Providing a list will increase donations. The length of the list does not decrease the number of donors. Furthermore, the drop-down treatment mitigates the treatment effect (List Long vs DD Long and List Short vs DD Short)*

Some of the features of the design, such as the CRT, the choices of the charities, and the opt-out option, call for additional predictions.

In line with the choice overload literature, the inverted U-shaped pattern should be more pronounced with participants that rely on fast heuristics. Relying on fast heuristics is a way to preclude the use of the frontal cortex, which is a strong energy consumer. Intuition is less demanding as it is faster and because the provision of a list demands less effort to the participants, those who tend to succumb to fast heuristics should respond more strongly to the treatment variation.

Participants could also specify a charity of their choice, but the provision of the list should increase the crowding-out effect of charities not listed. Following the same logic as above, finding a charity, not in the list, is cognitively demanding and the longer the list the more effort it requires to find a charity name that is not on the list. Charities that do not make it on the list have a very low chance of being chosen. Both the CRT and the crowding out of charities not listed are a replication of the study by Schulz et al. (2018).

Finally, the opportunity to opt-out at the end of the experiment offers the participants the possibility to revise the donation decision. In the opt-out stage, participants do not face the list of charities. Therefore, if participants felt forced to donate in the donation stage, they have the opportunity to opt-out, and without the list, they can preclude themselves from thinking about concrete needs. Since participants have social preferences and altruistic types should be equally prevalent in the no-list condition and the list conditions, the extra donors nudged by the “choice architecture” (Thaler and Sunstein 2008) in the list conditions should change their decision to donate. Therefore, I expect that more participants in the list conditions opt-out compared to those in the no-list condition.

2.4 Results

The charitable giving stage takes place at the end of a larger experiment, which consists of several tasks. I need to check whether the randomization into treatments of the charity stage succeeded. In Table 2.4, I report the

p -values from a multinomial logistic regression. I elicit social preferences by the behaviors of the participants in the previous tasks. I also compute their payoff or their expected payoff as they sometimes did not receive payoff feedback (indicated by E in the table). The category “Other” in the table is self-reported by the participants or coded by the experimenters. Social preferences, payoff, all tasks, and “other” are separate multinomial regressions to avoid possible confounds. As suggested by most of the reported p -values, the randomization succeeded (for further analyses on the randomization check, see section 2.7.1 in the supp. material).

Randomization check

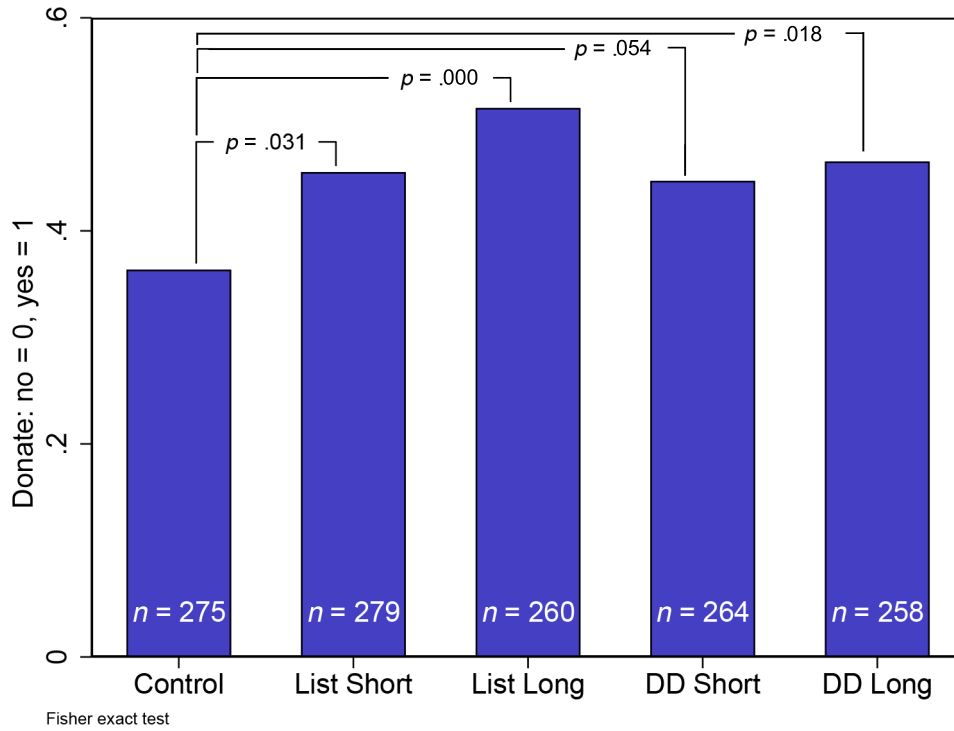
		Short List	Long List	Long List Drop-down	Short List Drop-down
Social Pref.	Risk-taking	0.199	0.053 ⁺	0.176	0.142
	Probability to lie	0.504	0.333	0.000**	0.305
	Contributing	0.828	0.206	0.046*	0.012*
	Property game - Steal	0.849	0.097 ⁺	0.433	0.291
	Property game - Plant	0.198	0.071 ⁺	0.609	0.126
Payoff	PGG E(Payoff)	0.813	0.945	0.143	0.058 ⁺
	Risk game payoff	0.051 ⁺	0.060 ⁺	0.513	0.398
	Honesty game payoff	0.898	0.339	0.479	0.301
	PG E(Payoff)	0.398	0.288	0.832	0.269
All Tasks	Expected Payoff	0.385	0.660	0.202	0.073 ⁺
Other	Location	0.543	0.150	0.160	0.981
	Female	0.855	0.711	0.208	0.683
	Survey Wave	0.955	0.332	n.a.	n.a.

This table reports the p -values given by a robust multinomial regression. The dependent variable is each treatment compared to the control group. The categories: social preferences, payoff, all tasks and other are independent multinomial logistic regressions. N.a. indicates that these cases are ruled out by design, i.e, drop-down treatments are only run in some waves. ⁺ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$

Table 2.1: Multinomial logistic regression on each treatment - base: Control group

I observe that in the baseline treatment 36.4 percent of respondents decided to donate some of their earnings, while in the four list treatments, on average 47.0 percent of respondents donated to a charity. The difference is significant ($p < 0.01$). All the p -values in this section are the results of Wald-tests in OLS models. Figure 2.1 shows the increase in the extensive margin in all the treatments compared to the control group.

The propensity to donate - Extensive margin

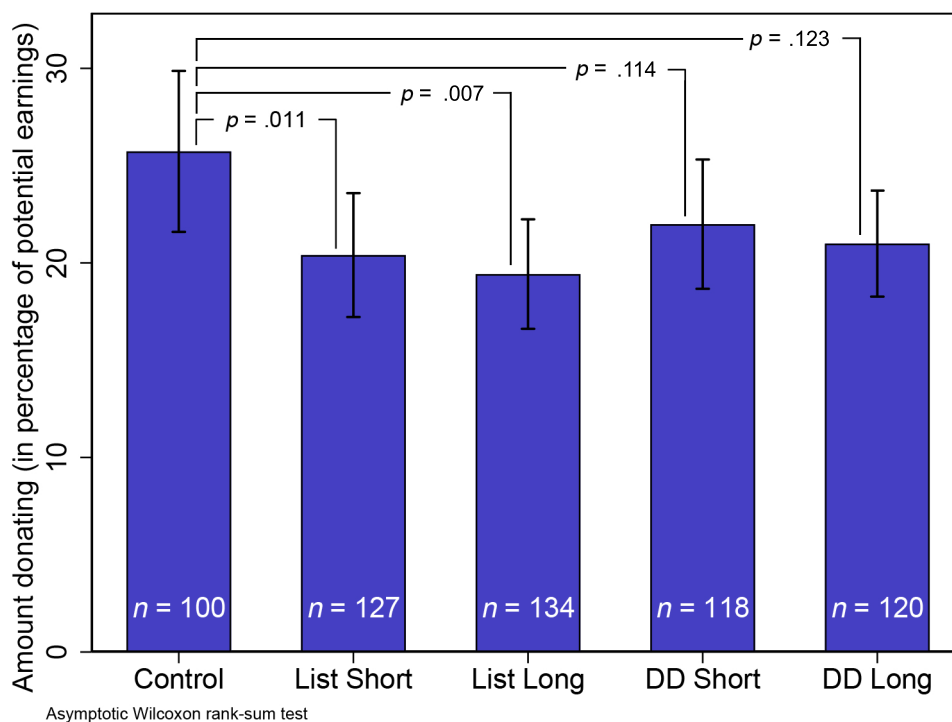


The y -axis is the fraction of participants that decide to donate to a charity. The p -values indicated between each bar are the result of a two-sided Fisher exact test.

Figure 2.1: Bar charts - Binary decision to donate

However, the donation amount conditional on donating shifts, significantly ($p < 0.05$), from 25.88% in the control group to 20.65% in treatments with the list. Figure 2.2 shows the decrease in the amount, the intensive margin, conditional on donating.

The donation amount - Intensive margin



The y -axis represents what percentage of potential earnings in the previous tasks a participant is willing to give conditional that she agreed to donate. The p -values indicated between each bar are the result of a two-sample Wilcoxon rank-sum (Mann-Whitney) test. Each bar displays the 95% confidence interval.

Figure 2.2: Bar charts - Donation amount conditional on donating

For the regression analysis, I pool the data from all waves and both locations.⁴² In Table 2.2 the dependent variable is either the binary decision to donate (Donate), the donation amount conditional on donating (Amount), or the realized level of donation, which is the propensity to donate times

⁴²I use OLS in my regression tables to interpret coefficients. Alternatively, I tested the hurdle model with two tiers proposed by Cragg (1971), plus tested a logistic regression for the binary decision to donate and a truncated model for the amount, all produce equivalent significance levels with less interpretable coefficients.

the amount (Aggregate).⁴³ The main coefficient of interest, *list*⁴⁴, is always positive and significant ($p < 0.05$) for the binary dependent variable *donate*. This means that the number of donors increases. However, at the same time, the main coefficient of interest is always negative and significant ($p < 0.05$) for the donation amount conditional on donating. Note that the test for joint significance of the regression “Amount” is insignificant, suggesting a possible type I error. However, in simpler linear regressions controlling independently for the different treatments, the test for joint significance is significant ($p < 0.01$). This suggests either a confound or an under-powered study. Therefore, the results of the list on the amount donated conditional on donating should be taken with caution. Assuming this is due to a lack of power, the double effect, propensity times the amount, sums as virtually no difference in the realized level of donation (Aggregate). The lack of an overall effect casts some doubt over the strong results reported in Schulz et al. (2018). This means that charities should not take it for granted that the list treatment increases donations. To conclude, I find no evidence in favor of H1, that the list increases donations.

As the number of donors is substantially higher in Wave 3 and as the donation decision takes place at the end of some other tasks, apart from the randomization checks, I run separate analyses per waves and per location and find that results are robust with respect to these extra analyses (see supplementary material, Tables S2.1, S2.2, S2.3, and S2.4).

I check whether the trust in the charitable organizations explains donations and find that trust in the charitable organization was positively correlated with the decision to donate and the donation amount.⁴⁵ Moreover, on average the trust in the charities was relatively high with means ranging from 2.85 to 3.34 on a 4-points Likert scale. Therefore, I can partially rule out the possibility that a charity in the list produced an adverse effect.

Figure 2.3 shows the distribution of amounts conditional on donating. In-

⁴³The different regressions are heteroskedasticity-robust. Due to the low number of possible clusters, I do not cluster standard errors per waves and locations. According to Cameron and Miller (2015), the minimum number of clusters should be 20 for this option to be reliable. However, I still include wave and locations fixed effects which captures most of the variability across clusters.

⁴⁴This dummy variable identifies all treatments with a list: Short List, Long List, DD Short, and DD Long.

⁴⁵For instance, the coefficients of correlations - R - with the binary decision to donate are from 0.08 to 0.17 for charities in the short list and from 0.08 to 0.25 for the charities in the long list.

Extensive Margin, Intensive Margin & Aggregate

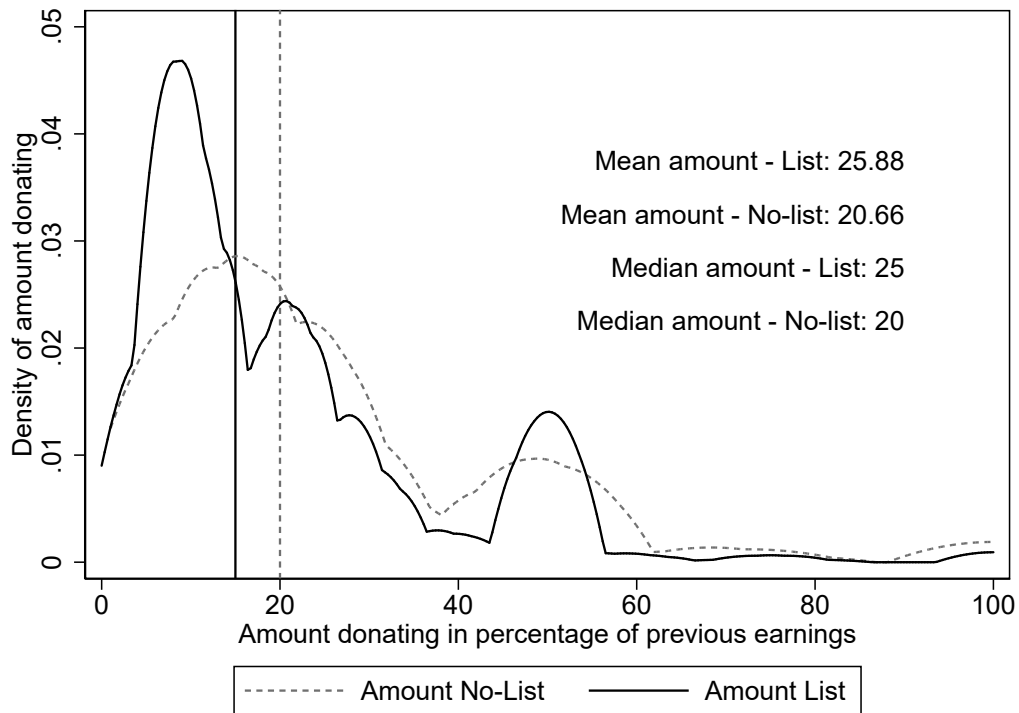
Dependent variable:	Donate	Amount	Donate	Amount	Aggregate
List	0.108** (0.039)	-5.483* (2.508)	0.093* (0.041)	-5.239* (2.614)	-0.011 (1.425)
Long list	0.029 (0.031)	-0.748 (1.570)	0.060 (0.043)	-1.226 (2.136)	0.657 (1.348)
Long list x Drop-down			-0.065 (0.063)	1.034 (3.112)	-0.855 (1.946)
Drop-down	-0.061+ (0.033)	1.538 (1.666)	-0.029 (0.045)	1.011 (2.424)	-0.151 (1.440)
Wave 2	0.056 (0.036)	-2.466 (2.015)	0.064+ (0.037)	-2.602 (2.048)	0.326 (1.171)
Wave 3	0.097** (0.035)	-0.884 (2.010)	0.100** (0.035)	-0.949 (2.037)	1.927+ (1.165)
Location	0.057* (0.029)	-1.639 (1.431)	0.057* (0.029)	-1.653 (1.432)	0.397 (0.894)
Female (dummy)	0.009 (0.027)	-0.347 (1.464)	0.010 (0.027)	-0.357 (1.468)	0.024 (0.880)
Constant	0.293** (0.037)	27.717** (2.836)	0.289** (0.037)	27.797** (2.864)	8.569** (1.443)
<i>F</i> -test	4.1	1.3	3.7	1.2	0.5
Prob > <i>F</i>	0.000	0.255	0.000	0.324	0.839
<i>R</i> ²	0.020	0.019	0.021	0.019	0.003
<i>N</i>	1,336	599	1,336	599	1,336

Notes: OLS estimates. I used a two-tiered model which consists of a binary decision to donate at the first level and a given amount at the second level. All the independent variables are dummy variables. Long list is a dummy indicating the long list feature (in List Long & DD Long treatments), Drop-down stands for the drop-down feature (in DD Short & DD Long treatments), Long-list x Drop-down controls for the interaction of the long list and the drop-down feature. The column aggregate is the decision times the amount. Robust standard errors in parentheses. + $p < 0.1$, * $p < 0.05$, ** $p < 0.01$.

Table 2.2: OLS estimates

terestingly, it shows that the median donor shifts, significantly, her donation from 25% of her potential earnings in the no-list condition to 20% in the list conditions ($p < 0.05$).⁴⁶

The distribution of amount - Intensive margin - List and no-list conditions



The grey dashed line represents the control treatment where no list was provided. The vertical dash grey line is the median amount donating conditional on donating. The black line groups all treatments in which the list was provided. The vertical black line is the median of the list treatments. The bandwidth of these kernel density figures is computed using the Silverman's rule of thumb (Silverman 1986).

Figure 2.3: Kernel density plot - The donation amount conditional on donating

This suggests that the treatment variation affects the propensity to do-

⁴⁶I run a quantile regression with the quantile at 0.5. The dependent variable is the donation amount conditional on donating and the independent variable is the dummy variable, list, which indicates the presence or absence of the list.

nate, but attracts only small donations. This is in line with Altmann et al. (2019) who find similar results in a comparable setting. They find that providing default options for the amount does not increase the realized level of donations. Their results and this study challenge previous empirical evidence that the donation decision is governed by two processes, one for the binary decision and one for the amount donated.

The length of the list, as identified by the “Long List” dummy variable in Table 2.2, produces a significant difference on neither the extensive nor the intensive margin. This result does not support the cognitive cost hypothesis. I thus reject H2.

On the other hand, Figure 2.1 shows the increase in the extensive margin and at the same time no significant difference between the short list and the long list treatments. This is in line with the predictions on the emotional arousal. Moreover, as Table 2.2 shows, the drop-down treatments produce a slight decrease, albeit marginally significant ($p < 0.1$), in the number of donors. I find evidence to confirm H3.⁴⁷

As mentioned in the hypothesis, additional explanatory variables of the design allow me to investigate further hypotheses. I expected participants that rely on fast heuristics to be more sensitive to the treatment variation. I report in Table 2.3 the effects of the treatment variation on deliberative participants, which were participants who reported 2 out of 3 correct answers in the CRT. I include interactions with the deliberative dummy and a dummy variable for the missing CRTs. Missing CRTs could be the sign of very diligent participants, that do not want to give wrong answers or of very lazy participants that do not bother to answer.

As Table 2.3 shows, the results are not in line with the prediction and are even qualitatively the opposite.⁴⁸ In this table, the independent variables are

⁴⁷Unfortunately, the number of times someone clicked on the button was only recorded for the long list drop-down treatment but not for short list drop-down treatment. Indeed, all the participants that donated clicked, as this is a prerequisite. From further analyses, once they clicked, they are slightly more willing to donate and donate slightly more compared to the long list treatment without the drop-down. This would suggest that clicking on the drop-down button nudge participants to donate more. Since there is no significant difference between these two treatments when I do not restrict to participants that clicked on the button, it would suggest that some participants strategically avoid the donation decision by not clicking. Nonetheless, I believe there are other reasons for not clicking which are difficult to disentangle from the strategic avoidance one, such as laziness or rush.

⁴⁸If I include the control variables, the only marginal significant effect vanishes, as the control variables are correlated with the deliberative variable and capture part of the effect. In the dataset, the variable men and deliberative are correlated which shows that a higher

the same as in Table 2.2. Schulz et al. (2018) find that deliberative participants are less sensitive to the provision of the list. However, the interaction with the list tends to show the opposite, although not significantly, to their result.

Cognitive Reflection Test

Dependent variable:	Donate	Amount	Donate	Amount
List	0.105** (0.039)	-5.266* (2.507)	0.094 (0.077)	-8.753+ (5.194)
Long list	0.040 (0.031)	-0.997 (1.529)	0.007 (0.060)	3.256 (2.642)
Drop-down	-0.032 (0.031)	1.512 (1.536)	-0.055 (0.060)	3.530 (2.742)
deliberate	0.050 (0.031)	0.993 (1.608)	0.012 (0.069)	0.411 (5.345)
deliberate x List			0.011 (0.090)	4.555 (5.944)
deliberate x Long list			0.047 (0.070)	-6.094+ (3.196)
deliberate x Drop-down			0.032 (0.070)	-2.396 (3.274)
CRT missing			-0.153 (0.184)	41.216 (28.534)
Constant	0.325** (0.038)	24.955** (2.480)	0.355** (0.061)	25.409** (4.821)
<i>F</i> -test	3.9	1.9	2.2	2.0
Prob > <i>F</i>	0.004	0.112	0.026	0.048
<i>R</i> ²	0.011	0.014	0.013	0.037
<i>N</i>	1,336	599	1,336	599

Notes: OLS estimates. I used a two tiers model that consists of the first tier as the binary decision to donate and the second to the amount donated. I do not include control variables in this regression. As mentioned, I classify a participant as deliberative if she gave two or more correct answers in the CRT. Robust standard errors in parentheses. + $p < 0.1$, * $p < 0.05$, ** $p < 0.01$.

Table 2.3: OLS estimates

proportion of men are deliberate compared to women, but this is not my concern here. My focus is on the deliberative characteristic and not on its distribution among the gender.

In a similar vein, U. Gneezy et al. (2009) find, in a review of the literature, that, while females and males are equally prosocial on average, female subjects are more sensitive to changes in the context. I test gender effects and interactions with the treatments in Tables 2.4 and 2.5. I find that both genders are equally prosocial. However, I find no interaction on the extensive margin between gender and the treatments variations and only a marginally significant ($p < 0.1$) interaction effect on the intensive margin. In particular, I find that female subjects react more strongly to the list and the long list treatments than male subjects. The effect is robust only for the long list when I control for deliberative participants. These results should be considered with caution since the tests for joint significance are marginally significant or insignificant.⁴⁹ Along the same line, DellaVigna, List, Malmendier, and Rao (2013) also find no particular gender effect on donation decisions but find that women donate less when they can strategically avoid being asked. However, the point estimates of the interaction between the drop-down and the female dummy have mixed signs (Female x Drop-down), suggesting that female subjects do not strategically avoid the ask to donate more than male subjects.

Along the same lines, I test whether social preferences predict donation decisions and whether there are heterogeneous responses to the treatment variations. I inferred participants' social preferences based on their behaviors in the experimental tasks they played before the charitable giving stage. In particular, I inferred participants' propensity to lie, to steal, to protect, their cooperative behavior, and their risk preference. I find that, indeed, social preferences matter for the donation decision. I find that the number of donors and the amount they donate is slightly higher among those who contributed more in the public goods game before the charitable giving stage. Furthermore, those who have a higher tendency to lie and steal, donate less often, and when they do, they donate a lower amount, albeit the effect is only either marginally significant or insignificant. However, I essentially find no heterogeneous responses to the treatment variations when testing interactions between treatments and social preferences. Among the few significant results, I find an interaction between the tendency to take risks and the drop-down treatments ($p < 0.05$) and a marginally significant interaction between con-

⁴⁹As there are slightly more intuitive women than men, I control that the effect was not explained by this characteristic. While, in the long list treatments, the marginal increase in the donation amount is robust to this extra specification, the effect is not robust in the short list treatments.

The propensity to donate - Extensive margin - Investigation of gender interactions

Dependent variable: Donate: yes = 1, no = 0					
List	0.107** (0.040)	0.104* (0.052)	0.108** (0.040)	0.107** (0.040)	0.100 ⁺ (0.057)
Long list	0.029 (0.031)	0.029 (0.031)	0.026 (0.043)	0.029 (0.031)	0.027 (0.044)
Drop-down	-0.052 (0.033)	-0.052 (0.033)	-0.052 (0.033)	-0.045 (0.043)	-0.041 (0.045)
Female (dummy)	0.007 (0.027)	0.002 (0.060)	0.005 (0.035)	0.013 (0.035)	0.002 (0.060)
Female x List		0.007 (0.067)			0.015 (0.079)
Female x Long list			0.006 (0.056)		0.005 (0.061)
Female x Drop-down				-0.015 (0.056)	-0.021 (0.061)
Wave 2	0.054 (0.036)	0.054 (0.036)	0.054 (0.036)	0.054 (0.036)	0.054 (0.036)
Wave 3	0.094** (0.035)	0.094** (0.035)	0.094** (0.035)	0.094** (0.035)	0.094** (0.035)
Location	0.057* (0.029)	0.057* (0.029)	0.057* (0.029)	0.057* (0.029)	0.057* (0.029)
Constant	0.294** (0.038)	0.297** (0.046)	0.295** (0.039)	0.291** (0.040)	0.297** (0.046)
<i>F</i> -test	3.7	3.3	3.3	3.3	2.6
Prob > <i>F</i>	0.001	0.001	0.001	0.001	0.004
<i>R</i> ²	0.019	0.019	0.019	0.019	0.019
<i>N</i>	1,334	1,334	1,334	1,334	1,334

Notes: OLS estimates. The dependent variable is always the binary decision to donate. Apart from those dummy variables for the treatment variation, I include the female dummy and the interactions with the different treatments. Standard errors in parentheses. ⁺ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$.

Table 2.4: OLS - Gender interaction

The donation amount - Intensive margin - Investigation of gender interactions

	Dependent variable: Amount donating (in percentage of potential earnings)				
List	-5.520*	-8.789**	-5.428*	-5.478*	-7.818*
	(2.249)	(2.993)	(2.246)	(2.251)	(3.235)
Long list	-0.794	-0.817	-3.399	-0.829	-2.690
	(1.601)	(1.599)	(2.246)	(1.603)	(2.338)
Drop-down	1.364	1.320	1.243	0.386	1.255
	(1.727)	(1.725)	(1.726)	(2.362)	(2.435)
Female (dummy)	-0.318	-5.703	-2.423	-1.046	-5.723
	(1.475)	(3.577)	(1.948)	(1.901)	(3.580)
Female x List		6.444 ⁺			4.659
		(3.901)			(4.494)
Female x Long list			4.858 ⁺		3.505
			(2.941)		(3.190)
Female x Drop-down				1.803	-0.020
				(2.969)	(3.185)
Wave 2	-2.374	-2.314	-2.214	-2.356	-2.215
	(1.986)	(1.983)	(1.985)	(1.987)	(1.986)
Wave 3	-0.810	-0.661	-0.643	-0.753	-0.582
	(1.903)	(1.902)	(1.902)	(1.906)	(1.906)
Location	-1.572	-1.575	-1.471	-1.594	-1.501
	(1.512)	(1.509)	(1.511)	(1.513)	(1.512)
Constant	27.754**	30.405**	28.668**	28.105**	30.327**
	(2.253)	(2.764)	(2.317)	(2.328)	(2.767)
<i>F</i> -test	1.6	1.7	1.7	1.4	1.5
Prob > <i>F</i>	0.133	0.085	0.085	0.175	0.130
<i>R</i> ²	0.019	0.023	0.023	0.019	0.025
<i>N</i>	598	598	598	598	598

Notes: OLS estimates. The dependent variable is the donation amount conditional on donating. Apart from those dummy variables for the treatment variation, I include the female dummy and the interactions with the different treatments. Standard errors in parentheses. ⁺ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$.

