CESifo WORKING PAPERS

10458
2023

May 2023

He, She, They? The Impact of Gendered Language on Economic Behavior

Paul M. Gorny, Petra Nieken, Karoline Ströhlein

CESifo

Impressum:

CESifo Working Papers
ISSN 2364-1428 (electronic version)
Publisher and distributor: Munich Society for the Promotion of Economic Research - CESifo GmbH
The international platform of Ludwigs-Maximilians University's Center for Economic Studies and the ifo Institute
Poschingerstr. 5, 81679 Munich, Germany
Telephone +49 (0)89 2180-2740, Telefax +49 (0)89 2180-17845, email office@cesifo.de
Editor: Clemens Fuest
https://www.cesifo.org/en/wp

An electronic version of the paper may be downloaded
· from the SSRN website: www.SSRN.com
· from the RePEc website: www.RePEc.org
· from the CESifo website: https://www.cesifo.org/en/wp

CESifo *Working Paper No. 10458*

He, She, They? The Impact of Gendered Language on Economic Behavior

Abstract

We conducted a controlled experiment to study how different gender frames used in the instructions affect economic behavior. In our experiment, we systematically varied the framing of the instructions, either using the male, the female, or a gender-inclusive form. Participants played three standard economic two-player games measuring prosocial behavior. In particular, we elicited the degree of sharing, reciprocal behavior, and honest reporting. We investigated if participants behaved differently if their self-reported gender matched the grammatical gender used in the instructions. The results reveal that the framing of instructions had the strongest impact on sharing, and the effects were mainly driven by participants self-identifying as men. In contrast, we observe only mild treatment differences, if any, regarding reciprocal behavior or honest reporting. We discuss the potential mechanisms and consequences of our findings.

JEL-Codes: C910, D010, J160, Z130.

Keywords: gender, gender inequality, gender stereotypes, grammatical gender, language, experimental methodology.

Paul M. Gorny
Institute of Management, Karlsruhe
Institute of Technology / Germany
paul.gorny@kit.edu

Petra Nieken
Institute of Management, Karlsruhe
Institute of Technology / Germany
petra.nieken@kit.edu

*Karoline Ströhlein**
Institute of Management, Karlsruhe
Institute of Technology / Germany
karoline.stroehlein@kit.edu

*corresponding author

May 23, 2023
We would like to thank Loukas Balafoutas, Simon Dato, Anna Dreber Almenberg, Eberhard Feess, Daniele Nosenzo, Arno Riedl, Eugenio Verrina, Ritchie Woodard, the audience of the ASFEE Conference 2021 and 2022, the SIOE annual meeting 2022, the SABE Pre-Conference Workshop 2022, the 2022 European ESA Meeting, LEOH 2022, and HeiKaMaX 2022 for helpful comments and discussion, and Behnud Mir Djawadi, Clara Hesse, Sergiu Panainte, Sabrina Schäfers, and Rebecca Zimmer for excellent research assistance. The authors gratefully acknowledge funding from "teamIn." The joint project "teamIn" is funded by the Federal Ministry of Education and Research (BMBF) and the European Social Fund (ESF) as part of the "Future of Work" program (funding reference number: 02L18A140) and supervised by the Project Management Agency Karlsruhe (PTKA). The responsibility for the content of this publication lies with the authors. Declarations of interest: none.

1 Introduction

Can language be sexist and contribute to gender differences in economic outcomes? The impact of language on gender differences, e.g., regarding leadership positions, civil rights, and economic outcomes, has garnered attention from researchers and the general public alike.[1] We need language to communicate and transmit complex information. However, natural language can also trigger gender stereotypes and influence behavior, potentially leading to discrimination and worse economic outcomes for the disadvantaged gender(s) (Mavisakalyan and Weber, 2018; Beblo et al., 2020). Thus, studying the potential impact of gendered language on economic behavior and gender differences is important (see, e.g., Niederle and Vesterlund, 2007; Borghans et al., 2009; Balafoutas and Sutter, 2012; Chen, 2013; Sutter and Glätzle-Rützler, 2015; Capraro, 2018; Wu, 2018; Chen and Houser, 2019; Card et al., 2021; Delfino, 2021).

In our paper, we address the following research questions by executing a controlled experiment: How do different written grammatical genders (male, gender-inclusive, female) affect economic behavior? Does the effect of grammatical gender depend on whether there is a match between the grammatical gender and the gender individuals associate with? Do men and women react differently to variations in grammatical gender?

Currently, we observe a heated and emotional public debate about the risks and benefits of using gender-inclusive language (see, e.g., United Nations, 2022a; Schuetze, 2020; Grullón Paz, 2021; Waters, 2021; Lankes, 2022). In the past, speakers in most languages typically used the *generic masculine* to address all genders (Sczesny et al., 2016). Recently, there has been a shift to using so-called "gender-inclusive" or "gender-fair language," to state preferred pronouns, and many organizations have established guidelines enforcing the usage of gender-inclusive language to address all individuals equally (see, e.g., United Nations, 2022b; Lu, 2015; European Parliament, 2018). The debate about the impact of gender-inclusive language on cognition and behavior has been ongoing in linguistics (Stahlberg et al., 2007). Previous research revealed that, indeed, linguistic structures can affect cognition and economic outcomes (Rubinstein, 2000; Wasserman and Weseley, 2009; Chen, 2013; Mavisakalyan and Weber, 2018). Given that most, if not even all, languages refer to gender in one form or the other, this is a general question, and experts are still discussing if and how language needs to be changed to address all individuals equally (see, e.g., Stahlberg et al.,

[1] See, e.g., Crawford and English (1984); Gabriel and Mellenberger (2004); Stahlberg et al. (2007); Gaucher et al. (2011); Vervecken and Hannover (2015); Horvath and Sczesny (2016); Sczesny et al. (2016); Hodel et al. (2017); Archer and Kam (2022) for research and May (2020); Schuetze (2020); Grullón Paz (2021); Lankes (2022) for the general public.

2007; Sczesny et al., 2016; Völkening, 2022).[2]

Even though in linguistics there is ample evidence that language can indeed influence perception and lead to biases, economic studies addressing the impact of (gendered) language on economic outcomes are still sparse (Rubinstein, 2000; Chen, 2013; Kricheli-Katz and Regev, 2021). To the best of our knowledge, our study is the first to systematically analyze the usage of male, gender-inclusive, or female forms in classical economic paradigms. The contribution of our study is twofold. First, we contribute to a better understanding of the impact of gendered language on economic behavior in general. Second, our findings have implications for scholars executing economic experiments. It is still an open question if and how the gender frame used in instructions can potentially affect experimental outcomes. On the one hand, using the generic masculine might be perceived as discriminating and outdated. On the other hand, deviations from the generic masculine could also trigger adverse reactions.

In our paper, we focus on a set of typical and highly relevant economic behaviors regarding prosociality: (1) sharing, (2) reciprocal behavior, and (3) honest reporting. In our study, we used standard economic games to measure these behaviors. In particular, we conducted a classical dictator game (Güth et al., 1982; Kahneman et al., 1986; Forsythe et al., 1994) to measure sharing, a sequential prisoner's dilemma (Bolle and Ockenfels, 1990; Dufwenberg and Kirchsteiger, 2000) to elicit reciprocal behavior, and a deception game (Gneezy et al., 2013) to measure honest reporting. To study if the gender frame of the instructions impacted behavior, we implemented three different types of instructions. We used the common generic male, a gender-inclusive, or the female formulation. Thus, we can study situations where the self-reported gender of a participant i) matched the gender frame of the instructions, ii) was "neutral" if the gender-inclusive frame had been used, or iii) did not match the gender frame of the instructions. The experiment was conducted in German. Given that German is a language with grammatical gender, the references to gender in the instructions were ubiquitous.

We base our hypotheses on social identity theory (Akerlof and Kranton, 2000) combined with previous findings from psycholinguistics (see Beblo et al., 2020, for a similar approach). Akerlof and Kranton (2000) postulated that social identities influence behavior through internalized prescrip-

[2] However, it is especially prevalent in languages with grammatical gender, such as Spanish, French, or German. In these languages, gender is coded as a grammatical category. Every noun has a gender that is either male, female, or (in some languages) neutral. Thus, articles, adjectives, and pronouns must fit with the gender of the respective noun they are referring to. At least 4.2 billion people live in countries where a language with grammatical gender is (one of) the official language(s). We used the languages identified as having grammatical gender in Haspelmath et al. (2005) and summed the inhabitants in countries where these are in the set of official languages. We researched population figures and official languages from the CIA World Factbook.

tions on how to behave. If an individual is associated with multiple identities, the situational context might determine the most salient one. Research found that gender stereotypes can have an influence on behavior and performance (Steele, 1997; D'Acunto et al., 2021). The grammatical gender used in texts might make the social categories "men" and "women" more or less salient depending on whether the male, a gender-inclusive, or the female form is used. Therefore, we hypothesize that a match between the grammatical gender and the self-chosen gender makes the social identity more salient. This might translate into different behavior across treatments. Indeed, previous results from psycholinguistics indicate that the usage of the generic male formulation can trigger gender stereotypes and mental images (Crawford and English, 1984; Gabriel and Mellenberger, 2004; Vervecken and Hannover, 2015; Sczesny et al., 2016). However, the findings are mixed. These gender stereotypes are triggered differently, depending on the formulation of instructions, where male instructions make male stereotypes more salient and female and inclusive instructions the female stereotypes. We systematically vary the formulation of instructions and analyze the influence on economic behavior. Since identity plays a role in this influence, we also analyze the influence of the different formulations of instructions depending on the self-stated gender of participants.

The data revealed the strongest effect of the gender-framed instructions in the dictator game measuring sharing, whereas the differences for reciprocity and honest reporting were less pronounced. In the dictator game, we observed a gender gap in amounts shared with the other participant when using the male frame, with almost 50% lower amounts shared by men than by women. This observation is in line with previous findings (Engel, 2011; Bilén et al., 2021). However, we neither economically nor statistically observed this gap if the participants were exposed to the female or gender-inclusive frame. The average amount shared by men was higher if they were exposed to the female or gender-inclusive frame instead of the generic male instructions. Thus, men reacted to the framing by changing their sharing behavior. The behavior of women was not prone to our treatment manipulation when looking at the intensive margin. However, a closer inspection of the extensive margin revealed that they had a significantly lower tendency to share strictly positive amounts if the gender-inclusive or the female frame had been used in the instructions. Regarding reciprocal behavior and honest reporting, we did not find changes in behavior due to our treatment manipulations. Nevertheless, when studying honest reporting we observed mild evidence that the reactions of men were qualitatively in line with our findings in the dictator game, meaning more honest reporting by men when female or gender-inclusive frames were used compared to the generic male frame.

