Stockholm School of Economics
Department of Economics
5350 Master's thesis in economics
Academic year 2020–2021

An Experimental Study on Gender Differences in Leading-by-Example in a Social Dilemma

Johanna Ohlin (23731)

Abstract

This study examines gender differences in leading-by-example in a social dilemma by the means of an online experiment with 350 UK participants recruited via Prolific Academic. A three-player sequential (leader-follower) public goods game is used to represent the social dilemma and gender differences in behavior related to the role of gender beliefs and the gender group composition are examined to investigate whether male or female leaders are better cooperative role models under free-riding incentives and thus more successful in crowding in. It is relevant for effectively organizing group interactions in situations where individual and group outcomes are in conflict. Using a mixed-design all participants are exposed to three gender group compositions: male-dominated groups, female-dominated groups and all-male or all-female groups. Participants play both roles (leader and follower) in all groups. Leaders' beliefs about followers' contributions are elicited to map to what extent unconditional leader contributions are related to the perceived cooperativeness of followers. The results indicate that there are no statistically significant gender differences in preferences for leading-by-example and in crowding-in effects, nor in the effect of the gender group composition or beliefs' about followers' support. Further studies with larger sample sizes or different experimental designs could confirm these patterns.

Keywords: leading-by-example, sequential public goods game, gender differences, voluntary contribution, cooperation

JEL: C72, C99, D91, H41

Supervisor: Anna Dreber Almenberg
Date submitted: May 16, 2021
Date examined: May 25, 2021
Discussant: Jonathan Grosin
Examiner: Robert Östling

Acknowledgements

This study is jointly and generously funded by Professor Anna Dreber Almenberg at Stockholm School of Economics (SSE) and Professor Rupert Sausgruber at Vienna School of Business and Economics (WU). I would like extend my sincerest gratitude to them for making this experiment feasible and enabling me to have this learning experience, and to Milos Fisar, lab manager at WU, for pointing me in the direction of Lioness and Prolific when few other options remained. I would like to extend this gratitude to my supervisor also for her continuous guidance and feedback throughout the development of this thesis. In addition, I would like to thank Gustav Karreskog for the valuable inputs during the PhD student feedback and Michelle Rudolph for the insightful comments during the mid-term seminar. All shortcomings and errors are my own.

Contents

1 Introduction

> *“A leader is one who knows the way, goes the way and shows the way.”*
>
> — John C. Maxwell (2018)

> *“As a leader, part of your job is to inspire the people around you to push themselves – and, in turn, the company – to greatness. To do this, you must show them the way by doing it yourself.”*
>
> — Bruna Martinuzzi, *The Leader as a Mensch - Become the Kind of Person Others Want to Follow.* (2009)

In a social dilemma (i.e. a collective action problem) people have an incentive to free-ride on others' efforts. It lies in the collective interest if everyone contributes to the public good (PG), but an individual has an incentive to contribute nothing as the PG is non-excludable and non-rivalrous. The PG is likely under-produced if contribution is voluntary due to free-riding incentives (Holcombe, 2008). In such a setting a challenge for a leader is how to induce followers to work for the group's best interest rather than prioritizing personal interests at the group's expense. Such PGs dilemmas are common in the workplace and political organizations (Gächter et al., 2012) and in voluntary contribution settings such as fund-raising events and community associations. When there is a public fund-raising for diabetes research (i.e. the PG is the cure against diabetes) individuals have an incentive not to contribute as they will likely not be prevented access to cure. However, if nobody contributes there will be no cure and everyone will be worse off than had all contributed (Weber et al., 2004).

Overcoming such free-riding is difficult without coordination (Moxnes and Van der Heijden, 2003) in voluntary settings, but a leader might influence followers' behavior and promote more efficient social outcomes through the mechanism of leading-by-example (Gächter et al., 2012). In the workplace a leader might be incentivized to stop working early, and the employees may follow suit, instead of working with something that benefits the entire organization. In terms of access to clean air (i.e. the PG) individuals have an incentive to ignore the collective welfare and commuting by car instead of taking public transportation (Van Vugt et al., 1995). A political leader might lead by example by taking public transportation instead of the car. A venerable question is whether women or men are better role models in terms of promoting more efficient social outcomes? The question is interesting because in social dilemmas where personal and collective interests are in conflict, leaders can be seen as first-movers in a sequential public goods game (PGG) who may shape the beliefs of those who come to follow by acting as role models.

A gender effect might occur due to heterogeneity in social preferences for cooperation e.g., according to sociocultural theory women are more communally oriented towards the welfare of the group, but according to evolutionary theory men are more adapt at coordinating. A gender effect may also occur due to different beliefs about others' cooperativeness (Greig and Bohnet, 2009) applying social structural and status character theories. The leadership context may create gender role-based interpretations (Sell, 1997) through the activation of gender stereotypes of the leadership role being more congruent with being a male. This may result in men and women having different perceptions of their influence as leaders, which may affect revealed preferences.

Previous research on cooperation in simultaneous versus sequential PGGs has highlighted the importance of leading-by-example as a means to influence followers' behavior. In the standard (linear) PGG all players make their contributions simultaneously and independently. Gächter and Renner (2004) introduce the leading-by-example setting by letting leaders be the first-movers in a sequential version of the game and followers the second-movers. They find that followers' and leaders' contributions are statistically significant and positively correlated. Leaders' and followers' average contributions are also higher than the average contributions of players in the simultaneous game. Gächter and Renner (2018) report similar results, also using randomly chosen leaders. Using a four-player voluntary contribution mechanism (VCM) game Güth et al. (2007) also find that leadership increases contributions to the PG compared to the standard game where all players move simultaneously. Since leaders' and followers' choices and payoff functions are identical in the sequential PGG, leaders may set an example to followers and promote higher cooperation rates (Gächter and Renner, 2004).

This is linked to the observed behavior in experimental PGGs that many people behave as conditional cooperators and cooperate if others do so (Fischbacher et al., 2001; Fischbacher et al., 2012). Gächter and Renner (2018) explore the link between leadership, beliefs and conditional cooperation in a sequential PGG. They find that leaders significantly affect followers' initial beliefs and contributions and are able to lead their groups to higher outcomes than groups without a leader. The same holds for the correlation between the average group leader contributions and the group average earnings. Thus, in a social dilemma a leader might influence followers by showing how to "walk-the-walk" through leading-by-example.

However, despite existing research on cooperation in simultaneous versus sequential PGGs, current research is inconclusive regarding gender differences in cooperation in social dilemmas (Ortmann and Tichy, 1999; Croson and Gneezy, 2009; Balliet et al., 2011). Gender differences appear to be sensitive to the situational context (Hyde, 2007; Croson and Gneezy, 2009; Balliet et al., 2011) such as the gender composition of the group. Albeit indications of women being marginally more cooperative exist (Balliet et al., 2011; Capraro, 2018), the results seem sensitive to the type of game, the particular design and the gender composition of the study sample (Ortmann and Tichy, 1999). Additionally, to the best of my knowledge, I have not found previous research on gender differences in leading-by-example in a group setting in a social dilemma where leadership is imposed. This paper addresses this research gap by focusing on gender differences in leading-by-example and conditional cooperation in a group setting where gender is announced.

I follow Gächter and Renner (2004), Gächter et al. (2012) and Gächter and Renner (2018) closely with respect to leading-by-example in a PGG, but also build upon previous research on gender beliefs linked to observed behavior in experiments. Vyrastekova et al. (2015) let both sexes make decisions in a three-player simultaneous PGG with female-dominated groups, male-dominated groups and same-sex groups. They find that contributions increase with the number of women in the group, but more so for men. Men also express more systematic beliefs that contributions increase with the number of women in the group. Studying whether a male-dominated environment (not a social dilemma) adversely affect women's willingness to lead four-person groups (male-majority or female-majority groups), Born et al. (2019) find that women are least willing to lead, and correctly anticipates least support, in male-dominated groups. Contrary, men are most willing to lead, and correctly anticipates most support in female-dominated groups. However, studying self-selected leadership in a three-player PGG Arbak and Villeval (2007) find that men on average are more willing to lead when gender is not announced. More charitable women (defined by the portion of their show-up fee donated to a charity before the game) are though most willing to lead. Yet, when gender is announced the gender effect disappears.

In contrast to Vyrastekova et al. (2015) I study gender differences in leadership in a sequential PGG. With respect to Born et al. (2019) I do not study willingness to lead, but gender differences when the leader role is imposed and in a PGs dilemma. The purpose is to explore whether such gender beliefs exist also in this experimental setting. My study is similar to parts of Arbak and Villeval (2007). However, in their imposed leadership treatment (i.e., leadership is randomly assigned within the group and groups randomly composed) gender is not known as gender differences is not their focus.

The overall aim of this paper is to understand gender differences in preferences (elicited by contributions to the PG) for leading-by-example and conditional cooperation in a social dilemma, related to the gender group composition and the role of gender beliefs. The primary research questions are *(i)* whether women or men would be better role models regarding leading-by-example, and thus *(ii)* if female or male leaders would be able to promote more efficient group outcomes by inducing higher cooperation rates amongst followers?

I address these research questions by the means of an online PGG with 175 men and 175 women of UK nationality recruited via Prolific Academic. I introduce leading-by-example using a sequential PGG with three-player groups and cooperativeness is elicited using a variant of Selten's strategy method (Selten, 1967) following Gächter et al. (2012). Participants play both roles to ascertain whether female or male leaders would be more successful in crowding in. To examine the role of beliefs for leader decisions I ask participants to guess their followers' contribution responses. To explore the role of the gender group composition on decisions, taking into account a possible gender heterogeneity effect, all participants play all roles in three different groups: male-dominated groups, female-dominated groups and all-male or all-female groups, without any feedback between decisions. Gender is introduced using female and male first names.

With this design I hope to contribute to current research on leading-by-example in VCMs by focusing on revealed and perceived gender differences in leadership and conditional cooperation. Understanding if gender affects individual leader and follower decisions in group interactions is interesting in order for organizations and associations to organize effectively in situations where individual and group outcomes are in conflict. When leadership is randomly imposed it is not clear whether gender would predict leader behavior, nor the leader's effectiveness. Sociocultural theory might predict women to be better role models as they would be more other-regarding than men, while evolutionary theory might predict the opposite as cooperation historically have been vital for men due to e.g., hunting and warfare. Apart from culture, individual preferences and age, gender is an important factor in the leadership context (see Balliet et al. (2011) and Eklund et al. (2017)). Gender identity may alter leader and follower decisions with the group context if activated.

Additionally, I hope to add insight to the inconclusiveness on gender differences in cooperation by examining if the gender group composition and gender beliefs affect decisions, building upon Vyrastekova et al. (2015) and Born et al. (2019). Inherent social preferences should not change with the gender group composition so if it matters it may indicate the presence of gender beliefs. Possible perceived gender differences in leadership status may affect men's and women's beliefs about their influence as leaders and thus revealed preferences. Applying social structural and status character theories men might be predicted to be more confident leaders as men have held leadership positions more often than women. The results indicate that male leaders' contributions and beliefs on average are insignificantly higher in all gender group compositions, and there are no significant indications that gender beliefs and the gender group composition affect leaders' and followers' decisions.

The remainder of this paper is organized as follows. Section 2 presents the PGG and discusses its relevance in this setting. It also discusses the literature most pertinent and relevant to the focus of this study. Section 3 outlines the research procedures and the experimental design, as well as the data collection and processing. Section 4 states the hypotheses to be tested, the associated analytical specifications and analyzes the results. Section 5 discusses the validity and implications of the results while Section 6 concludes.

2 Theory and previous research

This section first presents and justifies the PGG to model the social dilemma and studying leading-by-example. I then present and discuss previous research related to this study. Namely, research on leading-by-example, conditional cooperation and gender differences in cooperation in social dilemmas. Next, I present theoretical perspectives from social psychology as to why gender differences in leading-by-example and cooperation in social dilemmas might emerge. I first discuss gender differences in social (other-regarding) preferences, then gender differences in beliefs about others' cooperativeness. I do not empirically test these theories, they are intended as insights from a related research field. Specifically, they might explain how gender roles and beliefs create gender differences in behavior. Last, I briefly present gender differences in risk-aversion and discuss its relevance to gender differences in leading-by-example and cooperation in social dilemmas.

2.1 The public goods game

The PGG is one of the most used models for simulating a social dilemma and study cooperation under free-riding incentives. In the standard (linear) PGG I use, each player is endowed with an equal amount of tokens. The endowment is a proxy for their resources e.g., material means such as wealth, energy levels or knowledge/skills (Anderson et al., 2011). The PGG uses the VCM tool to elicit contributions to the PG. Within this game each player chooses how much of his/her endowment to contribute to a joint project and how much to keep for private use. The amount each player contributes is multiplied with a number greater than one and then distributed evenly amongst all players regardless of the individual contribution. Each player keeps the tokens not contributed to the project and they are translated into private earnings. All payoff structures are known beforehand to all players (Camerer, 1997).

I use a linear PGG where player i's payoff in a game with N players is given by

$$\pi_i = x - c_i + \frac{1}{N}\sum_{j=1}^{N}(1+r)c_j, \text{ where } 1 < r+1 < N. \tag{2.1}$$

x is the initial amount of tokens a player is endowed with to allocate between a personal fund and the group fund (Gächter and Renner, 2018). c_i is the individual contribution of player i to the PG. r represents the marginal return for each token contributed to the PG. I set $r = 0.5$ so that the multiplication factor equals 1.5. It is thus lower than the group size, but higher than one so the group's payoff is maximized if all players contribute all of their endowment. However, a profit-maximizing player has an incentive to free-ride on the contributions of others and to contribute nothing (Gächter et al., 2012) since $r < 1$. However, if all players defect the outcome is worse then had everyone contributed their entire endowment.

For mutual cooperation to be more beneficial than mutual defection $r + 1 > 1$ must hold (Hauert and Szabo, 2003), but since $r + 1 < N$ a *social dilemma* arises where personal and collective interests are in conflict. As such, the PGG makes it appropriate to study leadership incentives and motivation in situations where collective welfare is in conflict with individual self-interest (Gächter and Renner, 2018). This as leadership often involve making people to do something they otherwise would not do (Gächter et al., 2012). In addition, the PGG has provided a useful testing ground for gender differences in cooperation. If it would be true that one the sexes would emphasize the welfare of the group over that of the individual more, this would arguably translate into one of them contributing more to the PG (Ortmann and Tichy, 1999).

Empirical evidence is ambiguous regarding the group size's impact on cooperation. A common viewpoint is that cooperation is harder to sustain the larger the group as the more players, the lower is the individual gain of contributing to the project relative to the cost (Barcelo and Capraro, 2015). Thus justifying a higher marginal per capita return (MPCR), defined as $(1+r)/N$. However, in a meta-analysis using data from 27 PGGs Zelmer (2003) reports a positive effect of the group size, but only at the 10% significance level. When holding r fixed and increasing the group size Barcelo and Capraro (2015) find that larger groups cooperate more in a one-shot, simultaneous PGG, but the small sample size limits the generalizability of the findings. Isaac et al. (1994) do also not find any support that group size is inversely related to group contributions towards the PG.

However, Isaac and Walker (1988) do not find evidence that the group size (four or ten) per se affects cooperation levels when holding the MPCR constant. Group size makes a difference in groups with a low MPCR, but the same effect is not observed when the MPCR is high. In sum, when the group size is relatively small it appears the MPCR has a larger effect, but the group size's effect on cooperation is not evident.

2.2 Literature review

2.2.1 Evidence of leading-by-example

In the simultaneous PGG game all players make their contributions without observing others' contributions. Sequentially may be introduced in different forms. E.g., when Fischbacher et al. (2001) examine conditional cooperation they first let three players make their contributions without observing each other's contributions. This is described as the *unconditional* contribution. The fourth group member then makes his/her contribution upon observing the others' contributions. This is described as the *conditional* contribution.

Gächter and Renner (2004) introduce leadership to the standard (linear) PGG above by letting leaders be the first-movers who make an *unconditional* contribution before others, and followers the second-movers who upon observing the first mover's contribution make theirs simultaneously (i.e. without observing other followers' decisions). This sequential setup mimics a setting where a randomly chosen leader makes his/her contribution decision to the joint project first. The followers, upon viewing the leader's contribution, make their contributions simultaneously. Thus, the followers' contributions are not fully unconditional, an amount of uncertainty still exists as they cannot know whether the other followers will follow the leader's example.

Experiments in simultaneous versus sequential PGGs highlight the positive correlation between leaders' and followers' contributions (Gächter and Renner, 2004; Arbak and Villeval, 2007; Güth et al., 2007; Gächter et al., 2012, Gächter and Renner, 2018). Leading-by-example may promote more efficient social outcomes under free-riding incentives in the presence of *conditional cooperation*. I.e., the tendency to behave cooperatively if others do so, and to free-ride if others free-ride (Keser and Van Winden, 2000; Fischbacher et al., 2001; Fischbacher and Gachter, 2010; Gächter and Renner, 2018). Despite that the subgame perfect equilibrium for a rational, profit-maximizing agent in the PGG is to contribute zero as a best response to all possible contribution levels of the other players', free-riding is far from always observed in real-world experiments. People cooperate much more in VCM games than classical theory can predict (Fischbacher et al., 2001 and Fischbacher and Gachter, 2010). Even in one-shot PGGs (Dawes and Thaler, 1988).

For example, Dawes and Thaler (1988) find that players on average contribute 40% to 60% of their endowment to the PG. In a simultaneous game with 44 participants Fischbacher et al. (2001) observe that about 50% of the participants exhibit conditional behavior such that the own contribution increases in other group members' average contributions. Only 30% are characterized as free-riders. The conditional contribution patterns of another about 14% are positive up to a certain level, and then decreases (i.e. "hump-shaped"). Applying this player classification, Fischbacher and Gachter (2010) classify 55% of 140 participants as conditional cooperators. Only 23% free-ride and always contribute zero. 12% are “triangle contributors” (similar to "hump-shaped") whose contributions increase with others’ up to a point and then decrease the more others contribute. (10% are non-classifiable.)

Moreover, in the simultaneous PGG with 112 subjects Gächter and Renner (2004) observe on average 42% contribution rates without a leader. Gächter et al. (2012) classify 46% out of 102 participants as free-riders, 26% as strong conditional cooperators and 22% as weak conditional cooperators. (6% are non-classifiable.) Similarly, in a repeated, simultaneous PGG with 64 participants Nowell and Tinkler (1994) observe about 44% contribution rates initially, but they drop to less than 25% in the last period. In a repeated PGG, Güth et al. (2007) observe about 42% contribution rates in the non-leader treatment (i.e., in the simultaneous PGG). Albeit some of these studies have small sample sizes limiting the generalizability of the results, the evidence indicates that leaders might induce more efficient social outcomes.

Social (other-regarding) preferences might explain deviations from zero contributions. [1] This implies that people are not solely motivated by maximizing their own payoff, but also care positively (or negatively) for others’ material payoffs (Fehr and Fischbacher, 2002). This translates into having others’ utilities enter into the own utility function, which is not taken into account in the classical utility function in Equation 2.1 above. In the economic literature social preferences are typically modeled as altruism, inequity aversion, guilt aversion (positive preferences), or envy or reciprocity (negative social preferences) (Croson and Gneezy, 2009).

However, social preferences are complex and Fischbacher et al. (2001) highlight the equally important observation that cooperation in PGGs often tend to decline in repeated games. This behavior is observed across various parameters and participant pools. The authors argue that the notion of conditional cooperation may explain the fragility of cooperation. This as the motivation underlying conditional cooperation depends directly on how others behave or are believed to behave. Thus, players’ respond to others’ defecting behavior with reciprocity, implying that both positive and negative social preferences may interact.

Due to heterogeneous social preferences for cooperation people differ in their willingness to behave conditionally cooperative. Some players behave almost as perfect conditional cooperators. Others behave conditionally up to a certain contribution level of others and then start to "free-ride". Some behave as imperfect conditional cooperators which means that their own contribution increases with the contribution of others, but remains lower (Fischbacher et al., 2001 and Gächter and Renner, 2004). At the extreme opposite ends, players who contribute all of their endowment regardless of others’ contributions are found at one end, while free-riders who contribute zero regardless of others’ contributions are found at the opposite end.

Yet, the presence of conditional cooperation highlights the important role a leader may play in inducing higher cooperation rates amongst followers in social dilemmas. Leaders may shape followers’ beliefs and thus shift behavior (Gächter and Renner, 2018). Croson (2007) founds a highly positive and statistically significant correlation between beliefs and contributions. Gächter and Renner (2018) elicit both leaders’ and followers’ beliefs to observe how leaders’ actions influence followers’ beliefs in a four-player VCM game.

Gächter and Renner (2018) highlight two channels through which leaders may influence followers’ behavior –

[1]Indeed, non-zero contributions may also arise due to decision error, resulting in an upwards bias of cooperative behavior.

both relying on the assumption that many people act as conditional cooperators. The first is the direct channel according to which conditionally cooperative followers match the leader's contribution. The second, indirect channel operates by followers believing that other followers will follow the leader's example and consequently follow the leader themselves. In my setup this would translate into followers' contributions being influenced both by the leader's contribution and by beliefs that the other follower will follow the leader. Gächter and Renner (2018) find evidence for both channels. Gender beliefs, if present, may thus influence cooperation. The first channel appears most important in early rounds as the leader's contribution shapes followers' initial beliefs and actual contributions very strongly. Followers beliefs' are more impacted by the leader's contribution than the own belief about other followers' contributions. Yet, in later rounds followers put more emphasize on other followers' contributions than the leader's, indicating the stronger role of the second channel.

Cartwright and Lovett (2014) also highlight the role of leading-by-example. Applying the sequential version of the PGG used by Fischbacher et al. (2001) in which three players contribute before a follower, and the leader-follower variant used by Gächter and Renner (2004) in which one player leads and three follow, they observe that conditional cooperation remains relatively stable not only to changes in the MPCR, but also to both types of games. Taken together with the finding that the level of MPCR has a strong effect on unconditional contributions, the authors conclude that their results highlight the critical role played by leaders in PGGs. This as followers' choices are more uncertain in the leader-follower version than in the sequential version of Fischbacher et al. (2001) where the follower observes all other contributions.

Yet, Gächter and Renner (2004), Arbak and Villeval (2007) and Gächter and Renner (2018) find that followers exploit their leaders by contributing significantly less. Gächter and Renner (2004), Güth et al. (2007) and Arbak and Villeval (2007) yet find that both leaders and followers contribute more in the leader game in comparison to the simultaneous game. However, Gächter and Renner (2018) observe that followers exploit leaders to the extent that overall group outcomes are not higher than for groups without a leader. Thus, leadership appears costly. After 10 rounds Arbak and Villeval (2007) also observe that followers' responsiveness decline. Only when the leader is endowed with exclusion power the decline of cooperation over time is prevented (Güth et al., 2007). Leaders with exclusion power also induce substantially higher contributions. Gächter et al. (2012) mean that leaders need to handle free-riding (in voluntary settings).

In sum, albeit exclusion power of the leader might be necessary to sustain cooperation, most findings indicate that contributions are still higher with a leader (without exclusion power) than without a leader. Thus in PGGs, leading-by-example appears to promote more efficient social outcomes at least initially. The mechanism appears to work through both channels introduced by Gächter and Renner (2018), supporting the findings by Fischbacher et al. (2001), Fischbacher and Gachter (2010), and Fischbacher and Gachter (2010) that many people behave as conditional cooperators. Studying a VCM game with more than two players is important to understand the second channel. Albeit a PD may also be used to understand coordination problems, a PGG with more than two players more resembles a situation similar to coordination challenges in our society.

Nevertheless, the focus has not been on gender differences in leading-by-example. Using mixed-sex interactions Gächter et al. (2012) find that after controlling for cooperative preferences, male leaders contribute insignificantly more than females leaders, but gender is not announced. Arbak and Villeval (2007) do not find a significant gender effect on leader contributions in the imposed leadership treatment (where gender is not announced), nor in the voluntary leadership treatments where gender is either announced and not. However, their focus is on self-selected leadership. That gender is a consistent determinant of willingness to lead only when gender is not announced highlights the importance of the context. This as Born et al. (2019) found that men are most willing to lead in a male-dominated sector environment where gender was announced.

2.2.2 Evidence of gender differences in cooperation in social dilemmas

A great many studies have examined gender differences in cooperative behavior (Aguiar et al., 2009; Balliet et al., 2011). Experimental evidence suggests that social behavior varies with gender (Cox and Deck, 2006), and that gender affect contributions to the PGG (Anderson et al., 2011). Some PGG findings indicate that women cooperate more (Nowell and Tinkler, 1994; Belianin et al., 2005; Cadsby et al., 2007; Andersen et al., 2008), others that men cooperate more (Brown-Kruse and Hummels, 1993; Hardy and Van Vugt, 2006; Anderson et al., 2011; Gächter et al., 2012), and others no differences (Jackson, 2001; Eek and Biel, 2003; Croson et al., 2008; Kube and Traxler, 2011.) (See Balliet et al. (2011) for more references and examples.)

Social psychologists have studied gender differences in social behavior, but to a lesser extent since the 1990s without any clear theory to work from. During the last decades behavioral economists have studied gender differences in cooperation (Balliet et al., 2011). However, gender's impact on behavior and social decision-making is still subject to much debate and scrutiny (Cox and Deck, 2006) and inconclusive (Ortmann and Tichy, 1999; Aguiar et al., 2009). The hypothesis that women exhibit stronger social preferences than men has been tested in PGs, ultimatum and dictator games. However, different ultimatum, dictator and PGs experiments differ both in magnitude and direction of gender differences with regards to such preferences (Aguiar et al., 2009). Moreover, the gender composition of the study sample has been found to affect experimental results (Ortmann and Tichy, 1999). Kopelman et al. (2002) concluded that the mixed findings indicated that gender may influence cooperation in social dilemmas, but the gender effect may be small and variable.