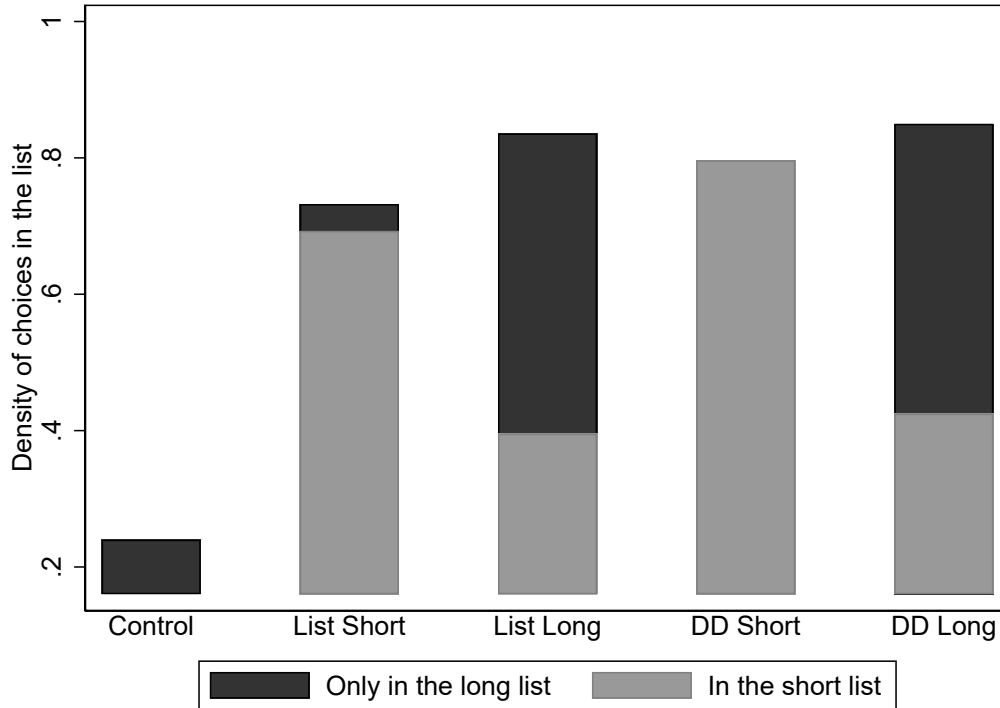
Table 2.5: OLS - Gender interaction

tributing in the public goods game and the list treatments ($p < 0.1$). Both of these interactions are solely significant for the donation amount conditional on donating and the point estimates are negative. In other words, those who tend to take risks donate less when they are in the drop-down treatments and those who contributed more in the public goods game also donate less when they are in a treatment with a list. Risk-takers might donate less in the drop-down treatments because of a lower emotional sensitivity (Nicholson et al. 2005). This explanation is in line with H3. As for those who contributed more in the public goods game, they might be more intrinsically motivated to donate, i.e., more than the lower contributors, and when they face the list, it withdraws more of their intrinsic motivation. I discuss this hypothesis in Section 2.5. I report the methodology for eliciting the behaviors in the supplementary material in Section 2.7.1, the regression outputs testing social preferences in Tables S2.6 and S2.5, and the tests for interactions in Tables S2.7 and S2.8.

The presence of the list increases the crowding out of charities not listed. Figure 2.4 shows the percentage of participants that choose a charity that was in either the short or the long list. I control that this shift could not be explained by the set of choices and this is partially the case, but the differences remain very significant ($p < 0.01$).⁵⁰ The crowding out effect of charities not listed is likely explained by the fact that the charities in the list conditions are the most well-known and it probably becomes difficult to find another charity name.

⁵⁰i.e., subjects, in the short list, entering a charity name that is in the long list, or subjects, in the control group, entering a charity that is in the short or long lists.

Choice of the charity from the short or the long list



The grey bars represent the charities that were in the short list and were chosen by participants. The black bars represent those that were in the long list. As the long list contains the names of the charities in the short list, the grey bars always overlap the black bars. The density of choices is conditional on donating.

Figure 2.4: Bar charts - Choice of a charity

At the end of the experiment, participants had the opportunity to opt-out of their charitable decision. This stage was exempt from anything that would trigger an emotional arousal. Therefore, I expected that those that were nudged by the list would modify their donation decision. I find that 10.3% of the participants who decided to donate revise their amount by reducing the amount they intended to give by 7.46% ($p < 0.05$, paired t-test) on average. Nevertheless, there is no difference across treatments. I do not observe significant differences in the opt-out stage between participants in the no-list condition and the list-conditions. In other words, participants possibly nudged by the list mostly stick with their previous choice. This is likely explained by the will to avoid cognitive dissonance or because it demands less reflection to not change the amount they intended to give.

Overall, the results suggest that cognitive cost is not the dominant mech-

anism in the binary decision to donate to a charity, but that the shift in the number of donors is mostly explained by the emotional arousal provoked by the list.

2.5 Discussion

In this experiment, I observe a backlash of using the nudge. By providing a list and therefore, shaping part of the decision, the donation amount conditional on donating decreases. The “warm-glow” theory states that it is intrinsic motivation that leads individuals to donate. As participants donate less, the provision of the list likely reduces their intrinsic motivation. The list likely renders donating cognitively easier but also crowds out the intrinsic motivation. Back in the 70s, Deci (1976) investigated the negative effect of external changes. He developed the theory of the hidden cost of rewards and showed that intrinsic motivation can be undermined by external rewards. More recently, Frey and Jegen (2001), reviewed the motivation crowding out theory⁵¹ and emphasized two linked effects in principal-agent settings, the relative price effect having an impact on the crowding out of intrinsic motivation. In general terms, a higher extrinsic motivation, such as a higher wage, can diminish intrinsic motivation. Intuitively, if an external change reduces the effort an individual needs to put into accomplishing a task, then the intrinsic motivation is no longer necessary. I postulate that a change in the choice architecture, with the provision of the list, is an external modification that simplifies the donation decision, at least cognitively. Therefore, the intrinsic motivation to donate is cut back.

Frey and Jegen (2001) also reviewed the existence of two channels that can undermine intrinsic motivation. Both are due to extrinsic control. The first channel, the impression of control, impairs subjects’ self-determination. The “choice architecture” possibly displays a willingness to control the decision of the participants. As, the locus of control shifts from inside - the participant - to outside - the experimenter - the subjects reduce their intrinsic motivation.

The second channel, it decreases subjects’ self-esteem. Due to the impression of being controlled, subjects perceive that their motivation is undervalued.

In both processes, the impression of being controlled is a key factor, and the provision of the list of charities is in line with these findings. Falk and Kosfeld (2006) confirm in a principal-agent setting the cost of control. They

⁵¹This theory is related to the cognitive evaluation theory, see Deci and Ryan (1985).

find that the increase of control decreases the effort the agents put into a task. Control is interpreted as a signal of distrust and therefore undermines intrinsic motivation. Along the same lines, the provision of a list of charities could be interpreted as pushing participants into a quasi forced donation, which would reduce their intrinsic motivation.

My results support the findings of a relative dependence between intrinsic and extrinsic motivation. The overall results indicate that an increase in extrinsic motivation leads to a crowding out of intrinsic motivation. As I observe no difference at the realized level of donation, one might suggest that the increase in the extrinsic motivation is approximately equal to the decrease in the intrinsic motivation. However, this does not mean that the lower cost of choosing a charity compensates for the decrease of intrinsic motivation. There might be other extrinsic mitigating factors. For instance, DellaVigna, List, and Malmendier (2012) show that peer pressure increases donations and that participants are concerned about their self-image, but care about the cost of signaling (Tonin and Vlassopoulos 2013). Since, I do not confirm the findings of Schulz et al. (2018), that the list increases donations, I suppose that the online context might undermine the extrinsic motivation compared to the offline context in their experiment. It could be that the will to enhance one's self-image is not sufficient to compensate for the crowding out of intrinsic motivation. In this online setting, participants, although not anonymous, might perceive their signaling to be too costly compared to the same payoff for their self-image in a face-to-face context. In short, I assume that participants' intrinsic motivation is reduced by the increase in extrinsic motivation. Nevertheless, in contrast to Schulz et al. (2018), who observe that the increase in extrinsic motivation does not produce a significant decrease in intrinsic motivation, in this online context, the will to enhance the self-image is likely insufficient to compensate for the decrease of intrinsic motivation. I find support for this hypothesis in Grossman (2015) and Tonin and Vlassopoulos (2013), who experimentally manipulate the cost of signaling to enhance the self-image possible payoff and investigate the impact on giving. They find that donation is correlated with the self-image. In this nudge experiment, the motivation to appear altruistic is already undermined by the context. Moreover, they implemented an architectural change that might induce an appearance of control. Thus, it is possible that it becomes even more costly to enhance one's self-image. Donating, from an external standpoint, could be attributed to the architectural change rather than an

altruistic act.⁵² Finally, I confirm previously cited findings that the donation decision is influenced by intrinsic, but also extrinsic motivation.

Before closing the discussion, I shall mention limitations. In this experiment, I cannot preclude that the cognitive cost plays a role in the propensity to donate, but it remains marginal with respect to the emotional response as I do not observe the decrease in donation expected in the long list treatments. Moreover, a between-subjects design offers limited information regarding the motivation behind the donation decisions. For instance, it does not allow me to investigate the constituting elements of intrinsic/extrinsic motivation, such as the respective weights of improve self-image or to feel the warm-glow of giving.

⁵²I postulate here a possible interaction of the treatment with the online context, which would be channeled through the motivation to increase one's self-image. In general terms, the signaling is even costlier in the list treatment and this might not be the case in face-to-face interactions.

Finally, this experiment has important external validity as the context is highly similar to online platforms. The generalization of the results might be limited quantitatively due to the subject pool which consists only of students, but not limited qualitatively. Quantitatively, it remains possible that students have a different propensity to donate compared to the general population. For instance, students might have lower wages and thus donate less or they might be more sensitive to donations and, thus donate more. These concerns are solely on the level of donation, but not the qualitative effect of the treatment variations. Firstly, according to Druckman and Kam (2011), students do not intrinsically pose a problem for the external validity of experimental inferences. Secondly, since this study focuses on the treatment variation, only an interaction between the subject pool and the treatment variation would bias the results. There seems no obvious reason why students should differ in their reaction to treatment variations discussed in this study.

Another aspect of the external validity is that online charitable donations are substantial and growing, 15% in 2016 and 17% in 2017.⁵³ However, with respect to policy implications I bring disappointing news. The provision of a list, although increasing the number of donors, likely crowds out the intrinsic motivation to donate resulting in virtually no effect on the realized level of donation. Moreover, charities that do not make it on the list have a very low chance of being selected. Finally, I believe nudges in the context of charitable giving should be carefully studied before they are implemented otherwise there is a considerable risk that the choice architecture undermines the intrinsic motivation to donate.

⁵³<https://www.charitynavigator.org>

References

- Altmann, S., Falk, A., Heidhues, P., Jayaraman, R., & Teirlinck, M. (2019). Defaults and donations: Evidence from a field experiment. *Review of Economics and Statistics*, *101*(5), 808–826.
- Andreoni, J. (1989). Giving with impure altruism: Applications to charity and ricardian equivalence. *Journal of Political Economy*, *97*(6), 1447–1458.
- Andreoni, J. (1990). Impure altruism and donations to public goods: A theory of warm-glow giving. *The Economic Journal*, *100*(401), 464–477.
- Andreoni, J., Rao, J. M., & Trachtman, H. (2017). Avoiding the ask: A field experiment on altruism, empathy, and charitable giving. *Journal of Political Economy*, *125*(3), 625–653.
- Blake, P. R., & Rand, D. G. (2010). Currency value moderates equity preference among young children. *Evolution and Human Behavior*, *31*(3), 210–218.
- Cameron, A. C., & Miller, D. L. (2015). A practitioners guide to cluster-robust inference. *Journal of Human Resources*, *50*(2), 317–372.
- Chernev, A. (2003). When more is less and less is more: The role of ideal point availability and assortment in consumer choice. *Journal of Consumer Research*, *30*(2), 170–183.
- Cragg, J. G. (1971). Some statistical models for limited dependent variables with application to the demand for durable goods. *Econometrica*, *39*(5), 829–844.
- Deci, E. L. (1976). The hidden costs of rewards. *Organizational Dynamics*, *4*(3), 61–72.
- Deci, E. L., & Ryan, R. M. (1985). Cognitive evaluation theory. *Intrinsic motivation and self-determination in human behavior* (pp. 43–85). Springer.
- DellaVigna, S., List, J. A., & Malmendier, U. (2012). Testing for altruism and social pressure in charitable giving. *The Quarterly Journal of Economics*, *127*(1), 1–56.
- DellaVigna, S., List, J. A., Malmendier, U., & Rao, G. (2013). The importance of being marginal: Gender differences in generosity. *American Economic Review*, *103*(3), 586–90.
- Druckman, J. N., & Kam, C. D. (2011). Students as experimental participants. *Cambridge handbook of experimental political science*, *1*, 41–57.
- Falk, A., & Kosfeld, M. (2006). The hidden costs of control. *American Economic Review*, *96*(5), 1611–1630.

- Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives*, *19*(4), 25–42.
- Frey, B. S., & Jegen, R. (2001). Motivation crowding theory. *Journal of Economic Surveys*, *15*(5), 589–611.
- Gneezy, A., Gneezy, U., Nelson, L. D., & Brown, A. (2010). Shared social responsibility: A field experiment in pay-what-you-want pricing and charitable giving. *Science*, *329*(5989), 325–327.
- Gneezy, U., Leonard, K. L., & List, J. A. (2009). Gender differences in competition: Evidence from a matrilineal and a patriarchal society. *Econometrica*, *77*(5), 1637–1664.
- Gourville, J. T., & Soman, D. (2005). Overchoice and assortment type: When and why variety backfires. *Marketing Science*, *24*(3), 382–395.
- Grossman, Z. (2014). Strategic ignorance and the robustness of social preferences. *Management Science*, *60*(11), 2659–2665.
- Grossman, Z. (2015). Self-signaling and social-signaling in giving. *Journal of Economic Behavior & Organization*, *117*, 26–39.
- Grossman, Z., & van der Weele, J. J. (2017). Self-image and willful ignorance in social decisions. *Journal of the European Economic Association*, *15*(1), 173–217.
- Iyengar, S. S., & Lepper, M. R. (2000). When choice is demotivating: Can one desire too much of a good thing? *Journal of Personality and Social Psychology*, *79*(6), 995–1006.
- Jenni, K., & Loewenstein, G. (1997). Explaining the identifiable victim effect. *Journal of Risk and Uncertainty*, *14*(3), 235–257.
- Karlan, D., & List, J. A. (2007). Does price matter in charitable giving? Evidence from a large-scale natural field experiment. *American Economic Review*, *97*(5), 1774–1793.
- Kistler, D., Thöni, C., & Welzel, C. (2017). Survey response and observed behavior: Emancipative and secular values predict prosocial behaviors. *Journal of Cross-Cultural Psychology*, *48*(4), 461–489.
- Kogut, T., & Ritov, I. (2005a). The identified victim effect: An identified group, or just a single individual? *Journal of Behavioral Decision Making*, *18*(3), 157–167.
- Kogut, T., & Ritov, I. (2005b). The singularity effect of identified victims in separate and joint evaluations. *Organizational Behavior and Human Decision Processes*, *97*(2), 106–116.
- Liu, B., Huang, Z., Xu, G., Jin, Y., Chen, Y., Li, X., Wang, Q., Song, S., & Jing, J. (2016). Altruistic sharing behavior in children: Role of theory of mind and inhibitory control. *Journal of Experimental Child Psychology*, *141*, 222–228.

- MacAskill, W. (2015). *Doing good better: Effective altruism and a radical new way to make a difference*. Guardian Faber Publishing.
- McManus, B., & Bennet, R. (2011). The demand for products linked to public goods: Evidence from an online field experiment. *Journal of Public Economics*, 95(5-6), 403–415.
- Meier, S. (2007). Do subsidies increase charitable giving in the long run? Matching donations in a field experiment. *Journal of the European Economic Association*, 5(6), 1203–1222.
- Morgan, J., & Sefton, M. (2000). Funding public goods with lotteries: Experimental evidence. *The Review of Economic Studies*, 67(4), 785–810.
- Nicholson, N., Soane, E., Fenton-O’Creevy, M., & Willman, P. (2005). Personality and domain-specific risk taking. *Journal of Risk Research*, 8(2), 157–176.
- Schulz, J. F., Thiemann, P., & Thöni, C. (2018). Nudging generosity: Choice architecture and cognitive factors in charitable giving. *Journal of Behavioral and Experimental Economics*, 74, 139–145.
- Schwartz, B. (2004). *The paradox of choice - why more is less*. Harper Perennial.
- Silverman, B. W. (1986). *Density estimation for statistics and data analysis* (Vol. 26). Chapman; Hall.
- Small, D. A., Loewenstein, G., & Slovic, P. (2007). Sympathy and callousness: The impact of deliberative thought on donations to identifiable and statistical victims. *Organizational Behavior and Human Decision Processes*, 102(2), 143–153.
- Thaler, R. H., & Sunstein, C. R. (2008). *Nudge: Improving decisions about health, wealth, and happiness*. Yale University Press.
- Tonin, M., & Vlassopoulos, M. (2013). Experimental evidence of self-image concerns as motivation for giving. *Journal of Economic Behavior & Organization*, 90, 19–27.
- Tonin, M., & Vlassopoulos, M. (2015). Corporate philanthropy and productivity: Evidence from an online real effort experiment. *Management Science*, 61(8), 1795–1811.

2.6 Appendices

Control treatment

*** Falls Sie für die Auszahlung ausgelost werden, wollen Sie einen Teil Ihres Gewinns für eine wohltätige Organisation spenden?**

Ja Nein

Falls ja, geben Sie an, wie viele Prozente (von Ihrem Gewinn) Sie spenden möchten.

Prozent

Falls ja, an welche Organisation möchten Sie spenden?

Translation from top to bottom: If you are selected and therefore get paid, would you be willing to give part of your earnings to a charitable organization? If you agree, what percents (of your earnings) would you be willing to give. If you agree, to which organization would you like to give that money to?

Figure A2.1: Screenshot - Control treatment

Short list treatment

The screenshot shows a survey form with three main sections, each separated by a yellow horizontal bar. The first section has a red asterisk and asks if the respondent wants to donate a portion of their profit to a charitable organization. It includes radio buttons for 'Ja' and 'Nein'. The second section asks for the percentage of profit to be donated, with a text input field followed by the word 'Prozent'. The third section asks for the recipient organization, listing WWF, Amnesty International, Rotes Kreuz, and Ärzte ohne Grenzen with radio buttons, and an 'Andere Organisation' option with a text input field.

*** Falls Sie für die Auszahlung ausgelost werden, wollen Sie einen Teil Ihres Gewinns für eine wohltätige Organisation spenden?**

Ja Nein

Falls ja, geben Sie an, wie viele Prozente (von Ihrem Gewinn) Sie spenden möchten.

Prozent

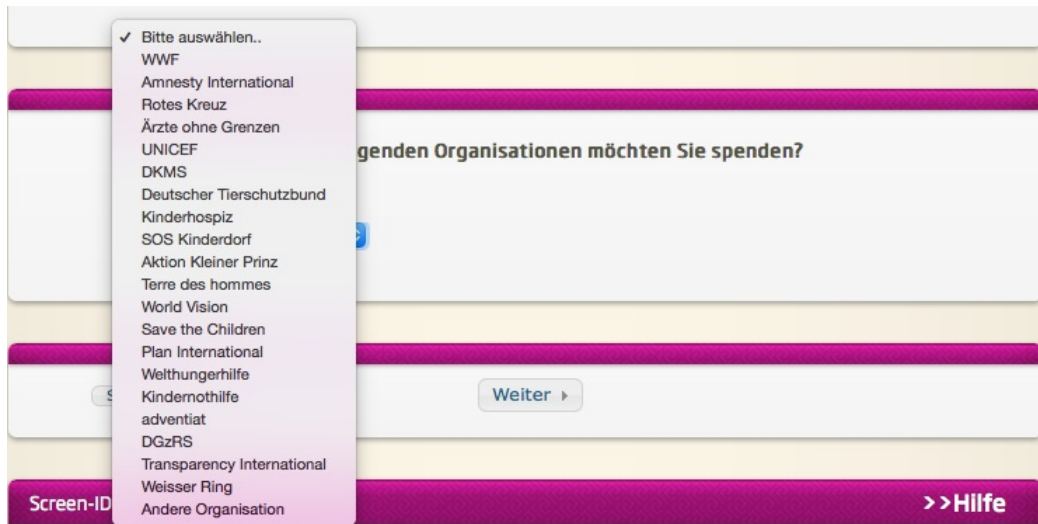
Falls ja, an welche der folgenden Organisationen möchten Sie spenden?

WWF Ärzte ohne Grenzen
 Amnesty International Andere Organisation
 Rotes Kreuz

Translation: same as in the control treatment, except that participants can also select between 4 options.

Figure A2.2: Screenshot - Short List

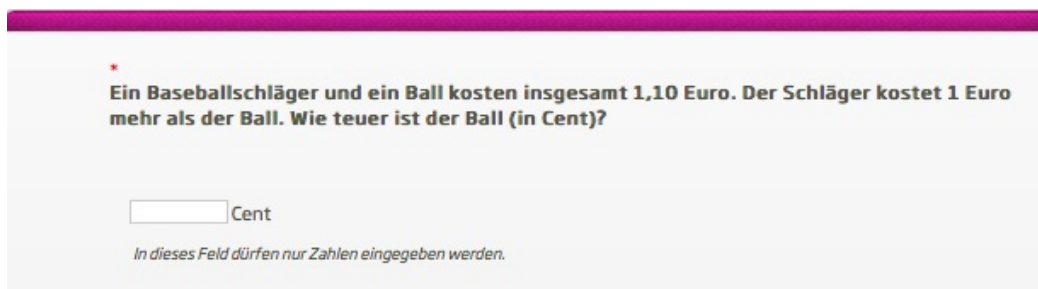
Drop-down long list treatment



Translation: same configuration as the control group. In the drop-down list, on the top: "please choose" and at the end: "Other organization". If the subject clicks on this last option, a blank space appears where she could enter another organization name.

Figure A2.3: Screenshot - DD Long

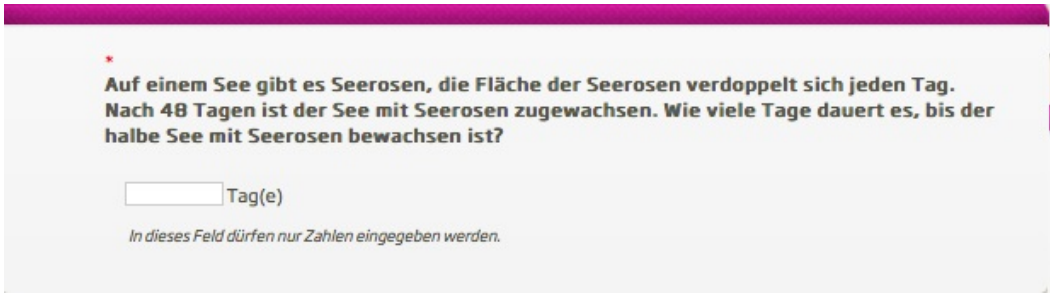
Cognitive Reflection Test 1



Translation: a baseball bat and a ball cost together 1.10 euros. The baseball bat costs 1 euro more than the ball. How much does the ball cost (in cents) ?

Figure A2.4: Screenshot - CRT 1

Cognitive Reflection Test 2



*
Auf einem See gibt es Seerosen, die Fläche der Seerosen verdoppelt sich jeden Tag. Nach 48 Tagen ist der See mit Seerosen zugewachsen. Wie viele Tage dauert es, bis der halbe See mit Seerosen bewachsen ist?

Tag(e)

In dieses Feld dürfen nur Zahlen eingegeben werden.

Translation: in a lake there are water lilies, the area of water lilies doubles every day. After 48 days, the lake is overgrown with water lilies. How many days does it take for half of the lake to be overgrown with water lilies?

Figure A2.5: Screenshot - CRT 2

2.7 Supplementary material

The supplementary material is structured as follows: (i) randomization checks, (ii) investigation of social preferences, and (iii) investigation of the income effect.

2.7.1 Randomization checks

Because the donation decision is the final stage of other previous games. I provide the p -values of a multinomial logistic regression in the manuscript in Table 2.4. As some of the variables are significant, I provide, here, additional analyses to confirm the independence of social preferences, payoffs, age, gender, and waves to the treatment variations of the experiment.

In the social preferences category, risk-taking is the percentage of potential earnings a participant decides to put on a lottery game with the possibility of gaining 2.5 times the amount put on the lottery with a probability of .5.

The probability to lie is derived from the honesty game, where a participant has to report the number of heads in a series of 16 flips. I assume that participants who reported a number of heads under or equal to 8, did not lie ($\rho_L = 0$). If they reported a number higher than 8, I calculated the probability that they were lying following the function:

$$1 - \frac{1}{\frac{P_B}{P_X}} = \rho_L$$

Where P_B is the probability of obtaining 8 heads out of 16 flips, P_X is the probability obtaining the number of heads reported, and ρ_L is the probability to lie. Another possible computation is to take the number reported, which is the exact same as the honesty game payoff, which is also tested in the following regressions.

In the property game, participants gather carrots. Each carrot gathered returns 1 experimental currency unit (ECU). Participants have three possibilities: plant, steal, or protect. A participant has to decide how to allocate 7 units between these 3 possibilities. She knows that every unit allocated to planting returns 1 ECU. However, another participant can steal carrots if she does not have sufficient protection. Each carrot, successfully stolen, also returns 1 ECU. In order to successfully steal a carrot, a participant has to invest more in stealing than the other participant in protecting. Every additional unit invested in stealing, than the other participant in protecting, returns 1 ECU.

Contributing is derived from the public goods game. It is the amount in ECU, from 0 to 100, a participant decides to donate to a public good.

The participant knows that every unit sent to the public account will be multiplied by 1.6 and then divided between the 4 members of the group.

The payoff is always expected, as it is only a participant or a group of participants that will be sorted out to receive the payment. Nevertheless, in some games, the participants receive feedback regarding their performance and in others, they do not.

In the public goods game (PGG), participants do not receive feedback, therefore, I indicate expected payoff (E(Payoff)). As the experimenters asked participants what is their belief about the average contribution of the other members of the group, I can calculate the expected payoff following this function:

$$E_i(\text{Payoff}) = 100 - c_i + \frac{2}{5} \sum_{j=1}^4 c_j + E_i(c_j)$$

The expected payoff of player i is a function of her contribution (c_i) to the public goods game and her belief about the others' contributions $E_i(c_j)$.

In the risk game, participants receive feedback, therefore, I indicate payoff. The payoff is given by the following function:

$$P_i = r_i \times 2.5 \times p$$

The payoff (P_i) is given by the amount invested in the lottery (r_i) multiplied by 2.5 and the binary outcome (p) whether the person wins or loses.

In the honesty game, the payoff is again not expected as the participant receives exactly the amount he reports.

In the property game, the payoff is expected, as it depends highly on the behavior of the other participant one is matched with. Assuming a uniform distribution of strategy, I simulated the average payoff dependent on a participant's strategy.

The overall payoff in all tasks is simply the addition of the payoffs and expected payoffs in each of the previous games.

Since there is some significance in Table 2.4, I run each regression from the main paper separately per wave and per location. In Table S2.1, I investigate the social preferences. I run each regression with the same dependent variable: the binary decision to donate. Each regression keeps only a subsample of the overall dataset. In other words, I keep only Wave 1 for the first regression, then only Wave 2, etc.