Overall, our results indicate that changing the gender frame of instructions does not uniformly impact participants' behavior across various domains where they can behave prosocially. Nevertheless, given the rather mild treatment manipulations in contrast to changing e.g. the incentive structure, our results suggest that language is indeed a decisive factor in certain behavioral processes.

Our paper relates to three strands of literature: gender in economic experiments (i), psycholinguistics and economic studies on (gendered) language (ii), and studies related to the effects of framing (iii).

First, differences in behavior between men and women are well-documented in experimental economic studies, for instance, regarding altruism, competitiveness, reciprocity, or honesty (Niederle and Vesterlund, 2007; Bertrand, 2011; Engel, 2011; Heinz et al., 2012; Capraro, 2018).[3] One explanation is gender stereotypes leading to or reinforcing gender inequalities in wages, career paths, and gender discrimination (Niederle and Vesterlund, 2007; Croson and Gneezy, 2009; Dato and Nieken, 2014). But there are also studies showing that economic behavior (competitiveness, risk-taking, and altruism) is not influenced to a great extent by gender (Fornwagner et al., 2022). In his meta-analysis of experiments using the dictator game, Engel (2011) found that women give more as dictators and receive more as recipients. However, when controlling for recipient gender, dictator gender becomes insignificant. Bilén et al. (2021) found similar effects in terms of their direction but smaller in size and with lower statistical power. In their meta-analysis, Doñate-Buendía et al. (2022) considered a range of experimental conditions and locations. They found that, on average, women give more as dictators than men. They analyzed these gender differences in more detail by considering several experimental conditions and locations. Women are more generous than men for moderate and large social distance, while they are less generous than men when playing with close friends or family members. Women give more than men in South America, North America, and Oceania, while they give less than men in South Africa. Brañas-Garza et al. (2018) found that women give more as dictators, and both, men and women, expect women to give more as dictators than men. Gender differences have also been documented in reciprocal behavior (see, e.g., Ortmann and Tichy, 1999; Ellingsen et al., 2013) and honest reporting (see, e.g., Houser et al., 2012; Conrads et al., 2014; Muehlheusser et al., 2015; Grosch and Rau, 2017; Gerlach et al., 2019). The meta-analysis on the prisoner's dilemma by Mengel (2018) suggests that gender gaps if they occur, are specific to the study design and are thus not a stylized finding. Furthermore, Ortmann

[3] For most of the references cited in this paragraph, the language used in the experimental instructions and their type (grammatical gender language or not) is unknown to us, and also if and how the used formulations included gender. We, therefore, need to be careful when comparing our results to the literature.

and Tichy (1999) found gender differences in cooperation in a prisoner's dilemma-type game only in the first round but not in subsequent rounds. They found that women cooperated significantly more than men in the first round, but this difference disappeared in the last round. Dreber and Johannesson (2008) found that men are significantly more likely than women to lie, using the sender-receiver game introduced by Gneezy (2005). The study by Gylfason et al. (2013) could not replicate this finding using smaller stakes. Rosenbaum et al. (2014) found in their meta-analysis of honesty experiments that in the majority of studies that found gender differences in honesty, women were more likely to tell the truth than men. In his meta-analysis of honesty experiments, Capraro (2018) found that men were significantly more likely than women to tell black lies and altruistic white lies, and results were inconclusive in the case of Pareto white lies.

A second explanation for differences in behavior between men and women can be found in language. As we have argued earlier, information on gender is often embedded in language. Depending on the language family this can happen in different ways. A broad distinction is between *natural gender languages* and languages with *grammatical gender*. Natural gender languages, like English or Scandinavian languages, use gendered pronouns like "he" and "she," but verbs, adjectives, and articles do not carry a grammatical marking indicating gender. Such grammatical markings are present in languages with grammatical gender, like Spanish, French, or German (Stahlberg et al., 2007; Prewitt-Freilino et al., 2012).[4] Since gender is encoded in more words across sentences and in the grammatical structure, it is most salient in languages with grammatical gender. Psycholinguistics postulates that language affects cognition and perceptions (Hunt and Agnoli, 1991; Majid et al., 2004; Semin, 2013; Houston, 2019). One strand of this literature studies how gender in language–and gendered language in specific–influences, for example, the categorization of objects and attitudes toward men and women in recruitment processes and labor participation (Cubelli et al., 2011; Perszyk and Waxman, 2018; Lindqvist et al., 2019; Jakiela and Ozier, 2021). In particular, there is comprehensive evidence that the generic male formulation fosters a so-called *male bias*–a preferential behavior toward men–and sex-stereotyping, some of which can be mitigated by the use of gender-inclusive language (Crawford and English, 1984; Stahlberg and Sczesny, 2001; Gabriel and Mellenberger, 2004; Mavisakalyan, 2015; Vervecken and Hannover, 2015; Sczesny et al., 2016). However, the usage of neutral forms such as "person" lead to ambiguous effects with respect to associations and seems to be more context-dependent (Stahlberg and Sczesny, 2001). On a more aggregate level, the usage of gender in languages correlates with economic phenomena like the gender wage gap (van der Velde et al., 2015), differences in human capital formation (Galor et al.,

[4] There are also a few genderless languages, like Finnish and Turkish, in which even the pronouns are genderless.

2020), gender differences in educational attainment (Davis and Reynolds, 2018), female participation on corporate boards and senior management positions (Santacreu-Vasut et al., 2014), labor force participation (Gay et al., 2018; Mavisakalyan and Weber, 2018), and the division of labor (Hicks et al., 2015). Proponents of gender-inclusive language thus argue that these phenomena are, at least partially, due to the predominant use of the generic male. The evidence discussed so far has inspired research in experimental economics on the effects of language on behavior, such as intertemporal choices (Sutter et al., 2015). Even more recently, experimental economists became interested in the nexus of gender and language. Closest to our paper are Balafoutas et al. (2023) and Gorny et al. (2023). Balafoutas et al. (2023) investigated the effect of gender-inclusive language on competitive and leadership behaviors and feelings of inclusion and belongingness to their group in the experiment. The study involved three treatments, including a masculine baseline condition, a condition with feminine and masculine (pro)nouns, and a condition with non-gendered (pro)nouns. The results showed that participants who self-reported to be men behaved similarly to those who self-reported to be women across all treatments, regardless of whether or not gender-inclusive language was used. Additionally, there was no difference in participants' feelings of inclusion or entitlement to compete or become a group leader in any of the treatments. Gorny et al. (2023) studied if the gendered framing of norms impacts if people comply with the norm. They do not find strong evidence that a match between the participant's self-reported gender and the norm formulation led to a higher increase in norm compliance compared to the differences in a mismatch or gender-inclusive frame. Yet, among men in a dictator game, a match led to a higher increase in norm compliance compared to the increase if gender-inclusive formulations were used.

Thirdly, our study also relates to the literature on framing effects in economic experiments. Varying the generic use of gender in the language of the experiment can be seen as a way to frame the instructions. Early on, Tversky and Kahneman (1981, 1989) and Kahneman et al. (1986) argued that framing, as an alternative way to describe a decision problem, influences the perception of that decision problem and hereby the preferences of people (see, e.g., Levin et al. (1998); Kahneman and Tversky (2013); Fiedler and Hillenbrand (2020)). However, the findings are mixed (Abbink and Hennig-Schmidt, 2006; Huber and Kirchler, 2012). Regarding the dictator game, framing has been shown to shift sharing considerably (Hoffman et al., 1994; Brañas-Garza, 2007; Capraro and Vanzo, 2019; Chang et al., 2019). Similarly, framing the prisoner's dilemma as a cooperative rather than a competitive game or referring to it as the "Community Game" as opposed to the "Wall Street Game" can substantially increase the cooperation rate (Deutsch, 1960; Liberman et al., 2004). Huber and Huber (2020) studied the effect of framing for truthful reporting by varying the description of the situational context as either abstract, neutral, or finance-related. While there were no differences for a student sample, they found that financial professionals acted more honestly in the financial and neutral context than in an abstract situation. Balafoutas et al. (2018) let subjects write about a situation in which someone else had control over them or they had control over someone else. Thus, subjects either received a low-power or a high-power prime. The authors studied the impact of this priming on the gender gap in competitiveness. Without priming and in the low-power prime men were more likely than women to choose competition; this gap vanished in the high-power prime. Boggio et al. (2020) conducted a field experiment to study the influence of gender-specific conceptual frames on performance in a financial task. They recruited participants from elementary school children and varied the framing of the task, either using a masculine frame (emphasis on competitiveness and physical abilities), a feminine frame (emphasis on cooperation and empathy), or a neutral frame (no special emphasis). They found that the exposure of girls to the feminine frame increases the probability of providing consistent answers in the financial task when combined with a workshop on the utility of saving.

The paper is organized as follows. In Section 2, we describe the experimental design and procedures, derive our hypotheses and explain our data preparation and estimation strategy. Section 4 contains our results. In Section 5, we discuss our results in light of a series of behavioral mechanisms that may drive them. Section 6 concludes.

2 Experimental Design and Hypotheses

2.1 General Description

To investigate the impact of gender frames on economic behavior, we conducted an online experiment implementing a 3×2 design.[5] First, we systematically varied the framing of the instructions using either the (generic) male, a gender-inclusive, or the female form. Second, we exogenously varied the share of participants referring to themselves as women by recruiting equal shares of men and women based on the data available in the recruiting system.[6]

In the following, we describe the general setup before providing details on the treatments and the procedures. Note that the treatments differed in the grammatical gender of the instructions and the

[5] We preregistered our study prior to data collection at aspredicted.org.

[6] We base our analysis on the self-reported gender, which we elicited at the end of the experiment, to avoid potential confounds. The share of participants referring to themselves as women was 50.56 percent

self-reported participant's gender, but the games and economic incentives were identical across all treatments. The original German and translated instructions are provided in the online appendix. The participants played three different standard economic one-shot games in groups of two in all treatments. Thus, in each game, there were two roles: player A and player B. We implemented the strategy method (Selten, 1965) if the respective game involved more than one active player. This allowed us to collect data from all participants in all roles of each game. The participants received no information on the game outcomes, the other participants' actions, or anyone's self-reported gender during the experiment. We implemented a perfect-stranger matching protocol to avoid moral balancing (Ploner and Regner, 2013) or perceived reciprocal behavior across games. To mute potential income effects, we randomly selected one game at the end of the experiment that determined the payment. In addition, the role of each participant as either player A or B was randomly chosen. In the general instructions, we informed participants about the experimental currency unit (ECU) and the exchange rate of 1 ECU = €0.40.

The participants first played the classical dictator game (DG) (Güth et al., 1982; Kahneman et al., 1986; Forsythe et al., 1994) in all treatments.[7] All participants were in the role of player A, the decision maker, and had to allocate 20 ECU between themselves and another participant. Given that player B is passive in this game, the participants did not have to make any decisions in the role of player B.