This is supported by recent meta-analyzes which offer new insight to the overall direction of gender differences. Balliet et al. (2011) review 272 studies, covering over 50 years of research, on gender differences in cooperation in social dilemmas. The overall effect is small, with women being marginally more cooperative, but they observe that male–male interactions are more cooperative than female–female interactions while women cooperate more than men in mixed-sex interactions. In a meta-analysis of gender differences in dictator games with 15,016 unique individual observations Bilén et al. (2020) report that women on average give 4 percentage points more than men. Finally, in a review of gender differences in cooperation in simultaneous PDs on Amazon Mechanical Turk, with 7,322 men and women in the US Capraro (2018) finds a small effect size (about 4 percentage points, but highly significant) indicating that women overall, marginally, cooperate more.

Albeit not reporting standardized effect sizes [2] Croson and Gneezy (2009) reviewed the economics literature on gender differences in cooperation. They report that women's social preferences were more sensitive to the experimental context than men's. This is why Balliet et al. (2011) focus on gender differences in cooperation with respect to contextual features. Referring to findings (see Weber et al., 2004; Hyde, 2005; Croson and Gneezy, 2009; Simpson and Van Vugt, 2009) they mean it is more productive to investigate the environment in which men and women make cooperative decisions rather than focusing on a main effect of gender. Hyde (2005) finds evidence against the dominating view that men and women are different psychologically. The author argues that men and women are similar on most, but not all, psychological traits. Results from a review of 46 meta-analyzes support this view, and that context (and age) matters. Hyde (2007) argues that not only the magnitude but also the direction of psychological gender differences depends on the context. Thus, possibly on the leading-by-example context.

[2] One advantage in meta-analyzes of using effect sizes is that one does not need to consider the varying sample sizes.

Balliet et al. (2011) also examine simultaneous versus sequential games. In repeated interactions men are more cooperative than women. However, some studies find that women contribute significantly more initially in repeated games (Mason et al., 1991; Cadsby and Maynes, 1998; Ortmann and Tichy, 1999). Balliet et al. (2011) find that women cooperate more in larger groups and in more recent studies, but these differences disappeared after controlling for several study characteristics. Thus, other contextual factors matter as well.

Moreover, results from the PGGs reviewed by Balliet et al. (2011) vary across same-sex and mixed-sex interactions. Gender is not always announced and many of the studies only involve same- or mixed-sex groups. An exception is Sell (1997) who examines the impact of the gender group composition in a simultaneous PGG with 254 participants with single-sex and female- or male-minority groups (a between-design with respect to the group). While women in same-sex groups on average contribute more than women in male-majority groups, men contribute more in female-majority groups than with other men only. However, when money is the resource in a similar setup with 99 participants Sell et al. (1993) instead observe that women in male-majority groups on average contribute more than women in same-sex groups. Men in same-sex groups also contribute more than men in female-majority groups. However, the gender composition of the group has no effect. Thus, also within studies the direction of gender differences vary across same-sex and mixed-sex groups.

In sum, the review in this section has shown indications of small gender differences in social preferences, with women on average being marginally more cooperative, but as gender differences in cooperation appears sensitive to contextual factors it is not clear whether women or men will be more prone to lead or/and follow in the PGG. Nor is not clear when women versus men cooperate more in same-sex versus mixed-sex interactions and thus how preferences for leading and following others might be effected by the gender group composition. One might expect women to also on average be better role models regarding leading-by-example, maybe due to stronger other-regarding preferences. Yet, findings by Balliet et al. (2011) of role of the gender group composition in social dilemmas also indicate the presence of gender identity being activated which may influence men and women to behave differently. As gender identity may also be important in the leadership context, gender stereotypes might be activated and affect beliefs which in turn may affect leader and follower behavior. In the next sections I present psychological theories on why gender differences in cooperation and leadership may emerge.

2.2.3 Gender preferences, roles and leading-by-example

Recently, sociocultural and evolutionary psychology theories have deepen the understanding of gender differences in social behavior. Applying evolutionary theory one might predict men to be better role models and more cooperative followers, at least in same-sex groups, as biological differences would have led our ancestors to keep a sexual division of labor where men hunted in groups and women were responsible for gathering food. Men would be more adapt to coordinate due to pressure to solve social dilemmas such as hunting and warfare in order to survive (Balliet et al., 2011). It is less clear regarding gender differences in mixed-sex groups. In contrast, applying sociocultural theory one might predict women to be better role models and followers in both same-sex and mixed-sex groups (Balliet et al., 2011) as women are perceived as less selfish and more caring (Eagly, 2009), while men typically assuming social roles of high status and power are perceived as more independent, assertive, ambitious, and dominant. Due to other social experiences and occupational roles, women would have developed more communal skills towards the welfare of the group. In mixed-sex groups, women would conform more to this expected role behavior and be even more cooperative.

The finding that women on average are more cooperative supports sociocultural theory. Yet, these theories cannot alone explain that women are more cooperative in mixed-groups and men in same-sex groups (Balliet et al., 2011) and introducing leadership adds to the complexity. However, gender differences in behavior might only appear if gender stereotypes/norms are activated (Balliet et al., 2011). Gender norms are social norms defining acceptable and appropriate actions for women and men in a given group or society (Cislaghi and Heise, 2020). This is interesting as Sell et al. (1993) note that the literature on sex differences in social dilemmas involves experimental evidence that consider the sex of participants rather than gender identity. Gender identity is defined as the extent to which an individual feels masculine or feminine (Lurye et al., 2008).

Gender identity may possibly be activated by the gender composition of the group, and/or by the leadership context which might induce gender specific role-based behavior and create gender differences in cooperative behavior. The leadership context may evoke gender role interpretations (Sell, 1993) influencing women to behave in stereotypically less masculine ways, and men in stereotypically more masculine ways (Van Anders et al., 2015). I.e., if the social dilemma setting would evoke gender role interpretations such that women are more communally oriented while men are more selfish according to sociocultural theory, women might confirm to the helping-female stereotype and set a good example both as leaders and followers. Alternatively, the social dilemma may induce men to behave as better role models and followers according to evolutionary theory, as coordination historically have been more vital for men.

While the meta-analyzes in the previous section indicated that women exhibit marginally more other-regarding preferences than men, possibly supporting the sociocultural perspective above rather than the evolutionary one, it is not evident what impact, if any, a leader-follower context will have on gender differences on preferences for cooperation in the PGG. Ben-Ner et al. (2004b) observe that women are more strongly influenced by the first-mover's decision than men in a sequential dictator game where participants play both roles and were anonymous to one another. However, Arbak and Villeval (2007) found that when self-selected male leader candidates were refused the role, due to multiple leaders, they became significantly more likely to revise their contributions downwards, regardless of the actual leader's contribution. Eliminated female leaders, on the contrary, only revised their contributions downwards when the actual leader's contribution was inferior to their own. Social-role theory argues that while there might be small gender differences in pro-social preferences, men and women differ as to why they engage in pro-social behavior as a result of different gender roles (Eagly and Crowley, 1986). This might explain why men and women react differently when "rejected" the leader role, and that they might react differently when playing both the leader and follower role in the PGG.

Furthermore, behavior in the social dilemma setting is closely linked to beliefs about others' cooperativeness. The leader-follower context may influence men's and women's beliefs about their influence as leaders differently, or gender stereotypes of cooperative behavior may be activated and affect leader and follower decisions with the gender group composition. I discuss this in the next section.

2.2.4 Gender beliefs, roles and leading-by-example

Social structural theory has been applied to understand gender differences in leadership (Eklund et al., 2017). The theory considers a society's gender division of labor to cause gender differences in behavior as these structures define the social constraints for men and women (Eagly and Wood, 1999), and posits that the differences in men's and women's normative roles affect their leadership behavior and outcomes. Gender identity may be important if the view that being a man is more in agreement with the leadership role than being a female (Eklund et al., 2017) persists in my setup. Applying this theory, one might hypothesize men to be better role models as leaders, but the opposite is possible, especially in male-dominated groups where gender identity may be more salient, if women wish to "defy" such gender views on leadership. On the other hand, the leadership context might also create gender heterogeneity in beliefs of the sexes potential "success" as leader, which may impact behavior if activated.

Historically, men have had higher status and may therefore have greater legitimacy as influence agents (Eklund et al., 2017). Applying status character theory it is therefore possible that men would be more confident leaders, and women possibly at risk of confirming a negative stereotype of women being regarded as less influential leaders (i.e. the stereotype threat, see Steele and Aronson (1995)). This as "gender differences in status are important determinants of influence because it relates to the perceived competency of the individual" (Eklund et al. 2017, p.138). The role of beliefs for gender differences in cooperation compared to social preferences is relatively unknown (Greig and Bohnet, 2009). Yet, in a leadership position a leader's decision might depend on the expected cooperativeness of followers, and followers' decisions on whether they expect other followers to follow the leader's example. Moreover, if the gender beliefs patterns of Born et al. (2019) also persist in the PGG setting of leading-by-example, and if beliefs exist such that women are believed to be more communally oriented, male leaders might expect most support in female-dominated groups, and female leaders least support in male-dominated groups. Previous literature indicates that gender beliefs exist and affect behavior, but it appears to depend on the experimental and cultural context.

For example, Aguiar et al. (2009) find that women to a larger extent (almost 80%) choose to cooperate with a another woman in a dictator game, while men only choose to play against another man in about 50% of the cases. This may indicate that beliefs such that women are more communally oriented affect cooperative behavior or that women expect less support from men, but the sample (68 participants) is quite small. Yet, Vyrastekova et al. (2015) found heterogeneity in gender beliefs such that men's beliefs of the contributions of other group members systematically increase with the number of other women in the group. However, women's beliefs on average decreased with the number other women in the group. Thus, female leaders might expect lower influence also from other females in my setup, but the sample is again quite small.

Yet, Greig and Bohnet (2009) find that female Nairobi slum dwellers contribute as much as men in same-sex groups, but less than men in mixed-sex groups. Mainly due to women having more pessimistic expectations of others' contributions in mixed-sex groups in a simultaneous PGG. Men do not distinguish between these groups, either in their expectations of others' contributions or in their behavior. These results differ in direction, both for contributions and beliefs, to those of Vyrastekova et al. (2015), highlighting the possible role of learning behavior shapes beliefs. (Greig and Bohnet (2009) note that Nairobi is a society were the provision of PG often hinges upon voluntary coordination and women are used to cooperate more with other women in every-day life.) In such a setting female leaders would perhaps expect less influence on followers in mixed-sex groups only.

Boschini et al. (2012) find that the activation of gender stereotypes creates no gender difference in generosity single-sex groups in a dictator game. However, gender-primed men become less generous in mixed-sex interactions while women instead become more generous. Further, men are most impacted by this priming effect. However, Ben-Ner et al. (2004a) find that women in a dictator game sent less to known female recipients than they did to known male recipients and to persons about whom no information was provided. Information of the recipient's gender did not affect to male dictators. However, the sample size is quite small.

Nevertheless, gender stereotypes might impact follower behavior in the sequential PGG depending on the leader's gender and the gender group composition. If men believe women are more cooperative in the PGG, male followers might for a given leader contribution, contribute more if both the leader and the other follower are females, or if the leader is a male and the other follower a female, compared to a situation where the leader is a female and the other follower a male. Accordingly, if a female follower believes females cooperate less in same-sex groups, she might contribute more if the leader is a male and the other follower a female, than if both the leader and the other follower are females.

In a simultaneous PGG van Staveren et al. (2009) find that gender beliefs affect behavior in single-sex groups where the group composition was announced, controlling for risk-aversion and pro-social preferences. Despite a small sample (42 participants) this is consistent with the finding of Chermak and Krause (2002) that gender matters when individuals know the roles they are to play in a PGG modeling common pool resources. In those treatments women are more generous (take less) than men. However, their findings (and those of van Staveren et al. (2009)) on gender announcement contrast those of Arbak and Villeval (2007) and it is not clear if and when gender beliefs would be activated in this study as found by Vyrastekova et al. (2015).

2.2.5 Gender, risk-aversion and leading-by-example

Contribution decisions in the sequential leader-follower PGG are associated with uncertainty. The leader's decision situation is similar to the one in a simultaneous PGG. However, an amount of uncertainty prevails also for the followers as they do not observe the other follower's contribution. Therefore, more risk-averse individuals might, ceteris paribus, contribute less to the PG.

However, with regards to gender differences in risk attitude Nelson (2016) does not find strong evidence that women are more risk-averse than men when re-examining experiments on gender differences in risk taking. Instead, the author finds substantial similarity and overlap between the distributions of men and women in risk taking, and the difference in means is not substantively large. Earlier, Charness and Gneezy (2012) wrote that they found strong evidence that women are more risk-averse than men when reviewing 15 sets of experiments with one simple underlying investment game. However, using the same data with improved statistical techniques Nelson (2016) does not find this to be the case.

Furthermore, using a lottery with high or low risks, and varying the payoffs Holt and Laury (2002) find that women are less risk-averse than men regarding low payoffs, but equally risk-averse in the high payoffs domain. Schubert et al. (1999) find that women are equally risk-averse as men, or more risk-averse than men depending on the contextual frame. van Staveren et al. (2009) elicit risk preferences using the more complex lottery game by Holt and Laury (2002) and find a small difference in contributions to a simultaneous PGG between risk-averse women and men. However, the interaction coefficient of female and risk-averse is much smaller in magnitude and significance compared to those for pro-social preferences and beliefs about others' contributions. They find no main effect of either risk-aversion or gender.

One might overall expect risk-aversion to be negatively associated with contributions in the PGG. However, Gächter et al. (2012) find the opposite. When adding risk-aversion (measured according to the 11-point scale by Dohmen et al. (2005)) as a control in addition to pro-social preferences for cooperation and positive beliefs about the follower's contribution to a regression on leader contributions, the coefficient of risk-aversion is positive, close to zero and not statistically significant.

In sum, risk-aversion appears to have a small effect, if any, on contributions to the PG, as well as an interaction effect with gender. This, after controlling for pro-social preferences for cooperation and beliefs.

3 Research strategy and design

This section first describes scope for this study and the choice of study sample, then detailing the experimental setup of the game and describing the study information, specific tasks and payments to participants. It ends with presenting how the data was collected and processed. As the literature indicated that the gender group composition might impact the sexes differently, that perceived beliefs of others' behavior affect contributions and that a potential gender heterogeneity in gender beliefs may exist, I focus on these aspects in this study. It is out of the scope of this study to also examine repeated interactions albeit gender differences have been found in that context. By using the strategy method and elicit cooperative schedules in the follower role, I will arguably obtain some understanding of gender differences in response to others' behavior.

3.1 Sampling frame

Participants were recruited from the UK crowd-sourcing platform Prolific Academic. Eligible participants were those who reported "UK" as their nationality in Prolific's pre-screening indicators. I imposed this restriction as previous research has found that culture predicts contributions to the PG. E.g., Cadsby et al. (2007) find that Canadians and females are more successful at providing the PG than Japanese. Anderson et al. (2011) find that Czechs and men on average contribute more than Americans and women. This is expected since different countries not only are assumed to differ in gender beliefs/inequality, but also along the individualism–collectivism dimension. Including more countries might yield more noisy data. Indeed, this must be weighted against the loss of generalization of the results. If UK participants systematically have different social preferences from the wider population, the results cannot be generalized. However, arguably a sample population from Prolific still entails more variation than, for example, a student sample.

I impose no other restrictions on the study sample. Not including age and e.g., income levels will give more noisy results in form of omitted variable bias and confounding variables. However, I will not have sufficient statistical power to perform heterogeneous analyzes on subsamples of age groups.

3.1.1 Statistical power

The use of an within-design is based on previous findings of the between-gender effect size, and that I will consider an result to be statistically significant if the associated p value is less than 1% for all tests in this study. In the meta-analytic review of gender differences in cooperation by Balliet et al. (2011) (studying PD's, PGG, and other resource dilemmas) the reported overall effect size indicates that women are marginally more cooperative than men (Cohen's $d = -0.05$). (They write that many of the reviewed studies had too small sample sizes to detect such a small effect size.) The d value they use is the pooled standardized effect size, corrected for sample size bias. I.e., it is the difference in means between men versus women divided by the pooled standard deviation as in Cohen's standardised d, but corrected for small sample bias with (Hedges Olkin, 1985). However, when the sample size is 30 or more the two are approximately equal (Walker, 2015).

However, for male–male interactions Cohen's $d = 0.16$ indicating they are more cooperative than female–female interactions. For mixed-sex interactions Cohen's $d = -0.22$ indicating that women cooperate more than men in mixed-sex interactions. Bilén et al. (2020) report an overall Cohen's $d = 0.16$. To have sufficient statistical power to detect smaller effect sizes, I use a mixed-design with repeated measures for each individual. A priori I made an power analysis using an ANOVA with between-within interactions as a proxy for my design. [3] With an alpha level of 1%, power of 80%, two groups and three repeated measures and assuming a correlation (a priori) between the repeated measures around 0.5 and a nonsphericity correction around 1 to be conservative, I would be able to detect an overall difference around Cohen's $d = 0.16$. [4] For the some of the between effects I will be under-powered for such low effect sizes. Albeit Balliet et al. (2011) find much higher effect sizes for some of the reviewed experiments, the review of gender cooperation in simultaneous PDs by (Capraro, 2018) also indicates a small overall effect that women marginally cooperate more.

3.2 The experimental game

Participants play the sequential (leader-follower) variant of the linear PGG introduced in Equation 2.1. I set $N = 3$, and with $r = 0.5$ the MCPR equals 0.5. Each player is endowed with 5 tokens to allocate between their personal fund and a group fund. Thus player i's payoff π_i in the game is given by

$$\pi_i = 5 - c_i + \frac{1}{N}\sum_{j=1}^{N}(1+r)c_j = 5 - c_i + 0.5\sum_{j=1}^{3} c_j. \tag{3.1}$$

In this model the group's payoff is maximized if all players contribute all their endowment. However, if no player contributes to the PG all players are worse of than had all contributed their entire endowment. I.e., if all three players contribute 5 tokens each, each player earns 7.5 tokens. However, if none of the players contributes any of their endowment, they earn 5 tokens each. If two players contribute all 5 tokens to the project and one player nothing, the total payoff to be distributed evenly amongst the three players amounts to 5 tokens. Thus, the player who contributed nothing earns 10 tokens whereas the other two players 5 tokens each.

I follow the viewpoint that cooperation is harder to sustain with larger groups. For example, Gächter et al. (2012) set it to $r = 0$ with two players, Gächter and Renner (2018) set $r = 0.6$ so the MCPR equals 0.4 with four players. However I set $r = 0.5$ so that the MCPR equals 0.5 for simplicity for participants in the online experiment. By setting the MCPR to 0.5 I reduce the incentives to cooperate compared to Vyrastekova et al. (2015) who set it to 0.6 for three players.

Unconditional contributions to the PGG vary with the MPCR (Isaac and Walker, 1988; Isaac et al., 1994; Goeree et al., 2002; Cartwright and Lovett, 2014), whereas conditional contributions remains more stable to changes in the MPCR (Cartwright and Lovett, 2014). Varying the MPCR from 0.2 to 0.4 to 0.8 Cartwright and Lovett (2014) find that the level of MPCR has a strong effect on unconditional contributions. If these results would generalize into other settings one might suspect that the MPCR has most impact on the leader's contribution (the unconditional one) and in small groups. In sum, the findings imply that the MPCR level needs to be considered when evaluating leader contributions, but it might not affect women and men differently.

Leadership is introduced by letting one player make his/her contribution before the other members. Participants play both the leader and follower role. In the leader role participants are asked to make one direct contribution. Using the strategy method, participants in the follower role privately state their conditional contribution for each possible leader contribution level. This resembles a situation where one player (the leader) first make an unconditional contribution and the two followers simultaneously make their decisions upon observing the leader's decision. Following Gächter et al. (2012) the terms "leader" or "follower" are not mentioned to

[3] I am using the statistical software G*Power 3 by Faul et al. (2007).

[4] Due to a misspecification error on my part this value was wrongly stated in the pre-analysis plan.

participants anywhere in the experiment. Instead, "first-mover" or "second mover" are used to describe the different roles. Participants also complete a non-incentivized prediction task in the leader role. Specifically, participants were asked to guess the average contribution of the two followers in response to the their chosen leader contribution.

Incentivized beliefs have shown to more accurately match contributions in the PGG than non-incentivized ones (Gächter and Renner, 2010). Yet, incentivized beliefs affect contributions – positively for them, but negatively for Croson (2000) they note. Importantly, Gächter and Renner (2010) find no difference in contribution levels when beliefs are not incentivized compared to the non belief-treatment. Since I do not want to infer strategic considerations, I choose not do incentivize beliefs.

I use a 2x3 mixed-design where gender is the between-participant factor (two levels) and the within-participant factor is the gender group composition (three levels). Male-dominated groups are denoted "MMF" and female-dominated groups "FFM". All female or all male groups are denoted "FFF" or "MMM". Thus, each man and woman is paired with two men (MM), one man and one woman (FM) and two women (FF). Participants play all roles in all groups amounting to nine decision tasks corresponding to one leader decision and two follower decisions in each group (see Figure 3.1 below). With no feedback between decisions the observations can be treated as independent statistical units thus raising the statistical power.

These nine tasks were shown on separate decision screens. The order was randomized at the individual level to avoid potential ordering effects. For each decision, each player received 5 tokens which were translated into earnings as follows: 1 token = £0.1. At the end, a randomly chosen player was selected for which the first-mover contribution became the relevant contribution for payment. For the other two members their corresponding conditional contribution level as followers became the relevant decision for payment.

A participant i earned the *maximum* bonus amount if he/she contributed 0 tokens while the two group members both contributed 5 tokens. In this case, participant i earned 10 tokens (£1). Participant i earned the *minimum* amount if he/she contributed 5 tokens while the two group members both contributed 0 tokens. In this case, participant i earned 2.5 tokens (£0.25). The participation fee was set to £3.25. Thus, total earnings ranged from £3.5 to £4.25. The estimated completion time was 20 minutes.

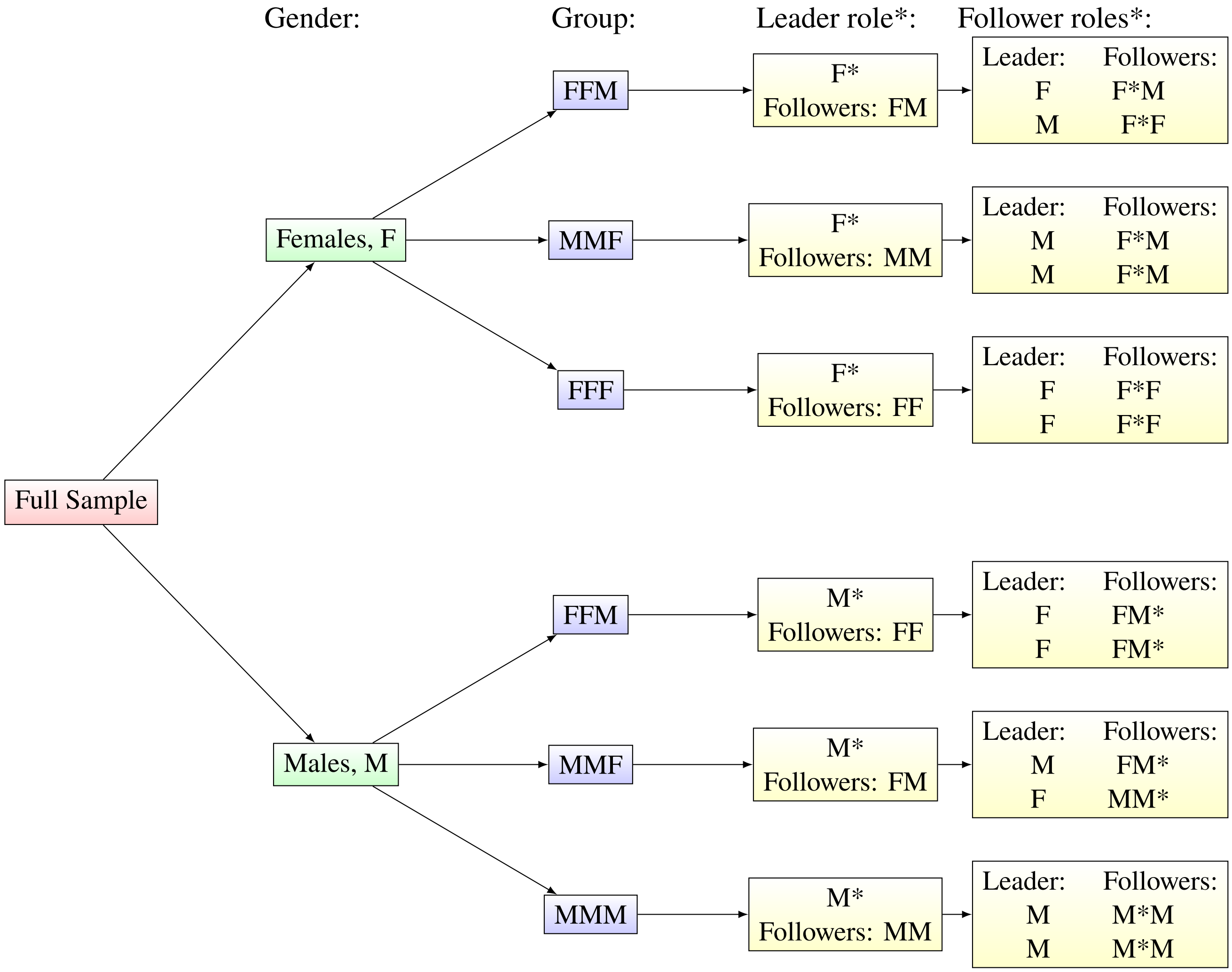
Two sessions, one for men and one for women, were launched simultaneously on Prolific. The experimental instructions and procedures were the same for men and women (see Figure 3.2 below for an overview of the game pages). They were launched under the same public name on Prolific ("Social decision-making study") but with internal, separate, names. Participants first saw a short introduction on Prolific and then clicked on the external survey link if they decided to participate.