I see that although I lose significance for the main coefficient of interest - the list - the treatment effect remains qualitatively the same. By taking a sub-sample, the study becomes under-powered to test the different treatment

The propensity to donate - Extensive margin - Robustness checks

	Dependent variable: Donate: yes = 1, no = 0				
	Wave 1	Wave 2	Wave 3	Magdeburg	Hamburg
List	0.076 (0.064)	0.085 (0.071)	0.108 (0.077)	0.092 ⁺ (0.049)	0.144* (0.066)
Long list	0.129* (0.063)	0.042 (0.072)	0.018 (0.046)	0.035 (0.039)	0.027 (0.051)
Drop-down	0.015 (0.064)	-0.136 ⁺ (0.075)	-0.044 (0.050)	-0.082* (0.041)	-0.033 (0.054)
Contributing	0.002* (0.001)	-0.001 (0.001)	0.000 (0.001)	0.001 ⁺ (0.001)	-0.001 (0.001)
Risk-taking	0.000 (0.001)	-0.000 (0.001)	0.001 (0.001)	0.000 (0.001)	0.001 (0.001)
Probability to lie	-0.158 (0.122)	-0.160* (0.072)	-0.199** (0.056)	-0.207** (0.052)	-0.116 ⁺ (0.066)
Property game - Steal	-0.044 ⁺ (0.024)	0.029 (0.026)	-0.060* (0.024)	-0.032 ⁺ (0.017)	-0.014 (0.024)
Property game - Plant	-0.015 (0.018)	0.040* (0.020)	-0.008 (0.018)	-0.001 (0.013)	0.020 (0.018)
Female (dummy)	-0.012 (0.050)	0.001 (0.054)	0.003 (0.045)	-0.020 (0.035)	0.050 (0.048)
Location	0.042 (0.053)	0.136* (0.057)	0.047 (0.043)		
Wave 2				0.085 ⁺ (0.045)	0.161* (0.069)
Wave 3				0.176** (0.047)	0.129* (0.063)
Constant	0.294* (0.130)	0.264 ⁺ (0.136)	0.486** (0.130)	0.299** (0.096)	0.187 (0.132)
<i>F</i> -test	2.2	2.2	3.1	4.1	2.0
Prob > <i>F</i>	0.018	0.018	0.001	0.000	0.024
<i>R</i> ²	0.052	0.058	0.057	0.052	0.046
<i>N</i>	410	369	530	827	482

Notes: OLS estimates. The dependent variable is always the binary decision to donate. Apart from the dummy variables for the treatment variation, I include variables for the behaviors in the previous games. Standard errors in parentheses. ⁺ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$.

Table S2.1: OLS - Social preferences

variations. Nevertheless, for none of the above coefficients, the sign goes in the opposite direction, suggesting that the results in the main paper is driven by a wave or a location.

Regarding social preferences, the probability to lie is significant in the majority of the regressions. This suggests that participants that lie also do not donate. Moreover, participants that decide to invest more in stealing tend to also donate less often. Overall, this shows that social preferences seem to matter in the decision to donate

In Table S2.2, I still investigate social preferences, however this time on the percentage of potential earnings a participant decides to donate conditional on donating.

The statement is essentially the same, as the main coefficient of interest goes in the expected direction. In these regressions, due to the low number of observations, the coefficients are often even less significant.

In Table S2.3, I investigate a possible income effect. Participants that earned more during the previous game might be more willing to donate as a form of reciprocity. But as this table shows, this is not the case here.

In Table S2.4, I still investigate the income effect, however, the dependent variable is always the donation amount in percentage of potential earnings conditional on donating. Again, the statement is essentially the same, as the coefficient goes in the expected direction.

As the significance level is dependent on the number of observations in the next section I run regressions with the whole dataset.

2.7.2 Social Preferences

One might be concerned with the low number of observations in the robustness checks. Therefore, I run the regression for social preferences with the overall dataset (Tables S2.5 and S2.6).

I find that the probability to lie in the honesty game and the decision to steal in the property game are both negative and significant ($p < 0.01$) in explaining the decision to donate. This confirms that some preferences matter for the decision to donate.

As for the donation amount conditional on donating, the amount contributed in the public goods game is a significant ($p < 0.01$) predictor for amount in the donation decision. The more a participant contributed in the public goods game, the more she donated. The probability to lie also predicts the donation amount, albeit the correlation is marginally significant ($p < 0.1$). A higher probability to lie is associated with lower amounts. In contrast to the claim in the paper about different processes in the donation decision, here, it seems that in the donation decision, the binary decision and

The donation amount - Intensive margin - Robustness checks

Dependent variable: Amount donating (in percentage of potential earnings)					
	Wave 1	Wave 2	Wave 3	Magdeburg	Hamburg
List	-4.157 (4.670)	-5.471 (3.405)	-6.124 (4.053)	-6.519* (3.117)	-4.716 (3.229)
Long list	-2.706 (3.889)	-2.433 (3.212)	-0.261 (2.361)	-1.090 (2.252)	0.387 (2.271)
Drop-down	-2.134 (4.132)	8.190* (3.520)	0.637 (2.541)	0.670 (2.452)	2.122 (2.361)
Contributing	0.123* (0.060)	0.033 (0.043)	0.103* (0.042)	0.097* (0.039)	0.082* (0.035)
Risk-taking	0.013 (0.062)	0.040 (0.043)	-0.034 (0.041)	0.003 (0.039)	0.008 (0.037)
Probability to lie	-10.154 (10.067)	-4.622 (3.727)	-0.904 (3.150)	-2.455 (3.648)	-2.605 (3.016)
Property game - Steal	1.436 (1.696)	0.530 (1.232)	-1.934 (1.447)	0.298 (1.141)	-1.059 (1.181)
Property game - Plant	1.121 (1.209)	1.180 (0.935)	0.203 (0.886)	0.868 (0.792)	0.411 (0.803)
Female (dummy)	1.831 (3.347)	-0.350 (2.478)	0.356 (2.275)	-0.232 (2.079)	1.378 (2.188)
Location	-6.236 ⁺ (3.650)	-0.424 (2.564)	0.444 (2.247)		
Wave 2				-3.063 (2.789)	3.167 (3.160)
Wave 3				-2.446 (2.761)	3.072 (2.976)
Constant	13.868 (9.463)	15.981* (7.123)	21.481** (6.538)	19.351** (6.113)	14.346* (6.251)
<i>F</i> -test	1.5	1.5	1.7	1.9	1.7
Prob > <i>F</i>	0.135	0.142	0.074	0.033	0.078
<i>R</i> ²	0.094	0.089	0.064	0.059	0.076
<i>N</i>	159	166	264	351	238

Notes: OLS estimates. The dependent variable is the donation amount conditional on donating. Apart from the variables for the treatment variation, I include variables for the behaviors in the previous games. Standard errors in parentheses.
⁺ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$.

Table S2.2: OLS - Social preferences

The propensity to donate - Extensive margin - Robustness checks

	Dependent variable: Donate: yes = 1, no = 0				
	Wave 1	Wave 2	Wave 3	Magdeburg	Hamburg
List	0.175 (0.185)	0.097 (0.071)	0.082 (0.078)	0.045 (0.064)	0.202* (0.078)
Long list	0.040 (0.214)	0.048 (0.072)	0.012 (0.047)	0.008 (0.048)	0.031 (0.057)
Drop-down	0.191 (0.216)	-0.152* (0.075)	-0.030 (0.051)	-0.041 (0.051)	-0.092 (0.061)
Overall E(Payoff)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)
Female (dummy)	0.042 (0.145)	-0.011 (0.053)	0.027 (0.045)	-0.002 (0.042)	0.036 (0.053)
Location	-0.073 (0.156)	0.138* (0.057)	0.034 (0.044)		
Wave 2				0.054 (0.091)	0.286* (0.141)
Wave 3				0.138 (0.093)	0.271+ (0.139)
Constant	-0.038 (0.261)	0.328** (0.093)	0.355** (0.098)	0.322** (0.114)	-0.010 (0.157)
<i>F</i> -test	1.0	2.0	0.5	0.7	2.4
Prob > <i>F</i>	0.420	0.062	0.814	0.641	0.022
<i>R</i> ²	0.131	0.032	0.006	0.009	0.043
<i>N</i>	48	369	530	570	377

Notes: OLS estimates. The dependent variable is always the binary decision to donate. Apart from the dummy variables for the treatment variation, I include the variable for the overall expected payoff. Standard errors in parentheses. + $p < 0.1$, * $p < 0.05$, ** $p < 0.01$.

Table S2.3: OLS - Income effect

The donation amount - Intensive margin - Robustness checks

	Dependent variable: Amount donating (in percentage of potential earnings)				
	Wave 1	Wave 2	Wave 3	Magdeburg	Hamburg
List	-22.958 (21.955)	-5.771 ⁺ (3.426)	-5.513 (4.068)	-5.459 (3.568)	-8.690* (3.869)
Long list	16.451 (18.011)	-1.508 (3.181)	-1.112 (2.381)	-0.507 (2.582)	-0.053 (2.532)
Drop-down	-15.266 (20.347)	7.087* (3.496)	0.791 (2.562)	0.757 (2.765)	4.610 ⁺ (2.687)
Overall E(Payoff)	0.049 (0.066)	0.007 (0.007)	0.011 (0.007)	0.007 (0.007)	0.014 ⁺ (0.007)
Female (dummy)	9.773 (14.492)	-0.723 (2.453)	0.814 (2.287)	0.317 (2.289)	1.424 (2.466)
Location	-19.025 (17.355)	-0.330 (2.565)	0.426 (2.232)		
Wave 2				-2.454 (5.516)	11.578 (9.565)
Wave 3				-1.161 (5.479)	11.983 (9.455)
Constant	11.163 (35.986)	21.287** (4.360)	20.594** (5.084)	23.647** (6.712)	7.381 (10.966)
<i>F</i> -test	0.8	1.4	0.9	0.6	1.9
Prob > <i>F</i>	0.619	0.220	0.486	0.742	0.076
<i>R</i> ²	0.364	0.050	0.021	0.017	0.066
<i>N</i>	15	166	264	252	193

Notes: OLS estimates. The dependent variable is donation amount conditional on donating. Apart from the dummy variables for the treatment variation, I include the variable for the overall expected payoff. Standard errors in parentheses. ⁺ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$.

Table S2.4: OLS - Income effect

The propensity to donate - Extensive margin - Investigation of Social Preferences

Dependent variable: Donate: yes = 1, no = 0					
List	0.106** (0.039)	0.110** (0.040)	0.105** (0.039)	0.112** (0.040)	0.111** (0.039)
Long list	0.028 (0.031)	0.028 (0.031)	0.035 (0.031)	0.028 (0.031)	0.033 (0.031)
Drop-down	-0.055 ⁺ (0.032)	-0.053 (0.033)	-0.053 (0.032)	-0.061 ⁺ (0.033)	-0.063 ⁺ (0.032)
Contributing	0.001** (0.000)				0.000 (0.000)
Risk-taking		0.001 (0.000)			0.001 (0.000)
Probability to lie			-0.181** (0.040)		-0.172** (0.041)
Property game - Steal				-0.031* (0.014)	-0.025 ⁺ (0.014)
Property game - Plant				0.008 (0.011)	0.005 (0.011)
Wave 2	0.056 (0.036)	0.058 (0.036)	0.102** (0.037)	0.059 ⁺ (0.036)	0.109** (0.037)
Wave 3	0.088* (0.035)	0.095** (0.035)	0.153** (0.037)	0.100** (0.035)	0.155** (0.037)
Location	0.058* (0.029)	0.057* (0.029)	0.063* (0.029)	0.061* (0.029)	0.068* (0.029)
Female (dummy)	0.012 (0.027)	0.013 (0.028)	0.007 (0.027)	-0.000 (0.028)	0.007 (0.028)
Constant	0.222** (0.047)	0.259** (0.046)	0.301** (0.038)	0.284** (0.072)	0.239** (0.077)
<i>F</i> -test	4.2	3.5	5.9	4.6	5.3
Prob > <i>F</i>	0.000	0.001	0.000	0.000	0.000
<i>R</i> ²	0.025	0.021	0.034	0.031	0.046
<i>N</i>	1,334	1,334	1,334	1,309	1,309

Notes: OLS estimates. The dependent variable is always the binary decision to donate. Apart from the dummy variables for the treatment variation, I include the variables for each derived social preference in the previous games. Standard errors in parentheses. ⁺ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$.

Table S2.5: OLS - Social preferences

The donation amount - Intensive margin - Investigation of Social Preferences

Dependent variable: Amount donating (in percentage of potential earnings)					
List	-5.843** (2.217)	-5.474* (2.251)	-5.451* (2.245)	-5.264* (2.258)	-5.613* (2.242)
Long list	-0.611 (1.578)	-0.757 (1.603)	-0.731 (1.599)	-0.647 (1.616)	-0.528 (1.603)
Drop-down	1.097 (1.703)	1.363 (1.728)	1.324 (1.724)	1.404 (1.732)	1.185 (1.718)
Contributing	0.109** (0.025)				0.091** (0.027)
Risk-taking		0.020 (0.027)			0.006 (0.027)
Probability to lie			-4.161+ (2.305)		-2.182 (2.374)
Property game - Steal				-0.731 (0.814)	-0.345 (0.817)
Property game - Plant				0.964+ (0.566)	0.683 (0.567)
Wave 2	-1.600 (1.964)	-2.219 (1.998)	-1.385 (2.056)	-2.036 (1.989)	-0.968 (2.055)
Wave 3	-1.193 (1.877)	-0.781 (1.904)	0.354 (2.005)	-0.932 (1.909)	-0.516 (2.011)
Location	-1.087 (1.493)	-1.584 (1.512)	-1.284 (1.517)	-1.575 (1.530)	-1.074 (1.527)
Female (dummy)	0.222 (1.458)	-0.150 (1.493)	-0.345 (1.472)	-0.192 (1.506)	0.243 (1.511)
Constant	20.622** (2.764)	26.673** (2.693)	27.759** (2.249)	23.160** (3.985)	18.211** (4.329)
<i>F</i> -test	3.8	1.5	1.8	2.4	2.9
Prob > <i>F</i>	0.000	0.167	0.072	0.012	0.001
<i>R</i> ²	0.049	0.020	0.024	0.035	0.058
<i>N</i>	598	598	598	589	589

Notes: OLS estimates. The dependent variable is the donation amount conditional on donating. Apart from the dummy variables for the treatment variation, I include the variables for each derived social preference in the previous games. Standard errors in parentheses. + $p < 0.1$, * $p < 0.05$, ** $p < 0.01$.

Table S2.6: OLS - Social preferences

the donation amount are two different processes. Here, the results tend to confirm that the decision to donate is more influenced by social preferences than the amount conditional on donating.

In Tables S2.7 and S2.8, I report the point estimates and the significance levels of interactions between social preferences and treatment variations. These tests for interaction effects are similar to the tests of interaction for gender in Tables 2.4 and 2.5. The main differences are the interaction tested and the female control variable.

The propensity to donate - Extensive margin - Social preferences interactions

	List	Long list	Drop-down
Contributing in PGG	0.001	-0.001	0.001
Probability to lie	-0.013	-0.072	-0.022
Risk-taking	0.000	-0.001	-0.000
Protect	0.031	0.004	0.023
Steal	0.007	0.023	-0.012

Notes: OLS estimates. I report the point estimates of all interactions for each derived social preference in the previous games with the treatment variations. The dependent variable is always the binary decision to donate. Each preference is an independent regression with control variables, i.e., treatments, waves, location, female, and the preference. + $p < 0.1$, * $p < 0.05$, ** $p < 0.01$.

Table S2.7: OLS estimates - Interactions

As mentioned in the result section, most of the interactions are insignificant except for two interactions on the donation amount conditional on donating. The first one is between risk-taking and the drop-down treatments, and the second one is between contribution in the public goods game and the list treatments. The first result suggests that risk-takers tend to donate less when they are in the drop-down treatments, possibly due to a lower emotional sensitivity (Nicholson et al. 2005). The second result suggests that higher contributors donate less when they are in a list treatments, possibly because they are intrinsically more motivated to donate in the first place compared to lower contributors and that facing the list withdraws more of their intrinsic motivation.

The donation amount - Intensive margin - Social preferences interactions

	List	Long list	Drop-down
Contributing in PGG	-0.118 ⁺	-0.038	-0.040
Probability to lie	6.258	-2.637	6.930
Risk-taking	-0.099	-0.079	-0.119*
Protect	2.230	0.924	1.485
Steal	0.006	0.082	0.341

Notes: OLS estimates. I report the point estimates of all interactions for each derived social preference in the previous games with the treatment variations. The dependent variable is always the donation amount conditional on donating. Each preference is an independent regression with control variables, i.e., treatments, waves, location, female, and the preference. ⁺ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$.

Table S2.8: OLS estimates - Interactions

2.7.3 Income effect

In Tables S2.9 and S2.10, I use expected payoffs from the previous games as explanatory variables for the binary decision to donate or the amount donated conditional on donating. I find that the (expected) payoff is playing a role. However, in the honesty game, the expected payoff probably also reflects social preferences, therefore, it is likely a confound, as it also goes in the same direction.

This might also be the case for the property game as planting returns the highest profit on average. Therefore, it goes in the same direction as the social preference to plant. In both cases, social preferences are hard to disentangle from income effects. As the overall expected payoff is dependent on these two games, the same might hold. The overall expected payoff is significantly ($p < 0.05$) correlated with the amount.

After controlling for social preferences and income effect, the main coefficient of interest, which is the list, remains significant ($p < 0.05$ or $p < 0.01$). I can, therefore, conclude, that although some of the behaviors and profits from previous games might explain the donation decision it remains often marginal with respect to the treatment manipulation of the list.

The propensity to donate - Extensive margin - Investigation of Income Effect

	Dependent variable: Donate: yes = 1, no = 0				
List	0.107** (0.040)	0.108** (0.040)	0.110* (0.048)	0.113** (0.040)	0.112* (0.049)
Long list	0.029 (0.031)	0.029 (0.031)	0.017 (0.036)	0.024 (0.031)	0.013 (0.037)
Drop-down	-0.054 ⁺ (0.033)	-0.051 (0.032)	-0.048 (0.038)	-0.062 ⁺ (0.033)	-0.058 (0.039)
PGG E(Payoff)	0.000 (0.000)				
Risk game E(Payoff)		0.000 (0.000)			
Honesty game payoff			-0.022** (0.005)		
PG E(Payoff)				0.060** (0.022)	
Overall E(Payoff)					0.000 (0.000)
Wave 2	0.055 (0.036)	0.055 (0.036)	0.115 (0.076)	0.059 (0.036)	0.129 ⁺ (0.076)
Wave 3	0.092** (0.035)	0.095** (0.035)	0.161* (0.076)	0.099** (0.035)	0.171* (0.077)
Location	0.057* (0.029)	0.056 ⁺ (0.029)	0.063 ⁺ (0.033)	0.056 ⁺ (0.029)	0.061 ⁺ (0.034)
Female (dummy)	0.010 (0.027)	0.011 (0.027)	0.012 (0.032)	0.013 (0.028)	0.017 (0.033)
Constant	0.254** (0.054)	0.276** (0.040)	0.446** (0.097)	0.054 (0.097)	0.172 ⁺ (0.092)
<i>F</i> -test	3.4	3.5	4.3	4.4	2.3
Prob > <i>F</i>	0.001	0.000	0.000	0.000	0.019
<i>R</i> ²	0.020	0.021	0.034	0.026	0.019
<i>N</i>	1,334	1,334	972	1,309	947

Notes: OLS estimates. The dependent variable is always the binary decision to donate. Apart from the dummy variables for the treatment variation, I include the variable for the expected payoff for each game. Standard errors in parentheses. ⁺ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$.

Table S2.9: OLS - Income effect

The donation amount - Intensive margin - Investigation of Income Effect

	Dependent variable: Amount donating (in percentage of potential earnings)				
List	-5.915** (2.227)	-5.465* (2.250)	-6.310* (2.563)	-5.238* (2.262)	-6.957** (2.602)
Long list	-0.495 (1.586)	-0.825 (1.602)	-0.763 (1.767)	-0.862 (1.615)	-0.332 (1.805)
Drop-down	1.042 (1.711)	1.375 (1.728)	2.807 (1.902)	1.350 (1.735)	2.560 (1.933)
PGG E(Payoff)	0.020** (0.005)				
Risk game E(Payoff)		-0.008 (0.009)			
Honesty game payoff			-0.718* (0.303)		
PG E(Payoff)				3.198** (1.172)	
Overall E(Payoff)					0.010* (0.005)
Wave 2	-2.043 (1.966)	-2.404 (1.986)	-0.076 (4.609)	-2.258 (1.985)	0.509 (4.639)
Wave 3	-0.916 (1.883)	-0.843 (1.903)	0.840 (4.556)	-0.867 (1.911)	1.357 (4.592)
Location	-1.368 (1.496)	-1.476 (1.516)	0.335 (1.648)	-1.692 (1.523)	-0.060 (1.668)
Female (dummy)	0.381 (1.471)	-0.412 (1.479)	0.101 (1.624)	-0.022 (1.491)	0.590 (1.669)
Constant	19.671** (3.117)	28.239** (2.322)	31.569** (5.746)	14.655** (5.328)	19.967** (5.487)
<i>F</i> -test	3.2	1.5	1.8	2.3	1.6
Prob > <i>F</i>	0.002	0.156	0.067	0.018	0.111
<i>R</i> ²	0.041	0.020	0.032	0.031	0.029
<i>N</i>	598	598	454	589	445

Notes: OLS estimates. The dependent variable is the donation amount conditional on donating. Apart from the dummy variables for the treatment variation, I include the variable for the expected payoff for each game. Standard errors in parentheses. + $p < 0.1$, * $p < 0.05$, ** $p < 0.01$.

Table S2.10: OLS - Income effect

3 Chapter 3

Gender equality and gender differences in values:

A puzzling relationship

Jason Wettstein

Abstract

Does gender equality increase or decrease gender differences in values? While some studies find that economic prosperity and more gender equality, as measured from gender differences in reproductive health, empowerment, and economic status, is associated with less gender differences in values, others find the opposite association. Using the World Value Survey and the European Value Survey over a period of 35 years, with 508,707 observations, I show, first, that more gender equality and economic growth are unambiguously associated with less gender differences in life-situations as measured in the surveys. Second, the effect of the evolution of gender equality and economic growth on differences in values depends on the level of the analysis. In the cross-country analysis, I show a robust divergence of values, while on the within-country analysis, more gender equality leads to fewer differences. This puzzle is robust to additional controls like ecological stress factors and several cultural differences. The results suggest that, despite numerous claims to the opposite, the causal link between gender equality and gender differences in values is not yet well understood in the literature.

3.1 Introduction

The development of gender policies in many countries has led to an increasing interest in understanding gender differences, their source, and variance. Whether women and men are inherently different in their values is of interest to fundamental research, as well as to policymakers and the general public. One interesting aspect is the question of whether an increase in gender equality leads to more or less differences in values between men and women. Using data from the World Values Survey (WVS) and the European Values Survey (EVS), I reanalyze the relationship between gender equality and gender differences in values.

In this paper, I contrast two hypotheses. The social role theory predicts that gender differences in values decrease in more gender-egalitarian countries (Eagly and Wood 1999; Gneezy et al. 2009). In this hypothesis, the division of labor is assumed to be the main cause of gender differences. This division is then supported by socialization, self-categorization, and self-stereotyping (Wood and Eagly 2012). Overall, the values are derived from social roles. As they are culturally constructed, the gender gap in values reflects the gap in roles.

The major competing theory, the resource hypothesis, predicts that access to sufficient material resources leads to more gender differences. The values are assumed to be intrinsic (Buss 1995; Geary 2010) and their manifestation is a matter of opportunity. Material resources are a prerequisite that removes the gender-neutral goal for subsistence, which leaves a scope for the expression of gender differences (Almås et al. 2016; Haushofer and Fehr 2014; McLoyd 1998; Tanaka et al. 2010). While economic growth indicates less overall competition for resources, greater gender equality might even be a better predictor for gender differences in values as it implies easier access to resources for women (Duflo 2012; Inglehart 2015; Inglehart and Norris 2003).

I study two dimensions: life-situations and values. I separate both these categories into two subcategories: for life-situation, whether items are subjectively or objectively measured; and for values, whether they point towards the respondent (self-centered) or towards others (general statements). I demonstrate that, while gender differences for life-situations unambiguously decrease with growing economic wealth and gender equality, the correlation with values is sensitive to the level of the analysis. In a cross-country investigation, the correlation is positive supporting the resource hypothesis (divergence of values between male and female values), whereas on a within-country level the correlation is negative, this time supporting the social role theory (convergence).

The results challenge studies supporting the increase of gender differences with more gender equality and economic prosperity (Falk and Hermle 2018; Mac Giolla and Kajonius 2019), but also those supporting convergence (Donnelly and Twenge 2017; Konrad et al. 2000). My findings particularly complements those of Connolly et al. (2019), who find both convergence, in the within-country analysis, and divergence, in the cross-country analysis, of gender differences in values. However, while they focus on the evolution of value priorities such as achievement, benevolence, power, stimulation, and universalism with respect to the evolution of gender equality from 2002 to 2016, I extend their research to a broader set of values⁵⁴ over a longer period of time (1981–2014). I also demonstrate that the indexes of economic growth and gender equality are valid proxies to measure life-situations, as these indexes systematically correlate with the categorization of this dimension in the WVS and the EVS.⁵⁵ I finally investigate the possibility of a sampling bias and tackle possible endogeneity issues in the model. I find that the divergence of values with respect to more gender equality and economic prosperity is robust to sub-sampling and alternative specifications in the cross-sectional analysis. I end up with a puzzle, where either the cross-country analysis still suffers from endogeneity or the within-country analysis is unable to capture the effect of the evolution of gender equality and economic prosperity due to a sampling bias or the limited variation of the indexes within countries.

3.2 Methods

The method section is structured in two parts: (i) the gender differences coefficients from the WVS and the EVS, and (ii) the indexes of emancipation and GDP.

3.2.1 Four measurements for gender differences

For my analysis, I use a combined dataset of the longitudinal World Values Survey (WVS) and the longitudinal European Values Survey (EVS). The

⁵⁴For instance, differences in religious beliefs, abortion perception, gender roles, national priorities, and child values priorities.

⁵⁵Life-situations measured in the WVS and the EVS correlate with gender equality indexes. In short, any increase reported in the gender equality index correlates with a convergence in life-situations between men and women in the surveys. While I show the validity of the WVS for investigating life-situations, other studies show the correlation of the WVS to preferences (Kistler et al. 2017) and for specific values, such as sexism (Brandt 2011).

WVS and the EVS associations started in 1981 to ask participants a large set of questions (henceforth: items) regarding, mostly, values to study their impact on social life. Since then, they have exceeded 500,000 participants, 110 countries, and the survey continues to accumulate observations. The surveys have evolved over the years. They have made major changes 6 times, named waves, but some items are recurrent, such as the following item:

- “Please tell me for each of the following actions whether you think it can always be justified, never be justified, or something in between [...]”⁵⁶

	Never justifiable										Always justifiable		
Homosexuality	1	2	3	4	5	6	7	8	9	10			

There is a large overlap between the WVS and the EVS. I focus on the WVS items as I seek to target comparable subjects across a wide distribution of growth domestic product (GDP) and gender equality indexes (see section 3.5.4 in the supp. material for summary statistics of the WVS and the EVS.).

Classification of the survey items. I develop a classification along two major dimensions: life-situations and values. The former category groups the survey items where the respondents report their life-situations in a broad sense (access to resources, perception of the quality of life, social mobility, etc.), the latter investigates values (including beliefs and preferences). In sum, the life-situations category answers the general question of “how is life?” and the values category answers “what do you think of life?”. Ultimately, I subdivided life-situations into two categories: objective and subjective measures; and values, also into two categories: self-centered value statements and general value statements. The final classification four categories can be summarized as follows:

- **Life situation, objective (LSO).** Questions about the circumstances in which the respondent lives and his habits. The qualifier “objective”

⁵⁶This question is item “V203” from Wave 6 of the WVS.

refers to items that could be easily verified by an external observer. Examples are questions about income, work hours, church going habits, or memberships in organizations.

- **Life situation, subjective (LSS)**. Questions about the circumstances in which the respondent lives, but which are open to subjective interpretation by the respondent. The answers are typically difficult or impossible to verify by an external observer. Examples are: “How satisfied are you with your life?” or “Are you satisfied with your job?”.
- **Self-centered value statements (SCVS)**. This category contains questions about self-reflexive value statements. In other words, the values refer to the respondent directly. Examples are: “Rate your confidence in the government” or “Are you a religious person?”.
- **General value statements (GVS)**. This category refers to value statements which are not directly linked to the respondent, but to society (or humanity) in general. In other words, this category contains items asking “How should the world be?”, or “How should one act?” Examples are: “Is it justifiable to cheat on taxes?” or “The government should reduce environmental pollution”.