Second, all participants played a sequential prisoner's dilemma (PD) (Bolle and Ockenfels, 1990; Dufwenberg and Kirchsteiger, 2000).

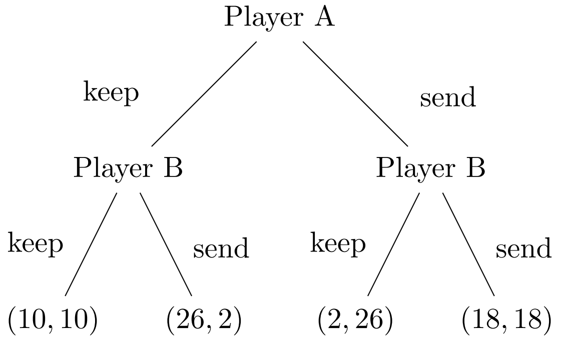


Figure 1: The sequential prisoner's dilemma.

Again, there were two player roles: A and B. All participants first were in the role of player A and second in the role of player B (see Figure 1 for an overview). Both players had an endowment of ten ECU. Player A could either send eight ECU from their endowment to Player B or keep the whole

[7] As we are interested in between-subject differences, we kept the order constant for all participants.

endowment. If they chose "send," the eight ECU were doubled, and the resulting 16 ECU were allocated to player B. As player A, player B could either keep the endowment or send 8 ECU to player A. If player B chose to send 8 ECU, this amount was doubled and allocated to player A. We used the strategy method (Selten, 1965) for player B to elicit a complete response function. Thus, player B had to make a decision for both possible decisions of player A. First, player B decided if they wanted to send the eight ECU if player A had sent their eight ECU. Second, they stated if they wanted to send the eight ECU to player A if player A had chosen to keep their endowment. If this game was selected for payment, player B's payment depended on their decision regarding the actual action of player A. If player A had chosen to send the eight ECU, player B's decision for this action determined the payment. If player A decided to keep the endowment, player B's decision for this action determined the payment.

The third game was the deception game (Dec)[8] introduced by (Gneezy et al., 2013). As before, there are two player roles: player A and player B, who form one group. A number between one and six was randomly assigned to each group. Both players knew that player B would not receive the information about the assigned number before making their decision. However, player A would send a message about the number to player B. Thus, first, all participants had the role of player A. They had to choose a pre-written message for each possible number. This message read "The assigned number is ...." and did not have to contain the true number. Player A's payment was 10 ECU plus twice the number sent, e.g., 12 ECU, in case player A sent the message that the assigned number was one, 14 ECU, in case player A sent the message that the assigned number was two, etc. irrespective of the action of player B or the true number. Again, we used the strategy method for player B. For every possible message from player A, player B decided whether to follow the message or not. The payment of player B was 10 ECU in case they followed the message of player A, and the message contained the true number and otherwise zero ECU. If player B did not follow player A's message, player B received three ECU.

After the three games were played, we elicited incentivized beliefs about fair sharing, unconditional cooperation, and honest reporting as well as norms for each game. See the online appendix for a more detailed description of the norm elicitation.

Next, all participants had to answer a brief survey containing questions on reciprocity (Dohmen et al., 2009), risk aversion (Dohmen et al., 2011; Kantar Public, 2020), moral values (Haerpfer et al., 2020), and questions regarding the comprehension of the instructions, and their attitude toward

[8] We use *Dec* here instead of an acronym to avoid confusion with the dictator game for which we use DG.

language.

We also collected demographic information, including the participants' age, study degree, field of study, and past participation in experiments. Importantly, we asked participants to report their gender. Precisely, we asked which gender they would "assign themselves to."[9] We asked for the participants' recall of the used gender frame throughout the instructions. Lastly, we included an optional text field in which we asked if participants had any comments on the experiment.

2.2 Treatments

The treatments differed regarding the grammatical gender used in the instructions and the participants' self-reported gender. Throughout the instructions and across treatments, we described the rules of the experiment referring to "a participant." This generic participant was described in either the (generic) male (Teilnehmer), the gender-inclusive (Teilnehmer*in), or the female frame (Teilnehmerin). There are two approaches to making language more inclusive: explicitly including women and gender-inclusive language. The first approach is operationalized with the help of male-female word pairs, or using the capital "I" in German. The second approach relies on gender neutral forms, the gender star (*), the tilde (∼), the underscore (_), or the colon (:) in German. The most prominent symbol is the gender star which makes other genders more salient, while using neutral forms still lets most people only think of men (Lindqvist et al., 2019; Völkening, 2022). This is why many of the people identifying as non-binary prefer this approach and equal opportunity officers as well as public authorities implemented its usage (Bendel, 2021; Antidiskriminierungsstelle des Bundes, 2023). Therefore, we chose to use this approach in our experiment. Our setup leads to six treatments labeled W-Match, W-Inclusive, W-Mismatch, M-Match, M-Inclusive, and M-Mismatch. The first letter refers to the participant's self-reported gender. The letter "W" indicates that the participant identifies as a "woman." In contrast, the letter "M" indicates that the participant identifies as a "man." The second part indicates if the grammatical gender used in the instructions matched the self-reported gender of the participant (Match) or not (Mismatch). If the instructions used the gender inclusive language, we label the treatments "Inclusive". Thus, the W-Match treatment encompasses all observations of women that were exposed to the female frame in the instructions. In contrast, the M-Match treatment refers to all men that participated in the treatment using the (generic) male frame. A treatment overview is provided in Table 1.

[9] The exact question we asked was "Which gender do you sort yourself into?" (German: "Welchem Geschlecht ordnen Sie sich zu?") with the options "Male" (German: "Männlich"), "Female" (German: "Weiblich"), and "Diverse" (German: "Divers"), which is equivalent to the non-binary option in English surveys.

		Congruence of gender frame and self-reported gender		
		Match	Inclusive	Mismatch
Self-reported gender	Women	W-Match	W-Inclusive	W-Mismatch
	Men	M-Match	M-Inclusive	M-Mismatch

Table 1: Treatments.

2.3 Procedures

The experiment was conducted online with a German-speaking participant pool from a large university in Germany.[10] We used ORSEE (Greiner, 2015) to recruit German-speaking student participants. To assess correct registration for the respective session and to allow participants to ask clarifying questions, the experiment was accompanied by a video call. Participants and experimenters were muted, their video feeds were disabled, and the lab rules were shown as screen-share throughout the session.[11] Participants received personalized links to the experimental software programmed in oTree (Chen et al., 2016). Participants read general instructions, played the three games, each preceded by game-specific instructions, stated their norms and beliefs, and answered a brief survey on demographics and attitudes. After the general instructions and before the battery of games, participants had to pass a short survey on the general understanding of the experiment. Before each game, we also conducted control questions on understanding the game rules. We did not provide any feedback during the experiment. Participants only learned about the realization of their choices of the randomly chosen game and role, which was relevant for the payoff, at the end of the experiment. The payoff consisted of the payment for one randomly chosen game and role, the payments for the belief and norm elicitation, and a show-up fee. Sessions lasted approximately 50 minutes, and participants received information on their accumulated earnings, on average €9.64, including a show-up fee of €2.50. We implemented an exchange rate of 1 ECU = €0.40.

2.4 Hypotheses

As stated in Section 1, we base our hypotheses on social identity theory (Akerlof and Kranton, 2000). Recall that our treatment variation lies in the frame of experimental instructions, either being formulated in a male, gender-inclusive, or female frame. This frame might make the social categories "men" and "women" more or less salient depending on which frame is used. There are

[10] Note that the data collected and used in this experiment is also used as a baseline for Gorny et al. (2023).

[11] Communication was limited to the text chat. Verbal communication was not used unless urgently necessary, e.g., if a participant went idle for longer than five minutes.

commonly stated typical gender roles and stereotypes in the literature. These include that men are more competitive, aggressive, and good at math tasks than women, and women are more caring, pro-social, and good at creative tasks (Cejka and Eagly, 1999; Rudman and Glick, 2001; Arias et al., 2023). These stereotypes are triggered by the frame of instructions, depending on which frame is used. Therefore, we hypothesize that the female frame of instructions results in more pro-social behavior than the other two frames because it triggers the female identity. The male frame results in the least pro-social behavior and the pro-social behavior in the gender-inclusive frame lies between these two extremes.

Hypothesis (Pro-social behavior)**.** Participants' prosocial behavior is highest in the female, followed by the gender-inclusive, and the lowest in the male frame of experimental instructions.

Recall that in the W-Match and M-Match treatments, participants' gender matched with the frame of instructions, in the W-Mismatch and M-Mismatch treatments, participants' gender did not match with the frame of instructions, and in the W-Inclusive M-Inclusive treatments, participants' gender neither matched nor mismatched the frame of instructions. We hypothesize that a match between the grammatical gender and the self-chosen gender makes the social identity more salient. This means, that when the self-chosen gender matches the frames of instructions, the impact of the frame is stronger.

Hypothesis (The gender match triggers identity more than Inclusive or Mismatch)**.** The female frame triggers higher prosocial behavior more strongly for women than for men, while the male frame triggers lower prosocial behavior more strongly for men than for women. We expect no difference between men and women for the inclusive frame.

3 Data Preparation and Estimation Strategy

This section describes our variables of interest and their use in our empirical strategy to test our hypotheses. We also provide information on the sample and the restrictions we applied based on our pre-registration.

3.1 Variables of Interest

Our main variables of interest concern the participants' behavior in each game. That is, the amount sent in the dictator game, keeping or sending in the prisoner's dilemma, and the number reported for each die roll in the deception game. We are interested in whether these differ when instructions are written in the male, gender-inclusive, or the female form. In particular, we are interested in whether participants behave differently when their self-reported gender matches the grammatical gender used in the instructions. Given that we analyze each of the three games separately, we describe the variables used in each game below. In the dictator game, we are interested in the amount sent by player A. Thus, we first analyze the effects at the intensive margin using the *Amount sent* measured in ECU. In a second step, we take into account that we observe a mass point at zero and use an indicator variable *Sent any* which is one if a participant has sent any positive amount and zero otherwise. In the prisoner's dilemma, we focus on analyzing reciprocity and thus concentrate on player B's conditional decisions. Following the literature (see, e.g., Miettinen et al., 2020), we classify participants as "selfish" if they keep their endowment irrespective of player A's decision. We classify participants as "conditional cooperators" if they sent 8 ECU in case player A also chose to send and kept their endowment if player A did the same. Participants who always send 8 ECU are classified as "altruistic." We classify participants as "antireciprocal" if they send 8 ECU in case player A chose to keep their endowment and keep their endowment if player A chose sent. In our analysis, we focus on the comparison between conditional cooperators and selfish types because of the very low shares of altruistic and antireciprocal types. Thus, we define an indicator variable *Reciprocal* that is one if a participant has been classified as a conditional cooperator and zero if the participant was classified as selfish. In the deception game, we are interested in the share of honest reports. Each player A had to select six messages and thus six opportunities to lie or to be honest. The variable *Share honest* refers to the share of honest reports ranging from zero (all lies) to one (all honest). Furthermore, we analyze an indicator variable $Honest_X$ $X \in \{1, 2, 3, 4, 5, 6\}$ that is one if the report was honest for the respective die result and zero otherwise.