3.2.1 Information and consent

Upon redirection to the experimental pages participants were asked to provide their Prolific ID and answer two background questions as follows: *Please fill in the following information.*

- Your Prolific ID:
- Your legal gender:
- Your age:

Figure 3.1: Flowchart over the experimental procedure and the different pages.



Upon answering these questions participants proceeded to a general information and consent page. Participants were informed that they would participate in a social-decision making study with the purpose of examining how individuals make decisions within a group context. In addition, participants were informed they would be randomly matched to three different groups with two other real Prolific players and make a total of nine decisions. At this stage gender was introduced to participants as follows:

Participation is anonymous. Your decisions and Prolific ID will never be revealed to the other participants. Instead, you will be given a random alias which will be shown to your group members. If you are a man, you will be given an alias like "Peter". If you are a woman, you will be given an alias like "Lisa".

To make gender a less salient feature, gender-neutral avatar pictures were shown above each alias. On each of the nine decision screens relevant aliases and gender-neutral avatar pictures of relevant group members and the individual player were shown. I used names over using gender avatars, or writing "female/male player", to make gender less salient given that participants made many decisions. Using names might introduce noise if certain names evoke positive/negative associations to participants. However, Holm (2000) finds no statistically significant difference when using female and male first and last names versus writing "female/male participant" in a two-player battle of the sexes game. I choose common UK names which may not encapsulate different cultural backgrounds in the UK. However, I would argue that it would not affect a gender effect unless I would have an disproportional larger number of men or women from a different cultural background.

Upon agreeing to the terms and conditions, participants continued to two pages containing information and examples of the game to be played. The first page informed of the general decision situation in the PGG and how earnings were determined. The second information page introduced the specific decision situations in the different roles, and the prediction task. To make sure all participants understood how their earnings were determined, they had to answer three control questions correctly before proceeding to the actual decision screens.

After the control questions participants were randomly assigned an alias they kept throughout the experiment. In addition, they were randomly assigned other aliases representing the other players. A woman was randomly assigned two female aliases for the FFF-group, one female and one male alias for the FFM-group and two male aliases for the MMF-group. A man was assigned two male aliases for the MMF-group, one female and one male alias for the MMF-group and two female aliases for the FFM-group. Aliases were unique for each gender group composition. The following set of aliases were used:

- **Male aliases:** David, William, Jack, Adam
- **Female aliases:** Jessica, Sara, Elizabeth, Emma

3.2.2 The decision tasks

After receiving an alias participants entered the decision loop with the nine screens. Images of these decision screens with example aliases are shown in Appendix A. For each first-mover role participants were asked to state:

You have 5 Tokens to allocate. Your contribution to the group project as the FIRST MOVER:

Your guess of the SECOND MOVERS average contribution to the project:
(Any whole number from 0 to 5 is possible to enter.)

For each of the second mover roles participants were asked to state:

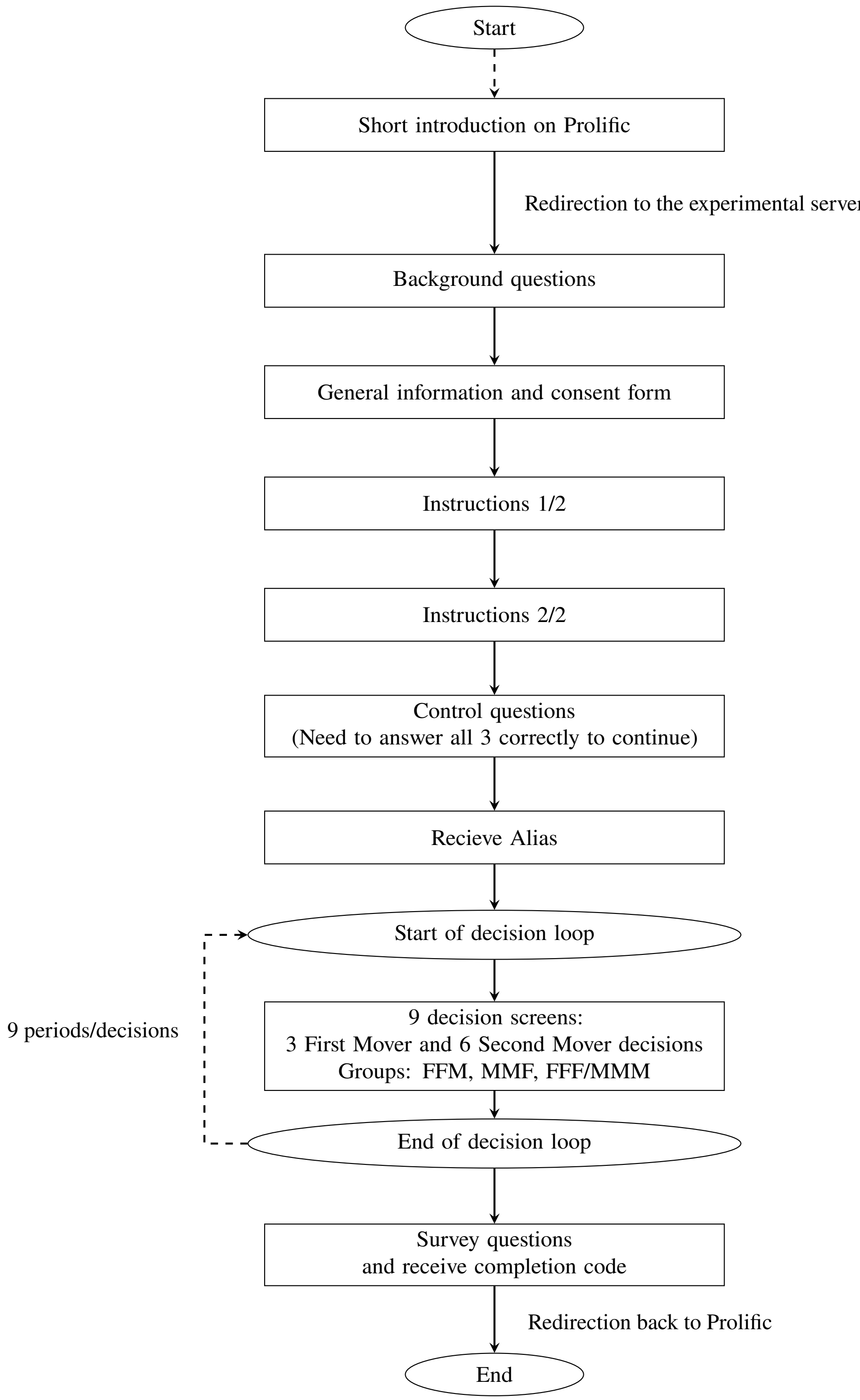
Please indicate, for each possible FIRST MOVER contribution below, how much you want to contribute to the group project:
(You can enter any whole number from 0 to 5 for each level.)

0:	**1:**	**2:**	**3:**	**4:**	**5:**

3.2.3 Survey questions

After having submitted their decisions, participants were asked to report what they believed the study was about. In addition, participants were asked to respond to a non-incentivized survey question about their attitude towards risk. This self-assessment of risk attitude was elicited using the question by Dohmen et al. (2005) which asks about willingness to take risks on an 11-point scale:

Figure 3.2: Overview of the experimental game flow for participants.



Question

You are now asked to indicate your attitude towards on the scale below. We ask you to think about: Are you generally a person who is fully prepared to take risks or do you try to avoid taking risks? 10 that you are fully prepared to take risks, 0 means that you are unwilling to take risks.

I am unwilling to take risks.										I am fully prepared to take risks.
0	1	2	3	4	5	6	7	8	9	10

3.2.4 Bonus payments

As stated in the pre-anaysis plan one of the groups was randomly drawn to determine payments according to the numbers (1:MMF, 2:FFM or 3:FFF/MMM). [5] I asked my supervisor to randomly draw one of the numbers one, two or three. Two was drawn. Participants were then matched according to their own aliases and those of their group members. Each participant was given two random numbers between zero and one in the Lioness software. Women and men were first sorted according the first random number in descending order within their alias group to form groups. The participant with the highest value of the second random number was then selected as the leader and thus the relevant role for payment was the first-mover decision. For the other two group members their second-mover decisions for the first mover's contribution level determined the earnings.

3.3 Data collection and processing

The main data collection took place on Prolific between February 27 and March 3, 2021. It stopped once I had 175 complete responses from men and women, respectively. The pre-analysis plan for this study was registered on the Open Science Framework (OSF) on February 26 (available at https://osf.io./3xqbh). A pilot-study was launched on Prolific on February 9 with four men and four women to assess the estimated completion time and verify that it did run smoothly. (Seven of eight participants completed the study in less than 20 minutes.) A second pilot was launched on Prolific on February 16 with two men and two women, as two participants from the first pilot could not see all images. This was not a problem in the second pilot and only two participants stated that they could not not see the images in the main data collection. The response data from the pilot studies was not analyzed. Participants in the pilot studies were excluded from taking part in the main study.

By using the third party Lioness, participants' response data cannot be linked to the personal data Prolific keeps of its participants as the response data cannot be accessed by Prolific, nor by Lioness. Prolific gives participants a 24-alphanumeric character ID which was used to pay participants on Prolific. After participants had been paid the Prolific ID was deleted from the response data before the statistical analysis. A new participant ID was instead added to the response data. To protect the Prolific ID an experimental server was rented and set up using a virtual server on the Google Cloud Platform. This after receiving verification from Lioness that Google, nor anyone else, could access the experimental database. (The enable setup function of the database was deleted for this purpose.) After the data collection on Prolific had finished and the database

[5] In the pilot studies I paid out the full bonus for practical reasons, but in the main study I calculated all bonuses manually as stated to avoid deception.

was downloaded, the server was shut down and deleted.

I did not impose a time restriction as I wanted participants to think through their decisions. Lioness prevents participants from the same household to take part in an experiment, but to further limit the risk of robot players I restricted the number of trials on the quiz to eight and used the two open-ended survey questions to verify the quality of the answers. Six men and 16 women were removed from the experiment in the quiz state or did not answer all questions. The participation fee was paid out to these participants for their time invested. This was not specified in the pre-analyis plan, but I did not allow re-entering to protect the integrity of the results. I therefore collected new, complete responses for the analysis. The average completion time was 18 minutes and participants earned on average £0.67 in bonus. Thus, average total earnings equaled £3.92 which is high on Prolific.

4 Empirical analysis

This section first presents some characteristics of the participant pool in Section 4.1 before proceeding to the main analysis. In the latter, I present the hypotheses being tested and the associated analytical specifications as specified in the pre-analysis plan, as well as the results. To address the first research question I examine gender differences in *leader behavior* in Section 4.2. I explore overall gender differences in contributions and compare how different gender group compositions affect women's and men's decisions. Secondly, the association between gender beliefs and leader contributions across different gender group compositions is investigated. Thirdly, I explore whether there is a gender heterogeneity effect of risk-aversion on leader contributions depending on the gender group composition. To address the second research question I turn to *follower behavior* in Section 4.3 and examine whether participants behave as conditional cooperators, as well as the degree of conditional cooperation amongst men and women in different gender group compositions with a female versus a male leader. This section ends with an exploratory analysis of *social preferences for cooperation* in Section 4.4 where I investigate whether more cooperative followers also make better role models as leaders.

4.0.1 Robustness checks

For all analytical specifications regarding leader and follower behavior, as well as for social preferences for cooperation, the robustness of the results are verified by a subgroup analysis only on those participants who did not mention "gender" or "sex" in the survey question regarding what they thought the study was about. 18 men and 18 women reported that they thought the study was about gender, of whom 10 men and nine women only stated gender. I also code one man as having stated gender as the participant stated gender as female. Two participants said that they saw no avatar images, but I code six additional ones as such as it was not clear from their responses whether they had seen all images.

Additionally, I exclude participants with short completion times for all specifications. Having a short completion time is classified as having completed the study below one standard deviation from the average time (sd = 9.29), i.e. in less than nine minutes. This is an addition to the pre-analysis plan. 14 men and 14 women in total completed the study in less than nine minutes, another six men and 12 women in less than 10 minutes which I exclude as well to check the robustness of the results. Four men and eight women completed it in less than nine minutes and another eight men and five women in less than eight minutes. One woman took less than seven minutes to complete it and two men less than six minutes. I do not think that these were robots, given their replies to the survey questions and their correct answers to the control questions, but the data quality is a concern. When testing the experiment on friends not familiar with the PGG or economics, it took them approximately 10 minutes to read all the instructions and answer the control questions correctly. They read the instructions carefully and asked questions.

4.1 Sample characteristics

I start the analysis by presenting sample characteristics with regards to age by gender. As depicted in Table 4.1, women in the sample were on average slightly younger than men. The age range is quite large, ranging from 18 to 75 for men, and for 18 to 69 for women. On average, men and women in the sample were 35–40 years old.

Table 4.1: Descriptive statistics for participants' age.

	mean	median	sd	min	max	n
full sample	37.85	34	14.3	18	75	350
men	39.6	35	15.0	18	75	175
women	36.1	32	13.5	18	69	175

4.2 Leader behavior

In this section I first present the primary hypotheses regarding gender and leader contributions. Next, I present the secondary hypotheses with regards to gender and leader beliefs. Last, I present the exploratory hypothesis with regards to the relationship between risk attitudes, gender and leader contributions.

4.2.1 Primary hypotheses – gender and leader contributions

I first test whether there is a main effect of gender on average leader contributions. Specifically, if the amount of tokens contributed in the leader role is either higher or lower across all gender group compositions for women than men. I then test whether men and women increase their leader contribution with the number of women in the group, such that there is a main effect of the gender group composition. Thirdly, I test whether men and women respond differently to the gender group composition. An interaction effect would imply that the effect of gender on leader contribution should not be interpreted without considering the gender group composition. Two-sided hypotheses tests are used since it was not clear a priori whether women would contribute more or whether there would be no gender differences at all.

Hypothesis 1 *There is no main effect of gender on the average leader contribution across all three gender group compositions (a between-subjects effect).*

Hypothesis 2 *There is no effect of the gender composition of the group on leader contributions for either men or women (a within-subjects effect).*

Hypothesis 3 *There is no interaction effect of gender and the gender composition of the group on leader contributions (a within-between subjects effect).*

If Hypothesis 3 is rejected I use pairwise t tests to test which groups that differ from each other. To correct for multiple hypothesis testing the Benjamini and Hochberg (1995) method is used. If the interaction is not significant I test if there is a main effect of gender, or the gender group composition. If there is a main effect of the gender group composition, pairwise t tests are used to test Hypothesis 2.

The Benjamini and Hochberg (1995) method controls the false-discovery rate (FDR), i.e. the proportion of false-positives whereas e.g., the Bonferroni and Bonferroni-Holm corrections control the family-wise error rate (FWER), i.e. the chance that one of the findings will be a false-positive. FWER methods have the disadvantage that the loss of power is greater than for FDR methods (Benjamini and Hochberg, 1995). E.g., "the cost of going from a single test with 80% power to two Bonferroni-adjusted tests is always a reduction of power to 71% for a given effect size and sample size. The number does not depend on the type of test, statistical model, effect size or sample size" (Lazzeroni and Ray, 2012, p.110).

The Bonferroni-Holm correction has higher power than the pure Bonferroni method, but Bonferroni methods are inflexible when hypotheses are highly correlated (Chen et al. (2017)), which is the case in my design. Albeit the FWER is more conservative on false-positives, I choose the FDR to avoid loss of power and since I want to obtain as many discoveries as possible with respect to gender and the gender group composition, but yet want to avoid too large a fraction of false-positives which could result in wrong overall conclusions (Benjamini and Hochberg, 1995). When the number of tests is not too large the difference between FWER and FDR corrections is however smaller due to a smaller loss of power associated with the FWER.

4.2.2 Secondary hypotheses – gender and leader beliefs

A secondary objective is to assess the presence of gender beliefs and their role in explaining contributions to the PG. If Hypothesis 1 is rejected I test for overall gender differences in beliefs about support from followers. If Hypothesis 2 is rejected I test whether gender beliefs exist such that women anticipate least support in male-dominated groups and men most support in female-dominated groups. Finally, if Hypothesis 3 is rejected I follow Vyrastekova et al. (2015) and test for heterogeneity in gender beliefs such that men's beliefs of the average contribution of the followers increase more with the number of women in the group than for women. Two-sided hypotheses tests are used since it was not clear a priori whether beliefs would increase with the number of women in the group, or if men would hold stronger gender beliefs than women.

Hypothesis 6a *There is no main effect of gender on leaders' belief such that female or male leaders on average expect lower support across all gender group compositions.*

Hypothesis 6b *Participants' reported belief of the average contribution of the followers do not systematically increase or decrease with the number of women in the group, for either men or women.*

Hypothesis 6c *There is no gender heterogeneity in gender beliefs such that men/women hold stronger gender beliefs than women/men.*

4.2.3 Analytical specifications

To formally analyze leader behavior, i.e., leader contributions and beliefs (as well as risk attitudes) I use OLS regressions as stated in the pre-analysis plan. Cluster (heteroskedasticity) robust standard errors (Wooldridge, 2012, p.501) are used with the individual as the independent clustering unit level. This assumes that observations and their residuals (as well as the residual variance) are correlated within an individual, but independent across individuals.

For leader contributions and beliefs I have three repeated measures per participant. Since the error terms are correlated the OLS normality assumption of independently distributed errors is violated. In terms of a pure autocorrelation (i.e. when the correlation between residuals is not caused by an omitted variable) the OLS estimates will still be unbiased and consistent, but the standard errors will be inconsistent and the confidence intervals will not be precise even asymptotically. Thus, OLS estimates are no longer efficient (Wooldridge, 2012). In case an omitted variable is correlated with any of the independent variables, OLS estimates are still consistent, but not necessarily unbiased under the presence of serial correlation between observations (Wooldridge, 2012, p.467). The latter case may occur if e.g., pro-social preferences for cooperation would be correlated with gender. Cluster-robust variance is one way to deal with the autocorrelation. With 350 clusters (350 participants) the number is enough to use this approach.

Mixed-effects models can also handle this dependence between observations. They include individual random effects or/and individual random slopes depending on the data structure. Thus, individual effects are controlled for directly in the model. However, mixed-effects models involve complex fitting procedures (Schielzeth et al., 2020). This involves estimating models with random intercepts and/or random individual slopes which may also vary across the group treatment, and with gender (since one might expect that the effect of the gender composition of the group is individual to some extent), to select the model that best fits the data.

Mixed-effects models also rely on strong assumptions, e.g. that the random effect coefficients are independent from the other covariates and that the random effect errors are identically distributed to ensure the validity of the model (Schielzeth et al., 2020). I.e., the individual-specific random effects cannot be correlated with any of the other covariates. Linear mixed-effects models provide unbiased estimates if data or individual observations in a treatment condition are missing at random, but this is not a problem in my data.

I therefore use cluster-robust inference to alter the covariance structure of the residuals to account for the within-cluster error correlation (Steiger (2010); Glen (2017)). I am interested in the population average effect for men and women, and those results are more complicated to interpret from mixed-effects models. Fischbacher and Gachter (2010) estimate OLS with clustered standard errors at the session level as the independent unit of observation, as well as models with random and fixed effects and with Tobit. However, the estimation results are very similar and they only report OLS for ease of interpretation.

However, linear models require continuous data. As contributions and beliefs take on a limited number of non-negative values (as the amount of tokens only ranges from 0 to 5), presumably with a non-trival number of observations left-censored at 0 due to free-riding, the data cannot be normalized taking the logarithm of the dependent variable and linear estimates will only have asymptotic justification (Wooldridge, 2012). In addition, due to left- and/or right-censoring contributions do not necessarily provide an accurate description of an individual's willingness to cooperate, especially as I have a small variance in the dependent variable. I.e., a willingness to contribute 4/5 tokens differs from a willingness to contribute 4/20 tokens.

OLS estimates will however be biased and inconsistent with a left- or/and right-censored dependent variable (Wooldridge, 2012). Tobit models can handle this data structure. Yet, Boschini et al. (2012) use both a Tobit regression with random effects and an interval regression with clustered standard errors at the individual unit in their dictator game on gender priming and find essentially the same effect (they have three repeated measures per individual using three price levels). Since the OLS bias will be small (also observed by Fischbacher and Gachter (2010)), I use OLS with contributions and beliefs in percentage of tokens (in decimal format) to render the variables quantitative meaning (Wooldrigde, 2012) albeit the residuals will not be normalized.

In the regressions below, I furthermore treat the gender group composition as a continuous variable rather than as categorical since continuous data generates higher power and thus lower p values as the precision of estimates increases. Thus, since the continuous model requires a smaller sample size it might detect statistically significant effects which a category model would miss. Continuous data also offers a more parsimonious model. Treating the gender group composition as continuous implies that I run an interval regression, i.e. I assume that increasing the number of females from 0 to 1 has the same effect as increasing the number from 1 to 2. Extrapolating to groups with higher values must be done cautiously (Lazic, 2008).

Treating the gender group composition as categorical would result in more estimates (two dummies for FM and FF with MM excluded as the control group) compared to it being treated as continuous. Since one degree of freedom is lost for each estimated parameter the model with the continuous variable has more degrees of freedom (and thus a higher F value) and higher power. (Thus the ANOVA-power analysis is conservative.) Treating it as continuous does not increase the likelihood of false-positives as this rate is set a prioi (Lazic, 2008).

4.2.3.1 Leader contributions

Specifically, Hypotheses 1-3 regarding leader contributions are tested by the following regression equation.

$$LC_i = \beta_0 + \beta_1 Female_i + \beta_2 NbFemales + \beta_3 Female_i \cdot NbFemales + \varepsilon_i, \tag{4.1}$$

where LC_i is the percentage of tokens (in decimal format) contributed to the PG in the leader role for participant i. $Female_i$ is an indicator equal to 1 if participant i is a female and 0 otherwise. $NbFemales$ [6] is a continuous variable equal to the number of other women in the group (0, 1 or 2). $Female_i * NbFemales$ is an interaction term equal to 0 if participant i is a male, or if participant i is a female and the number of other women in the group is 0 (MM). It is equal to 1 if if participant i is a female and the number of other women is 1 (FM). Finally, it equals 2 if participant i is a female and the number of other women is 2 (FF). e_i represents the individual random variation. A formal statement of Hypotheses 1–3 are provided in Appendix B.

Table 4.2: Specification of coefficients from regression (3.1).

Gender	MM	FM	FF
Male	β_0	$\beta_0 + \beta_2$	$2\beta_2$
Female	$\beta_0 + \beta_1$	$\beta_0 + \beta_1 + \beta_2 + \beta_3$	$\beta_0 + \beta_1 + 2\beta_2 + 2\beta_3$

4.2.3.2 Gender and leader beliefs

Next I assess the presence of gender beliefs and their role in explaining leader contributions to the PG.

[6] In the pre-analysis plan $NbFemales$ was denoted $NoFemales$. I changed the name to avoid confusion.

Hypotheses 6a–6c concerning leader beliefs are tested with the regression equation below. I stated in the pre-anaylsis plan that if the interaction term in Equation 4.1 would significant and/or if there is a main effect of the gender group composition I test the hypothesis that such effects are driven by beliefs as stated by

$$LB_i = \beta_0 + \beta_1 Female_i + \beta_2 NbFemales + \beta_3 Female_i \cdot NbFemales + \varepsilon_i, \tag{4.2}$$

where LB is participant i's reported belief in group j in the leader role of the average percentage of tokens (in decimal format) contributed by the two followers for the own leader contribution. (The formal statements of these hypotheses correspond to Hypotheses 1–3.) Since Hypotheses 1–3 were not rejected this is only an exploratory analysis as stated in the pre-analysis plan.

4.2.4 Results

4.2.4.1 Summary statistics

Before formally presenting the regression results I start with a visual inspection of leader contributions and beliefs. Table 4.3 below shows average the leader contributions (LC) and leader beliefs (LB) in each gender group composition, for men and women separately, and for the full sample. It shows small gender differences in leader contributions across all three group compositions with men, on average, contributing slightly more in each group than women. The variation in the dependent variable is also very similar. This suggests that there is no main effect of gender on leader contributions. Nor does it appear to be a main effect of the gender group composition such that either sex contributes systematically more/less with the number of other women in the group. As these variables have a limited amount of values the median might be more informative. The medians in all groups are the same apart being higher for men in same-sex groups (equal to 5). That the gender differences in mean and median differ across the gender group compositions may indicate a small interaction effect, but as seen in Figure 4.1 below the distributions by gender across all groups are similar.