While, in many cases, the allocation of the survey items to these four categories is fairly obvious, others leave room for interpretation. For instance, the item: “Rate your confidence in the government” is in the self-centered value statement category but might be interpreted as a general value statement or a life-situation subjective statement. Nevertheless, I deliberately add this item in the SCVS category, because it refers to the respondent. This was a strict criterion for the coders. The reason for this classification is that it is speculative to know what respondents may associate this item with. Particularly, there are reasons to suspect that the framing matters, i.e., whether the government can be trusted or the impact on someone’s life are different questions than whether you trust your government. The ambiguity of the classification bears the risk that the allocation of survey items to categories may be used to influence the outcome of the analysis. In order to bind my hands in this regard, I followed standard practices and hired two research assistants to code the survey items according to the above categorization (see section 3.5.3 in the supp. material for the complete instructions I gave them). Following the procedure of a coordination game proposed by Houser and Xiao (2011), I paid the coders a bonus depending on the consistency in

their categorization.⁵⁷ I randomly selected a set of questions from the overall survey items and took the categorization from the coder 1 and coder 2 of these randomly selected items and then I matched both coders' categorization. The research assistants received a bonus each time their categorizations corresponded. I calculate Cohen's Kappa coefficients to measure the inter-rater agreement. Overall, I obtain 79.6% of agreement ($p = .000$). The raters were not informed about the hypotheses and I ensured that they would not know each other. If the two raters agree on the classification of an item, then I use it as is. For the 20.4 percent of the items, the two coders did not agree. To resolve these cases, I hired a third research assistant to make the call between the two suggestions. I use this final coding for the analyses in the main paper (see section 3.5.10 in the supp. material for robustness checks). The following table shows the classification of the items in the different categories.

	LSO	LSS	SCVS	GVS
Number of items	137	61	335	320

The data types. In the WVS, the items are either ordinal or categorical. The type of data calls for different computations. I have a majority of ordinal items (69.5%)⁵⁸ and some categorical items (30.5%). For the ordinal items, I calculate a simple probabilistic measure of gender differences (Klotz 1966).

In particular, I calculate the probability that a randomly drawn female subject from the same country, the same wave, and for the same item chooses a higher option than a randomly drawn male subject, plus the probability at random that both gender choose the same option (henceforth: $P(\varphi \succ \sigma)$). If y_f (y_m) is a randomly drawn female (male) observation from the same wave, country and item, then $P(\varphi \succ \sigma) = P(y_f > y_m) + \frac{1}{2} P(y_f = y_m)$. If \mathbf{f} is the vector of relative frequencies of all female options in $y_i \in \{0, \dots, 1\}$, and \mathbf{m} for male options, then $P(\varphi \succ \sigma) = \mathbf{fQm}'$, with:

⁵⁷Houser and Xiao (2011) introduced a method to classify natural language messages based on a coordination game. In this game, participants have to classify messages according to some given instructions. The payoff is dependent on the coordination rate between participants. Since participants have incentive to coordinate, their best strategy is to follow the instructions.

⁵⁸This includes those that I rescaled to ordinal and binary items as I use the same computation. See section 3.5.4 in the supp. material.

$$\mathbf{Q} = \begin{pmatrix} \frac{1}{2} & 0 & 0 & \cdot & 0 \\ 1 & \frac{1}{2} & 0 & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ 1 & 1 & 1 & \cdot & \frac{1}{2} \end{pmatrix}$$

I end up with a probability ($P(\varphi \succ \sigma)$) that I use as a coefficient for gender differences. Because identical distribution in the options gives a probability of 0.5 (50%), any deviation from this value indicates that one gender has a higher (lower) occurrence of higher (lower) options than the other. As my main interest is not which gender is stochastically larger in a given item, I use the absolute deviation from 0.5 as the main coefficient (henceforth: $|C| = |P(\varphi \succ \sigma) - 0.5|$).

The computation of this coefficient follows the logic of the Wilcoxon rank-sum test, where monotonicity is assumed. The following artificial example of a 3 points Likert scale shows the computation.

Women\Men	Disagree 0.3	Indifferent 0.5	Agree 0.2
Disagree 0.5	$0.5 \times 0.3 \times 0.5 = 0.075$	$0.5 \times 0.5 \times 0 = 0$	$0.5 \times 0.2 \times 0 = 0$
Indifferent 0.2	$0.2 \times 0.3 \times 1 = 0.06$	$0.2 \times 0.5 \times 0.5 = 0.05$	$0.2 \times 0.2 \times 0 = 0$
Agree 0.3	$0.3 \times 0.3 \times 1 = 0.09$	$0.3 \times 0.5 \times 1 = 0.15$	$0.3 \times 0.2 \times 0.5 = 0.03$

The relative frequencies of each gender for each option of the ordinal responses appear in the first column and the first row. In each cell of the table, I calculate the product of the relative frequencies. As I choose to calculate the probability that women choose a higher or equal option than men, I sum all the cells where the women option is higher than the men option, respectively weighted 1 and the ties weighted 0.5. I get an overall probability of 0.455. Intuitively, the effect size provided by this computation is the probability that a randomly picked female respondent chooses a higher or equal option than another randomly picked male respondent in the respective country and wave for a given item. This example would give me a coefficient of gender difference of 0.045 ($|C|$) for the analysis ($|C| = |0.455 - 0.5|$). A major advantage is that this measure is not sensitive to any monotonic transformation (T. N. Bond and Lang 2019).

For the categorical items, I use Cramer's V (Cramér 1946) computation. The Cramer's V (ϕ_c) is given by the following formula:

$$\phi_c = \sqrt{\frac{\chi^2/n}{\min(k-1, r-1)}}$$

Where ϕ_c is the Cramer's V, χ^2 is the Pearson chi-square statistic, n is the number of observations, k the number of columns, and r the number of rows.

This coefficient has the advantage of being bounded between 0 (no association between the two subgroups) and 1 (perfect association). I choose Cramer's V over the chi2 statistic, because the asymptotic expectations of the χ^2 statistics is proportional to the sample size, and the estimator will be biased by the differences in n . Cramer's V is insensitive to sample sizes as it corrects the estimator by dividing it by a multiple of the sample size (Agresti and Kateri 2011).

The Coefficient $|C|$ from the ordinal data is bound between 0 (no difference) and 0.5 (completely different), I multiply this coefficient by 2. I made the above adjustments to obtain comparable coefficients. Nonetheless, I still control for the type of data in the regression model.⁵⁹ This multiplied coefficient has the advantage of being easy to interpret as it is a measure of gender differences where zero means no difference and one means maximum gender differences. In the paper, I report the gender differences in percentages (henceforth C_n).

For the four dependent variables (LSO, LSS, SCVS, and GVS), I calculate one gender coefficient C_n per survey item, per country, and per wave.

3.2.2 GDP and emancipation indexes

For the indexes, I draw inspiration from Falk and Hermle (2018) and gather the following indexes: the gender inequality index from the UN, the gender equality index from the World Economic Forum, and the gross domestic products from the UN. These are the main indexes in the analyses.

In the second part of the result section, I gather indexes to account for an omitted variable bias. I tackle three possible sources of omitted variables: ecological stress and two sources of cultural differences.

I first use indexes based on the ecological stress hypothesis proposed by Kaiser (2019) taken from Fincher et al. (2008). I specifically gathered the in-

⁵⁹As I use a fixed effect model for each survey item, it controls for the item type.

dex of pathogen prevalence (historical and contemporary), and the Hofstede individualism score. The historical pathogen prevalence indicators are estimates based on atlases published between 1944 and 1961. They are based on nine classes of pathogens detrimental to human reproductive fitness: leishmaniasis, trypanosomes, malaria, schistosomes, filariae, leprosy, dengue, typhus and tuberculosis. Contemporary pathogens indicators are based on the GIDEON database on infectious diseases and consist of the scores of seven classes of pathogens: leishmaniasis, trypanosomes, malaria, schistosomes, filariae, spirochetes, and leprosy. The Individualism score is based on a research project on cultural differences reported in Hofstede et al. (2005). Still in line with the ecological stress hypothesis, I also gather the historical food access score from the Food and Agriculture Organization (FAO). This indicator is the average food supply, in kcal/capita/day, from 1961 to 1984. These indicators are at the country-level. Lastly, I also gather contemporary food access which I match with the corresponding country and the year of the WVS/EVS surveys.

Second, I gather Hofstede's cultural dimensions, which includes the individualism index that I use in the ecological stress analysis. The others are power distance, masculinity, and uncertainty avoidance. I do not use the long-term orientation index, which is part of Hofstede's cultural dimensions, because it has been included only recently in the list of dimensions and is reported for a relatively low number of countries. Power distance defines societies with rather rigid hierarchical structures. Masculinity is typical of more competitive societies that favor material success. Lastly, uncertainty avoidance refers to the high density of rules and laws, such that most outcomes can be predicted. There is also a belief in a one and only truth on the religious and philosophical level. These variables are matched at the country level.

Third, following Welzel (2013), I compute means of different secular indicators from the WVS. Unlike most of the previous indexes of ecological stress factors and cultural dimensions, which are variables at the country level, these latter variables are averages at the country and wave level computed from the respondents of the WVS. I specifically compute means of defiance, disbelief, relativism, and skepticism. Defiance refers to a low level of respect for authority and a general defiance towards the government or the country. Disbelief indicates a low level of religious beliefs and practices. It is also indicative of the overall importance of religion. Relativism refers to a low level of conformism to the norms of the country. Lastly, Skepticism is an inverse trust indicator towards the government, the police, and the army. This last list of indicators should be considered with caution since they come from the same dataset as the one I use to compute the coefficient for gen-

der differences. The observations are not independent and therefore their explanatory power could be biased.

I then use these indexes in the different regressions to show the correlation with the WVS/EVS coefficient. I end up with four different regressions for each one of the categories per index.⁶⁰

As an example, the ordinal item given in the beginning of methods section on page 116 from the WVS on homosexuality gives on average a coefficient of 4.3% gender difference at the 25th percentile of the Gender Equality Index from the United Nations (UN) and 7.2% gender difference at the 75th percentile of the same index. There is a difference of 2.9 between the two percentiles, which is equal to an increase in gender differences for this item of more than 65% between the 25th and the 75th percentile.

3.2.3 Models

In line with Falk and Hermle (2018) and Mac Giolla and Kajonius (2019) who report a divergence of values with more gender equality, I (i) investigate the cross-sectional variation of gender equality on values and life-situations in the Model 1 (M1). Then, I (ii) analyze, in line with Donnelly and Twenge (2017) and Konrad et al. (2000) who report convergence of values with respect to more gender equality, the longitudinal effect of gender equality on values in Model 2 (M2). I also control for the longitudinal effect of these variables on life-situations. Both these investigations follow the work of Connolly et al. (2019), who report both divergence and convergence of values with respect to more gender equality. Finally, I (iii) investigate, following Hofstede et al. (2005), Kaiser (2019), and Welzel (2013), other country-specific variables to account for omitted variable biases in Model 3 (M3).

In a series of regressions, I estimate the effect of indexes of gender equality, economic prosperity (GDP), or other country-level variables on the four measures of gender differences: LSO, LSS, SCVS, and GVS.

In Model 1, I estimate the cross-sectional effect of gender equality or economic prosperity on gender differences in life-situations and values.

$$C_{nict} = \beta_0 + \beta_1 E_{tc} + \beta_2 Z_i + u_{ict} \quad (1)$$

Where $\beta_2 Z_i$ is the fixed effect term for each survey item i . It captures unobserved invariant heterogeneities across the survey items; $\beta_1 E_{tc}$ is the

⁶⁰See section 3.5.7 in the supp. material for the exact sources of these indexes.

index of either gender equality or economic prosperity across different years t and per country c . The dependent variable C_{nict} is the normalized coefficient C_n of gender differences. Each observation is an absolute gender difference per item, per wave, and per country; u is the error term.

Model 2 (M2) is an extension of Model 1, where I estimate the effect of the evolution of gender equality or economic prosperity over time. I include country dummies to control for the cross-sectional variability. Overall, this model analyzes the longitudinal effect of gender equality and economic prosperity.

$$C_{nict} = \beta_0 + \beta_1 E_{tc} + \beta_2 Z_i + \beta_3 D_c + u_{ict} \quad (2)$$

Where $\beta_3 D_c$ are dummies for each country c . It captures unobserved invariant heterogeneities across all countries.

Model 3 (M3) is an extension of Model 1, where I test multiple other country-level variables to test alternative explanations for the cross-sectional variability in gender differences on values.

$$C_{nict} = \beta_0 + \beta_1 E_{tc} + \beta_2 Z_i + \beta_3 E_c + \dots + \beta_n Z_c + u_{ict} \quad (3)$$

Where $\beta_3 E_c$ and $\beta_n Z_c$ are country-level variables of ecological stress, individualism, religiousness, etc. M3 aims to capture the unexplained variability in gender differences across countries. Note that for some explanatory variables we observe variation over time, such as the contemporary access to food supply, Welzel's cultural dimensions (thus: $\beta_3 E_{ct}$).⁶¹

⁶¹As a reminder, the matching procedure used for those variables is mentioned in the footnotes of all the tables of M3.

3.3 Results

The combined dataset from the WVS and the EVS contains 111 countries including very populated countries such as China, India, and the USA, to smaller States, such as Iceland, Malta, and Luxembourg. As measured by the UN, gender equality, on a scale from 0, no gender equality, to 1, perfect gender equality, ranges from 0.17, in Yemen, to 0.95, in the Netherlands and Sweden. Turkey has the highest increase in the dataset with 0.27 points, starting from 0.36 in Wave 2 to 0.64 in Wave 6. While some countries were part of the WVS only for one wave, such as Lebanon, others were present for each wave, such as the USA. Note that, we can only observe variation in the gender equality index for those countries which are present in the WVS in multiple waves.⁶²

As for gender differences in the categories, on average, there is 8.53 percentage points gender differences for objectively measured life-situations and 5.15 percentage points for subjectively measured life-situations. The averages in values are of a similar magnitude with a mean of 5.81 percentage points for self-centered value statements and 5.27 percentage points for general value statements. While the means of gender differences are relatively low, there is a significant amount of variability within a country. For instance, the Netherlands display gender differences, first in life-situations objective items from 0.02 percentage points to more than 66.93 percentage points, and second, from 0.01 percentage points to 22.88 percentage points for different items in general value statements. Note finally, that while unsurprisingly the greatest gender difference for a life-situation objective item is in Yemen with 94.88 percentage points, more surprisingly, the greatest gender difference in a general value statement item is in Jordan with 50.15 percentage points, a country which scores 0.47 on the UN Gender Equality Index (see Section 3.5.1 in the supp. material for a complete list of countries, their respective means, and variability).

My inferential investigation brings mixed results. They are dependent on whether I investigate on a cross-country level (M1) or a within-country level (M2). On a cross-country level, the correlation is in line with the resource

⁶²Although there is no guarantee regarding gender specific differences in response rate, which would lead to a sampling bias, there are reasons to suspect that the problem is of minor importance. Firstly, the WVS committee imposes sampling methods before data collection (see <http://www.worldvaluessurvey.org/WVSCContents.jsp>). Secondly, according to tables in Section 3.5.1 in the supplementary material, there seems to be no systematic gender differences or number of respondents in response rate across countries.

hypothesis, that access to sufficient material resources increases gender differences in values. On the other hand, on a within-country level, I confirm the social role theory, that gender differences are a construct of the cultural environment. Both these theories posit causal relationships. Given these mixed results, I tackle a possible selection bias and an endogeneity issue in the cross-sectional analysis (M3).

Specifically, I first report the cross-sectional variation of gender equality and economic prosperity in life-situations and values (M1). Second, I investigate their longitudinal effect (M2). Third, I investigate a possible selection bias, and finally, three possible sources of omitted variable bias in M1: ecological stress factors and two sources of cultural differences (M3).

Before discussing the effect of gender equality and economic prosperity on values, one might question the validity of these indexes to explain gender differences. To address this concern, I show with the joined measure of life-situations (LSO and LSS), that these indexes capture realities measured by the WVS/EVS well.⁶³ With one exception, all the following tables show the significant ($p < 0.01$) negative correlation of the life-situation with the indexes both on the cross-country level and the within-country level, even when adding additional controls in M3.⁶⁴ In other terms, more gender equality and economic growth unambiguously decrease differences in life-situations between males and females.

Moreover, there are reasons to suspect that changes in gender equality do not impact, at least not to the same extent, all types of values. For instance, items about gender roles in a society should be more influenced by the evolution of gender equality than other ones, such as whether it is justifiable to keep money found. This assumption serves as a falsification test to investigate whether gender equality or economic prosperity are good measures of the evolution of values.

I compute ratios of the evolution of values according to evolution in gender equality. I end up with the items that change the most on the UN Gender Equality Index and the ones that change the least. I find that items, such as “Is it justifiable to keep the money you have found”, “Whether you think

⁶³The systematic significant correlation also accounts for the viability of the WVS to capture gender equality.

⁶⁴The exception is in Table 3.6, where the LSS coefficient is not significant and with the opposite sign. I explain this result by the overlapping of the SCVS and the LSS categories. As reported by the coders, some items of the WVS were sometimes hard to classify in one or the other category. Another reason might be that these items tend to be subject to greater variability as they focus on the respondent. See also Footnote 70

we should put more emphasis on the development of technology”, and “Who would you not like to have as neighbors: immigrants” are the least impacted by changes in the UN Gender Equality Index. On the contrary, items such as “How often do you discuss political matters with friends”, “Is homosexuality justifiable”, and “Who would you not like to have as neighbors: heavy drinkers” are the most impacted. Based on these results, it seems reasonable to assume that the UN Gender Equality Index is a good proxy to investigate changes in values.

3.3.1 Cross-country analyses - M1

In Tables 3.1, 3.2, and 3.3, I report the results of regressions from M1 of the different main indexes on the absolute gender difference. The dependent variable is always the coefficient of gender differences (C_n). I run separate regressions for each of the four categories (LSO, LSS, SCVS, GVS). The independent variables UN and WEF gender equality indexes are bound between 0 and 1. The log GDP per capita ranges from 3.52 to 12.13.

As mentioned, I observe a systematic negative coefficient for both types of life-situations, but a systematic positive coefficient for both types of values. In a cross-country analysis, gender differences in values tend to increase with more gender equality and economic growth. The joined measure of values, SCVS and GVS, is significant ($p < 0.01$) and positively correlated with the indexes in all tables⁶⁵.

In all regressions in the main paper, I include fixed effects for survey items. The main reason for the fixed effects model for the survey items is that some items were only asked in some countries and therefore might bias the estimates.⁶⁶ The other reason is that some items have more options than others and therefore might be subject to greater heterogeneity.⁶⁷ This might drive the overall effect in one direction or another. Therefore, by accounting for item fixed effects I get a more consistent estimate.⁶⁸ I do not

⁶⁵Except again for the WEF index. As explained above, the reason is likely the ambiguity in the allocation of survey items to these two categories, see Footnote 64.

⁶⁶For instance, “Could you please mention any that you would not like to have as neighbors? Shia” was only asked in Guinea (wave 5, item V43₁₀).

⁶⁷For example, some items are on a ten-point scale and others are binary.

⁶⁸One might be concerned with the necessity to include fixed effects or random effects for years, as the computation of the indexes might vary over time, either in a fixed way or randomly. The result from a Durbin-Wu-Hausman test returns $p = 0.005$ which suggests the use of a fixed effects model, therefore in the supp. material, I include year fixed effects (see Table S3.18 and S3.19 in the robustness checks). I deliberately did not include year

Gender Equality Index (UN) - Cross-country

	Dependent variable: Coefficient: % of gender differences			
	LSO	LSS	SCVS	GVS
Gender Equality Index - UN	-4.747** (0.388)	-2.282** (0.422)	2.205** (0.202)	2.066** (0.156)
Constant	11.358** (0.275)	6.482** (0.302)	4.659** (0.142)	3.971** (0.110)
<i>F</i> -test	150.0	29.2	118.8	174.6
<i>p</i>	0.000	0.000	0.000	0.000
<i>R</i> ²	0.009	0.008	0.004	0.004
<i>N</i>	10006	3762	19031	23647

Notes: Fixed effects estimates. This table reports the effect sizes of the Gender Equality Index from the United Nations on the coefficient C_n . Each regression independently tests the effect of the independent variable on the different categories (LSO, LSS, SCVS, GVS). I include fixed effects for each survey item. Standard errors in parentheses. + $p < 0.1$, * $p < 0.05$, ** $p < 0.01$.

Table 3.1: OLS estimates - M1

GDP (UN) - Cross-country

	Dependent variable: Coefficient: % of gender differences			
	LSO	LSS	SCVS	GVS
Log GDP per capita in US dollars	-0.312** (0.041)	-0.156** (0.048)	0.135** (0.021)	0.302** (0.016)
Constant	10.620** (0.369)	6.336** (0.433)	4.869** (0.191)	2.667** (0.146)
<i>F</i> -test	57.9	10.3	40.0	338.0
<i>p</i>	0.000	0.001	0.000	0.000
<i>R</i> ²	0.006	0.001	0.001	0.007
<i>N</i>	13291	4415	24748	30825

Notes: Fixed effects estimates. This table reports the effect sizes of the log GDP from the United Nations on the coefficient C_n . Each regression independently tests the effect of the independent variable on the different categories (LSO, LSS, SCVS, GVS). I include fixed effects for each survey item. Standard errors in parentheses. + $p < 0.1$, * $p < 0.05$, ** $p < 0.01$.

Table 3.2: OLS estimates - M1

Gender Equality Index (WEF) - Cross-country

	Dependent variable: Coefficient: % of gender differences			
	LSO	LSS	SCVS	GVS
Gender Equality Index - WEF	-18.800** (1.391)	-12.583** (1.843)	-0.783 (0.778)	2.057** (0.642)
Constant	20.082** (0.958)	13.772** (1.260)	6.165** (0.535)	3.776** (0.441)
<i>F</i> -test	182.8	46.6	1.0	10.3
<i>p</i>	0.000	0.000	0.314	0.001
<i>R</i> ²	0.020	0.017	0.000	0.000
<i>N</i>	5870	1710	9423	11497

Notes: Fixed effects estimates. This table reports the effect sizes of the Gender Equality Index from the World Economic Forum on the coefficient C_n . Each regression independently tests the effect of the independent variable on the different categories (LSO, LSS, SCVS, GVS). As this index reports values only from 2006 to 2014, I only include observations from the WVS in the corresponding years. I include fixed effects for each survey item. Standard errors in parentheses. + $p < 0.1$, * $p < 0.05$, ** $p < 0.01$.

Table 3.3: OLS estimates - M1

cluster at the country, country and wave, or country and item level in the main paper as the use of clustering depends on different assumptions about the dependencies of the observations. Nevertheless, in Section 3.5.8 of the supplementary material, I investigate the stability of my results with respect to clustering and find that the results are robust to most of the clustering adjustments.

3.3.2 Within-country analyses - M2

The evidence in favor of the resource hypothesis is robust to these specifications, i.e., clustering and year fixed or random effects. However, as soon as I control for country fixed effects, the correlation between the index and the GVS or the SCVS category shifts sign. In Tables 3.4, 3.5, and 3.6, I report the regressions of M2 for the different indexes on the categories and include country fixed effects. The correlations are now negative and significant ($p < 0.05$ or $p < 0.01$) between the indexes and the GVS and the SCVS

fixed effects in the main paper as it certainly captures a substantial part of the real effect of the gender equality indexes or GDP index.

categories.⁶⁹ The life-situation dimension (LSO and LSS) remains negatively correlated with more gender equality or economic prosperity, and in most of the tables, the coefficients are significant at the 1% level.⁷⁰

Gender Equality Index (UN) - Within-country

	Dependent variable: Coefficient: % of gender differences			
	LSO	LSS	SCVS	GVS
Gender Equality Index - UN	-7.352** (1.845)	-8.087** (1.942)	-4.154** (0.868)	-3.490** (0.667)
Constant	15.042** (1.396)	8.966** (1.621)	10.065** (0.662)	6.634** (0.505)
Dummy Country	<i>Yes</i>	<i>Yes</i>	<i>Yes</i>	<i>Yes</i>
<i>F</i> -test	12.4	7.2	20.3	32.5
<i>p</i>	0.000	0.000	0.000	0.000
<i>R</i> ²	0.038	0.122	0.057	0.072
<i>N</i>	10006	3762	19031	23647

Notes: Fixed effects estimates. This table reports the effect sizes of the Gender Equality Index from the United Nations on the coefficient C_n . Each regression independently tests the effect of the independent variable on the different categories (LSO, LSS, SCVS, GVS). I include dummy variables for each country and fixed effects for each survey item. Standard errors in parentheses. + $p < 0.1$, * $p < 0.05$, ** $p < 0.01$.

Table 3.4: OLS estimates - M2

As there is a general upward trend of gender equality and the GDP in each country, I test in Table 3.7 that the variable year is also negatively correlated when I control for country fixed effects. I find that the correlation is again significant ($p < 0.01$) and negative.⁷¹

This suggests that the prior results showed above, in the cross-country analysis, are biased and that the coefficient estimates are inconsistent. The results likely suffer from a problem of endogeneity. Before tackling a possible

⁶⁹The loss of significance for some categories is related to footnote 64.

⁷⁰In Table 3.6, most of the coefficients are insignificant. The main reasons are likely to be the low number of observations in time, since this index only recorded values from 2006 to 2014, and the rather small variability of the values over time.

⁷¹The UN and WEF Gender Equality Index are positively correlated with time. This is also true for the GDP. In other words, every year, on average, countries tend to be more and more gender-equal and also richer. Note, that I do not account here for inflation, which might bias the GDP values. It is worth noticing that when I do not control for the country fixed effects I observe a negative coefficient between the indexes and years in the dataset. This is explained by the fact that in the early days of the WVS (wave 1,2), there was a high prevalence of rather gender-equal countries and rich countries compared to the last waves.

GDP (UN) - Within-country

	Dependent variable: Coefficient: % of gender differences			
	LSO	LSS	SCVS	GVS
Log GDP per capita in US dollars	-0.625** (0.108)	-0.518** (0.133)	-0.634** (0.056)	-0.295** (0.044)
Constant	13.491** (1.014)	8.326** (1.324)	11.496** (0.518)	6.643** (0.403)
Dummy Country	<i>Yes</i>	<i>Yes</i>	<i>Yes</i>	<i>Yes</i>
<i>F</i> -test	13.5	6.5	21.0	31.7
<i>p</i>	0.000	0.000	0.000	0.000
<i>R</i> ²	0.044	0.109	0.049	0.063
<i>N</i>	13291	4415	24748	30825

Notes: Fixed effects estimates. This table reports the effect sizes of the GDP from the United Nations on the coefficient C_n . Each regression independently tests the effect of the independent variable on the different categories (LSO, LSS, SCVS, GVS). I include dummy variables for each country and fixed effects for each survey item. Standard errors in parentheses. + $p < 0.1$, * $p < 0.05$, ** $p < 0.01$.

Table 3.5: OLS estimates - M2

Gender Equality Index (WEF) - Within-country

	Dependent variable: Coefficient: % of gender differences			
	LSO	LSS	SCVS	GVS
Gender Equality Index - WEF	-9.901 (8.183)	12.798 (11.751)	-3.623 (4.090)	-7.905* (3.483)
Constant	12.769* (5.465)	-2.855 (7.928)	7.706** (2.739)	9.841** (2.331)
Dummy Country	<i>Yes</i>	<i>Yes</i>	<i>Yes</i>	<i>Yes</i>
<i>F</i> -test	7.9	5.1	13.1	23.8
<i>p</i>	0.000	0.000	0.000	0.000
<i>R</i> ²	0.043	0.170	0.067	0.101
<i>N</i>	5870	1710	9423	11497

Notes: Fixed effects estimates. This table reports the effect sizes of the Gender Equality Index from the World Economic Forum on the coefficient C_n . Each regression independently tests the effect of the independent variable on the different categories (LSO, LSS, SCVS, GVS). As this index reports values only from 2006 to 2014, I only include observations from the WVS in the corresponding years. I include dummy variables for each country and fixed effects for each survey item. Standard errors in parentheses. + $p < 0.1$, * $p < 0.05$, ** $p < 0.01$.