We define four indicator variables to measure the impact of the self-reported gender and the gender frame of the instructions on economic behavior. The variable *Woman* is one if the participant self-reported to be a woman and zero if the participant self-reported to be a man. From hereon, we refer to a participant for whom *Woman* is equal to one as a woman and to a participant for whom *Woman* is equal to zero as a man. The variable *Match* is one if a participant's self-reported gender and the one used in the instructions were identical. This is the case for women in the W-Match treatment and men in the M-Match treatment. The variable *Inclusive* is one if the gender-inclusive form was used in the instructions and zero otherwise. This is the case for both men and women in the gender-inclusive treatments (W-Inclusive and M-Inclusive). The variable *Generic male*: This variable is one if the generic male form was used in the instructions and zero otherwise. This is the case for women in the W-Mismatch treatment and men in the M-Match treatment.

Next, we describe all additional control variables used in the analysis in detail. *Age* measures the participants' age in years. We asked participants for the current *Semester* they are in, including bachelor semesters, if the participant was in their master's. We asked participants for the subjects in which they major. We grouped those in majors related to *Business and Economics*, *Education*, and *Other majors*, with the latter category serving as a baseline unless otherwise mentioned. We asked a battery of 5 questions on participants' attitudes toward language change over time using a 7-point Likert scale. *Language attitude* is the mean reply with a high score indicating a more liberal position toward language change than a low score. At the very end of the experiment, we asked participants for the grammatical gender used throughout the experiment and if they had any comments. The variable *Remembered formulations* is one if a participant remembered the grammatical gender used correctly and zero otherwise. We coded free-text comments into three categories: *Language comments* is one whenever a free-text comment referred to the instructions and zero otherwise. *Other comments* is one whenever a comment was made that did not fall into the previous category and zero otherwise. *No comment* is one whenever the other two dummies are zero, and serves as a baseline in the regressions. Thus, the three dummies are mutually exclusive. We also asked participants to rate the clarity of the instructions on a 7-point Likert scale. We refer to the resulting variable as *Instructions clear.* After the general instructions and before the battery of games, participants had to pass a short survey on the general understanding of the experiment. Before each game, we also conducted control questions on understanding the game rules. *Failed attempts$_G$* is the number of failed attempts to answer the control questions asked before the respective game $G \in \{DG, PD, Dec\}$. *Failed attempts$_{all}$* is the sum of failed attempts across all questions asked in the experiment, including those for the questions of general understanding. Our risk measure *Risk aversion* is measured on an 11-point scale according to Dohmen et al. (2011) and Kantar Public (2020). Our measure for reciprocity is measured on a 7-point scale according to Dohmen et al. (2009) to measure *Positive reciprocity* and *Negative reciprocity.* We only include reciprocity in the regressions of the prisoner's dilemma because players can only reciprocate in this game. To elicit the variables *First-order belief* and *Second-order belief*, we first provided a brief summary of each game. Subsequently, we elicited beliefs relative to actions commonly viewed as moral in the respective game. Specifically, we phrased our belief elicitation around fair 50-50 sharing in the dictator game (giving 10 ECU from the 20 ECU endowment), unconditional cooperation in the prisoner's dilemma, and complete honesty (i.e., a true report for each possible outcome of the die roll) in the deception game. For first-order beliefs, we asked participants about their belief on the share of participants taking the respective action. In a second step, we asked for their belief about the average stated first-order belief among the other participants in their session. Every participant whose stated belief was strictly within ten percentage points off the true value received 2 ECU. If they were off by at least ten percentage points but less than twenty percentage points, they would receive 1 ECU. For the first and second-order beliefs, participants could thus earn between 0 and 4 ECU.

3.2 Sample Selection

Next, we describe our sample selection procedure and how the sample is balanced regarding demographic information.

In total, we gathered data from 109 participants. We conducted an attention check in our post-experimental survey, in which five participants failed. A single participant self-reported to be non-binary. In line with our preregistration, we excluded these observations from the data set. This leaves us with 103 observations in our analytical sample.

As can be seen in Table 9 in Appendix B, demographics are balanced across the different gender frames in the instructions in our analytical sample. Most importantly, the proportion of women is close to 50% between the differently framed instructions.[12]

3.3 Empirical Strategy

In the following, we describe our general empirical strategy. We analyzed the participants' behavior for each of the three games separately. We first reported descriptive statistics for the men's behavior before reporting the women's behavior across treatments. Then, we applied a conservative non-parametric approach and compared the results across treatments using a Jonckheere-Terpstra test for men and women separately. Furthermore, we want to test for differences between men and women across treatments and analyze the possible gender gap. However, to investigate the impact of a match and the potential interaction with the self-reported gender, we needed to apply an econometric approach. Thus, we estimated a series of linear regressions (OLS regressions with robust standard errors) for each game. In case the dependent variable was binary, we estimated a series of Probit regressions. The dependent variable varies for each game as described above, but in each game, we add the same independent variables and controls. First, we introduced the variables

[12] The shares of women are 53.13% for the male, 54.29% for the gender-inclusive, and 47.22% for the female frame. The differences are not statistically significant ($p = 0.818$, Kruskal-Wallis test).

Woman, *Match*, and *Inclusive* to study the impact of a match as well as the self-reported gender on the participants' behavior in the games. Then we added interaction terms for *Woman* and *Match* as well as *Woman* and *Inclusive* to be able to disentangle all treatment effects in a fully saturated specification. In the next step, we included demographics. Then we included controls for language and understanding. In a last step, we added various controls for attitudes and beliefs to show the robustness of our findings. We report the coefficients for the main effects and interactions in the paper. Full tables including the coefficients for all controls are reported in Appendix B.

As we include the variable *Woman*, our statistical baseline is the M-Mismatch treatment. The coefficient for *Woman* corresponds to the W-Mismatch treatment whereas the coefficient for *Match* corresponds to the M-Match treatment compared to the baseline. The treatment effect of the M-Inclusive treatment compared to the baseline is given by the coefficient of *Inclusive*. The sum of the interaction *Woman×Match, Match*, and *Woman* is equivalent to a dummy for the W-Match treatment. The effect of the W-Inclusive treatment can be calculated by summing up *Woman, Inclusive*, and the interaction *Woman×Inclusive*.

In a second step, we pooled the data from the gender-inclusive treatments (W-Inclusive and M-Inclusive) and the female treatments (W-Match and M-Mismatch) to study if a deviation from the generic male triggers behavioral differences. Again, we added the variable *Woman* but instead of using *Inclusive* and *Match*, we inserted *Generic male* as well as the interaction with *Woman* in our specifications. The additional control variables remained the same as reported above.

In case our dependent variable is binary, we applied Probit regressions. When interpreting the interaction terms Woman×Match, Woman×Inclusive, and Woman×Generic Male, we need to be careful interpreting their coefficients as effects (Ai and Norton, 2003). Thus, in the main part of the analysis, we discuss changes in the linear index of the nonlinear models under the respective specification. We also add subscript stars ($_\star$) to indicate the different levels of statistical significance of the *interaction effect* as opposed to the levels of statistical significance of the *interaction term*, which are indicated by superscript asterisks (*).[13]

4 Results

In the following, we analyze the participants' behavior for each of the three games separately.

[13] We thank Arno Riedl for pointing this out. See Appendix A for details on how we calculated the test statistics for the interaction effects based on Ai and Norton (2003).

4.1 Dictator Game

Recall our hypothesis on prosocial behavior. For the dictator game, this translates to the following.

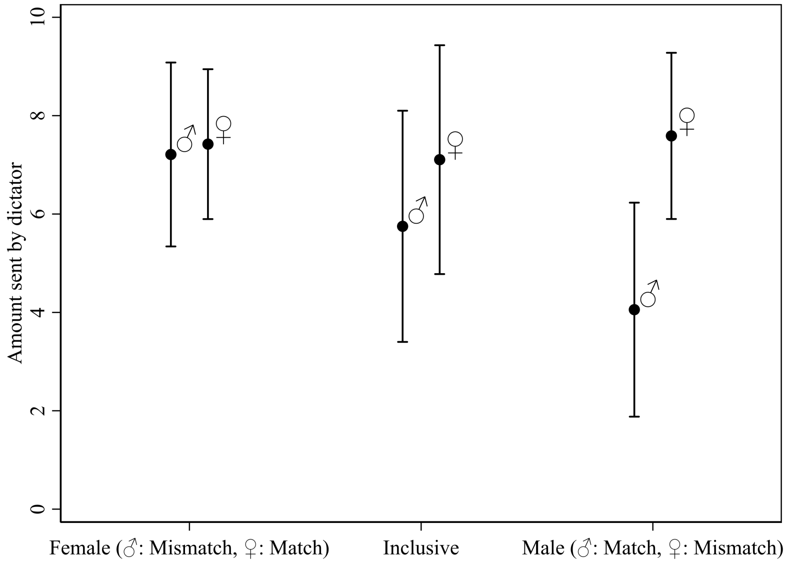
Hypothesis 1 (Sharing in the dictator game)**.** Participants share the highest amount as dictators in the female, followed by the gender-inclusive, and the lowest in the male frame of experimental instructions.

Further, recall our hypothesis on Match as a trigger of identity. For the dictator game, this translates to the following.

Hypothesis 2 (The gender match in the dictator game)**.** In the female frame, women's amounts shared as dictators are higher than men's amounts shared as dictators. In the male frame, men's amounts shared as dictators are lower than women's amounts shared as dictators. We expect no difference between men and women for the inclusive frame.

In the role of the dictator, men sent an average of 7.211 ECU in the M-Mismatch treatment, 5.750 ECU in the M-Inclusive treatment, and 4.133 ECU in the M-Match treatment. The amount shared was highest in the M-Mismatch and lowest in the M-Match treatment, indicating that a mismatch between the self-reported gender and the gender frame of the instructions increased the sharing by men. However, the differences are not statistically significant (Jonckheere-Terpstra test, $p = 0.104$). Women, on average, sent 7.588 ECU in the W-Mismatch treatment, 7.105 ECU in the W-Inclusive treatment, and 7.235 ECU in the W-Match treatment. Thus, we observe no statistically significant differences across the treatments for women (Jonckheere-Terpstra test, $p = 0.615$). See Figure 2 for a graphical illustration for men (♂) and women (♀).