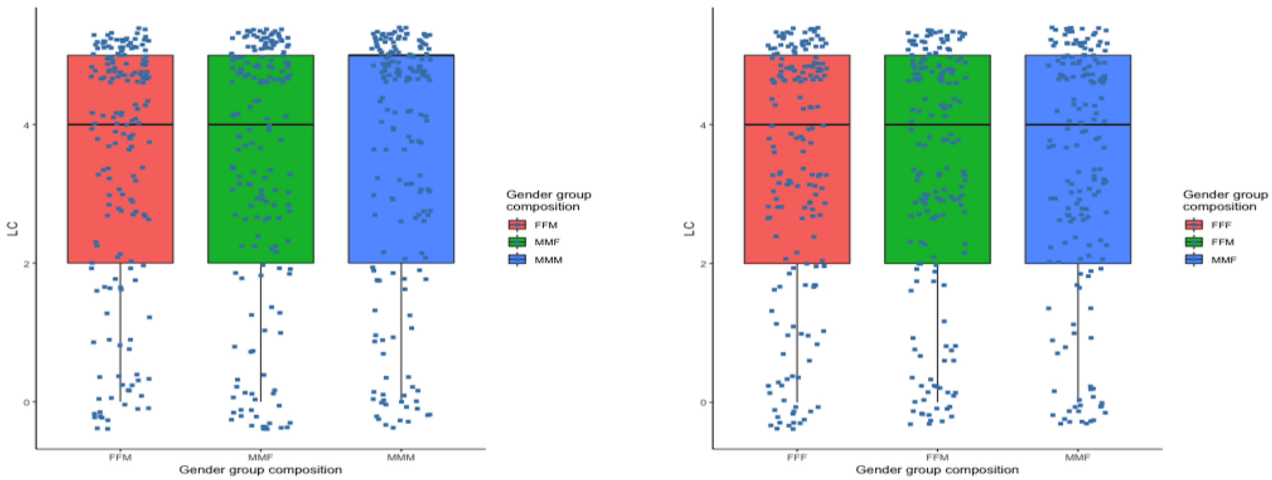
Table 4.3: Descriptive statistics – full sample.

		FFM		MMF		FFF/MMM	
		LC	LB	LC	LB	LC	LB
Men	mean	3.51	3.37	3.44	3.29	3.58	3.34
	median	4	4	4	3	5	3
	sd	1.78	1.64	1.83	1.59	1.82	1.60
Women	mean	3.32	3.06	3.40	3.11	3.36	3.20
	median	4	3	4	3	4	3
	sd	1.78	1.56	1.75	1.62	1.79	1.61
All	mean	3.41	3.21	3.42	3.20	3.47	3.27
	median	4	3	4	3	4	3
	sd	1.78	1.60	1.79	1.60	1.81	1.60

Figure 4.1: Leader contributions (LC) for men and women across gender group compositions - full sample.

(a) Men

(b) Women



Similarly, there appears to be no main effect of either gender or the gender group composition with respect to leaders' beliefs about followers' cooperativeness, albeit there is a larger difference in beliefs between the sexes in all groups than for contributions. There is a larger difference in beliefs in female-dominated groups (FFM) between the sexes as can be seen also by the difference in medians, possibly a sign of a small interaction effect. Similarly to leader contributions men's average beliefs of followers' cooperativeness are higher in all groups than for women. Averages for both leader contributions and beliefs are depicted below. In Figure 4.2 they are shown for the number of other females in the group (0, 1 or 2), and in Figure 4.3 for each gender group composition. The graphs show no systematic increase/decrease with the number of other women in the group. Both men and women contribute most when paired with two other men and least when paired with one man and one woman. As can be seen in Figure 4.3 this translates into men contributing most in same-sex groups (MMM) and least in male-dominated groups (MMF). Women contribute most in MMF and least in female-dominated teams (FFM). From the graphs it appears to be a level effect of gender with respect to both contributions and beliefs, but the scale of the y-axis is very narrow.

Figure 4.2: Leader contributions (LC) and beliefs (LB) by gender and the number of other females in the group – full sample.

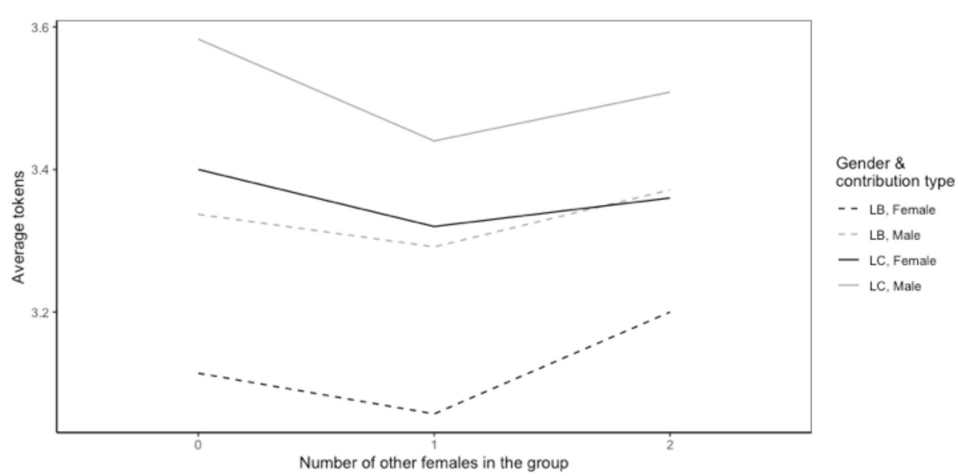
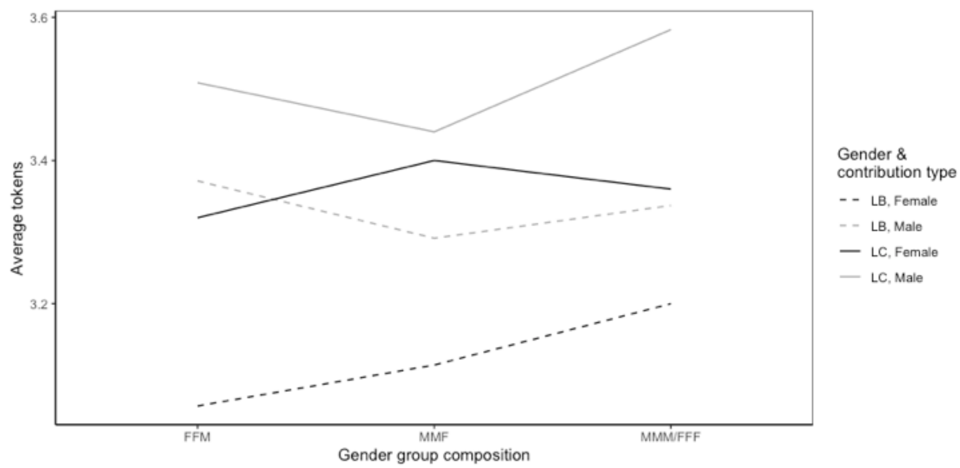
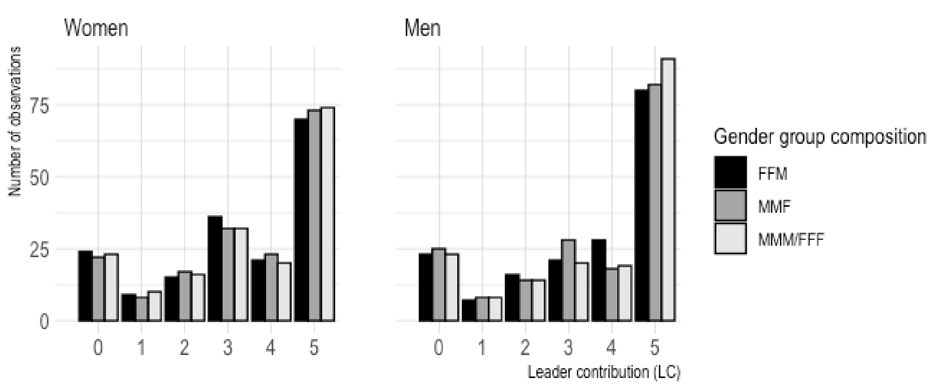


Figure 4.3: Leader contributions (LC) and beliefs (LB) by gender and gender group composition – full sample.



Notably, as seen in Figure 4.4 below, the most common leader contribution in all groups for men and women is the maximum number of tokens – 5. More men than women contribute 5 tokens in all groups. Women's contributions are slightly more distributed towards the middle around 3 tokens. The number of "free-rider leaders" are roughly the same for men and women in all groups, approximately 25. Leader contributions are thus more right-skewed towards 5 tokens than left-skewed towards 0 tokens.

Figure 4.4: Distribution of leader contributions (LC) by gender and gender group composition – full sample.



However, these are merely visual inspections. Next I formally analyze the results of the hypotheses tests. First I present the results for leader contributions, then for leader beliefs.

4.2.4.2 Results for leader contributions

The main results for Hypotheses 1–3 regarding leader contributions are shown in model (3) in Table 4.4 below (corresponding to Equation 4.1, p.24). They indicate that there is no significant main effect of gender, nor of the gender group composition or an gender heterogeneity effect of the gender group composition on leader contributions. The magnitudes of the variables are small and not statistically significantly different from zero. The low adjusted R squares indicate that the independent variables explain a low proportion of the variance in leader contributions, also due to omitted variables such as social preferences for cooperation. Adding $NbFemales$ in model (2) as well as the interaction term in model (3) $Female \cdot NbFemales$ compared to model (1) does not improve the fit of the model. An F test does not reject the hypothesis that all coefficients in model (3) are jointly equal to zero. The insignificant interaction $Female \cdot NbFemales$ implies that the effect of gender on average leader contribution does not depend on the number of other females in the group. An F test comparing model (2) and (3) show that adding the interaction term $Female \cdot NbFemales$ does not improve the model (F statistic = 0.016, p value = 0.899, df = 3) – Hypothesis 3 is thus not rejected. In accordance with the pre-analysis plan I do therefore not proceed to make pairwise t tests to compare in which group(s) men and women would differ. This positive interaction term indicates that increasing the number of other females in the group has a less negative effect for women than for men (-0.4% and -0.7% respectively), but it is not statistically significant different from zero.

The insignificant coefficient of $NbFemales$ in model (2) and (3) suggests that there was no main effect of the gender group composition for either men or women. Hypothesis 2 of no main effect of the gender group composition on leader contributions for either men or women is thus not rejected. I do therefore not proceed to test in which gender groups compositions a difference would occur.

Table 4.4: OLS regressions explaining leader contributions to the PG.

	Dependent variable: % of tokens contributed in the leader role (LC)						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Female (1=female, 0=male)	-0.030	-0.030	-0.034	-0.032	-0.045	-0.028	-0.016
	(0.032)	(0.032)	(0.037)	(0.040)	(0.038)	(0.038)	(0.039)
NbFemales (0,1 or 2 other females in the group)		-0.006	-0.007	-0.004	-0.010	-0.013	-0.013
		(0.010)	(0.012)	(0.013)	(0.012)	(0.011)	(0.012)
Female·NbFemales			0.003	-0.006	0.011	0.008	0.006
			(0.019)	(0.021)	(0.019)	(0.020)	(0.021)
Constant	0.702^{***}	0.708^{***}	0.710^{***}	0.699^{***}	0.715^{***}	0.709^{***}	0.705^{***}
	(0.024)	(0.026)	(0.026)	(0.028)	(0.026)	(0.027)	(0.028)
Total observations	1050	1050	1050	939	1026	966	912
Individual observations	350	350	350	313	342	322	304
R^2	0.002	0.002	0.002	0.003	0.003	0.001	0.001
Adj. R^2	0.001	0.000	-0.001	0.002	0.002	0.000	-0.000
F Statistic	0.90	0.64	0.46	0.60	0.64	0.60	0.51

Clustered (heteroskedasticity) robust standard errors at the individual level in parentheses.

Sig. codes: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

The F Statistic represents a joint coefficient test for all explanatory variables.

Similarly, the insignificant coefficient of $Female$ in models (1)–(3) implies that Hypothesis 1 of no main effect of gender is not rejected. Comparing the t statistics the magnitude of the main effect of gender is similar to those of Gächter et al. (2012) and Vyrastekova et al. (2015). (Their samples are relatively small in comparison to mine, indicating that my results may add some generalizability of these results. However, the low values of the t and F statistics, as well as the R^2 indicate that I am under-powered to detect such a small effect.)

The coefficients on $NbFemales$ and $Female{\cdot}NbFemales$ in model (3) indicate that the negative decrease associated with an additional female in the group in model (2) (equal to -0.6%) appears to not be meaningful. 62 women did make the same leader contributions in all three groups (of which 48 contributed 5 or 0 tokens). The corresponding number for men are 90 (of which 71 contributed 5 or 0 tokens). Model (3) indicates that increasing the number of other females in the group has a smaller negative effect on women than on men as suggested by Figure 4.3, but again, all results are insignificant. The results indicate that there is no statistically significant difference in preferences for leading-by-example between men and women, controlling for the gender group composition.

These results are robust to the sensitivity analyzes in models (4)–(7), both in terms of magnitude and statistical significance. Model (4) shows the results when removing those participants who thought the experiment was about gender. In model (5) those who could not see all avatar images are excluded. Finally, model (6) excludes participants with completion times below one standard deviation from the mean (i.e., less than nine minutes), and model (7) excludes participants with completion times less than 10 minutes. Albeit excluding these groups do not change the results of the hypotheses tests, removing participants who thought the experiment was about gender and those who could not see all avatar images changes the relative direction of average leader contributions and beliefs. E.g., the sign of the coefficient on $Female{\cdot}NbFemales$ in model (4) is now negative, but again insignificant. For summary statistics of leader contributions and beliefs corresponding to Table 4.3 for these robustness checks see Tables D.1–D.4 in Appendix D.0.1.

Since the leader contributions are not normally distributed, due to the right- and left-censoring, the OLS assumption of normally distributed error terms will be violated causing confidence levels and standard errors will to not be precise. (This can be seen in the histogram of the OLS residuals from model (3) in Figure C.1, Appendix C, or in the normal Q-Q graph in Figure C.2, Appendix C where observations are heavily right-tailed due to the right-censoring, but also quite heavily left-tailed.) However, the normality assumption of the residuals is not the strongest OLS assumption. [7]

Yet, as the OLS estimates may be biased with censored data I run the main regression 4.1 (OLS model (3) in Table 4.4) using censored Tobit and mixed effects (see Table D.5, Appendix D.0.2 for these regressions). This was not stated in the pre-analysis plan and is only intended as an additional sensitivity analysis. The statistical significance does not change, nor the direction of the coefficients between the OLS regression and the corresponding Tobit with clustered (heteroskedasticity) robust standard errors at the individual level, left-censored at zero and right-censored at one. To render the Tobit estimates approximately comparable to the OLS ones I calculate the average marginal effects of the explanatory variables in model (3) on the expected value of the actual tokens contributed (see Table D.6, Appendix D.0.2). The magnitude of the estimates are similar indicating that the OLS bias is small. However, as Tobit and OLS estimates are difficult to compare for binary variables this is only approximate measure.

[7] Since leader contribution (LC) is a limited, censored dependent variable testing the OLS assumptions will be misleading. I nevertheless visually inspect them for non-normality as it will render OLS estimates inefficient. Furthermore, there appears to be no multicollienarty between the explanatory variables (see Table C.1 in Appendix C).

To furthermore obtain an understanding of the effect of leaving the autocorrelation between observations entirely in the residuals, as with clustered standard errors in OLS, compared to partly when including individual random effects in the model, I also run a panel Tobit with individual random effects and heteroskedasticity robust standard errors (also in Table D.5, Appendix D.0.2.) The results are similar indicating that the bias is not large. Finally, I also run the panel Tobit model with individual random effects without heteroskedasticity robust standard errors since the Tobit model also is sensitive to heteroskedasticity (Wooldridge, 2012). However, the robust standard errors are very similar to the non-robust ones, indicating that the heteroskedasticity is small. In sum, the robustness checks in this section confirm the results from the OLS regressions in Table 4.4. Hypotheses 1–3 are thus not rejected.

To evaluate the magnitude of the effect and compare my results to those of Balliet et al. (2011) I compute Cohen's d [8] (not corrected for small sample size bias as my sample is large) for the between gender effect in each group using an independent t test (Lenhard and Lenhard, 2016), seen in Table 4.5. I also compute Cohen's d for a difference in means in leader contributions for men and women, respectively, in each gender group composition in Table 4.6. [9] Using Cohen's criteria for small, medium and large effect sizes (0.2, 0.5. and 0.8) (Cohen, 2013) the effect sizes are all negligible. In female-dominated groups and in same-sex groups the effect sizes are larger, while in male-dominated teams the between-gender effect size on leader contributions is almost zero as men's and women's contributions were most similar in this group. However, in same-sex groups the effect size of beliefs is no longer negligible, yet small, possibly indicating the presence of heterogeneity in gender beliefs. The within gender effect sizes in Table 5.5 are very small, as the gender group composition had a small impact on leader decisions.

Table 4.5: Cohen's d — full sample.

	FFM		MMF		FFF/MMM	
	LC	LB	LC	LB	LC	LB
(Diff. men–women)	0.12	0.11	0.02	0.14	0.14	0.24

Table 4.6: Cohen's d — men versus women.

	FFM–MMF	FFM–FFF/MMM	MMF–FFF/MMM
Men	0.05	0.05	0.08
Women	0.05	0.02	0.02

4.2.4.3 Results for leader beliefs

Next I turn to the formal results for Hypotheses 6a–6c regarding leader beliefs. These results are shown in

[8] Cohen's $d = \frac{Mean_{men} - Mean_{women}}{\text{Pooled } SD}$, where Pooled $SD = \sqrt{\frac{(SD^2_{men} + SD^2_{women})}{2}}$

[9] I calculated this Cohen's d from the dependent t test using the online software by Lenhard and Lenhard (2016), where $d = t_c \sqrt{\frac{2(1-r)}{n}}$. I calculated the correlation coefficients r in RStudio first.

model (3) in Table 4.7 below and show no significant results. Similarly to Hypotheses 1–3 the coefficients of the relevant variables all are small in magnitude and not statistically significant. Hypotheses 6a–6c are thus not rejected. The positive interaction term on $Female{\cdot}NbFemales$ indicates that increasing the number of other women in the group on average has a more positive effect on the leader's belief of followers' support for a woman than for a man (0.8% for a woman, 0.3% for a man). Thus, the average difference in leader beliefs appears to decrease with the number of other females in the group, but it is not significant. Furthermore, there is a small average main effect of $NbFemales$, but it appears to driven primarily by gender, comparing model (1) with just $Female$ as a regressor and model to with $NbFemales$ added to model (3) with the interaction term. The negative coefficient on $Female$ indicates a small, insignificantly negative main effect of gender on the leader's average belief of followers' cooperativeness.

Table 4.7: OLS regressions explaining leader beliefs (LB).

	Dependent variable: Leader's % belief (LB)						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Female (1=female, 0=male)	-0.042	-0.042	-0.047	-0.030	-0.054	-0.032	-0.010
	(0.028)	(0.028)	(0.034)	(0.035)	(0.034)	(0.035)	(0.036)
NbFemales (0,1 or 2 other females in the group)		0.006	0.003	0.005	0.006	0.006	0.008
		(0.009)	(0.012)	(0.013)	(0.012)	(0.013)	(0.013)
Female · NbFemales			0.005	-0.010	0.004	0.001	-0.003
			(0.018)	(0.019)	(0.018)	(0.019)	(0.020)
Constant	0.667***	0.661***	0.663***	0.660***	0.666***	0.655***	0.648***
	(0.020)	(0.022)	(0.023)	(0.024)	(0.024)	(0.025)	(0.025)
Total observations	1050	1050	1050	939	1026	966	912
Individual observations	350	350	350	313	342	322	304
R^2	0.004	0.005	0.005	0.004	0.006	0.003	0.003
Adj. R^2	0.003	0.004	0.004	0.004	0.006	0.001	-0.000

Heteroskedasticity robust standard errors, clustered individual, in parentheses.

Sig. codes: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

In sum, albeit being a female is associated with lower stated beliefs about followers' support in all gender group compositions, the difference is not statistically significant, nor the effect of the gender group composition on beliefs, or a gender heterogeneity effect of the gender group composition. The results suggest that gender differences in leading-by-example are not significantly driven by gender beliefs, albeit women on average express lower level of beliefs than men in all gender group compositions.

The robustness checks in models (4)–(7) do not alter the conclusion, albeit the results are marginally sensitive in magnitude and direction to the exclusion of certain groups. Model (4) shows the results when participants who thought the experiment was about gender are excluded. Model (5) shows the results when participants who could not see all avatar images are excluded. In model (6) participants with completion times below one standard deviation from the mean (i.e., less than nine minutes) are excluded, and in model (7) participants with completion times less than 10 minutes are excluded. Compared to model (3) the negative impact of being a female decreases, similarly as for leader contributions, expect in model (5). When excluding participants who thought the study was about gender and those who completed it under 10 minutes the interaction term $Female{\cdot}NbFemales$ becomes negative, but the estimates are not significant.

4.2.5 Heterogeneity analysis in risk-aversion

Last, I study how preferences for leading-by-example (i.e., leader contributions) are related to risk-aversion amongst men and women. The purpose is to test whether a treatment effect of the different gender group compositions on leader contributions for men and women are different for participants with various willingness to take risks. To formally test whether such a treatment effect of the gender group compositions for men and women is different depending on the willingness to take risks I use regression 4.3 below. RA_i is a continuous variable indicating participant i's willingness to take risks, ranging from 0 (unwilling to take risks) to 10 (fully willing to take risks). Being only an exploratory analysis no formal hypotheses statements are provided as stated in the pre-analysis plan.

$$\begin{aligned} LC_i = \beta_0 + \beta_1 Female_i + \beta_2 NbFemales + \beta_3 RA_i + \beta_5 Female_i \cdot NbFemales + \beta_6 Female_i \cdot RA_i \\ + \beta_7 RA_i \cdot NbFemales + \beta_8 Female_i \cdot RA_i \cdot NbFemales + \varepsilon_i \end{aligned} \tag{4.3}$$

Table 4.8 below shows descriptive statistics for stated risk attitude for men and women. As expected, men express a higher willingness to take risks than women. The variation is similar, but the mean and median is higher for men than women. The formal results are shown in model (3) in Table 4.9 below and indicate that there is no heterogeneity effect of risk attitude on leader contributions depending on the gender group composition on average leader contribution for either men or women.

Table 4.8: Descriptive statistics for risk attitude (RA).

	mean	median	sd	min	max	n
full sample	5.21	6	2.42	0	10	350
men	5.71	6	2.40	0	10	175
women	4.73	5	2.43	0	10	175

The insignificant interaction term $RA{\cdot}Female{\cdot}NbFemales$ implies that the marginal effect of gender on leader contributions do not differ for given high and low values of risk attitudes and the number of other females in the group. An F test does not reject the hypothesis that all coefficients jointly are equal to zero (F statistic= 1.25, df = 7). The signs of the coefficients indicate that for risk-averse men (i.e., for low values of RA) an increase in $NbFemales$ has a marginally negative effect on average leader contributions. However, for less risk-averse men (with high values of RA) increasing $NbFemales$ has a marginally positive effect on leader contributions. For risk-averse women an increase in $NbFemales$ similarly has a marginally negative effect on the average leader contribution. For less risk-averse women, increasing $NbFemales$ instead has a marginally more positive effect on leader contributions than for less risk-averse men. However, as the coefficients are not significant I cannot conclude that there is an heterogeneity effect of risk attitude depending on the gender group composition on average leader contribution for either men or women.

Table 4.9: OLS regressions of heterogeneity in risk attitude (RA).

	Dependent variable % of tokens contributed in the leader role (LC)						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Female (1=female, 0=male)	-0.014	-0.075	-0.069	-0.075	-0.070	-0.013	0.006
	(0.032)	(0.090)	(0.101)	(0.108)	(0.103)	(0.108)	(0.110)
NbFemales (0,1 or 2 other females in the group)			-0.031	-0.030	-0.030	-0.026	-0.025
			(0.031)	(0.034)	(0.032)	(0.035)	(0.037)
Risk attitude (RA)	0.017*	0.011	0.007	0.007	0.008	0.012	0.014
	(0.007)	(0.011)	(0.012)	(0.013)	(0.012)	(0.013)	(0.013)
RA · Female		0.012	0.009	0.010	0.007	-0.001	-0.001
		(0.015)	(0.017)	(0.018)	(0.017)	(0.018)	(0.018)
Female · NbFemales			-0.006	-0.020	-0.016	-0.012	-0.015
			(0.046)	(0.051)	(0.046)	(0.051)	(0.053)
RA · NbFemales			0.004	0.005	0.004	0.002	0.002
			(0.005)	(0.005)	(0.005)	(0.005)	(0.005)
RA·Female · NbFemales			0.003	0.004	0.006	0.005	0.005
			(0.008)	(0.009)	(0.008)	(0.009)	(0.009)
Constant	0.608***	0.641***	0.672***	0.656***	0.669***	0.642***	0.625***
	(0.050)	(0.070)	(0.077)	(0.082)	(0.078)	(0.085)	(0.087)
Total observations	1050	1050	1050	939	1026	966	912
Individual observations	350	350	350	313	342	322	304
R^2	0.014	0.015	0.016	0.022	0.020	0.013	0.015
Adj. R^2	0.013	0.014	0.015	0.021	0.019	0.012	0.014

Heteroskedasticity robust standard errors, clustered at the individual level, in parentheses.