Table 3.6: OLS estimates - M2

Time trend - Within-country

	Dependent variable: Coefficient: % of gender differences			
	LSO	LSS	SCVS	GVS
Year	-0.063** (0.008)	-0.041** (0.010)	-0.048** (0.004)	-0.024** (0.003)
Constant	134.695** (16.504)	86.349** (20.058)	102.526** (8.484)	51.632** (6.799)
Dummy Country	<i>Yes</i>	<i>Yes</i>	<i>Yes</i>	<i>Yes</i>
<i>F</i> -test	13.2	6.3	20.6	30.5
<i>p</i>	0.000	0.000	0.000	0.000
<i>R</i> ²	0.042	0.102	0.049	0.061
<i>N</i>	14320	4769	26686	33239

Notes: Fixed effects estimates. This table reports the effect sizes of each year on the coefficient C_n . Each regression independently tests the effect of the independent variable on the different categories (LSO, LSS, SCVS, GVS). I include dummy variables for each country and fixed effects for each survey item. Standard errors in parentheses. + $p < 0.1$, * $p < 0.05$, ** $p < 0.01$.

Table 3.7: OLS estimates - M2

omitted variable bias, I investigate the sampling bias by restricting the data.

The extreme case would be that I have in the data a vast majority of single observations from countries that display a very high positive correlation of gender equality/GDP and gender differences in values and only a minority of countries with multiple observations over time and with a wide range of values in the indexes that have a negative correlation. This would be a good example of a Simpson's paradox (Blyth 1972). This concern is fueled by the great difference in the participation rate of countries. For instance, the USA participated in all waves (six waves), while Tanzania participated only once.

3.3.3 Robustness of paradox to sub-samplings

To investigate the sampling bias, I restrict the data to a subsample of countries that have observations in the first wave (1981–1984) and the last wave (2010–2014) of the WVS. The cross-country analyses (M1) still show divergence of gender differences in values and the within-country analysis (M2) convergence (see Table S3.13 in the supp. material for the regression).

Figure 3.1 shows the relationship between the absolute gender differences in LSO and GVS categories and gender equality as measured by the UN Gender Equality Index. The figure contains only countries that were present in Wave 1 and 6. While the figure always shows a decrease in gender differences in life-situation (LSO) with an increase in gender equality or passing time, changes in gender differences in values are not systematic. The USA, Australia, Japan, Sweden, and Germany tend to support divergence, while Spain, Argentina, and very slightly the Netherlands support convergence.

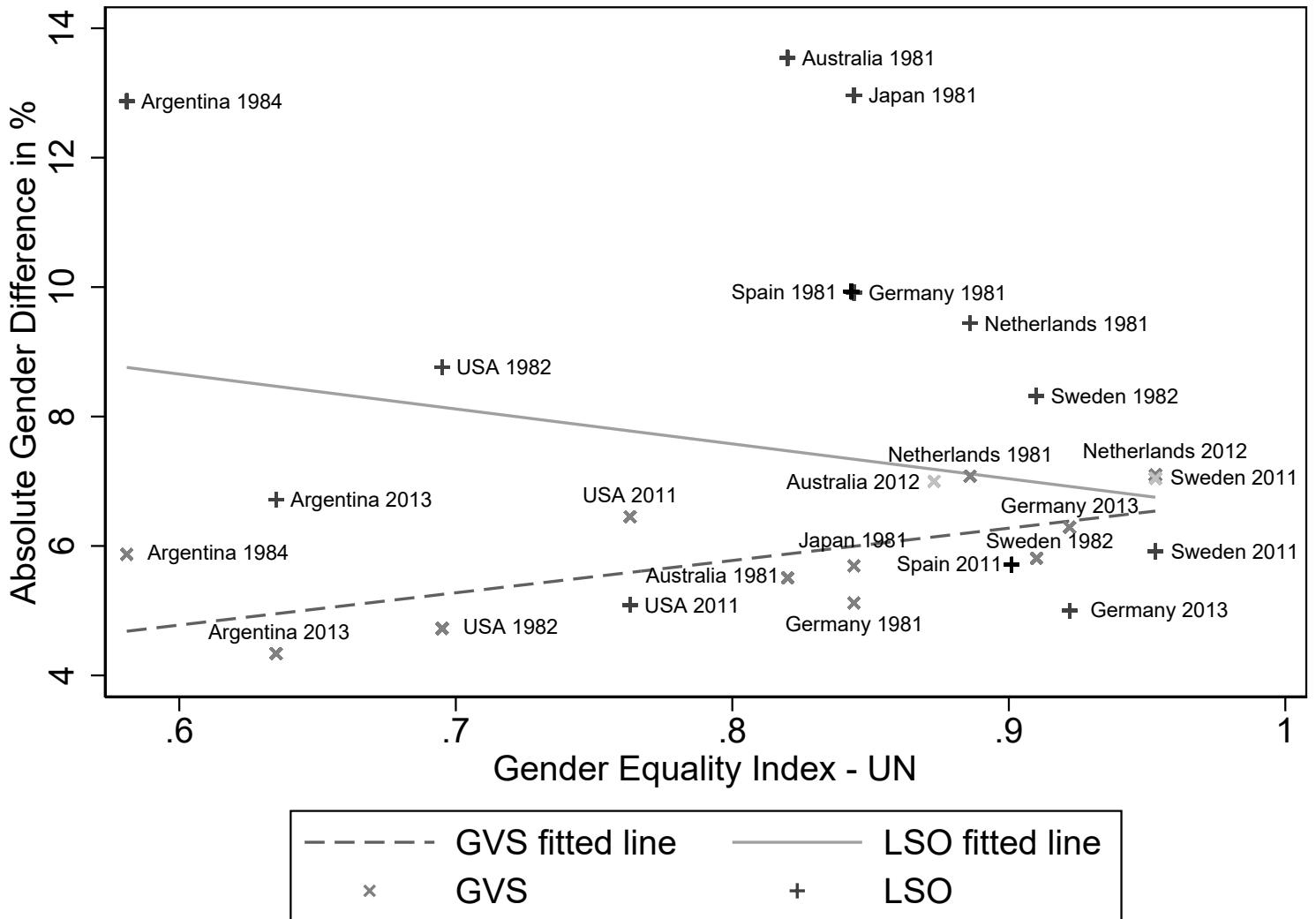
The robustness of the results - showing both divergence (M1) and convergence (M2) of gender differences in values - with this sub-sampling suggests that M1 is still biased by some omitted variables that the country fixed effect specification captures (M2). Another piece of evidence in the same vein is provided by restricting the sample to culturally similar countries. I restrict the data to specific areas provided by the UN⁷² and find that correlation coefficients (R) between the indexes and gender differences C_n tend to be larger. For instance, the correlation between the UN Gender Equality Index and the coefficient for gender differences is 0.13 if I restrict the sample to countries in Western Europe⁷³ and 0.14 for countries in North Africa⁷⁴. These

⁷²<https://unstats.un.org/unsd/methodology/m49/>

⁷³Western Europe is composed of Belgium, Austria, France, Germany, Liechtenstein, Luxembourg, the Netherlands, and Switzerland in the WVS/EVS dataset.

⁷⁴Northern Africa is composed of Algeria, Egypt, Libya, Morocco, and Tunisia in the

Wave 1 and Wave 6 - LSO and GVS



The scatter plot displays the absolute gender difference in percent on the y-axis and the values from the Gender Equality Index from the UN on the x-axis. The graph dots are means of the absolute gender differences per country and per year.

Figure 3.1: Gender equality and gender differences for selected countries

values are larger than correlation among all countries (0.07). This suggests that countries close to each other geographically share some similarities, such as cultural or historical origins, which otherwise likely increase noise in M1. Tables 3.8 and 3.9 report the regression outputs when the dataset contains only western European countries. While the loss of significance for some categories is likely attributed to the small variance and the low number of observations, the main story of convergence and divergence of values holds.

Gender Equality Index (UN) - Western Europe countries - Cross-country

	Dependent variable: Coefficient: % of gender differences			
	LSO	LSS	SCVS	GVS
Gender Equality Index - UN	0.599 (3.418)	1.365 (3.902)	-0.641 (1.989)	9.383** (1.701)
Constant	6.851* (2.966)	2.909 (3.354)	6.823** (1.724)	-2.404 (1.472)
<i>F</i> -test	0.0	0.1	0.1	30.4
<i>p</i>	0.861	0.727	0.747	0.000
<i>R</i> ²	0.000	0.009	0.000	0.017
<i>N</i>	975	428	1861	2273

Notes: Fixed effects estimates. This table reports the effect sizes of the Gender Equality Index from the United Nations on the coefficient C_n . This regression includes only countries from Western Europe. Each regression independently tests the effect of the independent variable on the different categories (LSO, LSS, SCVS, GVS). I include fixed effects for each survey item. Standard errors in parentheses. ⁺ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$.

Table 3.8: OLS estimates - M1

Interestingly, the coefficients in the regression models of GVS in the cross-country analysis and of the SCVS in the within-country analysis are substantially larger when I restrict the dataset to western European countries compared to the overall dataset. In line with the previously mentioned correlation coefficients, it suggests that changes in the UN Gender Equality Index produce more gender differences in values in culturally similar countries. Furthermore, although more coefficients are insignificant, the paradox of convergence and divergence of values is robust to this sub-sampling. I find similar results using other restrictions to culturally similar countries, such as North Africa, South America, and North America. Overall, these results

WVS dataset.

Gender Equality Index (UN) - Western Europe countries - Within-country

	Dependent variable: Coefficient: % of gender differences			
	LSO	LSS	SCVS	GVS
Gender Equality Index - UN	-35.835** (6.823)	-24.985** (8.813)	-15.749** (3.583)	1.086 (3.168)
Constant	38.646** (5.788)	24.765** (7.317)	19.842** (3.016)	4.804 ⁺ (2.659)
Dummy Country	<i>Yes</i>	<i>Yes</i>	<i>Yes</i>	<i>Yes</i>
<i>F</i> -test	7.7	3.0	6.7	7.4
<i>p</i>	0.000	0.004	0.000	0.000
<i>R</i> ²	0.005	0.009	0.010	0.017
<i>N</i>	975	428	1861	2273

Notes: Fixed effects estimates. This table reports the effect sizes of the Gender Equality Index from the United Nations on the coefficient C_n . This regression includes only countries from Western Europe. Each regression independently tests the effect of the independent variable on the different categories (LSO, LSS, SCVS, GVS). I include dummy variables for each country and fixed effects for each survey item. Standard errors in parentheses. ⁺ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$.

Table 3.9: OLS estimates - M2

show that cultural differences matter but that restricting the dataset does not allow to eliminate the paradox. In short, the issue of endogeneity in the models is still a concern. Therefore, in M3 I investigate three potential sources of this endogeneity: ecological stress and two different sources of cultural differences.

3.3.4 Investigation of omitted variables - M3

First, following the hypothesis and evidence of Kaiser (2019), I test whether ecological stress is the main predictor of gender differences. The causal relationship hypothesized by Kaiser is as follows: countries that face a high prevalence of pathogens tend to be more collectivist⁷⁵ and display a lower degree of gender differences. This causal relationship is in line with the resource hypothesis. As collectivism leaves less scope for deviant behaviors,

⁷⁵Collectivism refers to societies that display a high respect for traditions and a high degree of xenophobia. Both are means of protection. For instance, food traditions, such as the use of tannin, reduce the spread of pathogens. Collectivism is opposed to individualism where deviant behaviors are less reprimanded and the society is more open to others.

there are less opportunities for self-expression. Moreover, in more individualist countries, subjects tend to self-stereotype themselves through cross-gender comparisons, while in collectivist countries the comparison is made with the same gender. These social comparison processes increase gender differences (Guimond, Branscombe, et al. 2007; Guimond, Chatard, et al. 2006). I test if ecological stress predicts the differences in values.

In Table 3.10, I report the effect of ecological stress factors on absolute gender differences in values (M3). While, historical pathogen prevalence ranges from -1.29 in Canada, to 1.28 in Nigeria or Ghana, the contemporary measure ranges from 24 , in Albania and Germany to 46 in Burkina Faso. The historical values are z -scores and the contemporary ones are the sums of seven classes of pathogens. As for individualism, values range from 6 in Guatemala, to 91 in the USA, and values for the food supply from 1308 kcal/capita/day in Burkina Faso, to 3809 kcal/capita/day in the USA. The range is similar for the historical food supply. As an illustration, the most individualist country, as measured by this index, is the USA. This country had a relatively low prevalence of historical pathogens ($-.86$) and a relatively high historical food supply (3192.09). Nevertheless, on average, this country displays only slightly more gender differences, with 5.47% for GVS than the overall average at 5.27% .

In the estimates based on the UN Gender Equality Index, I find that even with these additional control variables, more gender equality is still associated with divergence of values between males and females. Ecological stress controls substantially increase the R^2 of the Model. The effects of the different ecological control variables are mixed in this multilinear regression. While a higher prevalence of historical pathogens tends to increase divergence, contemporary pathogens tend to decrease divergence. The effect of food supply is also rather mixed. However, taken separately, all these ecological stress variables are in line with the ecological stress hypothesis and are highly significant ($p < 0.01$, see section 3.5.9 in the supp.material for the independent regressions). The ecological variables are dependent and correlated, which might explain why I do not observe significant results for all of these variables in Table 3.10, that is societies tend to be more individualistic because of a lower prevalence of pathogens. This causality implies that that individualism is not independent of pathogen prevalence.

The robustness of divergence of values with respect to more gender equality including these additional variables suggests that the within-country level analysis captures more cross-country variability than the ecological stress factors. This suggests that other omitted variables explain the difference in values or that this is the true effect of gender equality, which points towards divergence of differences.

Ecological stress indicators - Cross-country

	Dependent variable: Coefficient: % of gender differences			
	LSO	LSS	SCVS	GVS
Gender Equality Index - UN	-13.662** (1.048)	-2.112+ (1.188)	0.170 (0.579)	2.316** (0.437)
Pathogen Prevalence (Historical)	1.574** (0.244)	0.562* (0.246)	0.651** (0.138)	0.173+ (0.102)
Pathogen Prevalence (Contemp.)	-0.381** (0.035)	-0.068+ (0.038)	-0.099** (0.019)	-0.027+ (0.014)
Individual. (Hofstede)	0.013+ (0.007)	-0.007 (0.007)	0.011** (0.004)	0.010** (0.003)
Food Supply	0.002** (0.000)	-0.000 (0.000)	-0.000 (0.000)	0.000 (0.000)
Historical Food Supply	-0.002** (0.000)	0.001* (0.000)	0.001** (0.000)	0.000 (0.000)
Constant	28.404** (1.623)	6.686** (1.809)	6.877** (0.889)	3.323** (0.676)
<i>F</i> -test	39.9	2.7	32.6	50.9
<i>p</i>	0.000	0.014	0.000	0.000
<i>R</i> ²	0.026	0.010	0.019	0.021
<i>N</i>	3073	1204	6168	7640

Notes: Fixed effects estimates. This table reports the effect sizes of the Gender Equality Index from the United Nations on the coefficient C_n . Each regression independently tests the effect of the independent variables on the different categories (LSO, LSS, SCVS, GVS). I include fixed effects for each survey item. Both indicators of pathogens and the historical food supply indicator are matched at the country level. The remaining control variables are matched with the corresponding country and wave. Standard errors in parentheses. + $p < 0.1$, * $p < 0.05$, ** $p < 0.01$.

Table 3.10: OLS estimates - M3

Second, I investigate cultural differences. I use Geert Hofstede's cultural dimensions to control for cultural differences that might explain gender differences in values (M3). While, there is some evidence for gender differences in competitive or cooperative behaviors (Croson and Gneezy 2009; Knight and Chao 1989; Niederle and Vesterlund 2007; Niederle and Vesterlund 2011; Peshkovskaya et al. 2018), these cultural dimensions are overall proxies of competition and cooperation across cultures (Hofstede et al. 2005). Thus, the true effect of more gender equality could be culturally influenced, such as, for instance, rising competition between men and women for the same job or, on the contrary, leading to more overall gender cooperation. I use these indexes to investigate these possible channels.

In Table 3.11, I report the OLS estimates. In terms of variation, for instance, power distance ranges from 11 in Australia, to 104 in Malaysia. All the other coefficients of this cultural source have similar variations. As an illustration, Japan has the highest score in masculinity (95), a relatively high score in uncertainty avoidance (92), and on average 6.04 gender differences in the GVS category.

Power distance and masculinity tend to decrease gender differences in values. In other words, societies that tend to have well defined hierarchical structures and that favor personal and material success tend to have lower gender differences in values. Both of these indicators would suggest that competition hampers the manifestation of gender differences. On the contrary, individualism and uncertainty avoidance are associated with more gender differences. Individualism is an indicator of societies that allow general differences between individuals, which might enhance gender differences. Uncertainty avoidance, because it is an indicator of societies that have a high degree of security, might lead to less competition between both genders. There might be a lower possibility for high competition if there is general security. I also assume that stronger rules settle more situations where competition would otherwise prevail. All these point estimates support the hypothesis that cooperation or competition explain gender differences in values.

However, in contrast to the previous hypothesis that ecological stress factors predict gender differences in values, the current hypothesis is that gender equality either raises competition or cooperation between genders. In other words, more gender equality should interact with indicators of competition and cooperation. In Table 3.12, I report the regression outputs of interactions between the UN Gender Equality Index and Geert Hofstede's cultural dimensions and find that gender equality tends to amplify competition or cooperation which would then explain gender differences in values. In other words, I find support for different channels with respect to more gender equality. On one hand, more gender equality leads in some cases

to more cooperation between both genders, and in other cases, it increases competition between both genders. The explanation of why some societies display an increase in cooperation or competition is still a matter of omitted variables.

In the estimates based on the UN Gender Equality Index, I find that this index is robust to these additional control variables in explaining gender differences in values, but the coefficients are lower, suggesting that some of the variation previously explained by gender equality is henceforth explained by cultural differences. Note finally, that the inclusion of these additional control variables increases, again, the R^2 .

Geert Hofstede's cultural dimensions - Cross-country

	Dependent variable: Coefficient: % of gender differences			
	LSO	LSS	SCVS	GVS
Gender Equality Index - UN	-4.687** (0.597)	-2.095** (0.638)	1.440** (0.323)	1.321** (0.253)
Power Distance	-0.028** (0.005)	-0.021** (0.006)	-0.025** (0.003)	-0.029** (0.002)
Individual. (Hofstede)	-0.003 (0.005)	-0.004 (0.005)	0.011** (0.003)	0.005* (0.002)
Masculinity	-0.004 (0.004)	-0.015** (0.004)	-0.006** (0.002)	-0.010** (0.002)
Uncertainty Avoidance	0.011** (0.003)	0.008* (0.004)	0.018** (0.002)	0.006** (0.002)
Constant	12.601** (0.686)	7.495** (0.714)	4.880** (0.378)	5.787** (0.294)
F -test	17.5	6.9	84.0	141.4
p	0.000	0.000	0.000	0.000
R^2	0.008	0.011	0.023	0.030
N	5958	2341	11615	14393

Notes: Fixed effects estimates. This table reports the effect sizes of the Gender Equality Index from the United Nations on the coefficient C_n . Each regression independently tests the effect of the independent variables on the different categories (LSO, LSS, SCVS, GVS). I include fixed effects for each survey item. All the control variables are matched at country level, except for the index of gender equality from the UN. Standard errors in parentheses. + $p < 0.1$, * $p < 0.05$, ** $p < 0.01$.

Table 3.11: OLS estimates - M3

Third, I investigate other cultural differences taken from Welzel (2013). In Table 3.13, I report the OLS estimates. While conformism is associated with the presence of stronger norms (Asch 1951) and that collectivist societies, where deviant behaviors are less tolerated, are also associated with

Geert Hofstede's cultural dimensions - Cross-country - Interactions

	Dependent variable: Coefficient: % of gender differences			
	LSO	LSS	SCVS	GVS
Gender Equality Index - UN	-3.562** (0.853)	-0.468 (0.891)	1.546** (0.477)	2.525** (0.372)
Power Distance x GE-UN	-0.039** (0.007)	-0.025** (0.008)	-0.037** (0.004)	-0.038** (0.003)
Individual. x GE-UN	-0.000 (0.006)	-0.001 (0.007)	0.014** (0.004)	0.014** (0.003)
Masculinity x GE-UN	-0.006 (0.004)	-0.016** (0.005)	-0.008** (0.002)	-0.012** (0.002)
Uncertainty Avoidance x GE-UN	0.020** (0.005)	0.011* (0.005)	0.028** (0.003)	0.012** (0.002)
Constant	11.360** (0.412)	5.836** (0.435)	4.649** (0.217)	4.224** (0.171)
<i>F</i> -test	19.1	5.7	88.8	151.9
<i>p</i>	0.000	0.000	0.000	0.000
<i>R</i> ²	0.009	0.009	0.024	0.032
<i>N</i>	5958	2341	11615	14393

Notes: Fixed effects estimates. This table reports the effect sizes of the Gender Equality Index from the United Nations on the coefficient C_n . GE-UN stands for the UN Gender Equality Index. This table includes control variables from Geert Hofstede's cultural dimensions and the interactions with the Gender Equality Index. Each regression independently tests the effect of the independent variables on the different categories (LSO, LSS, SCVS, GVS). I include fixed effects for each item of the WVS. All the control variables are matched at the country level, except for the index of gender equality from the UN. Standard errors in parentheses. ⁺ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$.

Table 3.12: OLS estimates - M3

conformism (R. Bond and Smith 1996), Welzel (2013) developed indicators of general emancipation to such norms. Thus, less pressure from the norms should be associated with more gender differences.

In Table 3.13, disbelief and defiance, which are indicative of a low level of religious belief and low respect for authority are associated with more gender differences in values. These indicators support the hypothesis that less pressure from the norms leads to more gender differences in values. However, skepticism, which is indicative of a general distrust, is ambiguously associated with more gender differences in self-centered values (SCVS) but with less gender differences in general values (GVS). Furthermore, albeit the coefficients are above significance levels, relativism, which is indicative of a lower level of conformism, is also associated with less gender differences in values. According to Welzel (2013), all of these variables are indicators of emancipation and, even though the effects of these indicators are ambiguous, they mostly suggest that emancipation is associated with more gender differences in values. However, these indicators should be taken with caution as they are not independent of the gender coefficient.

As for the effect of gender equality, the UN index is still significant ($p < 0.01$) and positively associated with gender differences in values in the GVS category. The loss of significance in LSS and SCVS is likely a similar issue of ambiguity in the classification (see Footnote 64).

3.3.5 General remarks

Overall the results show both convergence and divergence of gender differences in values with respect to gender equality. These correlations depend on whether I use country fixed effects or not. Moreover, the divergence of values with respect to more gender equality and economic prosperity is robust to alternative specifications. I end up not making a causal claim but suggest that further investigation is needed to tackle the issue of possible omitted variables.

This research is an empirical study with the limitations that arise in such cases. The analyses cannot reveal the complexity of the reality nor all the heterogeneity in values. It remains an investigation which approximates the underlying mechanism in the evolution of values.

As another limitation to this research, one should note that the evolution of within-country gender equality is often fairly small. For instance, the average evolution of the UN Gender Equality Index, in all the countries, is only of 6.14 percentage points. I also compute the range ratio of each index. The range of the UN Gender Equality Index is 2.89 times greater in the cross-country than in the within-country analysis. As for the two other indexes,

Welzel Cultural dimensions - Cross-country

	Dependent variable: Coefficient: % of gender differences			
	LSO	LSS	SCVS	GVS
Gender Equality Index - UN	-4.148** (1.461)	-2.203 (1.474)	1.019 (0.686)	2.433** (0.541)
Relativism	-6.880** (1.971)	-5.366* (2.125)	-1.095 (1.028)	-1.298+ (0.782)
Defiance	2.236 (1.972)	-3.274 (2.116)	1.516 (0.954)	2.178** (0.725)
Disbelief	-1.972* (1.000)	2.433* (1.106)	2.624** (0.502)	0.548 (0.381)
Skepticism	-6.050** (1.783)	-0.829 (1.942)	1.842* (0.906)	-2.335** (0.695)
Constant	18.038** (1.173)	10.274** (1.299)	3.806** (0.570)	4.641** (0.433)
<i>F</i> -test	15.6	4.9	26.7	43.9
<i>p</i>	0.000	0.000	0.000	0.000
<i>R</i> ²	0.014	0.026	0.016	0.023
<i>N</i>	2918	1015	5697	7069

Notes: Fixed effects estimates. This table reports the effect sizes of the Gender Equality Index from the United Nations on the coefficient C_n . Each regression independently tests the effect of the independent variables on the different categories (LSO, LSS, SCVS, GVS). I include fixed effects for each survey item. The index of gender equality from the UN is matched with the corresponding country and year of the survey. The remaining control variables are matched with the corresponding country and wave of the survey. Standard errors in parentheses. + $p < 0.1$, * $p < 0.05$, ** $p < 0.01$.

Table 3.13: OLS estimates - M3

the range ratios are 1.69 for the UN log GDP and 3.30 for the WEF Gender Equality Index. Due to this rather small variability of values in the indexes in the within-country analyses, it is also possible that the within-country results are not based on solid grounds.

3.4 Conclusion

I end up with a puzzling correlation between the effect of gender equality and values. I show both convergence and divergence of gender differences in values with respect to more equal social and economic life-situations.

Even if the correlation, suggesting divergence, in the cross country analysis is very convincing and robust to many alternative estimation methods, it still fails to make it through when I control for the country fixed effects. The question arises of whether gender equality indexes, such as those I use in this study, are able to capture a perceived reality or if the correlation is spurious. Nonetheless, this concern is partially rejected as the negative correlations between the life-situation dimension in the WVS/EVS and the indexes are almost always unambiguous and very significant.

The results finally support no particular causal relationship on the controversial link between values and social/economic life-situations. Nonetheless, I note that more variation in the within-country gender equality index would help to make a causal claim on gender differences in values, but finding an instrumental variable seems more promising since there is likely a latency between the evolution of gender equality and its possible impact on gender differences in values.

My contributions are the use of a very broad set of values, an analysis of longitudinal data over more than 30 years and I confirm, with the categorization of the life situations, that the WVS/EVS are attractive datasets for the elicitation of social differences and inequalities. I finally contribute to the growing literature that investigates gender differences in values and preferences (Connolly et al. 2019; Costa et al. 2001; Falk and Hermlle 2018; Kaiser 2019; Schmitt et al. 2008; Schwartz and Rubel 2005), but conclude with an unsolved puzzle of whether gender differences in values increase or decrease with respect to more gender equality or economic prosperity.