To evaluate if our results are in line with previous findings, we compared the amounts sent in the W-Mismatch and M-Match treatments where the instructions used the generic male frame. We observed that participants in the W-Mismatch treatment, on average, sent 3.455 ECU more than participants in the M-Match treatment (Mann-Whitney U test, $p = 0.029$). This gender gap in dictator games with women sharing more than men is well-documented in the literature (Engel, 2011; Bilén et al., 2021). This gap was reduced comparing the W-Inclusive and M-Inclusive treatments (1.355 ECU; Mann-Whitney U test, $p = 0.482$) and was almost zero when considering the W-Match and M-Mismatch treatments (0.025 ECU; Mann-Whitney U test, $p = 0.787$). Given that women did not vary the average amount shared across treatments, this reduction was driven by men who increased the average amounts shared in the M-Mismatch and M-Inclusive treatments



Note: Markers indicate means and whiskers indicate 95% confidence intervals.

Figure 2: Amount sent in dictator game by men (♂) and women (♀).

relative to the M-Match treatment. To investigate if the first impressions were robust, we executed a series of linear regressions reported in Table 2.

When reporting numerical differences, we focused our analysis on the fully saturated regression specification including all controls presented in column 5. The coefficient for *Match* is negative and marginally statistically significant. Thus, using the female frame instead of the male frame increased the sharing by men on average by 2.634 ECU, all other things equal. The coefficient for *Inclusive* reports the difference between the amount sent by participants in the M-Inclusive and the M-Mismatch treatment which is not statistically significant. Using the inclusive frame instead of the male frame increased the amount sent by men by $-1.535 - (-2.634) = 1.099$ ECU, but that difference is not statistically significant (F-test, $p = 0.504$). Thus, our results support the observation that men sent less if the gender frame of the instructions did not match their self-reported gender.

The greater amounts sent in the W-Mismatch treatment compared to the W-Inclusive treatment $(-(-1.535) - (0.821) = 0.714, F - test, p = 0.607)$, and in the W-Mismatch treatment compared to the W-Match treatment $(-(-2.634) - (2.378) = 0.256, F - test, p = 0.819)$, are not statistically significant. Also, the lower amount sent in the W-Inclusive treatment compared to the W-Match treatment $(-(-2.634) - (2.378) + (-1.535) + 0.821 = -0.458, F - test, p = 0.773)$ is not statistically significant. Thus, the regression specifications support the first impression that the gender frame

Dep. Var.: Amount sent	(1)	(2)	(3)	(4)	(5)
Woman	1.556*	0.378	0.064	-0.940	-0.288
	(0.862)	(1.288)	(1.384)	(1.644)	(1.567)
Match	-1.699*	-3.077*	-3.082*	-3.514**	-2.634*
	(0.986)	(1.551)	(1.667)	(1.688)	(1.579)
Inclusive	-1.013	-1.461	-1.254	-1.831	-1.535
	(1.057)	(1.532)	(1.573)	(1.657)	(1.630)
Woman × Match		2.724	2.831	3.302	2.378
		(1.970)	(2.131)	(2.207)	(1.976)
Woman × Inclusive		0.978	1.307	1.201	0.821
		(2.123)	(2.130)	(2.111)	(2.027)
Constant	6.654***	7.211***	6.803***	3.531	3.504
	(0.794)	(0.957)	(2.330)	(3.238)	(2.996)
Demographics	✗	✗	✓	✓	✓
Language & Understanding	✗	✗	✗	✓	✓
Attitudes & Beliefs	✗	✗	✗	✗	✓
R^2	0.053	0.069	0.116	0.187	0.290
Observations	103	103	103	103	103

Robust standard errors in parentheses, * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$
Note: For the complete table with all coefficients, see Table 10 in Appendix B.

Table 2: OLS regressions with the *Amount sent* in the dictator game as the dependent variable.

of the instructions did not impact the behavior of women.

To further investigate the impact of the generic male frame on gender differences in sharing, we need to compare the behavior in the W-Mismatch and the M-Match treatments (both using the male gender frame). Participants in the W-Mismatch treatment were more generous than participants in the M-Match treatment. They on average sent $-0.288 - (-2.634) = 2.346$ ECU more than participants in the M-Match treatment, which is marginally statistically significant (F-test, $p = 0.099$). Next, we compared the M-Inclusive and the W-Inclusive treatments. Here, we still observed that participants in the W-Inclusive treatment sent $-0.288 - 1.535 + 0.821 - (-1.535) = 1.273$ ECU more than participants in the M-Inclusive treatment, but this difference is not statistically significantly different from zero (F-test, $p = 0.800$). When analyzing differences between men and women exposed to the female frame, the data reveals that participants in the W-Match treatment sent $-0.288 - 2.634 + 2.378 = 0.544$ ECU less than participants in the M-Mismatch treatment, but this difference is not statistically significant (F-test, $p = 0.702$). Thus, we observed a difference between the M-Match and W-Mismatch treatment of roughly 3 ECU in the amounts sent from an initial 20 ECU, which is over 15% of the total budget. This difference was reduced and became statistically insignificant when comparing the W-Inclusive with the M-Inclusive treatment and was close to zero when comparing the W-Match and M-Mismatch treatments.

Dep. Var.: Amount sent	(1)	(2)	(3)	(4)	(5)
Woman	1.501*	0.624	0.545	-0.431	0.052
	(0.868)	(1.058)	(1.076)	(1.300)	(1.410)
Generic male	-0.927	-2.410*	-2.493*	-2.636*	-1.875
	(0.914)	(1.426)	(1.488)	(1.483)	(1.362)
Woman × Generic male		2.831	2.586	3.022*	2.321
		(1.819)	(1.856)	(1.794)	(1.704)
Constant	6.098***	6.543***	6.244***	2.851	2.832
	(0.691)	(0.757)	(2.198)	(2.999)	(2.711)
Demographics	✗	✗	✓	✓	✓
Language & Understanding	✗	✗	✗	✓	✓
Attitudes & Beliefs	✗	✗	✗	✗	✓
R^2	0.038	0.060	0.109	0.173	0.281
Observations	103	103	103	103	103

Robust standard errors in parentheses, * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$
Note: For the complete table with all coefficients, see Table 11 in Appendix B.

Table 3: OLS regressions with the *Amount sent* in the dictator game as the dependent variable when comparing the generic male with the "non-standard" female and gender-inclusive framed instructions.

Table 3 depicts the results of OLS regressions comparing the impact of the generic male frame against the "non-standard" female and gender-inclusive frame. This allows us to further investigate if a deviation from the generic male frame led to differences in behavior. The baseline in these regressions is the pooled data from the M-Mismatch and the M-Inclusive treatments. The negative coefficient of *Generic male*, albeit not or being only marginally statistically significant, also hints into the direction that men shared higher amounts if the instructions did not match with their self-reported gender. So far, it seems that the reduction of the gender gap was mainly driven by men reacting to the treatment manipulations.

However, considering the distribution of amounts sent, we observed a mass point at 0 ECU.[14] The share of participants sending 0 ECU was 21.05% in the M-Mismatch treatment, 37.50% in the M-Inclusive treatment, and 46.67% in the M-Match treatment. The differences are not statistically significant (Jonckheere-Terpstra test, $p = 0.115$). The share of participants sending 0 ECU was 0.00% in the W-Mismatch treatment, 15.79% in the W-Inclusive treatment, and 11.76% in the W-Match treatment. The differences are not statistically significant (Jonckheere-Terpstra test, $p = 0.245$). To analyze if gender frames triggered a reaction between purely selfish behavior (not sending anything) and sharing some positive amount, we again conducted a series of regressions. Table 4 contains the estimates from Probit regressions on *Sent any*–a dummy that is one whenever

[14] Note that there is also a mass point at 10 ECU, which was the point where the endowment was split into equal shares between the dictator and the recipient.

a participant sent any positive amount (1 to 20 ECU) and zero if they sent nothing (0 ECU).

Dep. Var.: Sent any	(1)	(2)	(3)	(4)	(5)
Woman	1.011***	4.614***	4.679***	4.505***	5.281***
	(0.303)	(0.329)	(0.374)	(0.509)	(0.759)
Match	-0.800**	-0.721	-0.807	-0.763	-0.690
	(0.374)	(0.460)	(0.513)	(0.568)	(0.598)
Inclusive	-0.747**	-0.486	-0.468	-0.698	-0.848
	(0.371)	(0.457)	(0.488)	(0.573)	(0.615)
Woman × Match		-3.511***	-3.450***	-3.520$^{***}_{\star\star}$	-4.425$^{***}_{\star\star\star}$
		(0.610)	(0.656)	(0.768)	(0.877)
Woman × Inclusive		-3.930$^{***}_{\star\star}$	-3.905$^{***}_{\star\star}$	-4.123$^{***}_{\star\star\star}$	-4.876$^{***}_{\star\star\star}$
		(0.577)	(0.605)	(0.664)	(0.915)
Constant	0.921***	0.805**	-0.146	-2.302	-2.241
	(0.302)	(0.326)	(1.257)	(1.572)	(1.499)
Demographics	✗	✗	✓	✓	✓
Language & Understanding	✗	✗	✗	✓	✓
Attitudes & Beliefs	✗	✗	✗	✗	✓
Pseudo R^2	0.140	0.154	0.192	0.292	0.346
Observations	103	103	103	103	103

Robust standard errors in parentheses, * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

p-values for interaction *effects* based on Ai and Norton (2003), $_{\star}$ 0.10 $_{\star\star}$ 0.05 $_{\star\star\star}$ 0.01

Note: *Failed attempts*$_{DG}$ was excluded from the controls as it perfectly predicts the outcome. For the complete table with all coefficients, see Table 12 in Appendix B.

Table 4: Probit regressions with the binary decision to send any positive amount (*Sent any*) in the dictator game as the dependent variable.

Whereas effects in the intensive margin were primarily driven by men, women reacted with a significantly lower probability of sending a strictly positive amount when instructions did not use the generic male frame. The significantly negative effects on interactions between *Woman* and *Match* or *Woman* and *Inclusive*, respectively, indicate that women in the W-Match and W-Inclusive treatments had a significantly lower tendency to send strictly positive amounts compared to the W-Mismatch treatment.

Result 1 (Sharing in the dictator game)**.** We do not find differences in the amount shared as dictators by men or women across treatments. The share of women who shared 0 ECU was lower if the generic male frame was used compared to the gender-inclusive or female frame.

Result 2 (The gender match in the dictator game)**.** On average, the amount shared was significantly higher in the W-Mismatch than in the M-Match treatment. We observed no differences between the M-Inclusive and the W-Inclusive or the M-Mismatch and the W-Match treatment. The differences in the average amounts shared were driven by men reacting to the treatment manipulations.