Sig. codes: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

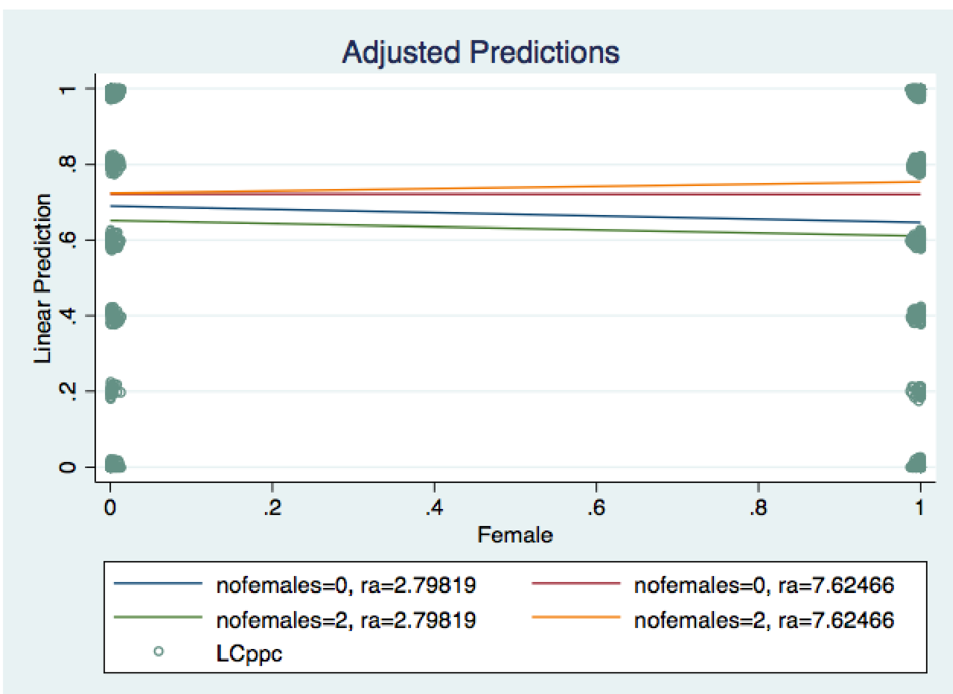
Furthermore, that the coefficient of RA loses its 5% significance level and decreases in magnitude from model (1) to model (2) when adding the interaction term would suggest that RA has an positive effect on leader contributions primarily for females, given by the positive coefficient of $RA{\cdot}Female$, albeit being more risk prone (having a high RA) is positively correlated with leader contributions for men as well. I.e., while RA has a small effect for men, the results would indicate that more risk prone women contribute marginally more than risk prone men, while more risk-averse female contribute marginally less than risk-averse men on average.

However, the results are not significant. Thus it appears that the positive effect of willingness to take risks on average leader contributions are primarily driven by less risk-averse women, but the non-significant results do not permit any conclusions to be drawn.

Figure 4.5 below shows the relationship between the percentage leader contribution ($LCppc$) and gender at high and low values of risk attitude and the number of other women in the group (low=0, high=2). (A low value of risk attitude is measured as one standard deviation below the mean i.e., RA = 2.798. A high value of risk attitude is measured as one standard deviation above the mean i.e., RA = 7.625. The graphs confirm that risk attitude marginally affect leader contributions for women, with less risk-averse women contributing more as leaders, while the gender group compositions has a very small effect for both men and women.

However, in sum the results show no statistically significant correlation between risk attitudes and preferences for leading-by-example across gender and the gender group composition. The conclusions do not change with the robustness checks in models (4)–(7) in Table 4.9 albeit the coefficients change slightly in magnitude. In model (4) participants who thought the experiment was about gender are excluded. In model (5) participants who could not see all avatar images are excluded. In model (6) participants with completion times below one standard deviation from the mean (i.e., below nine minutes) are excluded, and in model (7) participants with completion times below 10 minutes are excluded. The sign of the coefficient on $RA{\cdot}Female$ is now negative in models (7) and (8), but again the results are not significant.

Figure 4.5: Heterogeneity in correlation between risk attitude (RA) and LC



4.2.6 Summary of gender and leader behavior

To sum up this section on leader behavior I include controls for beliefs, risk attitude and age to the main OLS regressions for Hypotheses 1–3 explaining leader contributions to the PG in Table 4.4 (p.28), as a last robustness check. It was not stated in the pre-analysis plan as it is not the main focus of this study. The results are shown in Table 4.10 below and show that the effect of gender, the gender group composition and a gender heterogeneity effect of the gender composition of the group are robust to the inclusions of these controls.

Table 4.10: OLS regressions explaining leader contributions to the PG – full sample.

	Dependent variable: % of tokens contributed in the leader role (LC)					
	(1)	(2)	(3)	(4)	(5)	(6)
Female (1=female, 0=male)	-0.030	-0.007	-0.008	-0.001	-0.010	-0.004
	(0.032)	(0.026)	(0.033)	(0.033)	(0.033)	(0.033)
NbFemales (0,1 or 2 other females in the group)			-0.009	-0.009	-0.009	-0.009
			(0.013)	(0.013)	(0.013)	(0.013)
Female · NbFemales			0.001	0.001	0.001	0.001
			(0.019)	(0.019)	(0.019)	(0.019)
% leader's belief (LB)		0.550^{***}	0.550^{***}	0.543^{***}	0.551^{***}	0.544^{***}
		(0.045)	(0.045)	(0.045)	(0.045)	(0.045)
Risk attitude (RA)				0.007		0.006
				(0.006)		(0.006)
Age					-0.001	-0.001
					(0.001)	(0.001)
Constant	0.702^{***}	0.336^{***}	0.345^{***}	0.311^{***}	0.369^{***}	0.332^{***}
	(0.024)	(0.040)	(0.042)	(0.0513)	(0.056)	(0.068)
Total observations	1050	1050	1050	1050	1050	1050
Individual observations	350	350	350	350	350	350
R^2	0.002	0.243	0.243	0.245	0.244	0.246
Adj. R^2	0.001	0.242	0.241	0.242	0.240	0.241
F Statistic	0.90	77.55^{***}	38.99^{***}	31.68^{***}	32.27^{***}	27.40^{***}

Heteroskedasticity robust standard errors, clustered at the individual level, in parentheses.
Sig. codes: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

The results in Table 4.10 suggest that leader contributions to the PG are mainly explained by stated beliefs of followers' support (amongst the included regressors). Age is negatively correlated with leader contributions, but very small in magnitude and not significant. Comparing models (1) and (2) indicates that the negative impact on average leader contributions of being a female was mediated by women's, on average, less positive beliefs about followers' support. Notably, the leader's belief is the only significant predictor in explaining leader contributions. Model (3) corresponds to model (3) in Table 4.4, but with beliefs added. [10]

[10]As a robustness check I run model (3) in Table 4.10 as a Tobit regression with clustered (heteroskedasticity) robust standard

Albeit the impact of risk attitude on the average leader contribution is positive, it is similar in magnitude to that of Gächter et al. (2012), further indicating that risk attitude is not a meaningful predictor of leader contribution after controlling for beliefs and gender. The marginal decrease in the coefficient for risk attitude, RA, from model (4) to (6) indicates that some of the positive effect of risk attitude on leader contributions was driven by younger participants. This could be expected by the negative correlation between age and risk attitude (see Table C.1, Appendix C). Similarly, the marginally increased negative effect on average leader contributions for females appears to be partly reflected by women in the sample being younger than the men. However, as only the leader's belief is significant, no conclusions can be drawn from the results.

4.3 Follower behavior

4.3.1 Primary hypotheses – conditional cooperation

Next, I turn to follower behavior. This analysis is performed separately for men and women. The purpose is to explore whether male or female leaders would be more successful in leading followers to higher cooperation levels and is related to the role of gender beliefs. I test if participants indeed behave as conditional cooperators, and whether men and women change conditional cooperative behavior depending on the leader's gender and the gender group composition. Each man and woman fill out a conditional cooperative schedule for four different group compositions (see the right column in Figure 3.1). Specifically I test the following hypotheses.

Hypothesis 4 *There are no differences in conditional cooperation depending on the sex of the leader and the gender group composition for either men or women.*

Hypothesis 5 *Participants on average behave as conditional cooperators and increase their contribution to the PG with the leader's contribution.*

Note. In MMF and FFF each woman fills out two conditional follower schedules when the leader is a woman, and when the leader is a man respectively. Similarly, in MMM and FFM each man fills out two conditional follower schedules when the leader is a man, and when the leader is a man respectively. These are coded in as "LisF1" or "LisF2" for a woman, and "LisM1" or "LisM2" in Lioness. I therefore specified a priori in the pre-analysis plan that I would use measure "1" in the rest of the analysis. I choose this over using weighted means.

4.3.2 Analytical specifications and results

4.3.2.1 Hypothesis 4

To test Hypothesis 4 I construct two measures of conditional cooperation. The first is the mean average conditional contribution, calculated as the total sum of tokens contributed to the PG for all possible leader contribution levels, divided by six for the possible leader contribution levels (0–5). This variable is denoted CC_g in group g. g refers to the four follower measurements (see Table 4.11 below). The second measure is the slope coefficient from a linear regression of conditional cooperation on each possible leader contribution level, in group and leader constellation g. This variable is denoted $CCslope_g$. I calculate it by first constructing a variable LCL (the leader contribution level) ranging from 0 to 5. Then, for each participant, in each of the

errors at the individual level, left-censored at zero and right-censored at one, and with a panel Tobit model with individual random effects and heteroskedasticity robust standard errors (see Table D.7, Appendix D.0.2). The results are similar, and thus robust.

four groups in Table 4.11 below, I regress the conditional contribution for each leader contribution on the LCL variable. E.g., a free-rider has a mean (CC_g) and a slope coefficient $CCslope_g$ equal to 0. A perfect conditional cooperator has a mean equal to 15/6 = 2.5, (0+1+2+3+4+5 = 15), and a slope coefficient equal to 1.

Table 4.11: The 4 follower measurements for men and women.

Participant's gender:	FFM	MMF	FFF/MMM
Female	LeaderF; LeaderM	LeaderM	LeaderF
Male	LeaderF	LeaderM; LeaderF	LeaderM

Friedman's Chi-Square test is first performed on all four measurements. If the Q statistic from the test is significant, I use pairwise t tests investigate which measures that differ from each other for men and women. To correct for multiple hypothesis testing I use the Benjamini-Hochberg (1995) correction. I use the non-parametric Friedman test as it does not require the sample to be normally distributed as do the one-way ANOVA with repeated measures. The Friedman test requires that the sample groups are measured at three or more occasions (I have four), and that the group is a random sample from the population (which I have). Finally, it requires the data to be ordinal or continuous. Therefore, I the requirements are fulfilled.

I stated in the pre-analysis plan that if Hypotheses 1–3 would not rejected the presence of gender beliefs as stated by Hypotheses 4a–4b are only be explored as an possible indication of their existence. However, the results below show no such indications as there are no significant differences for the two measures for either men and women. Table 4.12 presents the average means and standard deviations for the two variables CC_g and $CCslope_g$ for the full sample for women, and Table 4.13 the corresponding measures for men. As can be seen the measures are similar in magnitude. I first present the main tests results and robustness checks for women, then for men.

Table 4.12: Hypothesis 4 – women.

group	CC_g	$CCslope_g$
FFM:LeaderF	2.440	0.344
	(1.23)	(0.545)
FFM:LeaderM	2.367	0.348
	(1.29)	(0.541)
MMF:LeaderM	2.327	0.381
	(1.28)	(0.520)
FFF:LeaderF	2.430	0.372
	(1.35)	(0.521)
observations	175	175

Mean values with standard errors within parentheses.

Table 4.13: Hypothesis 4 – men.

group	CC_g	$CCslope_g$
FFM:LeaderF	2.506	0.413
	(1.330)	(0.597)
MMF:LeaderM	2.470	0.443
	(1.250)	(0.595)
MMF:LeaderF	2.390	0.411
	(1.330)	(0.591)
MMM:LeaderM	2.411	0.416
	(1.300)	(0.574)
observations	175	175

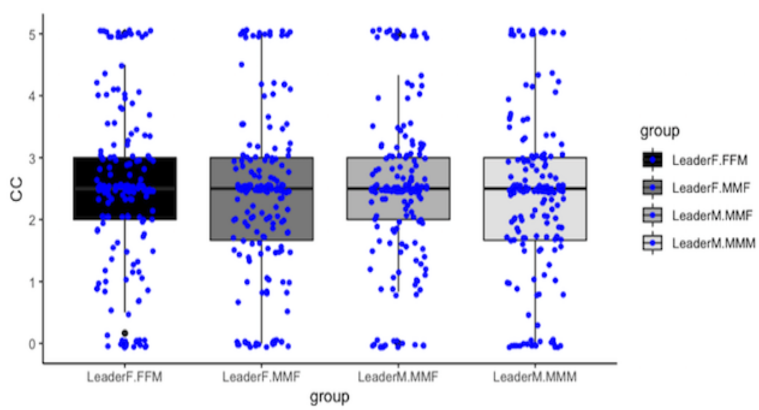
Mean values with standard errors within parentheses.

For women, the value of Friedman's chi-square statistic, Q, for CC_g equals 2.805 with three degrees of freedom. The associated p value is 0.4218. Thus, the hypothesis of no differences between the group measures is not rejected. Similarly, for $CCslope_g$ Friedman's chi-square statistic Q equals 1.4941 ($df = 3$). The associated p value is 0.6836. Again the hypothesis of no differences between the group measures is not rejected. The sensitivity analysis shows that these results are robust to the exclusion of participants who thought the study was about gender, those who could not see all avatar images and those with short completion times. The corresponding tables to Table 4.12 and the Friedman tests for these robustness checks are shown in Tables D.8–D.11, Appendix D.0.3. The average CC_g is still lower in with male leaders, but no clear pattern again arises from $CCslope_g$ and the measures are similar.

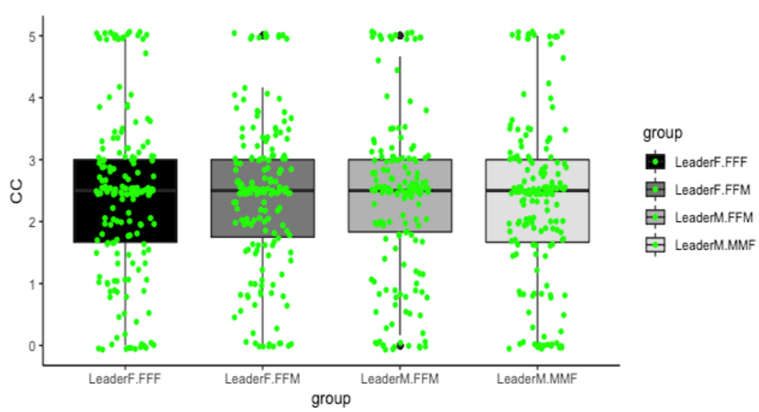
For men, the Friedman's rank sum test for differences between the paired groups yields a statistically significant result for CC_g, but at only at the 5% level ($Q = 8.4911$, $df = 3$ and p value = 0.0369). Figures 4.6 (a) and (b) below suggest that the distribution for CC_g across the different group and leader constellations vary more for men than for women. However, the hypothesis of no differences between the groups is not rejected at the pre-specified 1% level. I will therefore not consider this difference significant, and do not proceed with paired t tests to investigate which groups that differ from each other. Moreover, the Friedman's test detect no significant difference in the average $CCslope_g$, (Q statistic = 1.5746, $df = 3$, p value = 0.6652). Thus, again the hypothesis of no differences between the group measures is not rejected.

Figure 4.6: Average conditional cooperation CC_g, by gender, across group and leader constellations.

(a) Men



(b) Women



The sensitivity analysis show that these results for CC_g and $CCslope_g$ are robust to the exclusion of participants who thought the study was about gender, those who could not see all avatar images and those with short completion times. The corresponding tables to Table 4.13 and the Friedman tests for these robustness checks are shown in Tables D.12–D.15, Appendix D.0.3. The average CC_g is still lower with male leaders, but no clear pattern again arises from $CCslope_g$. When excluding those participants who thought the experiment was about gender the associated p value for the Friedman's test for CC_g is higher (= 0.0809) compared to the full sample and when excluding the other groups. In sum, all results show that the average values of CC_g both $CCslope_g$ are generally higher for men than women (also the robustness checks), but the results are not significant indicating that nor male nor female leaders would be more successful in crowding in. Hypothesis 4 is thus not rejected.

This could be expected from Tables 4.12 and 4.13 as there are no large discrepancies across the different gender group compositions and depending on the sex of the leader. As a last point of reference regarding $CCslope_g$ – in FFM with a female leader 34 participants had a slope equal to zero. In FFM with a male leader 39 participants had a slope equal to zero. The corresponding numbers for MMF where the leader is a male is 35 and in FFF 42. Having a zero slope does not necessarily mean that these participants contributed zero tokens, merely that they contributed the same amount for all leader contribution levels. I therefore coded them as having a zero slope as they are treated missing observations in the linear regression.

4.3.2.2 Hypothesis 5

To test Hypothesis 5 that *participants on average behave as conditional cooperators and increase their contribution to the PG with the leader's contribution*, I calculate Spearman's rank order correlation coefficient, ρ, as follows. The average conditional contribution for each possible leader contribution (six levels, 0–5) is calculated and stored in a variable CC (for conditional contribution). I.e., for participant i, for leader contribution level 0, I sum up the conditional contributions for all four follower roles in Table 4.11 and divide this sum by four to obtain the average (i.e. CC). For the same participant I then calculate CC for leader level contribution 3, for leader level contribution 3, and etc. Next, for participant $i + 1$ I similarly calculate CC and so forth for all participants. I then calculate the correlation coefficient ρ between CC and the LCL variable (the leader contribution level). See Table 4.14 below for an overview of how CC is constructed.

Table 4.14: Calculation of CC – e.g. for the four follower roles for a woman

id	LCL	CC	LisF:FFM	LisM:FFM	LisM:MMF	LisF:FFF
1	0	4=	0 +	0 +	3 +	5
1	1	5=	1 +	1 +	3 +	5
1	2	3=	2 +	2 +	3 +	5
1	3	3=	3 +	3 +	1 +	5
1	4	4=	4 +	4 +	3 +	5
1	5	4=	5 +	5 +	1 +	5
2	0	2=	0 +	0 +	3 +	5
2	1	2=	1 +	1 +	1 +	5
2	2	3=	2 +	2 +	3 +	5
.	.	.	.	.	.	.
.	.	.	.	.	.	.

The null hypothesis that participants do not behave as conditional cooperators is rejected if ρ is positive and statistically different from zero. This test is also performed separately from men and women.

The results show a positive and statistically significant correlation for both men and women, indicating that many participants behave as conditional cooperators and that leading-by-example might lead to more efficient social outcomes. For women, Spearman's rank sum correlation coefficient ρ equals 0.400 (p value < 0.0000). Thus the correlation is highly significant but not overly strong. For men, the correlation is marginally higher with $\rho = 0.4220$. The associated p value is again < 0.0000. The sensitivity analyzes (see Appendix D.0.4) for both men and women show that the results are robust to excluding participants who thought the experiment was about gender, participants who could not see all avatar images, and those with short completion times. The strength of the correlation is similar to that of the full sample. For women, the correlation marginally is lower, but there is no clear pattern for men. Thus, both men and women, on average, appear to behave as conditional cooperators and increase their contribution with the leader's contribution. Leading-by-example may be fruitful, yet the magnitude of ρ for both men and women suggest that players on average do not behave as perfect conditional cooperators (in which case $\rho = 1$).

To further explore whether it pays to lead by example I report the average conditional contribution for each possible leader contribution (six levels) when the leader is a man versus woman in all three gender group compositions, as stated in the pre-analysis plan. I.e., for leader contribution level 0 a participant's conditional contribution is averaged across all follower roles and etc. for leader contribution levels 1, 2,...,5. The purpose is to give an overview of the potential success for a male/female leader in different gender group compositions. The results are shown in Table 4.15 below.

Table 4.15: Average follower contributions by group and leader constellation.

	FFM		MMF		FFF	MMM
LC level	F	M	F	M	F	M
0	1.51	1.51	1.38	1.35	1.50	1.35
1	1.92	1.83	1.78	1.81	1.84	1.78
2	2.17	2.19	2.13	2.16	2.23	2.22
3	2.72	2.49	2.62	2.61	2.64	2.64
4	3.08	2.98	3.03	3.12	3.11	3.05
5	3.43	3.20	3.41	3.35	3.26	3.42
observations	350	175	175	350	175	175

For leader contribution levels between 0 and 2 tokens followers on average contribute more than the leader. However, for leader contribution levels of 3–5 tokens leadership is costly. Treating these values as a proxy for the two followers' average contribution and plugging in this in Equation 3.1 (p.14) indicate that leaders would earn less for contributions above 2 tokens, than had they contributed nothing (in which case they had earned 5 tokens). In the sample, average leader contributions for men and women across the different gender group compositions all ranged from 3.32 to 3.58 tokens (see Table 4.3, p.25). These are only averages, but the patterns are consistent with the findings of Gächter and Renner (2004), Güth et al. (2007), Gächter et al. (2012) and Gächter and Renner (2018) that it is costly to be a "good/bold" leader.

4.4 Social preferences for cooperation

4.4.1 Exploratory hypothesis of pro-social behavior

Albeit not being the main objective of this study I lastly explore the role of pro-social behavior for leader contributions as stated in the pre-analysis plan. I.e., if more cooperative followers also would be better role models when it comes to leading-by-example. Specifically, I test if *on average more conditionally cooperative players contribute on average more as leaders.* This as Gächter et al. (2012) find that more pro-social leaders contribute more even after controlling for positive beliefs about followers' support. The conditional contribution is arguably used to measure social preferences for cooperation as this decision determines the actual outcome as compared to the unconditional leader contribution (Gächter et al., 2012). My setup does not allow me to measure complete conditional behavior as followers do not observe the other follower's contribution. Thus, this analysis is only exploratory as stated in the pre-analysis plan. However, if there are no underlying gender differences in social preferences, there should be no gender effect after controlling for beliefs and the gender group composition as found by Vyrastekova et al. (2015).

4.4.2 Analytical specifications and results

To test the correlation between average leader contributions and average conditional contributions, I construct a variable denoted ALC, measured as the average leader contribution across all groups. The average conditional contribution will be stored in a variable denoted ACC. It is the slope coefficient from a linear regression of the average the conditional contribution across all four follower measurements (i.e. the CC variable in Hypothesis 5) on each leader contribution level (LCL). I.e., I average the conditional contributions, for each possible leader contribution, across all four follower measurements and then calculated the slope coefficient from an linear regression on each leader contribution level. (See Table 4.14 (p.39) where I obtain ACC by regressing CC on LCL.) I then calculate Spearman's rank sum correlation coefficient ρ to test if the correlation between ALC and ACC is statistically significant (and positive). The results in the left column of Table 4.16 below indicate that leader and follower behavior is statistically significantly correlated at the less than 1% level.

Additionally, I constructed a variable, MCC, calculated by first summing up all conditional contributions for all four types of follower roles (see Table 4.2, p.25) over all leader contribution levels (six levels) – yielding a total of 24 contributions. This number was then divided by 24 to obtain an average. E.g., a follower who contributes 5 tokens for all leader contribution levels, in all follower roles, will have MCC = 5. An MCC=2.5 depicts a perfect conditional cooperator, while an MCC=0 implies that the participant contributed 0 tokens for all leader contribution levels, in all follower roles. I then calculated Spearman's rank sum correlation coefficient ρ between MCC and ALC to test if the correlation was positive and significant. This was not stated in the pre-analysis plan. I added this as an individual might have a slope coefficient equal to zero because they contributed the same amount of tokens for each leader level contribution, across all follower measurements, which would not be reflected in the linear regression slope coefficient.

Table 4.16: Social preferences for cooperation.

	Spearman's rank sum correlation coefficient:	
	$\rho(ALC - ACC)$	$\rho(ALC - MCC)$
Men	0.3225	0.3939
	(0.0000)	(0.0000)
Women	0.2669	0.3711
	(0.0004)	(0.0000)
Full sample	0.2925	0.3878
	(0.0000)	(0.0000)

Sig. codes: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

p-value of the Spearman rank sum correlation coefficient test within parentheses.

The results in the right column in Table 4.16 indicate that this correlation is also statistically significantly and positive at the less than 1% level. The results are robust to excluding participants who thought the study was about gender, those with short completion times and those who could not see all avatar images (see Tables D.16–D.19, Appendix D.0.5). The MCC correlation is stronger than for ACC as a number of participants contributed the same amount for each leader level contribution, yielding a slope coefficient equal to zero.

Figure 4.7 below plots the average leader contribution, ALC against the average conditional contribution slope coefficient, ACC, for the full sample, by gender. Quite many observations to the right of zero for ACC confirms that more cooperative followers on average are better role models as leaders. A few participants have an ACC=0 and a high ALC indicating that they contribute much as leaders, but decrease their contributions as followers with the leader's contribution. An ACC=0 with an ALC=0 might imply that the participant on average contributed 0 tokens for each leader level contribution or e.g., 5 tokens. Some players are perfect conditional cooperators with an ACC=1, but most contribute less than the leader making it costly to lead. The scatter plot (with a fitted linear regression line) shows quite a variation albeit a positive trend. (I use the more conservative Spearman's rank sum correlation test as opposed to a t test of the linear regression coefficient.) Figure 4.8 plotting ALC against the mean conditional cooperation, MCC, indeed shows a stronger positive correlation, indicating that some of the zero ACC's did not reflect free-riders.

The graphs indicate that more cooperative followers contribute more in the leader role, consistent with the findings of Gächter et al. (2012), but there is no clear gender pattern. The correlations in Table 4.15 are stronger for men for both measures, but the distributions in Figures 4.7–8.8 are quite similar. It appears that social preferences for cooperation matter more than gender, but it is merely a visual inspection. Interestingly, the graphs show quite a few player types which on average contribute much in the leader role, but less as followers, possibly indicating a stronger leader effect on contributions in general.

Figure 4.7: The average leader contribution (ALC) against the slope coefficient of a linear regression of the average conditional contribution (ACC) for each leader contribution level (LCL).