References

- Agresti, A., & Kateri, M. (2011). *Categorical data analysis*. Springer.
- Allison, R. A., & Foster, J. E. (2004). Measuring health inequality using qualitative data. *Journal of Health Economics*, *23*(3), 505–524.
- Almås, I., Cappelen, A. W., Salvanes, K. G., Sørensen, E. Ø., & Tungodden, B. (2016). Willingness to compete: Family matters. *Management Science*, *62*(8), 2149–2162.
- Asch, S. E. (1951). Effects of group pressure upon the modification and distortion of judgments. *Organizational Influence Processes*, 295–303.
- Blyth, C. R. (1972). On Simpson’s paradox and the sure-thing principle. *Journal of the American Statistical Association*, *67*(338), 364–366.
- Bond, R., & Smith, P. B. (1996). Culture and conformity: A meta-analysis of studies using Asch’s (1952b, 1956) line judgment task. *Psychological Bulletin*, *119*(1), 111–137.
- Bond, T. N., & Lang, K. (2019). The sad truth about happiness scales. *Journal of Political Economy*, *127*(4), 1629–1640.
- Brandt, M. J. (2011). Sexism and gender inequality across 57 societies. *Psychological Science*, *22*(11), 1413–1418.
- Buss, D. (1995). Psychological sex-differences - Origins through sexual selection. *American Psychologist*, *50*(3), 164–168.
- Chen, L.-Y., Oparina, E., Powdthavee, N., & Srisuma, S. (2019). Have econometric analyses of happiness data been futile? A simple truth about happiness scales. *IZA Discussion Paper*.
- Connolly, F. F., Goossen, M., & Hjerm, M. (2019). Does gender equality cause gender differences in values? Reassessing the gender-equality-personality paradox. *Sex Roles*, 1–13.
- Costa, P. T., Terracciano, A., & McCrae, R. R. (2001). Gender differences in personality traits across cultures: Robust and surprising findings. *Journal of Personality and Social Psychology*, *81*(2), 322–331.
- Cramér, H. (1946). *Mathematical methods of statistics*. Princeton University Press.
- Crosby, R., & Gneezy, U. (2009). Gender differences in preferences. *Journal of Economic Literature*, *47*(2), 448–474.
- Donnelly, K., & Twenge, J. M. (2017). Masculine and feminine traits on the Bem Sex-Role Inventory, 1993–2012: A cross-temporal meta-analysis. *Sex Roles*, *76*(9), 556–565.
- Duflo, E. (2012). Womens empowerment and economic development. *Journal of Economic Literature*, *50*(4), 1051–79.

- Eagly, A. H., & Wood, W. (1999). The origins of sex differences in human behavior - Evolved dispositions versus social roles. *American Psychologist*, *54*(6), 408–423.
- Falk, A., & Hermle, J. (2018). Relationship of gender differences in preferences to economic development and gender equality. *Science*, *362*(6412).
- Fincher, C. L., Thornhill, R., Murray, D. R., & Schaller, M. (2008). Pathogen prevalence predicts human cross-cultural variability in individualism/collectivism. *Proceedings of the Royal Society B: Biological Sciences*, *275*(1640), 1279–1285.
- Geary, D. C. (2010). Male, female: The evolution of human sex differences. *American Psychological Association*, 397.
- Gneezy, U., Leonard, K. L., & List, J. A. (2009). Gender differences in competition: Evidence from a matrilineal and a patriarchal society. *Econometrica*, *77*(5), 1637–1664.
- Guimond, S., Branscombe, N. R., Brunot, S., Buunk, A. P., Chatard, A., Désert, M., Garcia, D. M., Haque, S., Martinot, D., & Yzerbyt, V. (2007). Culture, gender, and the self: Variations and impact of social comparison processes. *Journal of Personality and Social Psychology*, *92*(6), 1118–1134.
- Guimond, S., Chatard, A., Martinot, D., Crisp, R. J., & Redersdorff, S. (2006). Social comparison, self-stereotyping, and gender differences in self-construals. *Journal of Personality and Social Psychology*, *90*(2), 221–242.
- Haushofer, J., & Fehr, E. (2014). On the psychology of poverty. *Science*, *344*(6186), 862–867.
- Hofstede, G. H., Hofstede, G. J., & Minkov, M. (2005). *Cultures and organizations: Software of the mind* (Vol. 2). New York: McGrawHill.
- Houser, D., & Xiao, E. (2011). Classification of natural language messages using a coordination game. *Experimental Economics*, *14*(1), 1–14.
- Inglehart, R. (2015). *The silent revolution: Changing values and political styles among western publics*. Princeton University Press.
- Inglehart, R., & Norris, P. (2003). Gender equality & cultural change around the world. *Cambridge University Press*, 12.
- Kaiser, T. (2019). Nature and evoked culture: Sex differences in personality are uniquely correlated with ecological stress. *Personality and Individual Differences*, *148*, 67–72.
- Kistler, D., Thöni, C., & Welzel, C. (2017). Survey response and observed behavior: Emancipative and secular values predict prosocial behaviors. *Journal of Cross-Cultural Psychology*, *48*(4), 461–489.

- Klotz, J. H. (1966). The Wilcoxon, ties, and the computer. *Journal of the American Statistical Association*, 61(315), 772–787.
- Knight, G. P., & Chao, C.-C. (1989). Gender differences in the cooperative, competitive, and individualistic social values of children. *Motivation and Emotion*, 13(2), 125–141.
- Konrad, A. M., Ritchie Jr, J. E., Lieb, P., & Corrigan, E. (2000). Sex differences and similarities in job attribute preferences: A meta-analysis. *Psychological Bulletin*, 126(4), 593–641.
- Mac Giolla, E., & Kajonius, P. J. (2019). Sex differences in personality are larger in gender equal countries: Replicating and extending a surprising finding. *International Journal of Psychology*, 54(6), 705–711.
- McLoyd, V. C. (1998). Socioeconomic disadvantage and child development. *American Psychologist*, 53(2), 185–204.
- Niederle, M., & Vesterlund, L. (2007). Do women shy away from competition? Do men compete too much? *The Quarterly Journal of Economics*, 122(3), 1067–1101.
- Niederle, M., & Vesterlund, L. (2011). Gender and competition. *Annual Review of Economics*, 3(1), 601–630.
- Peshkovskaya, A., Babkina, T., & Myagkov, M. (2018). Social context reveals gender differences in cooperative behavior. *Journal of Bioeconomics*, 20(2), 213–225.
- Schmitt, D. P., Realo, A., Voracek, M., & Allik, J. (2008). Why cant a man be more like a woman? Sex differences in big five personality traits across 55 cultures. *Sex Differences*, 16.
- Schroeder, C., & Yitzhaki, S. (2017). Revisiting the evidence for cardinal treatment of ordinal variables. *European Economic Review*, 92, 337–358.
- Schwartz, S. H., & Rubel, T. (2005). Sex differences in value priorities: Cross-cultural and multimethod studies. *Journal of Personality and Social Psychology*, 89(6), 1010.
- Tanaka, T., Camerer, C. F., & Nguyen, Q. (2010). Risk and time preferences: Experimental and household survey data from Vietnam. *American Economic Review*, 100(1), 557–571.
- Welzel, C. (2013). *Freedom rising: Human empowerment and the quest for emancipation*. Cambridge University Press.
- Wood, W., & Eagly, A. H. (2012). Biosocial construction of sex differences and similarities in behavior. *Advances in experimental social psychology* (pp. 55–123). Elsevier.

3.5 Supplementary material

The supplementary material is structured as follows: (i) tables of descriptive statistics, (ii) the instructions I gave to the research assistants, (iii) the computation of the coefficient from the WVS and the EVS, (iv) the classification of the items, (v) methodological concerns with the computation of ordinal reported items, (vi) the different indexes, (vii) the different clustering, (viii) additional analyses, (ix) and robustness checks.

3.5.1 Descriptive statistics

In this section, I include descriptive tables of the dataset from the WVS/EVS. Precisely, I report the means for the main indexes for each country: the UN log GDP in million US\$ (log GDP), the UN Gender Equality Index (GE-UN), and the one from the WEF (GE-WEF). Then, I report the means of gender differences for each category (LSO, LSS, SCVS, GVS). I also report the average percentage of female respondents, the maximum variability for each index (Log GDP, GE-UN, GE-WEF), and the categories (LSO, LSS, SCVS, GVS). The maximum variability is given by taking the maximum value of the variable in the dataset minus the minimum value observed in the same category/index. Finally, I report the average number of respondents/observations per wave and the number of times a country was present in the dataset (Waves).

Countries, means, and their variability

Country	Mean US\$			Means in %						Variability in US\$						Maximum variability in %			Number	
	log GDP	GE-UN	GE-WEF	LSO	LSS	SCVS	GVS	female	Log GDP	GE-UN	GE-WEF	LSO	LSS	SCVS	GVS	Av. Obs/Wave	Waves	Number		
																		GE-UN	GE-WEF	LSO
Albania	745	62.0	65.9	8.41	4.90	6.63	4.31	50.4	165	0.00	0.000	51.0	12.9	34.5	25.9	1107	3			
Algeria	808	47.4	63.2	9.50	6.23	5.78	5.90	48.6	112	19.00	0.000	62.2	17.1	32.6	39.7	1127	2			
Andorra	1063			8.31	6.45	6.51	6.00	49.7	0			33.5	11.1	23.3	18.4	981	1			
Argentina	892	60.0	70.2	6.57	3.62	5.77	4.85	52.8	147	5.40	3.660	52.5	17.1	25.1	24.5	1005	6			
Armenia	767	59.8	66.6	6.16	6.14	6.14	4.76	58.8	204	13.50	0.230	44.6	27.8	24.9	26.5	1440	3			
Australia	1046	85.1	72.9	8.12	4.87	6.63	6.51	53.7	163	6.30	0.000	65.5	12.8	22.4	34.6	1518	4			
Austria	1029	82.9	71.5	7.41	3.96	6.39	5.36	53.3	87	3.60	0.000	52.6	11.7	29.0	18.6	1414	3			
Azerbaijan	766	68.9	65.8	8.58	7.41	6.25	4.63	50.2	267	0.00	0.000	47.8	30.4	32.2	26.1	1380	2			
Bangladesh	582	30.6	58.2	7.18	5.55	5.96	4.29	43.7	5	3.00	0.000	80.7	11.8	38.2	22.0	1442	2			
Belarus	790	85.4	71.7	6.80	3.44	5.74	6.48	56.3	181	4.00	0.000	47.3	18.7	32.6	29.7	1369	5			
Belgium	1002	85.0		6.23	4.42	5.16	4.27	49.7	148	4.00	0.000	51.4	15.5	23.6	18.5	1839	4			
Bosnia and Herzegovina	765			7.40	5.71	4.89	5.31	54.8	157	10.10	3.170	44.1	11.3	29.3	30.1	1139	3			
Brazil	857	52.1	67.1	6.23	4.42	5.16	4.27	49.7	148	4.00	0.000	51.4	15.5	23.6	18.5	1839	4			
Bulgaria	769	66.8	69.7	5.88	6.14	6.53	4.68	54.4	202	13.90	0.000	56.7	20.2	27.8	34.8	1020	4			
Burkina Faso	616			8.79	3.03	5.19	3.79	47.5	0			66.7	10.2	20.3	19.6	1367	1			
Burkina Faso	1004	83.5	71.7	7.31	3.27	5.74	5.69	54.2	117	6.30	0.000	51.6	11.6	26.1	20.3	1701	4			
Canada	871	57.8	65.7	8.24	5.17	5.49	4.91	52.6	177	13.20	2.210	57.9	15.1	24.7	31.6	1116	5			
China	845	54.4	69.0	8.52	4.23	4.62	4.34	49.4	97	9.00	2.650	53.9	15.5	25.6	29.8	1380	5			
Colombia	893	76.0	69.7	7.38	4.56	5.58	5.15	56.6	114	7.30	0.000	57.0	14.2	19.0	18.1	2319	3			
Croatia	1036	87.0	65.6	8.05	7.68	6.35	5.56	53.1	114	0.70	2.640	45.3	11.3	24.0	27.4	1171	3			
Cyprus				6.72	4.91	5.63	4.92	45.5	25			51.1	9.5	25.0	30.7	978	2			
Cyprus (T)				6.72	4.91	5.63	4.92	45.5	25			51.1	9.5	25.0	30.7	978	2			
Czech Republic	1028	90.6	67.7	6.65	3.45	6.70	5.63	53.1	167	2.60	0.000	54.8	11.7	22.7	18.6	480	1			
Denmark	766			6.95	4.60	5.73	5.00	50.2	0			53.9	21.6	26.4	28.0	1863	4			
Dominican Republic	871	59.9	73.9	9.33	7.09	5.73	5.46	59.1	0	0.00	0.000	51.2	11.0	14.7	27.2	1061	4			
Ecuador	765	39.9	59.1	13.65	4.04	6.30	4.85	59.5	91	3.60	1.430	79.0	20.0	45.3	22.7	2346	3			
Egypt	737			10.31	6.33	3.76	3.51	52.4	0			62.1	14.3	10.2	12.6	2346	3			
El Salvador	882	69.6	70.3	5.22	7.18	6.04	56.8	48.1	168	22.40	0.930	31.6	27.7	29.5	22.4	1144	5			
Estonia	842	42.8	59.9	9.66	3.00	3.23	3.04	48.1	0	0.00	0.000	41.1	10.1	13.4	21.6	1411	1			
Ethiopia	1036	91.0	82.5	6.79	5.18	7.42	7.44	50.4	67	3.00	0.000	44.9	17.9	32.7	48.9	872	4			
Finland	1010	80.4	69.9	6.06	3.50	5.74	5.17	52.1	141	9.60	8.210	65.3	11.5	24.4	20.1	1207	4			
France	769	60.4	67.4	5.78	5.84	5.03	4.08	55.9	190	5.80	2.160	41.5	20.8	28.7	22.4	1448	3			
Georgia	1025	87.3	75.1	6.90	4.25	6.32	5.80	53.5	151	7.80	1.890	52.3	26.3	31.0	25.3	2100	6			
Germany	749	42.8	67.5	8.85	3.78	4.00	3.83	49.2	40	0.00	0.330	29.0	14.5	17.5	25.4	1495	2			
Ghana	991	77.7	67.3	6.44	6.59	5.85	5.75	57.3	84	0.00	0.000	43.7	11.9	22.8	26.8	1274	2			
Greece	754	46.6		19.96	2.81	4.84	5.05	51.1	0	0.00	0.000	55.1	5.4	15.5	19.9	985	1			
Guatemala	655	25.0		2.85	2.70	2.33	4.89	50.8	0	0.00	0.000	19.9	16.8	26.1	47.6	1901	1			
Haiti	1048			7.42	5.50	5.64	5.57	53.7	42			49.4	20.0	20.8	21.7	1044	2			
Hong Kong SAR, China	872	69.5	68.7	6.94	3.82	6.64	5.24	52.6	185	3.40	0.120	45.3	21.2	29.3	21.5	1010	5			
Hungary	1015	82.7	82.8	7.08	4.56	7.21	6.91	49.7	126	3.60	0.000	53.8	18.6	28.4	34.7	806	4			
Iceland	637	36.7	62.2	11.77	4.70	5.96	4.34	42.9	138	12.00	4.310	71.7	18.8	34.4	22.1	2303	5			
India	709	48.0	65.4	10.46	3.40	4.26	5.18	48.2	75	7.80	0.000	71.2	7.8	19.7	33.4	1472	2			
Indonesia				11.93	6.14	3.75	4.47	47.2	182	7.50	0.000	67.5	14.6	17.2	30.4	2412	2			
Iran	803	48.4	59.0	15.26	5.60	5.89	4.88	48.5	0	0.00	0.000	82.9	23.2	43.9	35.2	1693	3			
Iraq	984			7.52	4.44	5.54	5.52	54.9	233	0.20	0.000	60.4	20.8	22.5	31.0	970	4			
Ireland	997	80.3	75.2	7.52	4.44	5.54	5.52	54.9	233	0.20	0.000	60.4	20.8	22.5	31.0	970	4			
Israel	997	80.3	75.2	7.52	4.44	5.54	5.52	54.9	233	0.20	0.000	60.4	20.8	22.5	31.0	970	4			
Italy	995	82.3	68.0	9.23	4.70	7.14	5.88	51.1	157	7.60	0.000	34.0	15.1	50.7	31.8	1532	4			
Japan	1037	85.6	65.2	8.61	5.54	7.44	6.04	51.6	146	3.80	0.000	76.0	24.6	32.3	28.1	1197	6			
Jordan	796	47.4	60.5	14.23	9.57	6.76	7.20	49.6	85	14.60	2.730	80.0	30.0	45.9	50.1	1151	3			
Kazakhstan	936	74.0	70.1	5.98	4.29	4.61	3.75	60.1	0	0.00	0.000	33.4	17.0	17.5	20.7	1441	1			
Kazakhstan	936	74.0	70.1	5.98	4.29	4.61	3.75	60.1	0	0.00	0.000	33.4	17.0	17.5	20.7	1441	1			
Kosovo	809			3.76	6.55	2.74	2.82	49.4	0			88.2	16.2	18.2	9.0	1522	1			
Kuwait	1067	67.5	64.6	11.39	11.87	9.32	8.65	35.7	0	0.00	0.000	43.4	28.3	26.9	43.6	1168	1			
Kyrgyzstan	652	61.9	70.4	5.55	3.52	3.74	3.73	53.2	108	0.40	0.000	41.5	13.0	18.1	22.8	1245	2			
Latvia	841	63.1	74.0	5.99	4.51	7.32	5.93	58.2	192	13.50	0.000	43.1	11.0	33.7	27.9	1052	4			
Lebanon	909	61.0	60.3	6.21	3.60	4.01	3.36	50.9	0	0.00	0.000	45.4	16.4	16.8	15.8	1135	1			

From left to right, this table reports the means for the main indexes. GE-UN is the Gender Equality Index from the United Nations. GE-WEF is the Gender Equality Index from the WEF. Maximum variability is the maximum observed difference in the dataset, i.e., the maximum gender difference observed in a given country for a given wave in a given category minus the minimum gender difference observed in the same country, the same category, in either the same wave or a different one. The last columns are the average number of observations per wave in a given country and the number of waves a country was present in the WVS.

Table S3.1: Descriptive statistics - 1 out of 2

Countries, means, and their variability

Country	Mean US\$			Means in %						Variability in US\$						Maximum variability in %						Number	
	log GDP	GE-UN	GE-WEF	LSO	LES	SCVS	GVS	female	Log GDP	GE-UN	GE-WEF	LSO	LES	SCVS	GVS	Av. Obs/Wave	Waves	Maximum variability in %					
																		GE-UN	GE-WEF	LSO	LES	SCVS	GVS
Libyan Arab Jamahiriya	860	82.7	72.2	10.34	8.65	8.02	7.53	47.6	0.0	0.00	54.8	21.9	32.1	38.5	1985	1							
Lithuania	834	68.9	51.9	5.44	5.67	7.15	5.44	51.9	168.1	4.50	38.6	17.8	28.3	23.4	1032	4							
Luxembourg	1123	82.5	68.0	5.65	4.57	4.32	5.16	51.1	80.3	4.00	0.00	46.7	8.0	19.0	22.5	2							
Macedonia, FYR	784	69.1	63.2	6.89	5.64	5.43	4.72	46.3	97.1	0.00	0.00	43.0	11.6	28.0	25.6	1105	3						
Malaysia	902	69.4	65.2	6.49	4.36	4.36	4.36	49.1	54.9	0.70	0.30	43.7	9.9	16.7	25.2	1213	2						
Mali	639	32.2	60.2	10.57	3.42	2.93	2.72	48.9	0.0	0.00	0.00	66.1	8.5	16.8	11.7	1374	1						
Malta	910	67.3	66.3	8.00	5.63	6.95	5.53	55.0	182.3	0.70	0.00	68.2	18.9	36.1	29.5	772	4						
Mexico	869	54.4	67.1	7.76	3.13	5.07	4.65	48.7	104.0	14.10	0.00	86.7	13.5	22.3	17.2	1553	5						
Moldova	649	60.8	71.8	6.34	5.14	5.66	4.34	53.5	132.2	19.60	1.16	30.5	18.9	22.9	26.1	1093	3						
Montenegro	791	7.38	5.35	5.98	5.70	5.23	5.70	52.3	164.7														
Morocco	772	40.4	57.4	6.75	4.61	5.11	4.99	49.7	82.0	17.00	1.28	48.7	18.0	43.6	41.9	1094	3						
Netherlands	1032	91.3	74.7	7.29	5.09	6.52	6.60	54.1	160.1	6.70	4.09	66.9	21.6	24.5	22.9	1233	5						
New Zealand	1012	82.1	78.1	8.85	6.27	7.09	6.50	55.7	92.6	2.00	0.00	40.3	15.6	25.1	26.8	901	3						
Nigeria	680	63.2	63.2	10.85	4.07	4.60	3.73	46.3	180.3														
North Ireland																							
Norway	1064	89.3	81.4	7.75	4.60	7.63	6.92	49.0	185.0	5.60	1.80	48.7	14.0	29.2	39.2	895	3						
Pakistan	671	39.3	54.8	21.05	8.28	8.57	7.93	47.2	84.7	20.60	0.00	91.2	36.5	37.1	37.9	1078	4						
Palestine, State of	795	55.5	66.9	12.25	4.78	5.87	4.81	50.0	0.0														
Peru	807	53.2	77.6	8.24	4.12	4.86	4.28	50.2	117.0	15.00	1.23	52.2	13.5	23.2	25.9	29.4	945	1					
Philippines	732	53.2	69.9	6.31	4.34	3.65	3.41	49.9	99.2	3.40	0.00	60.6	10.7	16.2	20.3	1286	4						
Poland	856	78.8	69.9	6.31	4.81	6.18	4.78	52.9	208.1	10.80	0.64	51.2	32.7	34.9	21.5	1183	3						
Portugal	943	77.1	70.5	7.47	6.70	7.27	5.33	56.5	113.5	3.50	0.00	46.1	24.0	33.9	31.5	1158	3						
Puerto Rico	960			7.84	8.38	4.99	4.98	64.6	45.5														
Qatar	1116		60.6	11.40	4.27	4.70	4.28	54.1	0.0														
Republic of Korea	933		63.4	8.10	5.05	6.85	6.35	51.0	240.9														
Romania	808	60.3	68.1	6.47	6.17	6.04	4.82	52.8	214.3	12.90	0.96	58.7	26.0	25.9	30.2	1166	6						
Russian Federation	848	59.7	69.5	6.13	6.62	7.66	5.93	58.2	236.7	16.10	2.67	57.7	24.5	31.0	27.6	1268	5						
Rwanda	629	57.2		7.02	3.18	3.09	2.80	50.3	0.0	1.80	74.9	9.7	14.5	10.1	1447	2							
Saudi Arabia	917			24.29	4.55	5.14	5.88	49.9	0.0														
Serbia	807	78.1		7.29	4.62	6.00	4.84	51.1	141.1	0.00	68.4	7.3	22.0	31.4	1435	1							
Singapore	1055	83.9	69.9	5.98	3.95	3.89	3.75	53.9	90.7	18.90	0.00	38.6	14.0	14.2	13.8	1768	2						
Slovakia	837	76.5	68.2	7.67	4.63	6.89	5.45	52.8	209.0	2.90	0.00	60.5	18.2	25.9	22.3	1052	4						
Slovenia	954	82.3	69.9	6.65	6.47	5.97	5.50	53.9	144.0	17.10	1.04	33.5	22.6	23.7	24.0	1020	5						
South Africa	835	58.2	73.3	7.72	2.56	4.35	3.35	50.3	0.0	2.00	3.85	46.5	6.9	24.3	26.9	2922	5						
Spain	977	86.9	74.4	7.98	3.48	6.93	5.26	51.8	190.1	5.80	2.99	67.8	14.8	30.9	21.1	1726	6						
Sri Lanka																							
Sweden	1040	92.8	81.0	8.86	4.83	5.92	5.09	43.3	148.0	4.30	0.95	52.0	31.6	31.1	26.5	997	6						
Switzerland	1087	90.6	71.4	9.06	3.53	6.39	5.69	53.1	86.5	6.00	4.36	61.2	14.2	23.3	23.8	1218	2						
Taiwan																							
Thailand	852	67.9	68.8	4.84	3.14	3.50	3.38	49.2	44.0	0.90	1.13	48.9	12.2	14.0	14.2	1293	2						
Trinidad and Tobago	964	65.1	70.8	7.37	4.53	4.68	4.88	54.6	16.7	0.00	5.56	66.7	10.4	36.0	21.5	952	2						
Tunisia	834	69.2	63.2	7.17	4.53	8.74	6.92	45.6	0.0	0.00	57.7	19.6	29.4	33.6	1100	1							
Turkey	874	48.5	58.7	9.96	6.73	5.34	4.05	50.4	132.4	27.10	2.47	77.4	18.4	26.0	24.5	1823	5						
Ukraine	565	35.0	68.4	6.24	6.30	6.85	5.45	61.9	0.0	0.00	40.8	14.2	38.3	39.4	968	1							
Uganda	751	78.0		7.32	4.87	7.13	5.87	54.1	181.5	15.00	0.64	57.5	22.1	28.2	32.0	1456	4						
United Kingdom										6.50	0.00	35.7	10.5	17.6	25.8	1204	5						
United States of America	1027	71.9	72.3	6.79	5.26	4.27	5.34	44.0	128.5	6.80	3.70	66.8	18.2	21.0	30.0	1691	6						
Uruguay	905	64.1	67.4	8.05	5.53	5.55	5.47	52.2	88.0	11.40	3.58	69.9	16.7	25.9	14.7	944	3						
Uzbekistan	537			6.67	5.10	4.85	4.19	61.0	46.2	0.00	51.1	13.3	18.1	18.1	1443	1							
Venezuela	822	46.1		10.41	3.97	5.27	3.64	49.8	66.1	0.00	46.3	13.1	14.7	22.1	1163	2							
Yvet Nam	653	96.1		9.07	6.08	13.97	15.31	49.1	0.0	0.00	66.3	49.1	31.6	23.3	1893	2							
Yemen	714			13.09	1.79	13.74	15.31	46.6	0.0	0.00	94.9	48.1	39.4	43.2	4894	1							
Zambia	701	43.2	62.9	6.32	2.91	3.72	3.32	61.0	46.8	6.40	0.00	38.4	17.7	17.7	17.6	1439	1						
Zimbabwe	665	40.2		10.30	4.70	5.54	5.20	51.6	46.8	0.00	38.9	13.3	30.4	26.1	1220	2							

From left to right, this table reports the means for the main indexes. GE-UN is the Gender Equality Index from the United Nations. GE-WEF is the Gender Equality Index from the WEF. Maximum variability is the maximum observed difference in the dataset, i.e., the maximum gender difference observed in a given country for a given wave in a given category minus the minimum gender difference observed in the same country, the same category, the same wave or a different one. The last columns are the average number of observations per wave in a given country and the number of waves a country was present in the WVS.

Table S3.2: Descriptive statistics - 2 out of 2

3.5.2 Instructions

Hired research assistants received the following instructions for the categorization of the WVS items:

For this task you will have to code a large number of survey questions into categories. The questions stem mainly from the World Values Survey, one of the largest international survey studies. For each questionnaire item you have to indicate (i) whether the item refers to the life-situation of the respondent or whether it is a value statement and (ii) indicate whether the question refers to a list of categories such as religion.

The data source for the coding is the official documentation of the World Values Survey, waves 1 to 6. We provide the list of survey items in the excel sheet QuestionsSetsCoding.xls. It contains a unique question code (under the column VARCODE), which is stable across waves, a short label and the precoded “Group 1”, “Group 1 - Second Choice”, “Group 1 - Comments” and other columns you will have to fill in with 0 (negative, no) and 1 (positive, yes). You should code all variables visible in the sheet according to the coding scheme which we explain in more detail below.

The “Labels” column will provide you with a short description of the question. However, this is not sufficient to code the items. That is, you will need to check the question in the printed PDF provided. The questions are in the same order as in the excel file (QuestionsSetsCoding.xls). The variable SOURCE indicates the wave number of the World Values Survey (W6 to W1), the variable VARCODE indicates the variable number in the respective document.

3.5.3 Coding the variables

You will have to code the questions in the File QuestionsSetsCoding.xls. The different columns contains the categories you have to use. The categories are divided into two groups.

Group 1: The column Group 1 contain a drop-down list with the possibilities described below. *All* items must be specified into *one* of the following categories.

- **Technical.** This category refers to questions which do not measure respondents’ answers. For instance, the wave number, the respondent ID, or the date of the interview. Moreover, every observation made by the interviewer is in this category.