The gender gap in the amounts sent is a stylized fact in the dictator game (Engel, 2011; Bilén et al., 2021). Overall, in our experiment, the gender gap in the amount sent was reproduced when the default, generic male frame, was used. However, we observed no such gender gap when the other grammatical gender forms were used. The gap closed solely because men increased their amounts sent in the inclusive and female frames, suggesting that men reacted more strongly to deviations from conventional frames.

4.2 Prisoner's Dilemma

In the prisoner's dilemma, we concentrated on player B's conditional decisions measuring reciprocal behavior. Recall that we classified participants as "selfish", "conditional cooperators", "altruistic", or "antireciprocal" according to Miettinen et al. (2020). The shares of these different types across treatments are shown in Figure 6 in Appendix B. Due to the very low shares of altruistic and antireciprocal types, we focused on the comparison between conditional cooperators and selfish types in our analysis. Thus, excluding the four participants classified as altruist and the one participant classified as antireciprocal, leads to a sample with 98 instead of 103 observations.[15] Recall that we defined the indicator variable *Reciprocal* as one if a participant has been classified as a conditional cooperator and zero if the participant was classified as selfish.

Recall our hypothesis on prosocial behavior. For the prisoner's dilemma, this translates to the following.

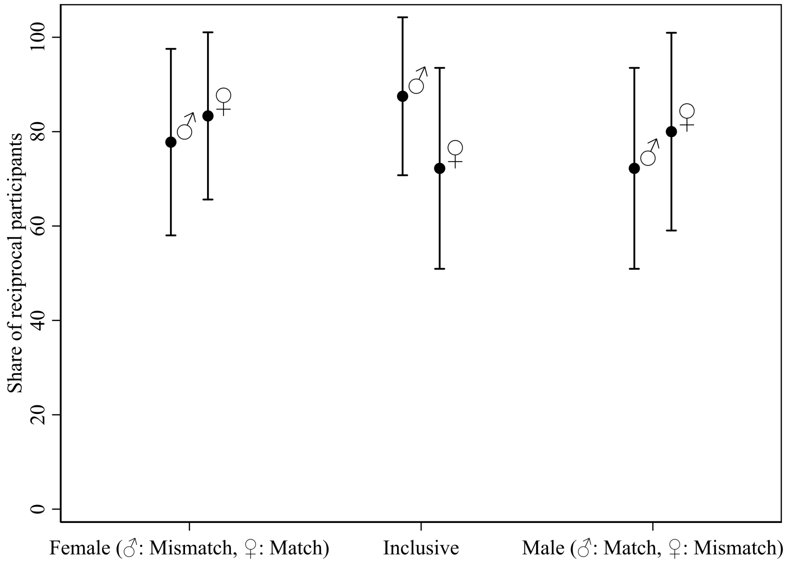
Hypothesis 3 (Reciprocal behavior in the prisoner's dilemma)**.** The share of participants classified as reciprocal is highest in the female followed by the gender-inclusive, and the lowest in the male frame of experimental instructions.

Further, recall our hypothesis on Match as a trigger of identity. For the prisoner's dilemma, this translates to the following.

Hypothesis 4 (The gender match in the prisoner's dilemma)**.** In the female frame, the share of women classified as reciprocal is higher than the share of men classified as reciprocal. In the male frame, the share of men classified as reciprocal is lower than the share of women classified as reciprocal. We expect no difference between men and women for the inclusive frame.

The share of participants classified as reciprocal was 77.78% in the M-Mismatch treatment, 87.50% in the M-Inclusive treatment, and 73.33% in the M-Match treatment. The differences between the treatments are not statistically significant (Jonckheere-Terpstra test, $p = 0.813$ and Fisher's exact

[15] To be precise, there was one altruist man, three altruist woman and one antireciprocal woman.



Note: Markers indicate means and whiskers indicate 95% confidence intervals.

Figure 3: Share of reciprocal men (♂) and women (♀).

test, $p = 0.615$).[16] The share of participants classified as reciprocal was 80.00% in the W-Mismatch treatment, 72.22% in the W-Inclusive treatment, and 81.25% in the W-Match treatment. The differences are not statistically significant (Jonckheere-Terpstra test, $p = 0.919$ and Fisher's exact test, $p = 0.833$). See Figure 3 for a graphical illustration for men (♂) and women (♀).

While the shares of reciprocal women and men looked rather similar if the male or female frame had been used, comparing the shares between the M-Inclusive and the W-Inclusive treatments indicated a difference. The share of reciprocal participants was higher in the M-Inclusive than in the M-Match or M-Mismatch treatment whereas we observed the reverse pattern for women (W-Inclusive compared to W-Match or W-Mismatch). To analyze the robustness and significance of these observations, we again ran a series of Probit regressions reported in Table 5. The results revealed no statistically significant effects of the gender frames in either specification. Also, there was no strong evidence for a gender effect (Wald-test, $p = 0.106$ for Woman). However, if we compare the M-Inclusive and the W-Inclusive treatment, the men's linear index was $-0.037 - (-1.331 - 0.037 - 0.571) = 1.902$ units higher than that of women (Wald-test, $p = 0.003$). This was also visible when we pooled the treatments using gender-inclusive and female frames. The results are reported in Table 6. The negative and significant coefficient for *Woman* in the saturated specification in column 5 reveals that women were less likely to be reciprocal and thus more likely

[16] Fisher's exact test is more suitable here due to the binary dependent variable. Since we pre-registered to use Jonckheere-Terpstra tests for our analysis, we report both test results here.

to be selfish than men when the gender-inclusive frame was used. Using this specification, we also find slightly stronger evidence for a gender gap between men and women in the generic male frame in terms of men being less reciprocal than women in the male frame. The sum of *Woman*, *Generic male*, and their interaction is $-1.270 - 0.531 + 0.620 = -1.181$ which is marginally statistically significant (Wald-test, $p = 0.099$).

Dep. Var.: Reciprocal	(1)	(2)	(3)	(4)	(5)
Woman	-0.072	0.077	0.152	-0.266	-1.331
	(0.284)	(0.497)	(0.539)	(0.601)	(0.824)
Match	-0.043	-0.142	-0.136	0.017	-0.565
	(0.353)	(0.481)	(0.504)	(0.547)	(0.608)
Inclusive	0.029	0.386	0.386	0.498	-0.037
	(0.345)	(0.522)	(0.530)	(0.565)	(0.634)
Woman × Match		0.187	0.023	0.060	1.084
		(0.708)	(0.770)	(0.822)	(0.945)
Woman × Inclusive		-0.638	-0.747	-0.804	-0.571
		(0.714)	(0.756)	(0.752)	(0.990)
Constant	0.832^{***}	0.765^{**}	1.760^{*}	-0.507	-5.460^{**}
	(0.283)	(0.331)	(1.007)	(1.271)	(2.424)
Demographics	✗	✗	✓	✓	✓
Language & Understanding	✗	✗	✗	✓	✓
Attitudes & Beliefs	✗	✗	✗	✗	✓
Pseudo R^2	0.001	0.016	0.047	0.155	0.444
Observations	98	98	98	98	98

Robust standard errors in parentheses, * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

p-values for interaction *effects* based on Ai and Norton (2003), $_{\star}$ 0.10 $_{\star\star}$ 0.05 $_{\star\star\star}$ 0.01

Note: Given the small number of observations in each category, we exclude altruistic or anti-reciprocal types in our regressions. This leads to a sample with 98 instead of 103 observations. For the complete table with all coefficients, see Table 13 in Appendix B.

Table 5: Probit regressions with *Reciprocal* as the dependent variable in the prisoner's dilemma.

Result 3 (Reciprocal behavior in the prisoner's dilemma)**.** Neither men nor women reacted to changes in the framing of instructions.

Result 4 (The gender match in the prisoner's dilemma)**.** There is mild and statistically marginally significant evidence for a gender gap in reciprocal behavior under the inclusive gender frame.

In line with the meta-analysis by Mengel (2018) our results suggest that gender gaps, if they occur, are specific to the study design and are thus not a stylized finding. In our analysis, there was mild evidence of a gender gap when the gender-inclusive frame was used. Apparently, changing the gender frame did not constitute enough of a change to the strategic interaction environment to affect behavior.

Dep. Var.: Reciprocal	(1)	(2)	(3)	(4)	(5)
Woman	-0.072	-0.207	-0.263	-0.563	-1.270***
	(0.286)	(0.348)	(0.351)	(0.384)	(0.467)
Generic male	-0.094	-0.306	-0.305	-0.217	-0.531
	(0.307)	(0.432)	(0.443)	(0.480)	(0.532)
Woman × Generic male		0.426	0.551	0.353	0.620
		(0.617)	(0.642)	(0.713)	(0.879)
Constant	0.858***	0.929***	1.911**	-0.220	-4.380*
	(0.222)	(0.254)	(0.969)	(1.216)	(2.291)
Demographics	✗	✗	✓	✓	✓
Language & Understanding	✗	✗	✗	✓	✓
Attitudes & Beliefs	✗	✗	✗	✗	✓
Pseudo R^2	0.002	0.006	0.039	0.143	0.421
Observations	98	98	98	98	98

Robust standard errors in parentheses, * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

p-values for interaction *effects* based on Ai and Norton (2003), $_{\star}$ 0.10 $_{\star\star}$ 0.05 $_{\star\star\star}$ 0.01

Note: Given the small number of observations in each category, we exclude altruistic or anti-reciprocal types in our regressions. This leads to a sample with 98 instead of 103 observations. For the complete table with all coefficients, see Table 14 in Appendix B.

Table 6: Probit regressions with *Reciprocal* as the dependent variable in the prisoner's dilemma when comparing the generic male with the "non-standard" female and gender-inclusive framed instructions.

4.3 Deception Game

Given that we focus on honest reporting, we concentrate our analysis on A players who could send a message to B players.[17] Recall our hypothesis on prosocial behavior. For the deception game, this translates to the following.

Hypothesis 5 (Honest reporting in the deception game)**.** The share of honest reports is highest in the female followed by the gender-inclusive, and the lowest in the male frame of experimental instructions.