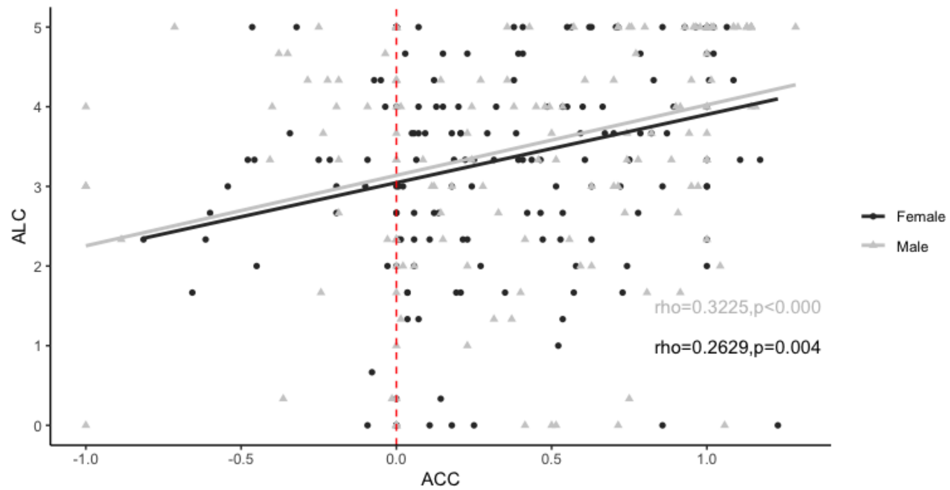
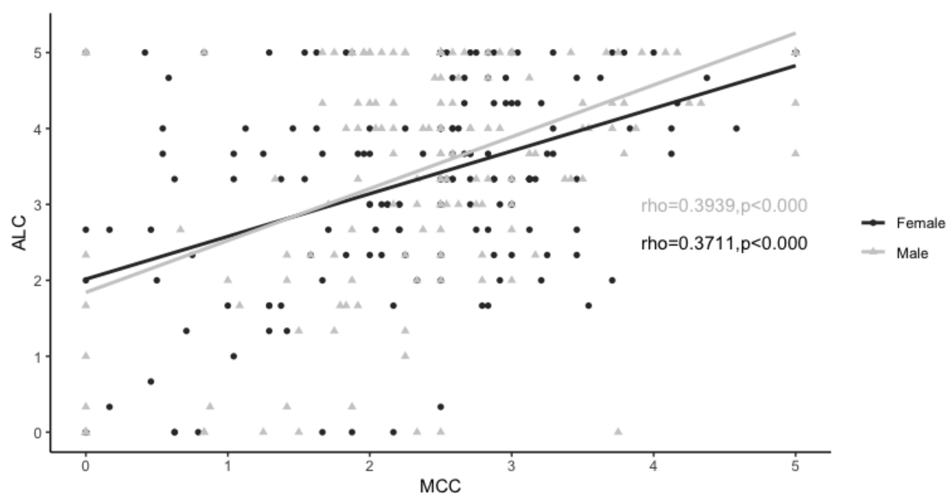


Figure 4.8: The average leader contribution (ALC) against the average conditional contribution for all leader contribution levels, across the four follower roles (MCC).



5 Discussion

The results would indicate that there are no gender differences in preferences for leading-by-example as there were no significant gender differences in leader contributions, nor that male or female leaders would be more or less successful in crowding in in the PGs dilemma setting as neither men or women significantly altered their follower contributions with the gender of the leader and the gender group composition. Albeit men contribute more than women in the leader role and report higher beliefs about followers' contributions in all gender group compositions, the main effect of gender is not statistically different from zero. Similarly, while the results would indicate a small heterogeneity effect of risk attitude on leader contributions so that less risk-averse women contribute more as leaders than less risk-averse men, the results are not significant. The interpretation of these findings might be that such gender differences do not exist, or that the variance of the response data was too large. Or that I was under-powered to detect such small effect sizes. The latter since women's systematically lower leader contributions were accompanied by lower stated beliefs of influence on followers, and that this negative effect appeared to be mitigated by higher willingness to take risks. Indeed, men might exhibit stronger preferences for leading-by-example in line with evolutionary theory as men on average also contributed more as followers. However, with insignificant results I cannot make any conclusions.

Moreover, the results showed that more cooperative followers on average were better role models as leaders, as the correlation between individual leader and follower contributions was statistically significant for both men and women and similarly positive in magnitude. This possibly suggests that individual social preferences for cooperation is a stronger predictor than gender in this setting. However, as this was only an exploratory analysis formal testing is required. Controlling for social preferences in explaining leader contribution might contribute to the understanding of what is driving possible gender differences in leading-by-example. However, previous research from meta-analyzes also indicates that there are small gender differences in pro-social preferences for cooperation in social dilemmas.

In line with previous research on cooperation in PGs dilemmas the results also indicate that many of the men and women behave as conditional cooperators and increase their contribution with the leader's contribution. This indicates that a leader may play an important role in inducing more efficient social outcomes, at least in one-shot interactions (I do not study repeated interactions with feedback between decisions). This as evidence suggests that people who behave as conditional cooperators when they observe all others' contributions, also behave as conditional cooperators when they only observe the leader's contribution (see Fischbacher et al. (2012)).

The small gender effect on contributions is consistent with the findings by Balliet et al. (2011) and Capraro (2018) of small gender differences in cooperation in social dilemmas, but the direction differ as men contributed more in the studies by Gächter et al. (2012) and me. This may be a result of the leadership setting creating gender-role interpretations rendering men more confident as leaders even in an experimental PGG, or merely due to a small sample deviation. I have a larger sample size than Gächter et al. (2012) which may render these findings more reliable. Yet, a limitation to the internal validity of my findings is that I cannot be sure whether participants did notice the different names indicating gender.

This limits the reliability of my findings, and they contrast those of Vyrastekova et al. (2015) that men's stated beliefs about others' contributions, and contributions, increase systematically with the number of other women in the group in a simultaneous PGG. They use the same three-player gender groups. One plausible explanation might be that gender stereotypes and beliefs were not activated in this setting. Albeit the rather small sample size limits the generalizability of their results, my results may also differ because the leader effect may have been stronger on men than on women. This as men's contributions were not systematically followed by their beliefs. Men's beliefs about followers' contributions were higher in female-dominated groups than in all-male groups similar to the findings by Born et al. (2019), but the contributions were on average highest in all-male groups, possibly indicating that men wanted to set an example as leaders. There is also no systematic pattern of men's beliefs similar to that of Vyrastekova et al. (2015) as beliefs were more positive in all-male groups than in male-dominated groups for men. Nor for females' beliefs as they were more positive in male-dominated groups in Vyrastekova et al. (2015).

Instead, similar to Born et al. (2019) female leaders' reported beliefs are highest in female-dominated groups, but there is no clear pattern in my data. Contrary to Born et al. (2019) the smallest difference between male and female leader contributions occurs in male-dominated groups. However, it is possible that the male-dominated sector setting in Born et al. (2019) did activate gender stereotypes which the PGs dilemma did not. This as sociocultural theory argues gender differences in cooperation might only emerge if gender stereotypes are activated. Therefore, future research might also benefit from altering the treatment groups and introducing gender differently. I had participants make decisions in all groups to increase the power and as I was interested in individual effects of the gender group composition. However, this may have created an experimental demand effect such that once participants had decided on a leader and follower strategy, there was not much reason to change strategy as only the names of the players changed. Indeed, the fact that 62 women and 90 men made the same decisions in all leader decisions may reflect this.

Using a within-design with respect to the gender group composition has the advantage that a smaller sample is needed than for a between-design. Thus, the probability of detecting an effect given that it exists in the population increases substantially. Without any feedback between the decisions the risk of ordering effects in a repeated measures design is decreased. Had I used a pure between-design I would have needed to detect a higher effect size and as the literature review showed mixed findings I chose to have participants make decisions in all gender group compositions. Further studies with higher power on my setup may confirm my findings.

Nevertheless, since men on average contributed more and expressed higher beliefs about followers' support it may indicate that men's higher leader contributions were driven by higher confidence in accordance with status character and social structural theories. It could also be consistent with an evolutionary theory at least in same-sex groups. However, as Balliet et al. (2011) concluded evolutionary theory is less clear in mixed groups, and it cannot solely explain that men contribute more in all groups. My findings are also not consistent with a sociocultural perspective that women would be more cooperative than men (and thus maybe be better role models and more cooperative followers). However, women do contribute more in male-dominated groups than female-dominated which is consistent with this theory. Had gender stereotypes been activated these patterns might have been stronger. Yet, these theories are not fully applicable but it may be due to my results being insignificant which prevents any conclusions to be drawn from the data. A different experimental setup to further explore the role of beliefs and gender roles might be beneficial, as the literature on gender differences in cooperation has focused on more on revealed preferences than on perceived preferences.

To investigate the role of gender stereotypes it may be beneficial to introduce gender differently. For example as Born et al. (2019) by letting participants see each other, using a between-design with respect to the gender group composition, and/or to randomly assign participants to either the leader or follower role. Then the follower's belief about other followers' contributions could be elicited following Gächter and Renner (2018). This to examine if female or male leaders are regarded as more or less influential. Such a design is out of the scope of this study, but it would perhaps yield less noisy data and be the least salient way to introduce gender. Vyrastekova et al. (2015) introduced gender by writing "female participant" or "male participant" and Holm (2000) find no significant difference when using this method as opposed to using female or male names. Yet, as mentioned, I cannot be sure that participants saw the names. When first testing the experiment on friends I afterwards first asked what they thought the study was about. If they did not mention gender I asked them if they did see the different names. All said yes, but I only had about ten participants.

There are weak indications of participants noticing the names as men changed follower behavior with the gender group composition. However, as a significant difference only occurred at the 5% level, and only for men, further research is required to confirm or reject these patterns. Men were more cooperative in male-dominated groups with a male leader than with a female leader. However, average follower contributions were highest in female-dominated groups with a female leader. Without significant results I cannot interpret if this indicates that female leaders are possibly seen as less influential leaders by men, but as more cooperative followers.

Nor did a systematic or significant pattern occur for women. Average follower contributions were higher in groups with female leaders than with male leaders. They were also on average higher in female-dominated groups than in male-dominated groups. Yet, women increased their contribution more with the leader's contribution in groups lead by a man than in groups lead by a woman, possibly indicating that they expect lower contributions from men. This is supported by female leaders' stated beliefs of followers support being higher in female-dominated groups than in male-dominated or female-dominated groups. However, it is partly contradicted as leader beliefs are higher in male-dominated groups than female-dominated groups. Without significant results these are only theories. In sum, my findings would indicate that in the presence of free-riding incentives the gender of the leader and the gender group composition would have a small impact, if any, on group outcomes using this experimental design.

Moreover, regarding conditional cooperation no pattern emerged of females being more sensitive to the first mover's contribution as found in some dictator games. Instead, men on average increase their follower contributions more with the leader's contribution than women. This may be a sample deviation, or further indicate that the type of game and number of players in the group may be important when evaluating gender differences in social preferences for cooperation. However, participants did not receive feedback between decisions in my study and future studies may benefit from study gender differences in leading-by-example and conditional cooperation in repeated interactions with feedback to understand what is driving the cooperative behavior.

That the gender group composition did not significantly moderate the relationship between gender and contributions is contrary to Balliet et al. (2011) and Vyrastekova et al. (2015), possibly due to gender stereotypes not being activated. The direction of the impact of the gender group composition on decisions in my data agree with those of Sell et al. (1993) and Boschini et al. (2012) indicating that these possible patterns of gender beliefs affect leader contributions similarly. Yet, similar to Sell et al. (1993) the gender group composition has no significant effect. However, the findings by Boschini et al. (2012) that the activation of gender stereotypes lead gender-primed men to become less generous in mixed-sex interactions, while women instead became more generous, indicate that if gender stereotypes are activated, the gender group composition might matter also in a leader-follower PGG, albeit somewhat contradicting findings by Ben-Ner et al. (2004a).

Furthermore, despite the contradicting evidence by Arbak and Villeval (2007) that no gender differences in willingness to lead persist when gender is announced, van Staveren et al. (2009) found that gender beliefs affected contributions in same-sex groups where the group composition was announced, controlling for risk-aversion and pro-social preferences. Despite a small sample this is consistent with the findings by Chermak and Krause (2002) that gender matters when individuals know the roles they are to play. This indicates that gender beliefs might persist even in "gender-neutral" experimental settings. Albeit my insignificant results would suggest that gender identity and roles are not taken into consideration when deciding how to lead and follow others in the PGs dilemma, weak signs of gender beliefs being present and affecting leader and follower behavior exist. Thus, further studies may benefit from higher power and/or altering the design to yield less noisy data.

In line with the literature review women on average expressed more risk-aversion than men and higher risk-aversion was negatively correlated with leader contributions to the PG after controlling for gender, but it was not significant at the pre-specified 1% level. Contrary to Gächter et al. (2012) risk-aversion was negatively correlated with leader contributions, but similarly very small in magnitude, possibly indicating that it is not an important predictor in this social dilemma setting. However, albeit not significant, less risk-averse females contributed more in the leader role than less risk-averse men. If beliefs do not change with the degree of risk-aversion for women, this may indicate that women's lower levels of beliefs are driven by men historically having had leader roles more often than women rendering women susceptible to the stereotype threat in line with status character theory. However, I do not have sufficient power to conclude this.

Another potential concern regarding the gender effect is the low stakes. I used Prolific Academic as I otherwise could not pay participants and set the reward high with Prolific standards. However, they are generally lower than those used in the laboratories by van Staveren et al. (2009), Gächter et al. (2012), Vyrastekova et al. (2015) and Gächter and Renner (2018). Would the gender effect be larger had the stakes been higher? However, van Staveren et al. (2009) find a significant effect of gender beliefs and the gender group composition with stakes not that higher than mine. Yet, higher stakes make cooperation more risky and women's stated risk-aversion were on average higher than men's. The role of gender beliefs (e.g., that women are more cooperative following a sociocultural theory) might then have had larger impact on decisions. Could it also have lead to a significant effect of the gender group composition? For higher stakes to have a gender effect they would need to impact men's and women's decisions differently, which is not evident.

I cannot be sure, but cooperation and defection levels in PGGs appear to be quite stable to the stake size, albeit cooperation appears to decrease with higher stakes. Kocher et al. (2008) find that an increase in stake size does not significantly affect cooperation, nor the level of punishment in the PGG. Camerer and Hogarth (1999) find that when incentives are low, participants state they would be more risk-preferring and generous than they actually are when incentives are increased. However, higher or lower incentives do not change average behavior substantively. Amir et al. (2012) find no differences in cooperation in the PGG, or trust or trustworthiness in the trust game, when comparing no stakes to $1 stakes on MTurk. The range of stakes is arguably quite small. Yet, conducting a meta-analysis on the effects stake size using 31 ultimatum and dictator games Larney et al. (2019) find that higher stakes reduce donations in the dictator game, albeit not by much, and have little to no effect in the ultimatum game. (A difference the authors attribute to proposers not wanting to risk their offer being rejected in the ultimatum game.)

Finally, most of the studies in the literature review were conducted with students and having a broader UK sample not restricted to younger students render the results more generalizable, even though my sample size still limits the generalizability of the results. As culture has been found to affect cooperation in PGGs I cannot easily generalize my findings to other cultures, despite the broader sample range. It is possible that a main effect of gender on leadership and cooperation, or on the gender group composition and beliefs, would be more pronounceable in cultures where gender norms and gender inequality are more pronounced in accordance with social structural, status character and sociocultural theories. Additionally, the contradicting findings of gender differences with respect to the gender group composition by Sell (1997) when money is no longer the resource, might indicate that also gender differences in preferences for leading-by-example and the impact of the gender group composition might be sensitive to the type resource used in the social dilemma.

6 Conclusion

This thesis aimed to investigate gender differences in preferences for leading-by-example and conditional cooperation in a PGs dilemma. To obtain an understanding of both revealed and perceived gender differences in cooperative behavior in a leadership context, I examined the role of the gender group composition and beliefs about followers' support in explaining leader contributions. Based on a quantitative analysis of leader and follower contributions to the PG, where men and women played both roles in male-dominated, female-dominated and same-sex groups, it cannot be concluded that women or men are better role models in leading-by-example, controlling for risk-aversion, age and beliefs. Nor can it be concluded that female or male leaders would be more successful in crowding in without significant results of either the gender group composition affecting leader or follower decisions, or of gender affecting leaders' reported beliefs about followers' support.

While this study clearly illustrates no significant gender effect on leading-by-example or beliefs, it raises the question whether women's on average lower leader contributions in all gender group compositions are driven by beliefs as they were followed by lower reported beliefs about followers' support. This while the negative effect on leader contributions by being a female appeared to be weakly mitigated by women with higher stated willingness to take risks. I cannot conclude it with insignificant results, but since Gächter et al. (2012) also reported that male leaders contributed more in a VCM game it may be worthwhile investigating if women exhibit more negative beliefs about their influence as leaders which may affect their revealed preferences. This, as the meta-analyzes indicated that women are marginally more cooperative in PGGs, PDs and dictator games, while when introducing leadership the contrary was observed. Previous findings on gender differences in co-operation related to the gender group composition and gender heterogeneity in gender beliefs also indicate that gender role interpretations may shape behavior in the leader-follower context had gender identity been activated.

As the results are insignificant further research on a larger study sample is required to establish whether or not the greater leader effect on men is significant. Albeit the effect appears to be small, it might be meaningful to investigate whether introducing leadership into the experimental context creates a stronger leader effect on men. It is related to status character theory and social structural theory as gender identity is an important factor in the leadership context. Albeit, the effect size appears to be small it might still have practical implication if such gender beliefs exist even in a simple PG experimental setting.

This study is relevant to research on leading-by-example and gender differences in cooperative behavior. While previous research has highlighted the important role a leader may play in inducing higher cooperation rates, less focus have been on gender differences in leading-by-example when leadership is imposed and gender announced. It is related to perceived gender differences in cooperation in social dilemmas as revealed differences have been found to depend on the social context such as the gender group composition, and inherent pro-social preferences should not change with the context. I extended on this to an imposed leadership context. While the generalizability of the results is limited to an experimental PGs dilemma, with a UK participant pool and using money as the resource, this approach provides insight into the potential fact that there might be no significant gender differences in preferences for leading-by-example and that men and women might thus be equally successful in leading groups to more efficient social outcomes. This as the correlation between individual leader and follower contributions was statistically significant for both men and women and similarly positive in magnitude, while gender differences in leader contributions appeared to be followed by stated beliefs.

Further studies may confirm if these findings hold when altering the experimental design to obtain less noisy measures. My findings hinge upon participants noticing the names indicating gender and that is was not the within-between design which lead participants to not alter strategies between the different leader and follower decisions. This as sociocultural theory argues that gender differences in social behavior may only occur if gender stereotypes are activated. This is a limitation and I am therefore cautious not to interpret my findings such that there would be no gender differences in leading-by-example in a setting where gender identity had been activated. Future research may address the role of gender identity in preferences for leading-by-example by introducing gender differently, alter the treatment groups and decision roles and/or using another type of resource than money. Or implementing it in another cultural setting. Nevertheless, my findings indicate that the gender effect on leading-by-example appears to be small. As leadership roles becomes more common for women, gender roles and stereotypes may disappear and change beliefs and revealed preferences.

References

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

Amir, O., Rand, D. G., et al. (2012). Economic games on the internet: The effect of $1 stakes. *PloS one*, 7(2):e31461.

Andersen, S., Bulte, E., Gneezy, U., and List, J. A. (2008). Do women supply more public goods than men? Preliminary experimental evidence from matrilineal and patriarchal societies. *American Economic Review*, 98(2):376–81.

Anderson, L. R., DiTraglia, F. J., and Gerlach, J. R. (2011). Measuring altruism in a public goods experiment: A comparison of US and Czech subjects. *Experimental Economics*, 14(3):426–437.

Arbak, E. and Villeval, M. C. (2007). Endogenous leadership selection and influence. Documents De Travail Working Paper No. 07-07, Available at SSRN: https://ssrn.com/abstract=988297.

Balliet, D., Li, N. P., Macfarlan, S. J., and Van Vugt, M. (2011). Sex differences in cooperation: A meta-analytic review of social dilemmas. *Psychological bulletin*, 137(6):881.

Barcelo, H. and Capraro, V. (2015). Group size effect on cooperation in one-shot social dilemmas. *Scientific reports*, 5:7937.

Belianin, A., Novarese, M., et al. (2005). Trust, communication and equlibrium behaviour in public goods. Experimental 0506001, University Library of Munich, Germany.

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

Ben-Ner, A., Putterman, L., Kong, F., and Magan, D. (2004b). Reciprocity in a two-part dictator game. *Journal of Economic Behavior & Organization*, 53(3):333–352.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300.

Bilén, D., Dreber, A., and Johannesson, M. (2020). Are women more generous than men? A meta-analysis, Available at SSRN: https://ssrn.com/abstract=3578038 or http://dx.doi.org/10.2139/ssrn.3578038.

Born, A., Ranehill, E., and Sandberg, A. (2019). A man's world? The impact of a male dominated environment on female leadership. Working Paper 744, Available at: http://hdl.handle.net/2077/58135, University of Gothenburg.

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

Brown-Kruse, J. and Hummels, D. (1993). Gender effects in laboratory public goods contribution: Do individuals put their money where their mouth is? *Journal of Economic Behavior & Organization*, 22(3):255–267.

Cadsby, C. B., Hamaguchi, Y., Kawagoe, T., Maynes, E., and Song, F. (2007). Cross-national gender differences in behavior in a threshold public goods game: Japan versus Canada. *Journal of Economic Psychology*, 28(2):242–260.

Cadsby, C. B. and Maynes, E. (1998). Gender and free riding in a threshold public goods game: Experimental evidence. *Journal of economic behavior & organization*, 34(4):603–620.

Camerer, C. F. (1997). Progress in behavioral game theory. *Journal of economic perspectives*, 11(4):167–188.

Camerer, C. F. and Hogarth, R. M. (1999). The effects of financial incentives in experiments: A review and capital-labor-production framework. *Journal of risk and uncertainty*, 19(1):7–42.

Capraro, V. (2018). Women are slightly more cooperative than men (in one-shot prisoner's dilemma games played online). *arXiv preprint arXiv:1805.08046, Cornell University*.

Cartwright, E. J. and Lovett, D. (2014). Conditional cooperation and the marginal per capita return in public good games. *Games*, 5(4):234–256.

Charness, G. and Gneezy, U. (2012). Strong evidence for gender differences in risk taking. *Journal of Economic Behavior & Organization*, 83(1):50–58.

Chen, S.-Y., Feng, Z., and Yi, X. (2017). A general introduction to adjustment for multiple comparisons. *Journal of thoracic disease*, 9(6):1725.

Chermak, J. M. and Krause, K. (2002). Individual response, information, and intergenerational common pool problems. *Journal of Environmental Economics and Management*, 43(1):47–70.

Cislaghi, B. and Heise, L. (2020). Gender norms and social norms: Differences, similarities and why they matter in prevention science. *Sociology of health & illness*, 42(2):407–422.

Cohen, J. (2013). *Statistical power analysis for the behavioral sciences*. Academic press.

Cox, J. C. and Deck, C. A. (2006). When are women more generous than men? *Economic Inquiry*, 44(4):587–598.

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

Croson, R., Marks, M., and Snyder, J. (2008). Groups work for women: Gender and group identity in social dilemmas. *Negotiation Journal*, 24(4):411–427.

Croson, R. T. (2000). Thinking like a game theorist: Factors affecting the frequency of equilibrium play. *Journal of economic behavior & organization*, 41(3):299–314.

Croson, R. T. (2007). Theories of commitment, altruism and reciprocity: Evidence from linear public goods games. *Economic Inquiry*, 45(2):199–216.

Dawes, R. M. and Thaler, R. H. (1988). Anomalies: Cooperation. *Journal of economic perspectives*, 2(3):187–197.

Dohmen, T. J., Falk, A., Huffman, D., Sunde, U., Schupp, J., and Wagner, G. G. (2005). Individual risk attitudes: New evidence from a large, representative, experimentally-validated survey. IZA Discussion Paper No 1730, Available at SSRN: https://ssrn.com/abstract=807408.

Eagly, A. H. (2009). The his and hers of prosocial behavior: An examination of the social psychology of gender. *American psychologist*, 64(8):644.

Eagly, A. H. and Crowley, M. (1986). Gender and helping behavior: A meta-analytic review of the social psychological literature. *Psychological bulletin*, 100(3):283.

Eagly, A. H. and Wood, W. (1999). The origins of sex differences in human behavior: Evolved dispositions versus social roles. *American psychologist*, 54(6):408.

Eek, D. and Biel, A. (2003). The interplay between greed, efficiency, and fairness in public-goods dilemmas. *Social Justice Research*, 16(3):195–215.

Eklund, K. E., Barry, E. S., and Grunberg, N. E. (2017). Gender and leadership. In Alvinius, Aida., editor, *Gender differences in different contexts* (pp. 129–150). BoD–Books on Demand. https://doi.org/10.5772/63040.

Faul, F., Erdfelder, E., Lang, A.-G., and Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, 39:175–191.

Fehr, E. and Fischbacher, U. (2002). Why social preferences matter–the impact of non-selfish motives on competition, cooperation and incentives. *The economic journal*, 112(478):C1–C33.

Fischbacher, U. and Gachter, S. (2010). Social preferences, beliefs, and the dynamics of free riding in public goods experiments. *American economic review*, 100(1):541–56.

Fischbacher, U., Gächter, S., and Fehr, E. (2001). Are people conditionally cooperative? Evidence from a public goods experiment. *Economics letters*, 71(3):397–404.

Fischbacher, U., Gächter, S., and Quercia, S. (2012). The behavioral validity of the strategy method in public good experiments. *Journal of Economic Psychology*, 33(4):897–913.

Gächter, S., Nosenzo, D., Renner, E., and Sefton, M. (2012). Who makes a good leader? Cooperativeness, optimism, and leading-by-example. *Economic Inquiry*, 50(4):953–967.

Gächter, S. and Renner, E. (2004). Leading by example in the presence of free riding incentives. Working paper, University of Nottingham.

Gächter, S. and Renner, E. (2010). The effects of (incentivized) belief elicitation in public goods experiments. *Experimental Economics*, 13(3):364–377.