- **Life situation, objective.** Questions about the circumstances in which the respondent lives and his habits. The qualifier “objective” refers to items that could—in principle—be verified by an external observer. Examples are questions about income, work hours, church going habits, or memberships in organisations.
- **Life situation, subjective.** Questions about the circumstances in which the respondent lives, but which are open to subjective interpretation by the respondent. The answers are not easily verifiable for an external observer. Examples: “How satisfied are you with your life?”, “Are you satisfied with your job”. The answers to these questions are open to interpretation, such that an external observer might perceive them differently.
- **Self-centred value statement.** This category contains questions about self-reflexive value statements. In other words, the values are self-centred and directly concern the subject himself. Examples: “Rate your confidence in the government” or “Are you a religious person”.
- **General value statement.** This category refers to value statements which are not directly linked to the respondent, but to the society (or humanity) in general. In other words, this category contains items asking about “How should the world be”, or “How should one act?”. Examples: “Is it justifiable to cheat on taxes” or “Government should reduce environmental pollution”.

Excluding the technical category all the categories from group 1 are divided between two major groups: life circumstances (which includes: life-situations objective and subjective) and values (which includes: self-centred value statement and general value statement). The former group investigates the life quality on a broad sense (“How is life”) and the latter, preferences and values on a broad sense (“What do you think about life”).

For some questions time gives another clue on how to distinguish between life-situations and values. If the question is about past behaviors (did you. . .), then you should classify it as a life-situation. If it is conditional or about future behavior (would you. . .), then you should classify it as a self-centered value.

As some ambiguity remains for the categorisation of the questions in group 1, you can optionally specify a second choice. You may also leave a comment for the categorisation of group 1, under the column “Group 1 - Comments” if necessary. On the following page you will find a table with further examples for the categories. Please go carefully through these examples

and use the table as a reference when coding the variables. In many cases you will find similar questions to those you are coding in the table.

Table S3.3:

Categories	Example questions
Life situation, objective	People use different sources to learn what is going on in their country and the world. For each of the following sources, please indicate whether you used it last week or did not use it last week to obtain information newspaper
	During the past year, did your family, save money, just get by, spent some savings, spent savings and borrowed money
	Now I am going to read off a list of voluntary organizations. For each organization, could you tell me whether you are an active member, an inactive member or not a member of that type of organization? ... Political party ...
Life situation, subjective	And for which, if any, are you currently doing unpaid voluntary work?
	All things considered, how satisfied are you with your life as a whole these days? Using this card on which 1 means you are completely dissatisfied and 10 means you are completely satisfied where would you put your satisfaction with your life as a whole?
	To what degree are you worried about the following situations? Losing my job or not finding a job
self-centred value statement	In your view, how often do the following things occur in this countrys elections? Votes are counted fairly
	On this list are various groups of people. Could you please mention any that you would not like to have as neighbors?
	Now Id like you to look at this card. Im going to read out some forms of political action that people can take, and Id like you to tell me, for each one, whether you have done any of these things, whether you might do it or would never under any circumstances do it
	I d like to ask you how much you trust people from various groups. Could you tell me for each whether you trust people from this group completely, somewhat, not very much or not at all?
	I am going to name a number of organizations. For each one, could you tell me how much confidence you have in them: is it a great deal of confidence, quite a lot of confidence, not very much confidence or none at all?
	I see myself as someone who is reserved
	Now I will briefly describe some people. Using this card, would you please indicate for each description whether that person is very much like you, like you, somewhat like you, not like you, or not at all like you?
general value statement	For each of the following, indicate how important it is in your life. Would you say it is: Family
	How proud are you to be French?
	Now I would like to ask you something about the things which would seem to you, personally, most important if you were looking for a job. Here are some of the things many people take into account in relation to their work. Regardless of whether you're actually looking for a job, which one would you, personally, place first if you were looking for a job?
	A variety of characteristics are listed here. Could you take a look at them and select those which apply to you?
	Would you be willing to pay higher taxes in order to increase your countrys foreign aid to poor countries?
	Here is a list of qualities that children can be encouraged to learn at home. Which, if any, do you consider to be especially important?
general value statement	I would like you to indicate which of these problems you consider the most serious one for the world as a whole?
	Please tell me for each of the following actions whether you think it can always be justified, never be justified, or something in between, using this card.

Group 2: The second group contains six categories that are not mutually exclusive, i.e., an item can be coded in none, one, or more than one category. If you want to code the item in a category mark a 1 in the respective column.

- **Liberal/Authoritarian/Conservative/Progressive - G3:** refers to the questionnaire items that relate to the role of the state about individual freedom, openness to immigration, change of political systems, and democratic values.
- **Religion - G2:** refers to the items that relate to religion on a broad sense; spirituality is included.
- **Trust - G2:** refers to the items that relate to trust/confidence. For instance: “How much do you trust strangers?”
- **Altruism - G2:** refers to the items about altruism/generosity. For instance: “Do you donate money to charities”.
- **Cooperation - G2:** refers to all the items that relate to the relation between the individual and the society which have a cooperative character. This could be items about preferences or willingness to participate in social movements, volunteering and the like.
- **Redistribution - G2:** Items related to views about the importance of material wealth, the redistribution of wealth, and the justification of wealth differences in general. Examples “Do the rich deserve their wealth”, or “How important is it for you to be rich”.

IF YOU HAVE ANY QUESTIONS, PLEASE DO NOT HESITATE TO CONTACT ME.

3.5.4 The coefficient from the WVS and the EVS

I use the longitudinal World Value Survey (available here <http://www.worldvaluessurvey.org/WVSDocumentationWVL.jsp>). I combine it with the European Value Survey (available here: <https://dbk.gesis.org/dbksearch/GDESC2.asp?no=0009\&DB=E>, page 9), but from this data, I keep only the items that are the same as in the WVS. For a complete list of the items see <http://www.worldvaluessurvey.org/WVSContents.jsp>.

The WVS & EVS in numbers

Summary Statistics WVS & EVS

Number of participants	508,707
Years	1981-2016
Female	53%
Countries	112
N of items	1231
Av. N of questions per participant	243.24

This table reports summary statistics of the overall WVS and the EVS. It includes technical variables and items only present in the EVS that I do not use in the analyses.

Table S3.4: Descriptive statistics

Computation I need to disentangle the data type for analysis purposes. I identify the following types: categorical, binary, ordinal, and continuous. I also add another category for the items that could be rescaled to ordinal. For instance, the options in the item: “1 = approve, 2 = disapprove and 3 = depends” could be interpreted as ordinal in the following order: 1 = approve, 2 = depends, 3 = disapprove. I drop the continuous items as the vast majority of them are recorded by the interviewer. Apart from the item, “How many children do you have” where I do not expect differences between males and females, the remaining items are coded by the interviewer. In the main paper, I discuss only ordinal items (which include, binaries and rescaled to ordinal) and categorical items to facilitate the reading. The percentage of ordinal, binary, and rescaled to ordinal is 69.5%. The percentage of categorical items is 30.5%. These percentages include only the ones I use in the analysis, therefore, not the technical items.

Ordinal items: For the ordinal, binary, and rescaled to ordinal items, I use the computation proposed by Klotz. This computation provides a coefficient (henceforth: C_o) for gender differences. In the computation, I plot the relative frequencies (e.g., the percentage of women who choose option 1) of each gender for each option and per item in a matrix and then compute the probability that the sample from one gender will choose a higher option than the sample from the other gender. I, ultimately, compute the absolute difference as follows: $|C| = |C_o - 0.5|$ and end up with a coefficient that measures the absolute difference between males and females not sensitive to monotonic transformation. As a robustness check I compare the coefficient $|C|$ to Cohen’s d and obtain a correlation of 97.5% ($p = 0.000$, from a Pearson’s correlation test). I did not use Cohen’s

d as it computes differences in the mean and is subject to monotonic transformation (see the next section for all concerns with the computation of ordinal variables). The coefficient $|C|$ is normalized in the main paper and I use C_n in percent as the dependent variable in all the regressions.

Categorical items: For the categorical variables, I use Cramer's V (Cramér 1946) computation. The given coefficient has the advantage of being bound between 0, no association between the two subgroups, and 1, perfect association. As mentioned in the main paper, as the asymptotic expectation of the X^2 statistics is proportional to the sample size, the estimator will be biased by the differences in n . Cramer's V, on the other hand, is insensitive to sample sizes as it corrects the estimator by dividing it by a multiple of the sample size (Agresti and Kateri 2011).

As the computation for ordinal data is bounded between 0 and 0.5, I multiply the ordinal coefficient by 2. I end up with a coefficient that has the same interpretation regardless of the data type. Therefore, 0 = no gender difference, 1 = total gender difference. I always report the gender difference in percent.

3.5.5 Classification of the items from the WVS.

I mentioned in the main paper that I hired two research assistants for the coding of the items. Overall the kappa-statistic returns 79.6% of agreement between the two research assistants. When they agreed I used their coding as is, but in cases when they did not match I hired another research assistant to make the call between the two classifications of the research assistants. As I also had my own classification, I tested the agreement rate between each research assistant and my classification. The kappa-statistic returns 79.6% agreement between one of them and my original classification and 93.8% for the other one and my classification. I did the coding myself as a robustness check, but did not use it in the analysis. To assess the robustness of the categorization, I report in the robustness checks section the regressions with the categorization of both coders.

3.5.6 Ordinal analysis of data

Schroeder and Yitzhaki (2017) and T. N. Bond and Lang (2019) bring some interesting criticism to the literature about the analysis of ordinal data that I aim to answer in this section. Their argument mainly stems from strong

necessary assumptions made on the characteristics of the ordinal data. They argue that as long as the variance between the two subgroups is not equal and that I cannot observe a first-order stochastic dominance, there exists a monotonic transformation that reverses the conclusion. T. N. Bond and Lang focus on some of the major findings in the happiness literature where researchers investigate factors influencing the mean happiness. As researchers use mean computations to infer which groups display a higher level of happiness, they need to assume that the two subgroups have the same reporting function (same cardinality) and that the gaps between the different options are equal (interval interpretation of ordinal data). In their review of these major findings in the happiness literature, T. N. Bond and Lang apply some monotonic transformation to the happiness scale and reverse the primary conclusions. For instance, an exponential function can reverse the conclusion depending on the skewness of the data.

As a monotonic transformation influences the mean, Chen et al. (2019) proposed to restore most of the findings in the happiness literature by shifting the conclusion made on the mean happiness to the median happiness. A similar methodology was earlier proposed by Allison and Foster (2004) with a focus on self-reported health status. However, the proposition to shift from the mean to the median is not entirely appropriate in my case as most of the Likert scales used in the WVS have a restricted number of options. Therefore, most of the gender differences would then vanish. For instance, most of the variables are on a 3-point scales would result in a median of 2. Moreover, in some cases, the differences can be overestimated, such as when the median shifts from 1 to 2.

Nevertheless, I overcome the vast majority of these issues in my study. First, the computation I use (see methods section and Klotz 1966) is not sensitive to a monotonic transformation, as I do not compute means. The only concern that holds in my computation is the possibility of a different gender reporting function. For instance, men always report higher points in the Likert scales. Nevertheless, I am not interested in whether one group has a higher propensity or a lower propensity to choose higher or lower options than another group. I am interested in the absolute difference. Therefore, any difference in the reporting function needs to be systematic in order to bias the estimate and would only quantitatively influence the result, but not qualitatively. More precisely, it might change the absolute difference, but then in no case, whether absolute gender differences increase or decrease with more gender equality or economic prosperity.

As I am comparing the same subgroups in different countries, I am more concerned with a systematic interaction between the gender and the culture resulting in a difference in the reporting function. This may indicate that the

difference I observe between the countries is a matter of culture. However, a bias would occur if, and only if, there is a systematic correlation of one of the indexes I use, such as the GDP or gender equality indexes on the reporting of one gender and not the other. In the extreme case where the indexes only impact one gender, then the story I want to tell would be different. However in any more reasonable interpretation, even if there is a different interaction effect of the indexes and the gender, it would only sum up again as a level issue and then not change the story I want to tell.

One might also be concerned with the occurrence of bounded items correlated with the indexes I use as independent variables. For instance, both genders might report systematically more extreme options in low GDP countries and the opposite in high GDP countries. This would explain why I observe fewer or more gender differences in relation to indexes such as the GDP. However, according to the categorization (LSO, LSS, SCVS, and GVS) even in countries where I observe fewer differences in life-situations (LSO and LSS), I observe more differences in the values (SCVS and GVS) in the cross-country analyses. One should expect that if the reporting function is culturally influenced, it should also hold across the categorization.

Finally, criticism regarding a different reporting function is dependent on the number of measured items. As I combine many items from many different countries, the reporting function of males and females should be systematically different to bias the estimate. The more items and the more countries, the less the reporting function is a concern.

3.5.7 Indexes

In the regression models, the choice of the indexes is mostly inspired by Falk and Hermle (2018) and Hofstede et al. (2005), Kaiser (2019), and Welzel (2013). I use the following indexes as independent variables.

- United Nations - Gender Inequality Index, from the Human Development Report 2015. Values inverted to create an index of equality. <http://hdr.undp.org/en/composite/GII>
- WEF Global Gender Gap Index, from the World Economic Forum Global Gender Gap Report 2015 <http://reports.weforum.org/global-gender-gap-report-2015/rankings/> combined with the 2013 one, which reports values from 2006 to 2013 http://www3.weforum.org/docs/WEF_GenderGap_Report_2013.pdf.
- Gross Domestic Product (GDP) in US dollars and per year

from the United Nations. <https://data.un.org/Data.aspx?q=Per+capita+GDP&d=SNAAMA&f=grID\%3a101%3bcurrID%3aUSD%3bpcFlag%3a1>

- Historical and Contemporary Pathogen Prevalence. Taken from the supplementary material of Fincher et al. (2008). <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2602680/bin/rspb20080094s06.pdf>
- Hofstede Individualism score (2001). Taken from the supplementary material of Fincher et al. (2008). <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2602680/bin/rspb20080094s06.pdf>
- Geert Hofstede's cultural dimensions: power distance, masculinity, and uncertainty avoidance taken from: <http://clearlycultural.com/geert-hofstede-cultural-dimensions/power-distance-index/>.
- Food access taken from the Food and Agriculture Organization: <http://www.fao.org/faostat/en/#data/FBSH>.
- Historical food access taken from the Food and Agriculture Organization: <http://www.fao.org/faostat/en/#data/FBSH>. I calculate the mean from year 1961 to year 1984 per country.
- Welzel indicators of defiance, defiance, disbelief, relativism, and skepticism. For their computation, see <http://www.worldvaluessurvey.org/WVSContents.jsp?CMSID=welzelidx> I compute the means of each of these indicators per country and per wave.

I match the indexes of gender equality and economic growth with the corresponding country and the corresponding year. For some indexes, such as the UN Gender Inequality Index, there is no record for every year present in the WVS. Therefore, I match the years with the closest year in the indexes and when it was in between, I pick the above year.

For the index of pathological prevalence, individualism, food access, and Geert Hofstede's cultural dimensions, I matched them with the country only, as all of these variables have only one value in time.

For the cultural dimensions of Welzel, I compute the means per country and per wave and matched them accordingly.

3.5.8 Clustering

In the main paper, I do not include any clustering. As mentioned above the observations represent gender differences per item, per wave, and per

country. This computation already decreases biases due to individual cross-correlations. Nevertheless, I test different clusterings, which allows for other cross-correlations and serial correlations.

The general trend in the literature is to cluster at the country level. However, there are at least three possible clusterings in this situation: country level, country and wave, item and country. Different clustering implies different assumptions that I will discuss in this section.

The first clustering allows cross-correlations between the survey items and between the waves. Whenever I cluster at this level, I lose the significance level for some of the categories (GVS and LSS). The loss of significance might be due to some overestimation of the standard errors as the regression might overvalue the dependence between the survey items. Put differently, some items relate to a more general value and their error terms in the regressions are correlated. Even if, one can assume that some items might be related, such as “Who would you not like to have as neighbors” and “Is it justifiable to...”, some others are likely independent. For instance: “Belief about child qualities” and “Belief about religion” refer to different values.

This clustering also allows for serial correlations. There are two competing arguments for serial correlation. The first argument in favor is that, since values are likely derived from sources, these sources could evolve over time, but they would not likely change drastically. As an example, a country’s religious composition is likely to explain religious beliefs and the religious composition is unlikely to change drastically in a short period of time. The second argument in disfavor is related to different samplings over the years. For instance, one wave may have focused on the south of the USA and another wave on the north, where views differ. Note finally, that if this last argument makes sense in the USA, it might not be generalized to smaller countries. I report in Table S3.5 the regression of gender differences on the UN Gender Equality Index. I lose the significance level due to the possible limitations mentioned above.

The second clustering allows cross-correlations between the survey items. This clustering assumes that observations are independent over time, but not within the same wave. The main argument for this point is that within the same country all values are somehow derived from the same “sources”. I mentioned, above, limitations of this argument. This clustering also assumes that the respondents are different over time. I discuss above the pros and cons of this assumption. In Table S3.6, I report the results from the same regression as above, but with clustering at the wave and country-level. Unsurprisingly, the coefficients have slightly higher significance levels. This is unsurprising because smaller clusters allow for smaller correlations of the error terms.

The last possible clustering is at the country and item level. This re-

Gender Equality Index (UN) - Cross-country - Cluster at the country level

	Dependent variable: Coefficient: % of gender differences			
	LSO	LSS	SCVS	GVS
Gender Equality Index - UN	-4.747** (1.609)	-2.282 (1.476)	2.205* (0.914)	2.066+ (1.045)
Constant	8.321** (1.778)	5.970** (1.087)	3.032** (0.754)	1.410+ (0.716)
Item fixed effects	<i>Yes</i>	<i>Yes</i>	<i>Yes</i>	<i>Yes</i>
R^2	0.675	0.198	0.354	0.301
Clusters	89	89	89	89
N	10,006	3,762	19,031	23,647

Notes: OLS estimates. This table reports the effect sizes of the Gender Equality Index from the UN on the coefficient C_n . Each regression independently tests the effect of the independent variable on the different categories (LSO, LSS, SCVS, GVS). I include fixed effects for each item of the WVS. Robust standard errors, clustered on country, in parentheses. The F -test is not computed because the VCE is not of sufficient rank to perform the model test. + $p < 0.1$, * $p < 0.05$, ** $p < 0.01$.

Table S3.5: OLS estimates

Gender Equality Index (UN) - Cross-country - Cluster at the country and wave level

	Dependent variable: Coefficient: % of gender differences			
	LSO	LSS	SCVS	GVS
Gender Equality Index - UN	-4.747** (1.556)	-2.282+ (1.357)	2.205* (0.900)	2.066* (1.012)
Constant	8.321** (1.758)	5.970** (1.009)	3.032** (0.746)	1.410* (0.693)
Item fixed effects	<i>Yes</i>	<i>Yes</i>	<i>Yes</i>	<i>Yes</i>
R^2	0.675	0.198	0.354	0.301
Clusters	172	128	202	236
N	10,006	3,762	19,031	23,647

Notes: OLS estimates. This table reports the effect sizes of the Gender Equality Index from the UN on the coefficient C_n . Each regression independently tests the effect of the independent variable on the different categories (LSO, LSS, SCVS, GVS). I include fixed effects for each item of the WVS. Robust standard errors, clustered on country and wave level, in parentheses. The F -test is not computed because the VCE is not of sufficient rank to perform the model test. + $p < 0.1$, * $p < 0.05$, ** $p < 0.01$.

Table S3.6: OLS estimates

gression allows for serial correlations over time, but assumes that the items are independent from each other. For instance, the regression assumes that views on homosexuality are unrelated to views on drug consumption, but that views on homosexuality over time are correlated. I report in Table S3.7 this last clustering. All the coefficients of the UN Gender Equality Index are significant ($p < 0.01$) as in the main paper.

Gender Equality Index (UN) - Cross-country - Cluster at the country and item level

	Dependent variable: Coefficient: % of gender differences			
	LSO	LSS	SCVS	GVS
Gender Equality Index - UN	-4.747** (0.646)	-2.282** (0.600)	2.205** (0.310)	2.066** (0.232)
Constant	8.321** (1.268)	5.970** (0.509)	3.032** (0.433)	1.410** (0.253)
Item fixed effects	<i>Yes</i>	<i>Yes</i>	<i>Yes</i>	<i>Yes</i>
R^2	0.675	0.198	0.354	0.301
Clusters	6,378	2,490	10,044	13,558
N	10,006	3,762	19,031	23,647

Notes: OLS estimates. This table reports the effect sizes of the Gender Equality Index from the UN on the coefficient C_n . Each regression independently tests the effect of the independent variable on the different categories (LSO, LSS, SCVS, GVS). I include fixed effects for each item of the WVS. Robust standard errors, clustered on country and item level, in parentheses. The F -test is not computed because the VCE is not of sufficient rank to perform the model test. + $p < 0.1$, * $p < 0.05$, ** $p < 0.01$.

Table S3.7: OLS estimates

Overall, clustering has advantages and disadvantages, but the major issue would be if some serial correlation would not taken into account, such that one country is over-represented in the data and would drive the effect. I partially account for this issue in the main paper by restricting the data to a sub-sample. Overall, except for the first clustering, which I believe overestimates cross-correlations between the survey items, the significance levels are similar to regressions without clustering.

3.5.9 Additional analyses

As mentioned, the regressions use the C_n I compute from the WVS and the EVS as the dependent variable. For the independent variables, I use the different indexes mentioned above. I control for fixed effects for the different survey items. I include in this section only the regression tables not

present in the main paper. I test whether pathogen prevalence (historical and contemporary) explains gender differences in values independently. In Tables S3.8 and S3.9, I report the regression outputs and show that the correlation is negative and significant ($p < 0.01$). Countries that faced a high prevalence of pathogen display lower gender differences in values. In Table S3.10, I test if individualism explains gender differences in values and find a positive correlation. These results are in line with Kaiser (2019). Finally, I test whether historical and contemporary access to food supply explain gender differences in values and the correlation is positive and significant ($p < 0.01$). In short, countries that had and have better access to food tend to have more gender differences, which is still in line with the ecological stress hypothesis.

Although these correlations are significant, the evolution of gender differences remains difficult to explain. First, I lack data on the evolution of individualism across societies, and second, the effect of pathogen prevalence is not immediate. According to Kaiser (2019), there is an important latency between changes in pathogen prevalence and a possible change in values.

Pathogen prevalence - Historical

	Dependent variable: Coefficient: % of gender differences			
	LSO	LSS	SCVS	GVS
Pathogen Prevalence (Historical)	0.716** (0.088)	0.035 (0.097)	-0.902** (0.047)	-0.961** (0.036)
Constant	7.946** (0.058)	4.826** (0.065)	5.915** (0.031)	5.162** (0.024)
<i>F</i> -test	65.8	0.1	375.6	724.0
<i>p</i>	0.000	0.720	0.000	0.000
R^2	0.005	0.000	0.010	0.015
<i>N</i>	11920	4102	22478	28031

Notes: Fixed effects estimates. This table reports the effect sizes of the pathogen prevalence (historical) on the coefficient C_n . Each regression independently tests the effect of the independent variable on the different categories (LSO, LSS, SCVS, GVS). I include fixed effects for each item of the WVS. Standard errors in parentheses. + $p < 0.1$, * $p < 0.05$, ** $p < 0.01$.

Table S3.8: OLS estimates

I run correlation tests between these ecological factors and the different indexes and find that the UN Gender Equality Index correlates with the contemporary and historical pathogen prevalence, the Hofstede individualism index, the food supply historical, and the contemporary food supply access with the respective values, -0.73 , -0.66 , 0.59 , 0.52 , and 0.64 . I get very similar results when testing the correlation with the log GDP and slightly lower coefficients for the WEF Gender Equality Index. These results support the ecological stress hypothesis that differences in environment shaped

Pathogen prevalence - Contemporary

	Dependent variable: Coefficient: % of gender differences			
	LSO	LSS	SCVS	GVS
Pathogen Prevalence (Contemp.)	0.073** (0.014)	-0.013 (0.014)	-0.114** (0.007)	-0.101** (0.006)
Constant	5.657** (0.427)	5.092** (0.438)	9.564** (0.220)	8.402** (0.169)
<i>F</i> -test	26.3	0.8	247.3	326.9
<i>p</i>	0.000	0.359	0.000	0.000
<i>R</i> ²	0.006	0.000	0.013	0.013
<i>N</i>	6103	2147	11620	14349

Notes: Fixed effects estimates. This table reports the effect sizes of the pathogen prevalence (contemporary) on the coefficient C_n . Each regression independently tests the effect of the independent variable on the different categories (LSO, LSS, SCVS, GVS). I include fixed effects for each item of the WVS. Standard errors in parentheses. + $p < 0.1$, * $p < 0.05$, ** $p < 0.01$.

Table S3.9: OLS estimates

Individualism score - Hofstede

	Dependent variable: Coefficient: % of gender differences			
	LSO	LSS	SCVS	GVS
Individual. (Hofstede)	-0.008** (0.003)	-0.013** (0.003)	0.017** (0.001)	0.021** (0.001)
Constant	8.072** (0.137)	5.415** (0.164)	5.234** (0.076)	4.303** (0.060)
<i>F</i> -test	11.1	20.8	144.9	365.0
<i>p</i>	0.001	0.000	0.000	0.000
<i>R</i> ²	0.001	0.007	0.005	0.009
<i>N</i>	10153	3469	19141	23866

Notes: Fixed effects estimates. This table reports the effect sizes of the Hofstede individualism score on the coefficient C_n . Each regression independently tests the effect of the independent variable on the different categories (LSO, LSS, SCVS, GVS). I include fixed effects for each item of the WVS. Standard errors in parentheses. + $p < 0.1$, * $p < 0.05$, ** $p < 0.01$.

Table S3.10: OLS estimates

Food Supply - Historical

	Dependent variable: Coefficient: % of gender differences			
	LSO	LSS	SCVS	GVS
Historical Food Supply	-0.001** (0.000)	-0.000+ (0.000)	0.001** (0.000)	0.001** (0.000)
Constant	11.812** (0.419)	5.614** (0.443)	2.655** (0.205)	2.845** (0.162)
<i>F</i> -test	76.1	3.4	270.6	243.8
<i>p</i>	0.000	0.064	0.000	0.000
<i>R</i> ²	0.008	0.002	0.008	0.005
<i>N</i>	10027	3529	18985	23618

Notes: Fixed effects estimates. This table reports the effect sizes of the historical food supply on the coefficient C_n . Each regression independently tests the effect of the independent variable on the different categories (LSO, LSS, SCVS, GVS). I include fixed effects for each item of the WVS. Standard errors in parentheses. + $p < 0.1$, * $p < 0.05$, ** $p < 0.01$.

Table S3.11: OLS estimates

differences in values, however, due to the above-mentioned limitations (lack of data over time and latency) the recent evolution of differences in values cannot be attributed causally to these factors. In other words, gender differences at the outset could be the cause of these ecological factors, but the recent evolution (last 40 years) of these differences remains puzzling.

As in the main paper I show only the evolution of LSO and GVS with more gender equality for countries present in Wave 1 and 6 in a scatter plot, In Table S3.13, I report the regression of the UN Gender Equality Index on the different categories. Therefore, the sample is restricted to countries that were part of both Wave 1 and 6. This sub-sample allows me to investigate the effect of the evolution of the index over an important period of time (at least 26 years, before 1995 and after 2009). The convergence/divergence for gender differences in values is robust to this sub-sample.

3.5.10 Robustness checks

The robustness checks are structured as follows: (i) the robustness of the categorization, (ii) investigation of time inconsistencies of the indexes, and (iii) robustness to sub-samplings, such as number of countries or number of waves. The following regressions always report cross-country and within-country analyses.