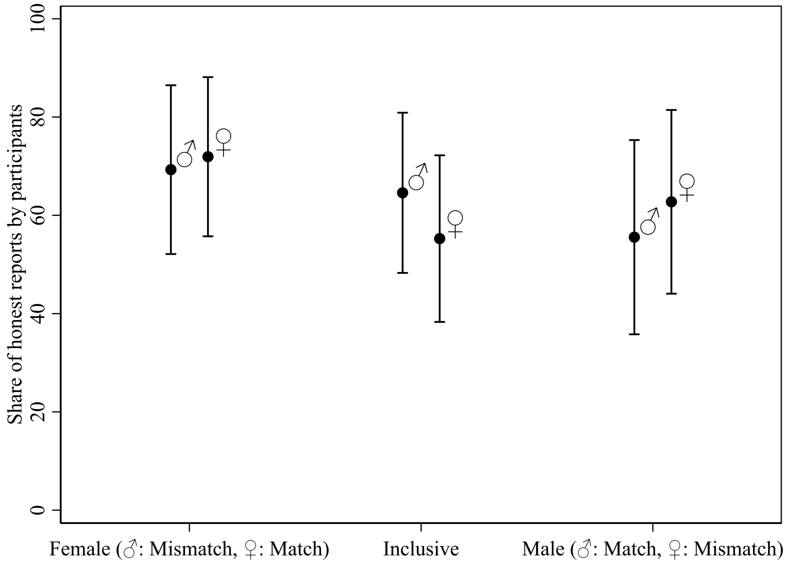
Further, recall our hypothesis on Match as a trigger of identity. For the prisoner's dilemma, this translates to the following.

Hypothesis 6 (The gender match in the deception game)**.** In the female frame, the share of honest reports by women is higher than the share of honest reports by men. In the male frame, the share of honest reports by men is lower than the share of honest reports by women. We expect no difference between men and women for the inclusive frame.

On average, 69.30% of messages sent by player A in the M-Mismatch treatment were honest. In the M-Inclusive treatment, an average of 64.58% of A players sent the honest message, whereas,

[17] Player B behavior and further analysis on player A behavior can be found in the online appendix.

in the M-Match treatment 57.78% of messages were honest. As depicted in Figure 4, men (♂)



Note: Markers indicate means and whiskers indicate 95% confidence intervals.

Figure 4: Means of the share of honest reports in the deception game among all messages by men (♂) and women (♀).

behaved more honestly when moving from the M-Match over the M-Inclusive to the M-Mismatch treatment. The increase when going from the M-Match to the M-Mismatch treatment is roughly 12 percentage points. However, statistically, we do not find a significant pattern (Jonckheere-Terpstra test, $p = 0.403$).

62.75% of A-players in the W-Mismatch treatment sent an honest message. In the W-Inclusive treatment, an average of 55.26% of messages sent were honest. Finally, 75.49% of all messages were honest in the W-Match treatment. There is no qualitative or statistically significant pattern when moving from the W-Mismatch over the W-Inclusive to the W-Match treatment (Jonckheere-Terpstra test, $p = 0.335$).

Again, we employed a series of linear regressions to analyze the behavioral patterns. The results are provided in Table 7 and Table 8. The data did not corroborate any statistically significant differences across treatments.

Result 5 (Honest reporting in the deception game)**.** There is no significant difference in the share of honest reports across treatments.

Result 6 (The gender match in the deception game)**.** There is no evidence for a gender gap in the share of honest reports across treatments.

Dep. Var.: Share honest	(1)	(2)	(3)	(4)	(5)
Woman	0.055	-6.553	-6.957	-7.114	-11.405
	(7.542)	(12.966)	(12.895)	(14.926)	(14.511)
Match	0.981	-11.520	-13.662	-11.195	-13.166
	(9.625)	(14.146)	(14.602)	(14.534)	(11.773)
Inclusive	-6.684	-4.715	-6.718	-4.368	-9.295
	(8.814)	(12.086)	(12.322)	(12.082)	(10.376)
Woman × Match		24.266	23.867	19.635	18.365
		(19.205)	(19.587)	(20.848)	(18.841)
Woman × Inclusive		-2.767	-2.330	-6.287	9.862
		(17.670)	(17.915)	(17.839)	(17.450)
Constant	66.178***	69.298***	97.930***	47.509	35.030
	(7.295)	(8.788)	(25.172)	(30.591)	(31.655)
Demographics	✗	✗	✓	✓	✓
Language & Understanding	✗	✗	✗	✓	✓
Attitudes & Beliefs	✗	✗	✗	✗	✓
R^2	0.008	0.033	0.076	0.145	0.361
Observations	103	103	103	103	103

Robust standard errors in parentheses, * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Note: For the complete table with all coefficients, see Table 15 in Appendix B.

Table 7: OLS regressions with *Share honest* as the dependent variable in the deception game.

Dep. Var.: Share honest	(1)	(2)	(3)	(4)	(5)
Woman	-0.067	-2.328	-4.048	-5.293	-3.410
	(7.507)	(8.778)	(8.996)	(9.459)	(9.436)
Generic male	-5.544	-9.365	-10.377	-9.370	-8.599
	(8.405)	(12.528)	(12.937)	(12.741)	(9.345)
Woman × Generic male		7.295	10.430	10.556	5.625
		(16.926)	(17.055)	(17.280)	(14.018)
Constant	65.997***	67.143***	100.773***	50.225*	30.637
	(5.646)	(6.047)	(23.303)	(29.867)	(30.047)
Demographics	✗	✗	✓	✓	✓
Language & Understanding	✗	✗	✗	✓	✓
Attitudes & Beliefs	✗	✗	✗	✗	✓
R^2	0.005	0.007	0.051	0.125	0.355
Observations	103	103	103	103	103

Robust standard errors in parentheses, * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Note: For the complete table with all coefficients, see Table 16 in Appendix B.

Table 8: OLS regressions with *Share honest* as the dependent variable in the deception game when comparing the generic male with the "non-standard" female and gender-inclusive framed instructions.

The pattern in behavior across treatments was reminiscent of the pattern we observed in the dictator game, but statistically not significant. Our results qualitatively indicated a tendency that "non-standard" gender frames increased honest reporting by men by up to roughly 12 percentage points. This is an economically relevant difference, but it is statistically insignificant.

5 Discussion

The goal of our study is to investigate if and how different gender frames used in economic experiments impact prosocial behavior. Following theories and evidence on social identity, we investigated if a match between the self-reported gender and the gender-framed language had a different impact on women and men. Overall, we observed mild effects of our treatment manipulations. Given that we only varied the frame and not the underlying game, the small effects are noteworthy. We observed the strongest effect in the first game, the dictator game, where we reproduced the well-documented finding that women share more than men *only* if the generic male frame was used. This effect was driven by men who shared more if the gender-inclusive or the female frame were used. For men, a clear mismatch between their self-reported gender and the gender frame in the instructions triggered more prosocial behavior in the dictator game. This observation is not pronounced in the other two games and the differences are far from being significant. However, the direction of the effect in the deception game is in line with the observations in the dictator game. Regarding women, we observed a low variance in behavior due to our treatment manipulations. A notable exception is the observation that all women sent a strictly positive amount in the dictator game under the male frame whereas the share was lower using the other two frames. This pattern might be due to a negative reaction to a higher salience of the female identity in these frames. From the women's perspective, their anyway salient female identity with the associated stereotypical roles and behavioral prescriptions was made even more salient in the "non-standard" gender frames. Some women reacted to this high salience of their female identity with what can be described as psychological reactance (Brehm, 1966; Rains, 2013). Sending nothing is the strongest possible reaction in the dictator game. However, in all treatments, women held their social identity of being female which related to "typically" female stereotypes such as being more caring, prosocial, or cooperative (Eckel and Grossman, 1998; Cejka and Eagly, 1999; Rudman and Glick, 2001; Azmat and Petrongolo, 2014).

Summing up, our results point in the direction that men reacted more strongly to variations in the gender frame of the instructions. Women, on the other hand, seemed to be more used to being addressed by different gender frames leading to fewer behavioral changes.

However, the effects might also be driven by differences in the beliefs about the other player in the game or due to differences in the comprehension of the instructions. Note that we already included incentivized beliefs and survey items relating to comprehension in our regression models.

Overall, including these controls did not hint at alternative behavioral mechanisms and the main findings are robust if we include these controls. Nevertheless, in the following, we present a series of additional robustness checks addressing these potential confounding factors. We subsequently discuss the limitations of our study and highlight avenues for future research.

Previous studies indicate that men behave in a more prosocial way when they are matched with women in certain economic situations (Eckel and Grossman, 2001). If men in the gender-inclusive and female treatments of our experiment assumed to interact with women (or at least assumed a higher likelihood that they did), this *chivalry* could explain our findings in the dictator game. We did not provide participants with any information on each other to trigger such an assumption. Nevertheless, participants might have perceived the gender frame in the instructions as a signal about the gender of the participants they were being matched with. If this was the case, it would also be reflected in the strategic beliefs we elicited at the end of the experiment. However, we did not find such differences in strategic beliefs between the gender frames in all games (smallest p-value is $p = 0.133$, Kruskal-Wallis test for women's second-order beliefs in the deception game).[18] Thus, we conjecture that assumptions about the self-chosen gender of the other participant were not the main drivers of our results.

Next, we focus on attention and comprehension which might differ due to the gender frame used in the instructions. As we argued earlier, there is ample evidence that language affects cognition (Hunt and Agnoli, 1991; Majid et al., 2004; Semin, 2013; Houston, 2019) which might result in different attention levels. Recall that we had an attention check built into the survey which we can use as a proxy for attention. If the gender frame of the experiment impacted overall attention, we should observe differences in this attention check. In total, however, only five participants failed the attention check rendering any statistical analysis on this variable not feasible. Three of the five failed attention checks occurred in the M-Match treatment, and the other two occurred in the W-Match treatment. Thus, the share of participants who failed the attention check was low across all treatments. Including these five observations in our analysis does thus not change our results.[19]

Another way in which the "non-standard" gender frames might have drawn the participants' cognitive resources is by lowering their comprehension of the underlying games. Though research in social psychology suggests that this is not the case (Friedrich and Heise, 2019), this could be different when texts are used as instructions for games in which readers have to engage in potentially complex strategic reasoning and deliberation over different economic and social motives.

[18] Recall that we elicited first and second-order beliefs for each of the three games.

[19] The results are available from the authors upon request.

We included control questions on the general instructions, the game-specific instructions, and for the belief elicitation. Whenever a participant provided a wrong answer to a control question, they could not proceed to the next page and they received a prompt.[20] We recorded the number of failed attempts at each of the questions and used failed attempts as a proxy to measure comprehension. First, we considered the failed attempts across all control questions in the entire experiment ($\textit{Failed attempts}_{all}$) and observed that the number of failed attempts did not vary significantly across gender frames ($p = 0.107$, Kruskal-Wallis test for men; $p = 0.287$, Kruskal-Wallis test for women). Second, we analyzed the number of failed attempts for each stage of the experiment (pre-game, each of the three games, and post-games). Across the three gender frames, there was a significant difference for the pre-game control questions for women ($p = 0.023$, Kruskal-Wallis test) but not for men ($p = 0.178$, Kruskal-Wallis test). For the other stages, we observed that the number of failed attempts did not constitute a significant pattern across gender frames.[21] In addition, we also elicited a subjective measure of comprehensibility which was also included in the controls for language and understanding in our regressions. Across the three gender frames, there was a marginally statistically significant difference for women ($p = 0.085$, Kruskal-Wallis test) but not for men ($p = 0.309$, Kruskal-Wallis test), which is in line with the differences in failed attempts at control questions in the pre-game stage. This was driven by the W-Inclusive treatment, in which participants stated a lower subjective comprehensibility ($p = 0.046$, for the comparison of the W-Inclusive treatment to the W-Match treatment and $p = 0.057$ for the W-Mismatch treatment, Mann-Whitney U tests). The difference between the W-Match with the W-Mismatch treatment is not statistically significant ($p = 0.951$, Mann-Whitney U test).