Gächter, S. and Renner, E. (2018). Leaders as role models and 'belief managers' in social dilemmas. *Journal of Economic Behavior & Organization*, 154:321–334.

Glen, S. (2017). Clustered standard errors: Definition. From StatisticsHowTo.com, Retrieved on February 16, 2021 from: https://www.statisticshowto. com/clustered-standard-errors/.

Goeree, J. K., Holt, C. A., and Laury, S. K. (2002). Private costs and public benefits: Unraveling the effects of altruism and noisy behavior. *Journal of Public Economics*, 83(2):255–276.

Greig, F. and Bohnet, I. (2009). Exploring gendered behavior in the field with experiments: Why public goods are provided by women in a nairobi slum. *Journal of Economic Behavior & Organization*, 70(1-2):1–9.

Güth, W., Levati, M. V., Sutter, M., and Van Der Heijden, E. (2007). Leading by example with and without exclusion power in voluntary contribution experiments. *Journal of Public Economics*, 91(5-6):1023–1042.

Hardy, C. L. and Van Vugt, M. (2006). Nice guys finish first: The competitive altruism hypothesis. *Personality and Social Psychology Bulletin*, 32(10):1402–1413.

Hauert, C. and Szabo, G. (2003). Prisoner's dilemma and public goods games in different geometries: Compulsory versus voluntary interactions. *Complexity*, 8(4):31–38.

Holcombe, R. G. (2008). Why does government produce national defense? *Public Choice*, 137(1-2):11–19.

Holm, H. J. (2000). Gender-based focal points. *Games and Economic Behavior*, 32(2):292–314.

Holt, C. A. and Laury, S. K. (2002). Risk aversion and incentive effects. *American economic review*, 92(5):1644–1655.

Hyde, J. S. (2005). The gender similarities hypothesis. *American psychologist*, 60(6):581.

Hyde, J. S. (2007). New directions in the study of gender similarities and differences. *Current Directions in Psychological Science*, 16(5):259–263.

Isaac, R. M. and Walker, J. M. (1988). Group size effects in public goods provision: The voluntary contributions mechanism. *The Quarterly Journal of Economics*, 103(1):179–199.

Isaac, R. M., Walker, J. M., and Williams, A. W. (1994). Group size and the voluntary provision of public goods: Experimental evidence utilizing large groups. *Journal of public Economics*, 54(1):1–36.

Jackson, J. W. (2001). Effects of endowment amount and attributions on responses to a subgroup social dilemma. *Group Dynamics: Theory, Research, and Practice*, 5(2):67.

Keser, C. and Van Winden, F. (2000). Conditional cooperation and voluntary contributions to public goods. *Scandinavian Journal of Economics*, 102(1):23–39.

Kocher, M. G., Martinsson, P., and Visser, M. (2008). Does stake size matter for cooperation and punishment? *Economics Letters*, 99(3):508–511.

Kopelman, S., Weber, J. M., and Messick, D. M. (2002). Factors influencing cooperation in commons dilemmas: A review of experimental psychological research. In E. Ostrom, T. Dietz, N. Dolšak, P. C. Stern, S. Stonich, E. U. Weber (Eds.) Committee on the Human Dimensions of Global Change, Division of Behavioral and Social Sciences and Education, *The drama of the commons* (pp. 113–156). National Academy Press.

Kube, S. and Traxler, C. (2011). The interaction of legal and social norm enforcement. *Journal of Public Economic Theory*, 13(5):639–660.

Larney, A., Rotella, A., and Barclay, P. (2019). Stake size effects in ultimatum game and dictator game offers: A meta-analysis. *Organizational Behavior and Human Decision Processes*, 151:61–72.

Lazic, S. E. (2008). Why we should use simpler models if the data allow this: Relevance for anova designs in experimental biology. *BMC physiology*, 8(1):1–7.

Lazzeroni, L. and Ray, A. (2012). The cost of large numbers of hypothesis tests on power, effect size and sample size. *Molecular psychiatry*, 17(1):108–114.

Lenhard, W. and Lenhard, A. (2016). Calculation of effect sizes. Dettelbach (Germany): Psychometrica., Retrieved on April 9, 2021 from: https://www.psychometrica.de/effect_size.html.

Lurye, L. E., Zosuls, K. M., and Ruble, D. N. (2008). Gender identity and adjustment: Understanding the impact of individual and normative differences in sex typing. *New directions for child and adolescent development*, 2008(120):31–46.

Mason, C. F., Phillips, O. R., and Redington, D. B. (1991). The role of gender in a non-cooperative game. *Journal of Economic Behavior & Organization*, 15(2):215–235.

Moxnes, E. and Van der Heijden, E. (2003). The effect of leadership in a public bad experiment. *Journal of Conflict Resolution*, 47(6):773–795.

Nelson, J. A. (2016). Not-so-strong evidence for gender differences in risk taking. *Feminist Economics*, 22(2):114–142.

Nowell, C. and Tinkler, S. (1994). The influence of gender on the provision of a public good. *Journal of Economic Behavior & Organization*, 25(1):25–36.

Ortmann, A. and Tichy, L. K. (1999). Gender differences in the laboratory: Evidence from prisoner's dilemma games. *Journal of Economic Behavior & Organization*, 39(3):327–339.

Schielzeth, H., Dingemanse, N. J., Nakagawa, S., Westneat, D. F., Allegue, H., Teplitsky, C., Réale, D., Dochtermann, N. A., Garamszegi, L. Z., and Araya-Ajoy, Y. G. (2020). Robustness of linear mixed-effects models to violations of distributional assumptions. *Methods in Ecology and Evolution*, 11(9):1141–1152.

Schubert, R., Brown, M., Gysler, M., and Brachinger, H. W. (1999). Financial decision-making: Are women really more risk-averse? *American economic review*, 89(2):381–385.

Sell, J. (1997). Gender, strategies, and contributions to public goods. *Social Psychology Quarterly*, 60(3):252–265.

Sell, J., Griffith, W. I., and Wilson, R. K. (1993). Are women more cooperative than men in social dilemmas? *Social Psychology Quarterly*, 56(3):211–222.

Selten, R. (1967). Die Strategiemethode zur Erforschung des Eingeschräncht Rationalen Verhaltens in Rahmen eines Oligopolexperiments', in Beitraägge zur Experimentellen Wirtschaftforschung, ed. by H. Sauermann, Tübingen, Germany: JCB Mohr.

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

Steiger, J. H. (2010). Modeling residual covariance structure. Vanderbilt University, Retrieved on February 16, 2021 from: http://www.statpower.net/ Content/GCM/Lectures/SW07.pdf.

Van Anders, S. M., Steiger, J., and Goldey, K. L. (2015). Effects of gendered behavior on testosterone in women and men. *Proceedings of the National Academy of Sciences*, 112(45):13805–13810.

van Staveren, I., Sent, E.-M., and Vyrastekova, J. (2009). Gender beliefs and cooperation in a public goods game experiment. *RePub*. Publications from Erasmus University, Rotterdam.

Van Vugt, M., Meertens, R. M., and Van Lange, P. A. (1995). Car versus public transportation? the role of social value orientations in a real-life social dilemma. *Journal of applied social psychology*, 25(3):258–278.

Vyrastekova, J., Sent, E.-M., and van Staveren, I. (2015). Gender beliefs and cooperation in a public goods game. *Economics Bulletin*, 35(2):1148–1153.

Walker, D. A. (2015). Jmasm34: Two Group Program for Cohen's d, Hedges'g, η2, Radj2, ω2, ϵ2, Confidence Intervals, and Power. *Journal of Modern Applied Statistical Methods*, 14(2):21.

Weber, J. M., Kopelman, S., and Messick, D. M. (2004). A conceptual review of decision making in social dilemmas: Applying a logic of appropriateness. *Personality and Social Psychology Review*, 8(3):281–307.

Wooldridge, J. M. (2012). *Introductory Econometrics: A Modern Approach, Fifth Edition*. Mason, Ohio: South-Western Cengage Learning.

Zelmer, J. (2003). Linear public goods experiments: A meta-analysis. *Experimental Economics*, 6(3):299–310.

A Experimental instructions

I use experimental instructions by Gächter et al. (2012) and Gächter and Renner (2018) available online, as well as instructions from example experiments made available by researchers in the Lioness repository. Specifically, I use "Public Goods Game with punishment" created by Lucas Molleman,University of Nottingham (UK), "Advisorsgroups lab" created by Marta Balode, Frele Universität Berlin (Germany), and "Multi-level PGG (4 persons)" created by Jered Abernathy, University of South Carolina (United States). Below follow an overview of the experimental screens.

Introduction description on Prolific:
This study takes on average 20 minutes to complete and is conducted as part of a master thesis project in economics. The purpose of this study is to examine how individuals make decisions within a social context. You will be asked to:

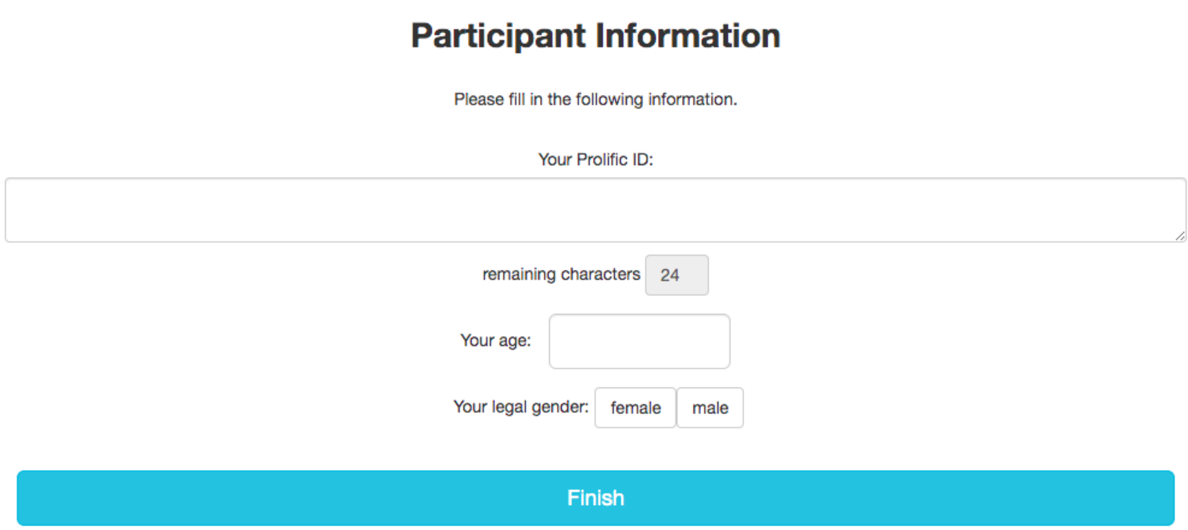
- Answer 2 brief demographic questions.
- Make individual and anonymous decisions within a group context.
- Answer 3 short survey questions.

In addition to the £3.5 you can earn up to £4.25 in bonus.

Important: use an updated web browser such as Google Chrome, Apple Safari, Mozilla Fire Fox, or Opera. JavaScript must be enabled. **Internet Explorer cannot** be used for this study.

Link to enabling JavaScript in common web browsers: https://www.enable-javascript.com

Screen 1: Participant information
Webpage name for subjects: ID



Screen 2: General information and consent from
Webpage name for subjects: Consent Form

Study Consent Form and General Information

Thank you for your interest in this study. Please read the information below.

1. Aim of the study

We want to examine how individuals in a group context make decisions that will affect both themselves and other group members.

2. Content of the study

The setup of this study is different from other Prolific studies you might be used to. It involves you and other **real** Prolific players deciding how to allocate your endowments, **called Tokens**, between your private funds and a joint fund for a group project. Your endowment is a proxy for your knowledge, means (monetary or other types), etc. You will be randomly matched to 3 groups. Your task is to make 9 **individual decisions**, 3 in each group.

3. Data Protection

Participation is **anonymous**. Your decisions and Prolific id will never be revealed to the other participants. Instead, you will be given a **random alias** which will be shown to your group members. If you are a man, you will be given an alias like 'Peter'. If you are a woman, you will be given an alias like 'Lisa'. Please **do not discuss** this study with other Prolific players.

The study data will be used for research purposes only and stored anonymously under an individual code number. In line with the research institution's data privacy policy, your Prolific id will be deleted from the study data so that there is no link to your response data. The study data (without your Prolific id) may also be made publicly accessible via research databases or scientific publications to enable verification and replication of the results. The study data may also be used for new research questions going beyond the purposes of this particular study.

4. Voluntary Participation

Participation is **voluntary** and you can withdraw from the study at any time.

5. Risks involved

None.

6. Payment

Upon completion all participants are paid **£3.5**. This consists of a participation fee of £3.25, and a **bonus** ranging from £0.25 to £1 depending on the decisions you and other participants make. Your earned **bonus** depends on the amount of Tokens you accumulate (1 Token = £0.1), and will be paid privately to you via Prolific. As a researcher, I commit myself to not deceive participants. Upon your completion, group members decisions will be retrieved which will determine your earnings. One **'round'/group** will be randomly selected to calculate payments.

7. Time for the study

The estimated completion time is 20 minutes. During the experiment, please **do not close the browser window** or leave the experiment's web pages. If you do close your browser or leave the experiment, you will not be able to re-enter and we will not be able to pay you.

Responsible researcher: Johanna Ohlin. For more information, please contact me at email: 23731@student.hhs.se

Would you like to participate? By clicking on 'I agree' you are indicating that you have read and understood the information above, that you consent to participate in the study, and agree to the collection, storage, and use of your response data as described above.

I agree

To Instructions 1/2

Screen 3: The general decision situation

Webpage name for subjects: Instructions 1/2

The General Decision Situation

Groups consists of 3 players. Each group member receives **5 Tokens**.
Each member decides individually and privately how many of the 5 Tokens to **contribute** to the **group project**.
Each Token you do not **contribute** you keep **for private use**.
All group members benefit **equally** from the Tokens allocated to the project,
regardless of how many Tokens each participant allocated to the project.

To see this:
After all members have made their decisions, all Tokens contributed to the **group project** are added up and **multiplied by 0.5**.
The resulting number of Tokens is then distributed to **all** members.

In summary
Your total earnings = private earnings + earnings from group project
= (5 - *your contribution*) + 0.5 × (*sum of all members' contribution)*

Group project - Example 1

- All 3 members contribute 5 Tokens to the group project.
- Sum of contributions is 15 Tokens.
- This amount is multiplied by 0.5, resulting in 7.5 Tokens.
- Each member receives 7.5 Tokens from the project.
- Private earnings are 0 for all members.

- Each member earns 7.5 Tokens in total.

Group project - Example 2

- All members contribute each 0 Tokens to the project.
- Sum of contributions is 0 Tokens.
- Private earnings equals 5 Tokens for all members.

- All members earn 5 Tokens in total.

Group project - Example 3

- Members A and B contribute each 5 Tokens to the group project.
- Member C contributes 0 Tokens.
- Sum of contributions is 10 Tokens.
- This amount is multiplied by 0.5, resulting in 5 Tokens.
- Each member earns 5 Tokens from the group project.
- Private earnings equals 0 Tokens for members A and B.
- Private earnings equals 5 Tokens for member C.

- Members A and B earn 5 Tokens.
- Member C earns 10 Tokens (5 from the private fund *plus* 5 from the group project).

To Instructions 2/2

Back to Consent Form

Screen 4: The specific decision situations

Webpage name for subjects: Instructions 2/2

**Avatar icon made by Icons8 from www.icons8.com*

The specific Decision Situations

Contribution Decisions

Your will make contribution decisions for **2 roles in the group:** FIRST MOVER and SECOND MOVER.
All group members play all roles. Thus in each group:
you will make **1** contribution decision as the FIRST MOVER and **2** as one of the SECOND MOVERS.
You will see these decisions on 9 different screens.

The order in which these appear is **randomized for each player** to make sure everyone is paying attention.

FIRST-MOVER role:
You will be asked to decide **first** how many Tokens to contribute to the project,
before any of your group members make theirs.
You can enter any whole number from 0 to 5 (inclusive).
For your chosen contribution, you are also asked to **guess** how many Tokens the other 2 members will contribute on average.
You have an incentive to try to make an accurate guess.

FIRST MOVER decision

FIRST MOVER : **You**

JESSICA

You decide **first** how much to contribute to the project.
You have **5 Tokens** to allocate.

SECOND MOVERS:

DAVID THOMAS

Your total earnings = (5 - *your contribution to the project*) + 0.5 × (*sum of all members contribution to the project*)

Your contribution to the group project as the FIRST MOVER:

Your **guess** of the SECOND MOVERS' average contribution to the project for your chosen contribution:
Any whole number from 0 to 5 is possible to enter.

SECOND-MOVER role:
You will be asked to make a contribution decision for **every possible contribution** by the FIRST MOVER. That is,
• what you would do if the FIRST MOVER contributed 0 Tokens?
• what you would do if the FIRST MOVER contributed 1 Token? etc.

SECOND MOVER decision

SECOND MOVERS:

You: JESSICA ELIZABETH

You and the other SECOND MOVER decide **simultaneously**, i.e. without observing each others' contributions.

FIRST MOVER:

ALICE

Your total earnings = (5 - *your contribution to the project*) + 0.5 × (*sum of all members contribution to the project*)

Please indicate, for each possible FIRST MOVER contribution below, how much you want to contribute to the project.
You can enter any whole number from 0 to 5 for each level.

0:	**1:**	**2:**
3:	**4:**	**5:**

No information of other members' contribution will be given.

How Earnings are determined

Within the group relevant roles for payment will be randomly selected; FIRST MOVER or SECOND MOVER.
• If your chosen role is **FIRST MOVER** your relevant contribution is what you entered in this role, in this group.
Your members' relevant contributions is determined by their decision for your FIRST MOVER contribution, in their role as SECOND MOVERS.
• If your chosen role is **SECOND MOVER**, yours, and the other SECOND MOVER's, relevant decisions are determined by the FIRST MOVER'S contribution.

All contributions are then added up to calculate group earnings.

To make sure everyone understands how their earnings are determined, you will be asked to answer 3 simple control questions next.
The computer will directly check your answers.

To the Quiz

Back to Instructions 1/2

Screen 5: Control questions
Webpage name for subjects: Quiz

Quiz

Your total earnings = (5 - *your contribution to the project*) + 0.5 × (*sum of all members contribution to the project)*
I.e. first sum all contributions, then multiply the sum by 0.5, equivalent to dividing the sum by 2.

1. Each group member is endowed with 5 Tokens.

Suppose both SECOND MOVERS contribute **5 Tokens** each to the project.
Suppose the FIRST MOVER contributes **0 Tokens** to the project.

How many Tokens will the FIRST MOVER earn in *total*?

How many Tokens would a SECOND MOVER earn in *total*?

2. Each group member is endowed with 5 Tokens.

Suppose the FIRST MOVER contributes **5 Tokens** and one of the SECOND MOVERS contributes **4 Tokens** to the project.
Suppose the other SECOND MOVER contributes **1 Token**.

How many Tokens will the SECOND MOVER who contributed **1 Token** earn *in total*?

Finish

Attempts left to answer the control questions: 8

Back to Instructions 2/2

Screen 6: Own alias
Webpage name for subjects: Alias
**Avatar icon made by Icons8 from www.icons8.com*

Are you ready?!

Your assigned alias is:

JESSICA

Press 'Start' to go to the decision screens.

Please note that on the 9 decision screens, you are asked to **confirm your choices**.
This as once you press the 'Submit' button at the bottom you **cannot** change them.

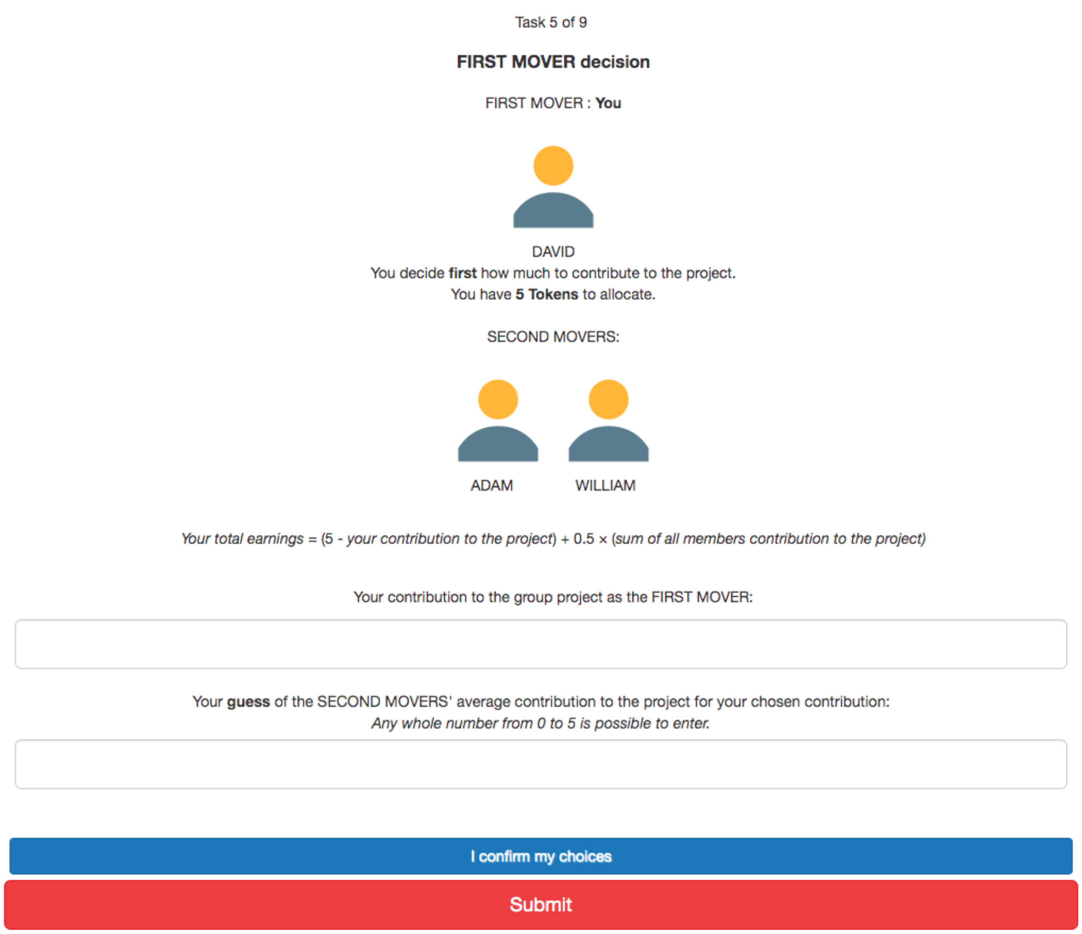
Start

Decision loop: first and second mover screens

Webpage name for subjects: FM1, SM1, SM2, FM2, SM3, SM4, FM3, SM5, SM6

**Avatar icon made by Icons8 from www.icons8.com*

First mover screen example:


Task 5 of 9

FIRST MOVER decision

FIRST MOVER : **You**

DAVID

You decide **first** how much to contribute to the project.

You have **5 Tokens** to allocate.

SECOND MOVERS:

ADAM WILLIAM

Your total earnings = (5 - *your contribution to the project*) + 0.5 × (*sum of all members contribution to the project*)

Your contribution to the group project as the FIRST MOVER:

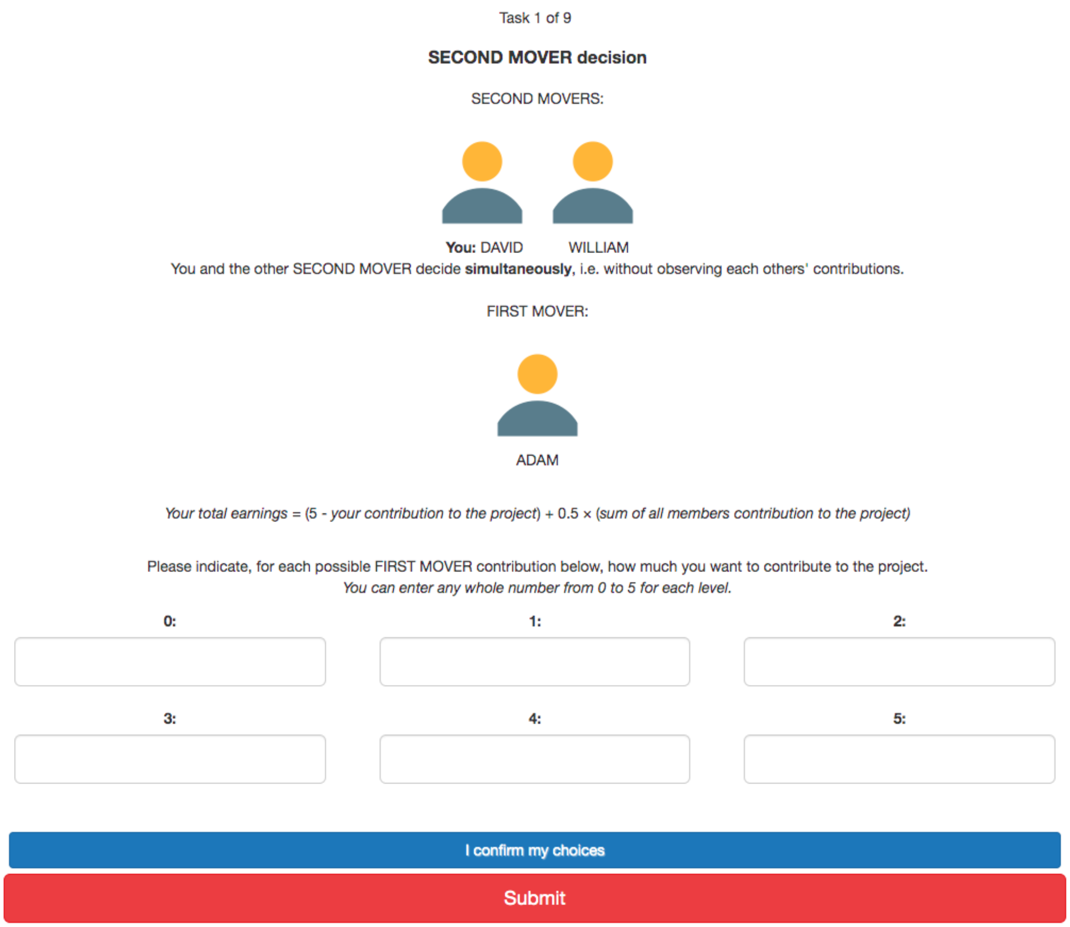
Your **guess** of the SECOND MOVERS' average contribution to the project for your chosen contribution:

Any whole number from 0 to 5 is possible to enter.