Food Supply - Contemporary

	Dependent variable: Coefficient: % of gender differences			
	LSO	LSS	SCVS	GVS
Food Supply	-0.001** (0.000)	-0.000* (0.000)	0.001** (0.000)	0.001** (0.000)
Constant	10.413** (0.352)	5.738** (0.386)	3.623** (0.180)	3.284** (0.138)
<i>F</i> -test	56.3	4.8	189.0	223.1
<i>p</i>	0.000	0.029	0.000	0.000
<i>R</i> ²	0.005	0.002	0.005	0.003
<i>N</i>	13247	4421	24821	30967

Notes: Fixed effects estimates. This table reports the effect sizes of the food supply (contemporary) on the coefficient C_n . Each regression independently tests the effect of the independent variable on the different categories (LSO, LSS, SCVS, GVS). I include fixed effects for each item of the WVS. Standard errors in parentheses. ⁺ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$.

Table S3.12: OLS estimates

Gender Equality Index (UN) - Wave 1 and 6

	Dependent variable: Coefficient: % of gender differences							
	LSO	LSS	SCVS	GVS	LSO	LSS	SCVS	GVS
Gender Equality Index - UN	-1.981 (1.749)	5.300** (1.965)	2.815* (1.249)	4.446** (1.054)	-82.828** (12.473)	-33.319** (11.759)	-33.097** (6.760)	-14.482* (6.034)
Constant	9.024** (1.481)	0.264 (1.659)	4.644** (1.052)	2.239* (0.889)	59.080** (7.661)	23.295** (7.115)	26.667** (4.144)	13.580** (3.699)
Dummy Country	<i>No</i>	<i>No</i>	<i>No</i>	<i>No</i>	<i>Yes</i>	<i>Yes</i>	<i>Yes</i>	<i>Yes</i>
<i>F</i> -test	1.3	7.3	5.1	17.8	7.3	4.3	6.3	5.3
<i>p</i>	0.258	0.007	0.024	0.000	0.000	0.000	0.000	0.000
<i>R</i> ²	0.003	0.020	0.002	0.012	0.037	0.003	0.026	0.012
<i>N</i>	632	326	1071	1167	632	326	1071	1167

Notes: Fixed effects estimates. This table reports the effect sizes of the scatter plot in the main paper. These regressions include only the countries that were asked both in wave 1 and 6. Each regression independently tests the effect of the independent variable on the different categories (LSO, LSS, SCVS, GVS). I include fixed effects for each item of the WVS. Standard errors in parentheses. ⁺ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$.

Table S3.13: OLS estimates

firstly, I test the sensitivity of the categorization by running the regression of the main paper with each research assistant categorization (see Tables S3.14 and S3.15 for coder 1 and Tables S3.16 and S3.17 for coder 2). I find that all results are robust to small changes in the categorization.

Robustness checks - Categorization from Coder 1 - Cross-country

	Dependent variable: Coefficient: % of gender differences											
	LSO	LSS	SCVS	GVS	LSO	LSS	SCVS	GVS	LSO	LSS	SCVS	GVS
Gender Equality Index - UN	-2.302** (0.340)	-1.509* (0.618)	1.441** (0.187)	2.021** (0.162)								
Log GDP per capita in US dollars			-0.155** (0.036)	-0.130+ (0.069)	0.096** (0.020)	0.314** (0.017)						
Gender Equality Index - WEF									-11.271** (1.247)	-5.349+ (2.797)	-3.094** (0.754)	1.392* (0.643)
Constant	8.763** (0.241)	6.523** (0.447)	4.861** (0.132)	3.929** (0.113)	8.340** (0.321)	6.646** (0.620)	4.913** (0.175)	2.507** (0.153)	14.193** (0.859)	9.450** (1.924)	7.600** (0.519)	4.135** (0.440)
F-test	45.8	6.0	59.7	155.9	18.7	3.5	24.2	332.7	81.7	3.7	16.9	4.7
p	0.000	0.015	0.000	0.000	0.000	0.061	0.000	0.000	0.000	0.056	0.000	0.030
R ²	0.004	0.004	0.002	0.004	0.004	0.001	0.001	0.007	0.013	0.003	0.002	0.000
N	11323	1889	21035	21520	15059	2276	27335	27705	6540	738	9732	11119

Notes: Fixed effects estimates. This table reports the effect sizes of the indexes present in the main paper. The coding of the categories is solely the one done by Coder 1. Each regression independently tests the effect of the independent variable on the different categories (LSO, LSS, SCVS, GVS). I include fixed effects for each item of the WVS. Standard errors in parentheses. + $p < 0.1$, * $p < 0.05$, ** $p < 0.01$.

Table S3.14: OLS estimates

Robustness checks - Categorization from Coder 1 - Within-country

	Dependent variable: Coefficient: % of gender differences											
	LSO	LSS	SCVS	GVS	LSO	LSS	SCVS	GVS	LSO	LSS	SCVS	GVS
Gender Equality Index - UN	-7.289** (1.533)	-10.981** (2.612)	-3.302** (0.840)	-3.600** (0.688)								
Log GDP per capita in US dollars					-0.559** (0.095)	-0.688** (0.1171)	-0.532** (0.051)	-0.325** (0.048)				
Gender Equality Index - WEF									-5.669 (7.262)	9.562 (15.998)	-6.084 (3.985)	-5.590 (3.508)
Constant	14.691** (1.170)	11.081** (2.150)	8.625** (0.632)	6.708** (0.525)	12.655** (0.886)	9.649** (1.693)	10.165** (0.470)	6.724** (0.436)	9.837* (4.852)	-0.855 (10.733)	9.514** (2.661)	7.817** (2.357)
Dummy Country	Y_{es}	Y_{es}	Y_{es}	Y_{es}	Y_{es}	Y_{es}	Y_{es}	Y_{es}	Y_{es}	Y_{es}	Y_{es}	Y_{es}
F-test	10.6	3.5	20.7	31.3	11.0	3.2	21.1	30.3	7.4	2.0	15.0	23.9
p	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
R ²	0.033	0.113	0.052	0.074	0.036	0.097	0.044	0.066	0.038	0.155	0.077	0.103
N	11323	1889	21035	21520	15059	2276	27335	27705	6540	738	9732	11119

Notes: Fixed effects estimates. This table reports the effect sizes of the indexes present in the main paper. The coding of the categories is solely the one done by Coder 1. Each regression independently tests the effect of our independent variable on the different categories (LSO, LSS, SCVS, GVS, LSO, LSS, SCVS, GVS). We include fixed effects for each item of the WVS. Standard errors in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table S3.15: OLS estimates

Robustness checks - Categorization from Coder 2 - Cross-country

	Dependent variable: Coefficient: % of gender differences											
	LSO	LSS	SCVS	GVS	LSO	LSS	SCVS	GVS	LSO	LSS	SCVS	GVS
Gender Equality Index - UN	-4.879** (0.389)	-1.511** (0.398)	1.931** (0.216)	2.186** (0.149)								
Log GDP per capita in US dollars			-0.312** (0.041)	-0.056 (0.045)	0.094** (0.023)	0.299** (0.016)						
Gender Equality Index - WEF									-19.243** (1.372)	-11.531** (1.711)	-1.444+ (0.802)	2.558** (0.629)
Constant	11.427** (0.276)	6.065** (0.283)	5.186** (0.152)	3.789** (0.105)	10.601** (0.371)	5.548** (0.405)	5.639** (0.206)	2.570** (0.139)	20.431** (0.944)	12.985** (1.168)	6.794** (0.550)	3.362** (0.433)
F-test	157.6	14.4	79.6	214.2	57.4	1.5	16.5	366.3	196.7	45.4	3.2	16.6
p	0.000	0.000	0.000	0.000	0.000	0.218	0.000	0.000	0.000	0.000	0.072	0.000
R ²	0.010	0.004	0.004	0.005	0.006	0.000	0.000	0.007	0.022	0.015	0.000	0.000
N	10087	4394	17491	25267	13328	5154	22592	33210	6004	1963	9035	11835

Notes: Fixed effects estimates. This table reports the effect sizes of the indexes present in the main paper. The coding of the categories is solely the one done by Coder 2. Each regression independently tests the effect of the independent variable on the different categories (LSO, LSS, SCVS, GVS). I include fixed effects for each item of the WVS. Standard errors in parentheses. + $p < 0.1$, * $p < 0.05$, ** $p < 0.01$.

Table S3.16: OLS estimates

Robustness checks - Categorization from Coder 2 - Within-country

	Dependent variable: Coefficient: % of gender differences											
	LSO	LSS	SCVS	GVS	LSO	LSS	SCVS	GVS	LSO	LSS	SCVS	GVS
Gender Equality Index - UN	-5.903** (1.803)	-8.819** (1.796)	-4.293** (0.913)	-3.633** (0.649)								
Log GDP per capita in US dollars					-0.573** (0.109)	-0.575** (0.127)	-0.657** (0.062)	-0.333** (0.041)				
Gender Equality Index - WEF									-7.746 (8.019)	8.581 (10.337)	-4.792 (4.213)	-8.807* (3.439)
Constant	13.825** (1.378)	9.753** (1.457)	10.959** (0.705)	6.677** (0.486)	12.718** (1.025)	8.876** (1.222)	12.422** (0.568)	6.859** (0.377)	10.989* (5.360)	-0.068 (7.006)	8.838** (2.827)	10.408** (2.297)
Dummy Country	Y_{es}	Y_{es}	Y_{es}	Y_{es}	Y_{es}	Y_{es}	Y_{es}	Y_{es}	Y_{es}	Y_{es}	Y_{es}	Y_{es}
F-test	13.0	7.7	19.5	32.8	14.1	7.0	19.8	32.9	8.5	6.0	12.8	23.4
p	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
R ²	0.040	0.109	0.067	0.065	0.046	0.098	0.060	0.058	0.045	0.175	0.075	0.092
N	10087	4394	17491	25267	13328	5154	22592	33210	6004	1963	9035	11835

Notes: Fixed effects estimates. This table reports the effect sizes of the indexes present in the main paper. The coding of the categories is solely the one done by Coder 2. Each regression independently tests the effect of my independent variable on the different categories (LSO, LSS, SCVS, GVS, LSO, LSS, SCVS, GVS). I include fixed effects for each item of the WVS. Standard errors in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table S3.17: OLS estimates

Second, I investigate if indexes might have inconsistencies in their computation. This test accounts for the possibility that the indexes could have a random component, such as the measure of the index taking into account different indicators one year and not the others. As I include dummy variables for each survey item in the fixed effect model, the R^2 increases substantially. However, this is an artificial increase, as it does not increase substantially the validity of the model. Tables S3.18 and S3.19 show that the convergence/divergence story holds even when I control for possible inconsistencies in the indexes. I see, however, that I get lower coefficients and partially lose the significance for some categories. This is probably due to part of the effect being captured by the year dummies. As mentioned in the main paper, every year, the values in these indexes increase for almost all countries. Therefore the increase can be captured by the dummies.

Robustness checks - Year fixed effects - Cross-country

	Dependent variable: Coefficient: % of gender differences											
	LSO	LSS	SCVS	GVS	LSO	LSS	SCVS	GVS	LSO	LSS	SCVS	GVS
Gender Equality Index - UN	-4.517** (0.454)	-1.899** (0.481)	1.573** (0.234)	1.995** (0.181)	-0.193** (0.047)	-0.147** (0.054)	0.101** (0.025)	0.335** (0.019)	-16.184** (1.501)	-7.566** (1.978)	-0.399 (0.839)	3.828** (0.687)
Log GDP per capita in US dollars									17.834** (1.076)	9.803** (1.411)	5.825** (0.587)	2.637** (0.482)
Gender Equality Index - WEF									Y _{es}	Y _{es}	Y _{es}	Y _{es}
Constant	13.698** (0.525)	6.309** (0.478)	6.330** (0.264)	4.941** (0.208)	11.378** (0.568)	6.251** (0.565)	6.538** (0.291)	3.372** (0.227)	Y _{es}	Y _{es}	Y _{es}	Y _{es}
Dummy Year									27.7	15.5	22.6	39.5
F-test	16.3	8.3	19.9	26.8	12.9	6.9	17.2	31.0	0.000	0.000	0.000	0.000
p	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.023	0.062	0.009	0.019
R ²	0.016	0.046	0.017	0.014	0.014	0.030	0.012	0.016	5870	1710	9423	11497
N	10006	3762	19031	23647	13291	4415	24748	30825				

Notes: Fixed effects estimates. This table reports the effect sizes of the indexes on our coefficient. Each regression independently tests the effect of our independent variable on the different categories (LSO, LSS, SCVS, GVS). I include fixed effects for each item of the WVS and for each year. Standard errors in parentheses. + $p < 0.1$, * $p < 0.05$, ** $p < 0.01$.

Table S3.18: OLS estimates

Robustness checks - Year fixed effects - Within-country

	Dependent variable: Coefficient: % of gender differences											
	LSO	LSS	SCVS	GVS	LSO	LSS	SCVS	GVS	LSO	LSS	SCVS	GVS
Gender Equality Index - UN	3.952 (3.009)	-16.855** (3.467)	0.835 (1.591)	-2.527* (1.211)	0.609* (0.256)	-0.928** (0.316)	-0.628** (0.135)	-0.202* (0.103)	-16.915 (11.455)	16.092 (16.522)	-1.156 (5.994)	-5.681 (4.946)
Log GDP per capita in US dollars									17.245* (7.398)	-5.025 (10.725)	5.925 (3.869)	8.561** (3.198)
Gender Equality Index - WEF									Y_{es}	Y_{es}	Y_{es}	Y_{es}
Constant	8.703** (2.164)	16.136** (2.545)	8.372** (1.109)	6.898** (0.842)	5.872** (1.781)	11.933** (2.220)	11.963** (0.933)	6.437** (0.709)	7.2 Y_{es}	4.7 Y_{es}	12.2 Y_{es}	21.9 Y_{es}
Dummy Year and Country									5870 Y_{es}	1710 Y_{es}	9423 Y_{es}	11497 Y_{es}
F-test	10.7	6.2	17.2	26.2	11.5	5.6	17.2	25.5	7.2	4.7	12.2	21.9
p	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
R ²	0.038	0.134	0.060	0.072	0.046	0.114	0.051	0.064	0.043	0.170	0.065	0.102
N	10006	3762	19031	23647	13291	4415	24748	30825	5870	1710	9423	11497

Notes: Fixed effects estimates. This table reports the effect sizes of our indexes on our coefficient. Each regression independently tests the effect of the independent variable on the different categories (LSO, LSS, SCVS, GVS). I include fixed effects for each item of the WVS, for each year and for each country. Standard errors in parentheses. + $p < 0.1$, * $p < 0.05$, ** $p < 0.01$.

Table S3.19: OLS estimates

Finally, I test whether the story holds if I restrict the data to a sub-sample of countries and a sub-sample of survey items.

Table S3.20 and S3.21 report the regression where I restrict the survey items only to those that were asked at least in 3 waves. This restriction allows me to control that the effect I observe is not driven by some survey items that were asked only once or twice. It allows me to investigate the time trend of the survey items and also their evolution over a wider range of values in the indexes since the more time passes the greater variance I observe in the indexes. The results are robust to this sub-sample.

Robustness checks - Sub-sampling - At least 3 waves - Cross-country

	Dependent variable: Coefficient: % of gender differences											
	LSO	LSS	SCVS	GVS	LSO	LSS	SCVS	GVS	LSO	LSS	SCVS	GVS
Gender Equality Index - UN	-5.424** (0.809)	-4.312** (0.686)	2.430** (0.244)	2.550** (0.202)	-0.379** (0.090)	-0.410** (0.074)	0.089** (0.025)	0.335** (0.021)	-25.582** (3.283)	-14.723** (2.824)	1.190 (0.971)	3.577** (0.873)
Log GDP per capita in US dollars									29.152** (2.242)	15.301** (1.935)	4.969** (0.671)	2.526** (0.600)
Gender Equality Index - WEF									60.7	27.2	1.5	16.8
Constant	17.358** (0.565)	8.246** (0.478)	4.774** (0.171)	3.516** (0.142)	17.099** (0.799)	8.982** (0.653)	5.508** (0.223)	2.261** (0.190)	1746	17782	6071	5922
<i>F</i> -test	44.9	39.5	99.2	158.9	17.6	30.9	12.4	243.4	0.000	0.000	0.221	0.000
<i>p</i>	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.031	0.000	0.001
<i>R</i> ²	0.001	0.025	0.005	0.006	0.001	0.014	0.000	0.008	0.006	0.000	0.000	0.001
<i>N</i>	3189	1354	13053	12981	4017	1746	17782	17146	1581	675	6071	5922

Notes: Fixed effects estimates. This table reports the effect sizes of the indexes present in the main paper. These regressions include only the survey items that were asked in at least 3 waves. Each regression independently tests the effect of the independent variable on the different categories (LSO, LSS, SCVS, GVS). I include fixed effects for each item of the WVS. Standard errors in parentheses. + $p < 0.1$, * $p < 0.05$, ** $p < 0.01$.

Table S3.20: OLS estimates

Robustness checks - Sub-sampling - At least in 3 waves - Within-country

	Dependent variable: Coefficient: % of gender differences											
	LSO	LSS	SCVS	GVS	LSO	LSS	SCVS	GVS	LSO	LSS	SCVS	GVS
Gender Equality Index - UN	-14.572** (3.094)	-11.120** (2.406)	-4.352** (0.975)	-3.494** (0.780)	-1.321** (0.231)	-0.786** (0.1175)	-0.666** (0.063)	-0.283** (0.054)	-38.367* (17.629)	3.265 (13.973)	-4.275 (5.146)	-9.722* (4.414)
Log GDP per capita in US dollars												
Gender Equality Index - WEF												
Constant	23.727** (2.400)	11.461** (1.883)	10.634** (0.726)	6.417** (0.591)	23.478** (2.111)	10.381** (1.619)	12.119** (0.569)	6.220** (0.494)	34.007** (11.959)	2.584 (9.419)	8.413* (3.432)	10.385** (2.957)
Dummy Country	Y_{es}	Y_{es}	Y_{es}	Y_{es}	Y_{es}	Y_{es}	Y_{es}	Y_{es}	Y_{es}	Y_{es}	Y_{es}	Y_{es}
F-test	4.7	4.9	13.5	20.8	5.1	4.5	15.0	19.4	2.4	3.3	6.7	14.4
p	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
R ²	0.048	0.231	0.052	0.084	0.042	0.190	0.046	0.068	0.025	0.277	0.047	0.116
N	3189	1354	13053	12981	4017	1746	17782	17146	1581	675	6071	5922

Notes: Fixed effects estimates. This table reports the effect sizes of the indexes present in the main paper. These regressions include only the survey items that were asked in at least 3 waves. Each regression independently tests the effect of the independent variable on the different categories (LSO, LSS, SCVS, GVS, LSO, LSS, SCVS, GVS). I include fixed effects for each item of the WVS. Standard errors in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table S3.21: OLS estimates

Tables S3.22 and S3.23 report the regressions when I restrict the survey items to those that were at least asked in 60 countries. This sub-sampling highly overlaps the previous sub-sampling, as most of the survey items that were asked in at least 3 waves tend to be asked also in a lot of countries. Nonetheless, this test allows me to get rid of some survey items that tend to be asked frequently, but only in some parts of the world. I observe that the distribution of gender equality in the world is not random. For instance, when a country displays a high degree of gender equality, as measured by the indexes, then an adjacent country usually displays a similar level of gender equality. As some survey items are region specific (concerning only a small number of adjacent countries with similar values in the indexes), I wanted to control that these items were not driving the effect. Overall the divergence/convergence story of gender differences in values with respect to an increase in gender equality or economic growth is robust to these additional tests.

Robustness checks - Sub-sampling - At least in 60 countries - Cross-country

	Dependent variable: Coefficient: % of gender differences											
	LSO	LSS	SCVS	GVS	LSO	LSS	SCVS	GVS	LSO	LSS	SCVS	GVS
Gender Equality Index - UN	-5.250** (0.429)	-3.248** (0.650)	2.383** (0.219)	2.120** (0.180)								
Log GDP per capita in US dollars					-0.326** (0.045)	-0.219** (0.070)	0.148** (0.023)	0.312** (0.019)				
Gender Equality Index - WEF									-18.666** (1.557)	-13.287** (2.550)	-0.608 (0.814)	2.053** (0.730)
Constant	12.151** (0.304)	7.769** (0.452)	4.643** (0.153)	4.059** (0.126)	11.125** (0.404)	7.496** (0.626)	4.881** (0.207)	2.676** (0.168)	20.335** (1.073)	14.853** (1.747)	6.140** (0.559)	3.940** (0.501)
F-test	149.8	24.9	118.9	138.9	52.8	9.8	40.5	272.8	143.7	27.1	0.6	7.9
p	0.000	0.000	0.000	0.000	0.000	0.002	0.000	0.000	0.000	0.000	0.455	0.005
R ²	0.011	0.012	0.005	0.004	0.007	0.002	0.001	0.008	0.020	0.022	0.000	0.000
N	8254	1762	16016	17670	11145	2266	21058	23320	4863	1057	8642	9216

Notes: Fixed effects estimates. This table reports the effect sizes of the indexes present in the main paper. These regressions include only the survey items that were asked in at least 60 countries. Each regression independently tests the effect of the independent variable on the different categories (LSO, LSS, SCVS, GVS). I include fixed effects for each item of the WVS. Standard errors in parentheses. + $p < 0.1$, * $p < 0.05$, ** $p < 0.01$.

Table S3.22: OLS estimates

Robustness checks - Sub-sampling - At least in 60 countries - Within-country

	Dependent variable: Coefficient: % of gender differences											
	LSO	LSS	SCVS	GVS	LSO	LSS	SCVS	GVS	LSO	LSS	SCVS	GVS
Gender Equality Index - UN	-8.058** (1.957)	-10.288** (2.463)	-4.120** (0.914)	-3.234** (0.730)								
Log GDP per capita in US dollars					-0.633** (0.116)	-0.522** (0.177)	-0.654** (0.061)	-0.313** (0.051)				
Gender Equality Index - WEF									-14.615 (9.043)	-2.136 (14.828)	-5.643 (4.287)	-6.770+ (3.836)
Constant	16.045** (1.473)	10.996** (1.970)	10.359** (0.697)	6.563** (0.552)	14.060** (1.078)	8.797** (1.669)	11.987** (0.561)	6.764** (0.460)	16.228** (6.038)	7.581 (10.016)	9.294** (2.877)	8.960** (2.576)
Dummy Country	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
F-test	11.5	4.8	18.0	27.5	12.7	4.5	18.8	26.7	6.7	3.4	11.6	19.8
p	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
R ²	0.040	0.177	0.062	0.078	0.047	0.150	0.054	0.066	0.042	0.192	0.067	0.099
N	8254	1762	16016	17670	11145	2266	21058	23320	4863	1057	8642	9216

Notes: Fixed effects estimates. This table reports the effect sizes of the indexes present in the main paper. These regressions include only the survey items that were asked in at least 60 countries. Each regression independently tests the effect of the independent variable on the different categories (LSO, LSS, SCVS, GVS, LSO, LSS, SCVS, GVS). I include fixed effects for each item of the WVS. Standard errors in parentheses. + $p < 0.1$, * $p < 0.05$, ** $p < 0.01$.

Table S3.23: OLS estimates

4 General conclusion

My thesis has finally three very distinct chapters, which are all completely different from the first proposal I had made. Although they all relate to decision making to some extent, they differ in research questions and methodologies. Therefore, Before I mention the meeting points of the research, I will highlight the key elements of my thesis.

In a joint work with my supervisor, we elaborate a precise stereotype elicitation mechanism in Chapter 1. While the measure of stereotypes helps investigate how stereotypes influence attitudes and behaviors, their measures are subject to biases. Assuming participants are aware of their stereotypes, an obvious bias in measuring them is the social desirability one. Apart from this bias, there is often room for measurement errors, such as due to unknown behaviors baselines or overestimation of differences. Our design gets partially rid of the former bias, since it is costly to express a socially desirable stereotype, and because we provide participants with baselines and outline estimated differences, we also reduce measurement errors. Finally, using our elicitation mechanism, we find no systematic gender stereotypes in cooperation, but a systematic overestimation of left-leaning individuals in cooperation compared to right-leaning ones.

In Chapter 2, I investigate different underlying mechanisms in donation decisions. Using an experiment on nudging (Thaler and Sunstein 2008), I am able to some extent to disentangle, which mechanism dominates the decision to donate. I find that this decision is mostly triggered by an emotional arousal in contrast to a lower cognitive cost. However, I do not exclude that the cognitive cost is playing a role, since, it likely decreases the intrinsic motivation of participants to donate jointly with the impression of control that this nudge sets. The joint mechanisms sum up to no difference on the realized level of donation. The overall experiment shows that this nudge is not a free lunch for charities, as suggested by previous research (Schulz et al. 2018), but shows that more research is needed before implementing this nudge in donation decisions.

In Chapter 3, I investigate the evolution of gender differences in values with respect to economic growth and the increase in gender equality. Apart from an obvious fundamental inquiry, this research has implications for policymakers. Assuming the intention is to reach gender equality, at least, in economic outcomes, whether gender differences in values increase or decrease with the economic/gender equality growth could lead to different policies. For instance, if women and men tend to differ more and more in their job choices, the “equal work, equal pay” policy is unlikely to reach gender equality in economic outcomes. Nevertheless, I do not reach a definite answer in this chap-

ter. While I unambiguously show that more gender equality and economic growth decrease differences in life-situations between men and women, I show conflicting evidence for differences in values. On a cross-country investigation, I find that gender differences in values increase, but on a within-country analysis, they decrease. While these results suggest an endogeneity issue in the cross-country analysis, the paradox is robust to additional specifications, such as ecological stress factors and cultural differences. I conclude on a puzzle and show that more research is needed, especially since it might have important policy implications.

One of the common themes of all three chapters is the focus on gender. While in Chapter 3, I use gender as a proxy to investigate differences in values, in contrast, in Chapters 1 and 2, I find no systematic gender differences in cooperative behavior nor in altruistic behavior. Moreover, I find no systematic gender stereotypes in cooperation. These results contrast a common social representation that gender is, first a good predictor of differences, and second that people hold strong gender stereotypes. In other words, although gender is a useful proxy to investigate behaviors, I find that gender differences in effective and perceived behaviors are to a lesser extent than expected.

Furthermore, my research always highlights the pragmatic implications. Precisely, the general study in behavioral economics goes often beyond fundamental research since it can have implications in public policies. In the present case, knowing stereotypes might help to address them, investigating the effect of choice architecture on donation can help increase the provision of public goods, and investigating the evolution of values is likely a necessary step before designing adequate public policies.

Finally, although I believe having a link to a possible policy implication is valuable, the contributions I bring, are not exempt from general concerns. The construction of knowledge in science is mostly incremental, such that new findings are based on other previous findings. However, without advocating deeply for a replication crisis⁷⁶, I end up with 2 out of 3 papers that relate partially to this issue. In Chapter 2, I suppose that I do not replicate most of the findings from Schulz et al. (2018), because of a different environment (in university courses vs online). This suggests that findings in one context cannot always be generalized. On the other hand, while in Chapter 3, I replicate the findings from Falk and Hermlé (2018), I show however that the estimate of the effect of gender equality on gender differences in values

⁷⁶The replication crisis is the observation that many previous research fail to be replicated (Pashler and Wagenmakers 2012).

is inconsistent. This shows that these relationships might be more complex than expected. Overall, both chapters advocate for a precautionary principle in research, especially if the findings have policy implications.

References

- Falk, A., & Hermle, J. (2018). Relationship of gender differences in preferences to economic development and gender equality. *Science*, *362*(6412).
- Pashler, H., & Wagenmakers, E.-J. (2012). Editors introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science*, *7*(6), 528–530.
- Schulz, J. F., Thiemann, P., & Thöni, C. (2018). Nudging generosity: Choice architecture and cognitive factors in charitable giving. *Journal of Behavioral and Experimental Economics*, *74*, 139–145.
- Thaler, R. H., & Sunstein, C. R. (2008). *Nudge: Improving decisions about health, wealth, and happiness*. Yale University Press.