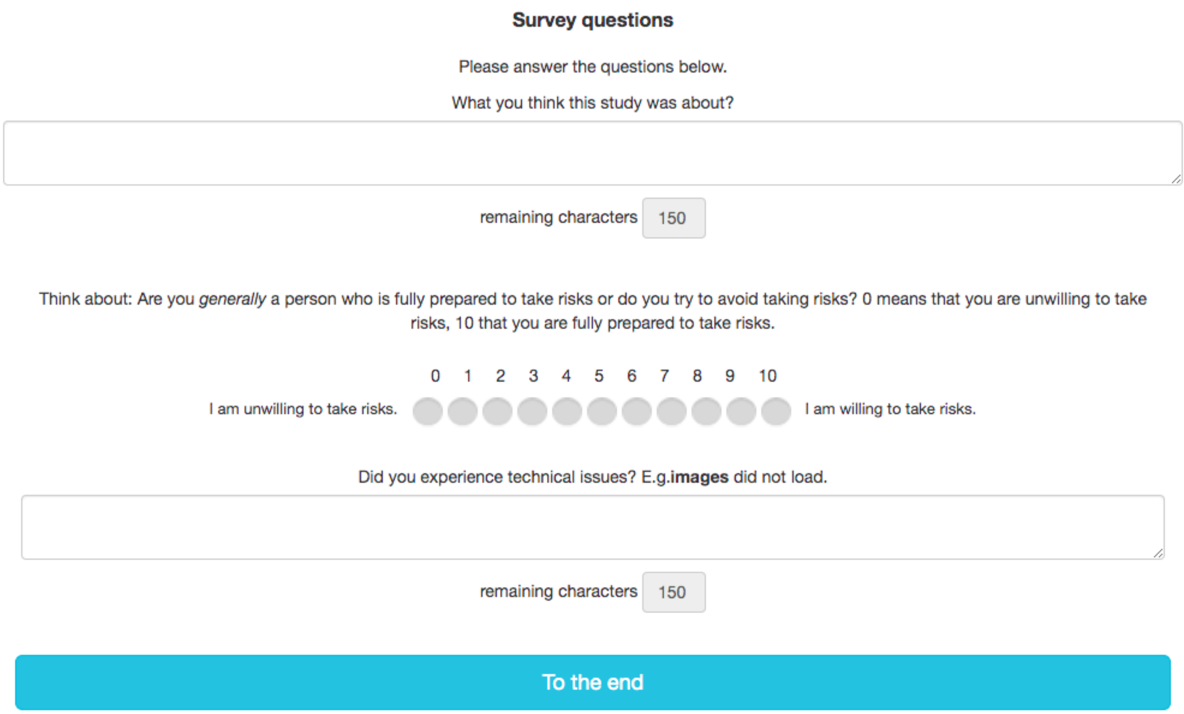
I confirm my choices

Submit

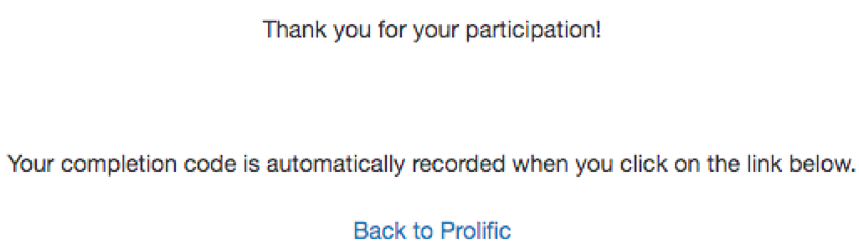
Second mover screen example:



Screen 16: Survey questions
Webpage name for subjects: Survey Questions



Screen 17: End of experiment
Webpage name for subjects: End



B Hypotheses

Hypothesis 1

$H0$:
$\beta_1 > 0$ and $\beta_1 + \beta_3 < 0$
or
$\beta_1 < 0$ and $\beta_1 + \beta_3 > 0$
or
$\beta_1 > 0$ and $\beta_3 < 0$
or
$\beta_1 < 0$ and $\beta_3 > 0$
or
$\beta_1 + \beta_3 > 0$ and $\beta_3 < 0$
or
$\beta_1 + \beta_3 < 0$ and $\beta_3 > 0$

$H1$:
$\beta_1 > 0$ and $\beta_1 + \beta_3 > 0$ and $\beta_3 > 0$
or
$\beta_1 < 0$ and $\beta_1 + \beta_3 < 0$ and $\beta_3 < 0$

Hypothesis 2

$H0 : \beta_2 = 0$ and $\beta_2 + \beta_3 = 0$

$H1 : \beta_2 \neq 0$ or $\beta_2 + \beta_3 \neq 0$

Hypothesis 3

$H0 : \beta_3 = 0$
$H1 : \beta_3 \neq 0$

C OLS diagnostics

Note. The regression residuals are not normally distributed. This can be seen in the histogram in Figure C.1 below of the OLS residuals from model (3) in Table 4.4, where a normal curve are plotted and density is on the y-axis, or in the normal Q-Q graph in Figure C.2 where observations are both right- and left-tailed.

Figure C.1: Residuals normality check for model (3) in Table 4.4 – full sample.

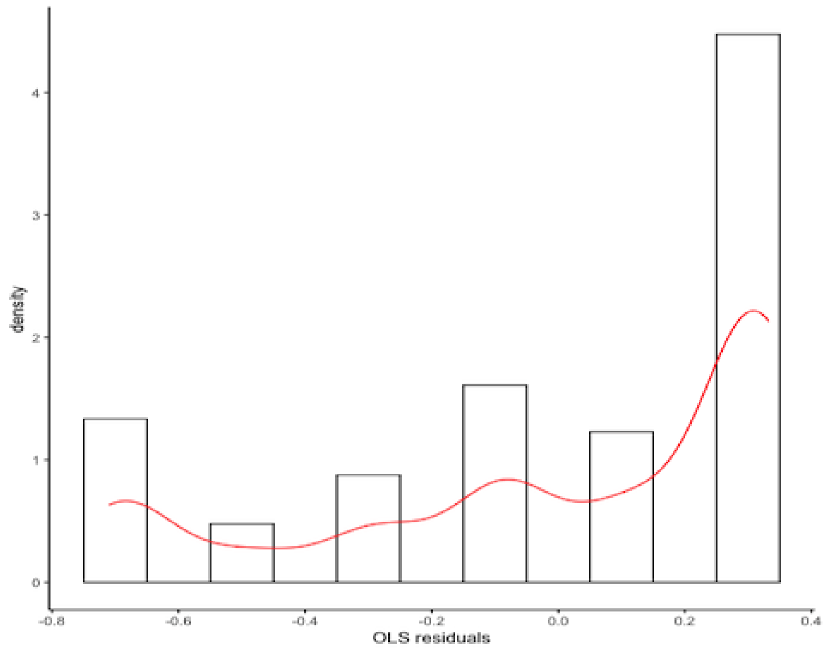


Figure C.2: Residuals normality QQ-check for model (3) in Table 4.4 – full sample.

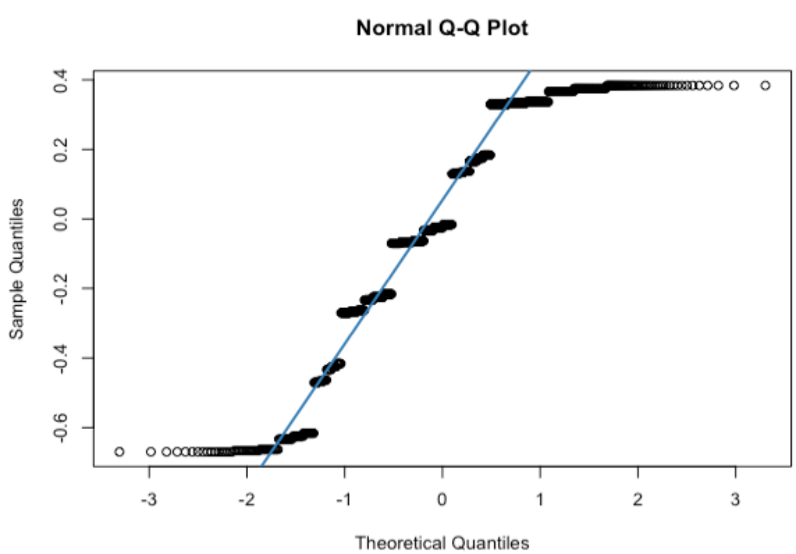


Table C.1: OLS correlation analysis – full sample.

	LCppc	LBppc	Female	noFemales	RA	age
LCppc	1.00	0.49	-0.04	-0.01	0.12	-0.01
LBppc	0.49	1.00	-0.07	0.02	0.14	0.03
Female	-0.04	-0.07	1.00	0.00	-0.20	-0.12
noFemales	-0.01	0.02	0.00	1.00	0.00	0.00
RA	0.12	0.14	-0.20	0.00	1.00	-0.07
age	-0.01	0.03	-0.12	0.00	-0.07	1.00

D Robustness checks

D.0.1 Hypotheses 1–3: Leader behavior

Table D.1: Summary statistics — without those who thought the study was about gender (157 women, 156 men left).

		FFM		MMF		FFF/MMM	
		LC	LB	LC	LB	LC	LB
Men	mean	3.48	3.37	3.42	3.28	3.52	3.32
	median	4	4	4	3	4	3
	sd	1.80	1.63	1.85	1.60	1.82	1.59
Women	mean	3.26	3.06	3.34	3.18	3.24	3.13
	median	4	3	4	3	4	3
	sd	1.81	1.60	1.78	1.62	1.83	1.65
All	mean	3.37	3.21	3.38	3.23	3.38	3.23
	median	4	3	4	3	4	3
	sd	1.80	1.62	1.81	1.61	1.83	1.62

Table D.2: Summary statistics — without those with completion time < 10 (149 women, 155 men left).

		FFM		MMF		FFF/MMM	
		LC	LB	LC	LB	LC	LB
Men	mean	3.43	3.33	3.39	3.26	3.56	3.25
	median	4	4	4	3	4	3
	sd	1.81	1.66	1.84	1.58	1.82	1.64
Women	mean	3.36	3.16	3.46	3.22	3.39	3.27
	median	4	3	4	3	3	3
	sd	1.76	1.55	1.71	1.62	1.80	1.61
All	mean	3.40	3.25	3.43	3.24	3.47	3.26
	median	4	3	4	3	4)	3
	sd	1.78	1.60	1.78	1.60	1.81	1.62

Table D.3: Summary statistics — without completion time <9) (161 women, 161 men left).

		FFM		MMF		FFF/MMM	
		LC	LB	LC	LB	LC	LB
Men	mean	3.45	3.35	3.40	3.27	3.58	3.29
	median	4	4	4	3	5	3
	sd	1.80	1.64	1.84	1.59	1.82	1.64
Women	mean	3.33	3.10	3.43	3.14	3.38	3.21
	median	4	3	4	3	4	3
	sd	1.77	1.57	1.74	1.62	1.79	1.60
All	mean	3.39	3.22	3.42	3.21	3.48	3.25
	median	4	3	4	3	4	3
	sd	1.78	1.61	1.79	1.61	1.80	1.62

Table D.4: Summary statistics — without those who could not see all images (170 women, 172 men left).

		FFM		MMF		FFF/MMM	
		LC	LB	LC	LB	LC	LB
Men	mean	3.51	3.40	3.46	3.33	3.60	3.34
	median	4	4	4	3	5	3
	sd	1.78	1.60	1.82	1.57	1.81	1.61
Women	mean	3.32	3.05	3.36	3.09	3.37	3.19
	median	4	3	4	3	4	3
	sd	1.78	1.56	1.76	1.62	1.78	1.61
All	mean	3.41	3.23	3.41	3.21	3.49	3.27
	median	4	3	4	3	4	3
	sd	1.78	1.59	1.79	1.60	1.80	1.61

D.0.2 Tobit regressions

Table D.5: Tobit regressions explaining leader contributions to the PG – full sample.

	Dependent variable: % of tokens contributed in the leader role (LC)			
	(1)	(2)	(3)	(4)
Female (1=female, 0=male)	-0.034	-0.110	-0.122	-0.122
	(0.037)	(0.085)	(0.091)	(0.089)
NoFemales (0,1 or 2 other females in the group)	-0.007	-0.032	-0.026	-0.026
	(0.012)	(0.049)	(0.029)	(0.031)
Female · NoFemales	0.003	0.028	0.024	0.024
	(0.019)	(0.064)	(0.043)	(0.042)
Constant	0.710***	0.965***	0.977***	0.977***
	(0.026)	(0.068)	(0.069)	(0.065)
Total observations	1050	1050	1050	1050
Individual observations	350	350	350	350
$Wald(chi)$ statistic	1.39	2.93	2.26	2.31

Sig. codes: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Note to Table D.7: Model (1) corresponds to the main OLS regression (3) in Table 4.4. Model (2) shows the corresponding Tobit regression with clustered (heteroskedasticity) robust standard errors at the individual level, left-censored at zero and right-censored at one. Model (3) shows the panel Tobit model with individual random effects and heteroskedasticity robust standard errors. Finally, Model (4) corresponds to Model (3), but now without heteroskedasticity robust standard errors.

Table D.6: Marginal effects for Tobit regressions in model (2) Table D.5 and model (2) in Table D.7.

	Dependent variable: Leader contribution (LC)	
	Model (3) Table in 5.6	Model (2) in Table D.7
Female (1=female, 0=male)	-0.048	-0.020
NoFemales (0,1 or 2 other females in the group)	-0.014	-0.017
Female · NoFemales	0.012	0.009
Leader's belief (LB)		0.628
Total observations	1050	1050
Individual observations	350	350

Table D.7: Tobit regressions explaining leader contributions to the PG – full sample.

	Dependent variable: % of tokens contributed as leader (LC)		
	(1)	(2)	(3)
Female (1=female, 0=male)	-0.008	-0.042	-0.064
	(0.033)	(0.080)	(0.076)
NoFemales (0,1 or 2 other females in the group)	-0.009	-0.035	-0.030
	(0.013)	(0.048)	(0.030)
Female · NoFemales	0.001	0.019	0.019
	(0.019)	(0.019)	(0.043)
% leader's belief (LB)	0.550^{***}	1.286^{***}	1.024^{***}
	(0.045)	(0.099)	(0.116)
Constant	0.345^{***}	0.105^{***}	0.282^{***}
	(0.042)	(0.097)	(0.092)
Total observations	1050	1050	1050
Individual observations	350	350	350
$Wald(chi)$ statistic	155.95^{***}	246.70^{***}	80.38^{***}

Sig. codes: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Note to Table D.7: Model (1) corresponds to the OLS regression with heteroskedasticity robust standard errors, clustered at the individual level, in model (3) in Table 4.10. Model (2) shows the corresponding Tobit regression with clustered (heteroskedasticity) robust standard errors at the individual level, left-censored at zero and right-censored at one. Caluclating the marginal effect for model (2) (see Table D.6 above) show that they are similar. Finally, model (3) shows a panel Tobit model with individual random effects and heteroskedasticity robust standard errors.

D.0.3 Hypotheses 4: Follower behavior

Table D.8: Hypothesis 4 - Women: without those who thought the study was about gender.

Group	CC_g	$CCslope_g$
FFM:LeaderF	2.41	0.344
	(1.23)	(0.545)
FFM:LeaderM	2.35	0.321
	(1.31)	(0.545)
MMF:LeaderM	2.30	0.360
	(1.29)	(0.526)
FFF:LeaderF	2.41	0.346
	(1.38)	(0.529)
observations	157	157

Sig. codes: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$
Mean values in each group. Standard errors within parentheses.
The median in all groups is 2.5.

Women CC_g: Friedman chi-squared = 1.9841, df = 3, p-value = 0.5757 Thus H0 of no differences between the group measures is not rejected.

Women $CCslope_g$: Friedman chi-squared = 1.4941, df = 3, p-value = 0.6836. Thus H0 of no differences between the group measures is not rejected.

Table D.9: Hypothesis 4 - Women: without those who could not see all images.

Group	CC_g	$CCslope_g$
FFM:LeaderF	2.45	0.366
	(1.25)	(0.538)
FFM:LeaderM	2.36	0.337
	(1.31)	(0.539)
MMF:LeaderM	2.33	0.378
	(1.30)	(0.517)
FFF:LeaderF	2.45	0.381
	(1.36)	(0.517)
observations	170	170

Sig. codes: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$
Mean values in each group. Standard errors within parentheses.
The median in all groups is 2.5.

Women CC_g: Friedman chi-square statistic = 4.2715, $df = 3$, p value = 0.2336. Thus H0 of no differences between the group measures is not rejected.

Women $CCslope_g$: Friedman chi-square statistic = 1.8519, $df = 3$, p value = 0.6037. Thus H0 of no differences between the group measures is not rejected.

Table D.10: Hypothesis 4 - Women: without completion times < 10 min.

Group	CC_g	$CCslope_g$
FFM:LeaderF	2.47	0.332
	(1.22)	(0.552)
FFM:LeaderM	2.41	0.321
	(1.28)	(0.551)
MMF:LeaderM	2.36	0.348
	(1.29)	(0.533)
FFF:LeaderF	2.45	0.343
	(1.36)	(0.528)
observations	149	149

Sig. codes: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$
Mean values in each group. Standard errors within parentheses.
The median in all groups is 2.5.

Women CC_g: Friedman chi-square statistic = 1.9356, $df = 3$, p value = 0.5859. Thus H0 of no differences between the group measures is not rejected.

Women $CCslope_g$: Friedman chi-square statistic = 0.52977, $df = 3$, p value = 0.9123 Thus H0 of no differences between the group measures is not rejected.

Table D.11: Hypothesis 4 - Women: without completion times < 9 min.

Group	CC_g	$CCslope_g$
FFM:LeaderF	2.44	0.349
	(1.22)	(0.548)
FFM:LeaderM	2.37	0.332
	(1.28)	(0.546)
MMF:LeaderM	2.34	0.365
	(1.29)	(0.527)
FFF:LeaderF	2.41	0.351
	(1.35)	(0.526)
observations	161	161

Sig. codes: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$
Mean values in each group. Standard errors within parentheses.
The median in all groups is 2.5.

Women CC_g: Friedman chi-square statistic = 1.3262, $df = 3$, p value = 0.7229 . Thus H0 of no differences between the group measures is not rejected.

Women $CCslope_g$: Friedman chi-square statistic = 0.47875, $df = 3$, p value = 0.9235. Thus H0 of no differences between the group measures is not rejected.

Table D.12: Hypothesis 4 - Men: without those who could not see all images.

group	CC_g	$CCslope_g$
FFM:LeaderF	2.52	0.421
	(1.33)	(0.597)
MMF:LeaderM	2.50	0.449
	(1.23)	(0.598)
MMF:LeaderF	2.41	0.411
	(1.33)	(0.594)
MMM:LeaderM	2.43	0.424
	(1.30)	(0.575)
observations	172	172

Sig. codes: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$
Mean values in each group. Standard deviation within parentheses.
(The median in all groups is 2.5.

Men CC_g: Friedman chi-square statistic = 8.6104, $df = 3$, p value = 0.03494. Thus H0 of no differences between the group measures is not rejected at the pre-specified 1% level.

Men $CCslope_g$: Friedman chi-square statistic = 1.3932, $df = 3$, p value = 0.7071. Thus H0 of no differences between the group measures is not rejected.

Table D.13: Hypothesis 4 - Men: without those who thought it was about gender.

group	CC_g	$CCslope_g$
FFM:LeaderF	2.50	0.379
	(1.34)	(0.607)
MMF:LeaderM	2.51	0.411
	(1.26)	(0.603)
MMF:LeaderF	2.40	0.379
	(1.34)	(0.600)
MMM:LeaderM	2.46	0.381
	(1.32)	(0.577)
observations	156	156

Sig. codes: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$
Mean values in each group. Standard deviation within parentheses.
(The median in all groups is 2.5.

Men CC_g: Friedman chi-square statistic = 6.7332, $df = 3$, p value = 0.08091. Thus H0 of no differences between the group measures is not rejected at the pre-specified 1% level.

Men $CCslope_g$: Friedman chi-square statistic = 1.7457, $df = 3$, p value = 0.6268. Thus H0 of no differences between the group measures is not rejected.

Table D.14: Hypothesis 4 - Men: without completion times < 10 min.

group	CC_g	$CCslope_g$
FFM:LeaderF	2.49	0.410
	(1.27)	(0.613)
MMF:LeaderM	2.46	0.429
	(1.21)	(0.609)
MMF:LeaderF	2.37	0.389
	(1.30)	(0.604)
MMM:LeaderM	2.39	0.407
	(1.26)	(0.585)
observations	155	155

Sig. codes: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$
Mean values in each group. Standard deviation within parentheses.
(The median in all groups is 2.5.

Men CC_g: Friedman chi-square statistic = 8.6187, $df = 3$, p value = 0.03481. Thus H0 of no differences between the group measures is not rejected at the pre-specified 1% level.

Men $CCslope_g$: Friedman chi-square statistic = 0.27771, $df = 3$, p value = 0.9642. Thus H0 of no differences between the group measures is not rejected.

Table D.15: Hypothesis 4 - Men: without completion times < 9 min.

group	CC_g	$CCslope_g$
FFM:LeaderF	2.51	0.417
	(1.29)	(0.608)
MMF:LeaderM	2.47	0.440
	(1.22)	(0.608)
MMF:LeaderF	2.39	0.399
	(1.32)	(0.601)
MMM:LeaderM	2.41	0.414
	(1.28)	(0.581)
observations	161	161

Sig. codes: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$
Mean values in each group. Standard deviation within parentheses.
(The median in all groups is 2.5.

Men CC_g: Friedman chi-square statistic = 8.0491, $df = 3$, p value = 0.04501. Thus H0 of no differences between the group measures is not rejected at the pre-specified 1% level.

Men $CCslope_g$: Friedman chi-square statistic = 0.93728, $df = 3$, p value = 0.8164. Thus H0 of no differences between the group measures is not rejected.

D.0.4 Hypothesis 5

Without those who thought the study was about gender:
Women – CC on LCL: Spearman's $\rho = 0.3766062$, p value < 0.0000.
Men – CC on LCL: Spearman's $\rho = 0.3934739$, p value < 0.0000.

Without those who could not see all images:
Women – CC on LCL: Spearman's $\rho = 0.3974239$, p value < 0.0000.
Men – CC on LCL: Spearman's $\rho = 0.4274284$, p value < 0.0000.

Without those with completion times < 10 min:
Women – CC on LCL: Spearman's $\rho = 0.3712543$, p value < 0.0000.
Men – CC on LCL: Spearman's $\rho = 0.4208496$, p value < 0.0000.

Without those with completion times < 9 min:
Women – CC on LCL: Spearman's $\rho = 0.3836426$, p value < 0.0000.
Men – CC on LCL: Spearman's $\rho = 0.4236219$, p value < 0.0000.

D.0.5 Social preferences for cooperation

Table D.16: Social preferences for cooperation – without those who could not see all images.

	Measurements:	
	$\rho(ALC - ACC)$	$\rho(ALC - MCC)$
Men	0.3247	0.4048
	(0.0000)	(0.0000)
Women	0.2535	0.3757
	(0.0009)	(0.0000)
Full sample	0.2882	0.3947
	(0.0000)	(0.0000)

Sig. codes: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

p-value of the Spearman rank sum correlation coefficient test within parentheses

Table D.17: Social preferences for cooperation – without those who thought the study was about gender.

	Measurements:	
	$\rho(ALC - ACC)$	$\rho(ALC - MCC)$
Men	0.2942	0.3819
	(0.0002)	(0.0000)
Women	0.2626	0.3976
	(0.0009)	(0.0000)
Full sample	0.2739	0.3950
	(0.0000)	(0.0000)

Sig. codes: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

p-value of the Spearman rank sum correlation coefficient test within parentheses.

Table D.18: Social preferences for cooperation – without those with completion times < 10 min.

	Measurements:	
	$\rho(ALC - ACC)$	$\rho(ALC - MCC)$
Men	0.3556	0.4084
	(0.0000)	(0.0000)
Women	0.2738	0.3265
	(0.0007)	(0.0000)
Full sample	0.3139	0.3728
	(0.0000)	(0.0000)

Sig. codes: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

p-value of the Spearman rank sum correlation coefficient test within parentheses.

Table D.19: Social preferences for cooperation – without those with completion times < 9 min.

	Measurements:	
	$\rho(ALC - ACC)$	$\rho(ALC - MCC)$
Men	0.3234	0.4088
	(0.0000)	(0.0000)
Women	0.2674	0.3427
	(0.0006)	(0.0000)
Full sample	0.2915	0.3788
	(0.0000)	(0.0000)

Sig. codes: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

p-value of the Spearman rank sum correlation coefficient test within parentheses.