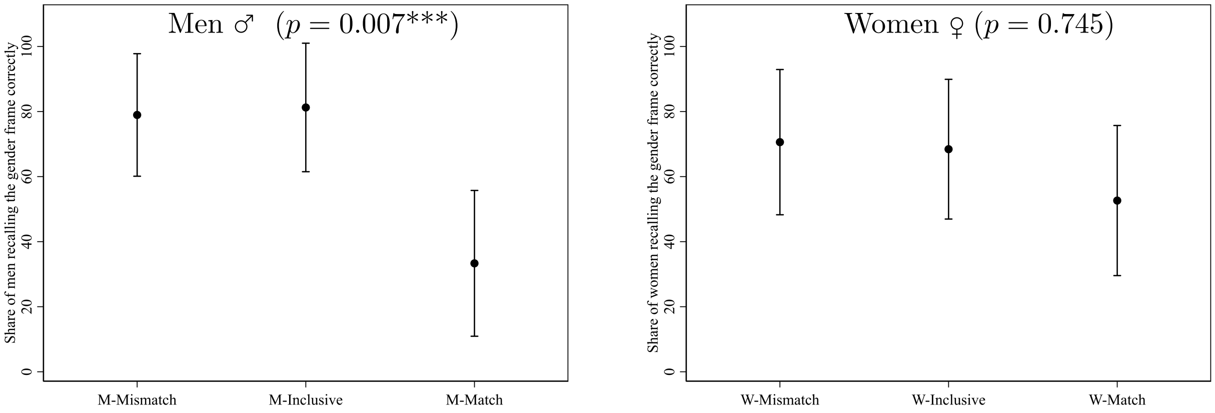
At the end of the experiment, we asked participants which grammatical gender was used throughout the experiment.[22] This can serve as a proxy for the salience of the gender frame. Consider Figure 5 which shows the share of participants who remembered the correct grammatical gender used in their instructions. Men's recall of the correct grammatical gender was significantly lower if the male frame was used compared to the female and gender-inclusive frame ($p = 0.007$, Kruskal-Wallis test; $p = 0.018$, Mann-Whitney U test for male versus inclusive instructions; $p = 0.019$, Mann-Whitney

[20] If it was their first failed attempt and there were more than two options, the prompt would read "Unfortunately, your answer to this question is wrong. Please review the instructions at the bottom of the page and try again". If there were n options provided for the control question, the prompt would read "Unfortunately, you have repeatedly answered the question incorrectly. The correct answer is: X.," with X being the respective correct answer, whenever a participant failed to provide a correct answer for at least $n - 1$ times.

[21] We also conducted regressions where we reduced our analytical sample to only those participants with less than four failed attempts. The results did not differ from those presented earlier and are available from the authors upon request.

[22] On the page where we asked this question, there was no occurrence of gendered nouns or pronouns that would give away the correct answer and participants could not return to previous pages.

U test for male versus female instructions). For women, there was no significant difference between the treatments ($p = 0.745$, Kruskal-Wallis test). This suggests that gender is a much more salient feature to men in the "non-standard" gender frames than to women.



(a) (b)

Note: Markers indicate means and whiskers indicate 95% confidence intervals.

Figure 5: Men's recall of the grammatical gender is significantly better in non-standard formulations (a) while women's recall does not differ (b).

Recall that we found that men behaved in a more prosocial way (sharing more in the dictator game) in the female gender frame compared to the generic male frame. According to identity theory, this can be explained by the female identity being more salient. This is in line with the finding that participants in the M-Match treatment failed the attention check more often than in the other two treatments and remembered the gender frame of instructions significantly better. The increase in the amounts sent when comparing the M-Inclusive treatment to the amounts sent in the M-Match treatment is statistically insignificant but the direction is in line with this explanation as well. For women, we found that they behaved slightly less prosocial by giving zero more often in the dictator game in the "non-standard" gender frames compared to the generic male frame and acted less reciprocal in the prisoner's dilemma in the W-Inclusive treatment compared to the W-Mismatch treatment which can be described as psychological reactance (Brehm, 1966; Rains, 2013).

In the following, we discuss further explanations and limitations of our findings. First, we find the strongest effect of gendered language on economic behavior in our first game, the dictator game. Recall that we did not randomize the order of the three games played. We chose to not randomize the order of games since randomization would add another layer of complexity to the analysis. However, this choice leaves the question if the battery of games influences the salience of the gender frames and the effect on economic behavior for further research. However, our controls for salience have been elicited at the end of the experiment. The fact that we did find differences for men in recall of the correct gender frame across treatments is one piece of evidence indicating that some treatment differences remained until the conclusion of the experiment.

Second, we deliberately wrote our instructions with a high frequency of grammatical gender formulations, like articles, pronouns, and the word "participant". Also, we employed a gender-inclusive form (the gender star) that was rather salient while reading. Indeed, we recorded most of the language-related free-text comments in the M-Inclusive and W-Inclusive treatments. Other forms of gender-inclusive language are less conspicuous, rendering our results a potential upper bound of the effects of gendered language on economic behavior.

Third, we ran our study in a fairly small, homogeneous group of students. As we have argued earlier in the paper, there is a heated debate over gender-inclusive language, which might reach some groups in society easier than others. In the general public, effects are likely to be more pronounced as older people grew up before gender-inclusive forms were introduced.

Fourth, given the small sample size, the effects need to be rather large to be picked up by statistical tests. Our experiment employed a very light frame from an economist's perspective. Typically in framing, the different formulations across treatments change the perspective of the inherent trade-off in the decision to be made (Fiedler and Hillenbrand, 2020) or the externalities imposed on others (Cartwright and Ramalingam, 2019). Our gender frames did not change any incentives and thus the effect sizes could be expected to be rather small.

Our main result, that the gender difference in dictator game giving can only be found between the W-Mismatch and M-Match treatments, resulted in a 64.04% power with a posthoc power analysis for the Mann-Whitney U test.

Yet, also with respect to correcting for multiple comparisons (List et al., 2019) and the statistical implications that this would have for the results of this paper, our experiment should be seen as a starting point, investigating the influence of grammatical gender in texts on economic behavior.

Finally, we only investigated a small range of economic domains, namely those of sharing, reciprocity, and honesty. Domains like competition (Niederle and Vesterlund, 2007), individual decision-making, and leadership behavior (Chen and Houser, 2019), to only name a few, are domains where gender effects have been documented, largely using generic male instructions.

Beyond these points, we want to emphasize an important issue for experimental economics. Experiments typically encompass a baseline and one or more treatments, varying factors like the incentive structure, the available information, or the group size. As just mentioned, in the games we investigated, we held these factors constant and only varied the gender frame of personal nouns and pronouns. On the one hand, this is a very mild treatment variation for an economic experiment. On the other hand, though, we did not actually investigate how the changes of gendered language affect treatment effects induced by varying any of the aforementioned, more "traditional" factors. Yet, this interaction between the effects of gendered language and treatment effects in economic experiments is an interesting and important future area for research.

Taken together, our results indicate that there is no immediate reason to doubt stylized experimental findings in economics in general. In the specific context of the dictator game, the well-known gender gap was closed when gender-inclusive and female frames were used, exclusively due to men increasing the amounts sent. The tendencies in the other games were in line with this finding, even though they were not statistically significant.

Understanding whether these findings are only a result of our specific and comparably small sample remains an empirical question for future studies. Furthermore, a study run in several countries with languages with and without grammatical gender would be an interesting extension of our experiment. Another variation could be to reveal the gender of the partner and test for differences between the behavior of men and women when interacting in same-gender and mixed-gender dyads.

6 Conclusion

We reported results from a controlled experiment in which we varied the grammatical gender in the instructions such that it could either match the self-reported gender of participants or not or was gender-inclusive and did neither explicitly include nor exclude any gender. In the dictator game, we observed the well-known gender gap in the amount sent if the instructions used the generic male frame. In the other two frames, the gender gap attenuated or vanished completely. The results regarding reciprocal behavior and honest reporting were less pronounced.

In a narrow sense, our experiment helps to shed light on the question if a gender framing of instructions affects experimental results in the laboratory. Our experiment is only a first step toward a better understanding of this topic. We need more experiments with different subject pools and focus on different economic behaviors to deepen our insights. From a broader perspective, our results are also informative for the ongoing debate about the risks and benefits of gendered language. Many of the well-documented differences between men and women can be attributed to a feedback loop of an existing, structural inequality between men and women that itself leads to differences in behavior. These differences in behavior, in turn, are known to lead to those structural differences over time. While our stylized experiment needs to be evaluated outside the lab, our results indicate, at least to some extent, that language is a potential tool to help break this loop. What is more, this tool, in comparison to other tools like quotas and affirmative action, is a relatively inexpensive, if yet controversial, intervention. Furthermore, this intervention does not only favor one group, such as quotas but optimally results in the inclusion of all humans alike (Balafoutas and Sutter, 2012; Niederle et al., 2013).

References

Abbink, K. and H. Hennig-Schmidt (2006): "Neutral versus loaded instructions in a bribery experiment," *Experimental Economics*, 9, 103–121.

Ai, C. and E. C. Norton (2003): "Interaction terms in logit and probit models," *Economics Letters*, 80, 123–129.

Akerlof, G. A. and R. E. Kranton (2000): "Economics and identity," *The Quarterly Journal of Economics*, 115, 715–753.

Antidiskriminierungsstelle des Bundes (2023): "Warum die Antidiskriminierungsstelle des Bundes den Genderstern nutzt," https://www.antidiskriminierungsstelle.de/SharedDocs/aktuelles/DE/2023/20230227_Genderstern.html, Last accessed: 31/03/2023.

Archer, A. M. and C. D. Kam (2022): "She is the chair (man): Gender, language, and leadership," *The Leadership Quarterly*, Article 101610.

Arias, O., C. Canals, A. Mizala, and F. Meneses (2023): "Gender gaps in Mathematics and Language: The bias of competitive achievement tests," *PloS ONE*, 18, e0283384.

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

Balafoutas, L., H. Fornwagner, E. Hauser, and O. Hauser (2023): "Gender-Inclusive Language and Economic Decision-Making," SSRN Working Paper 4411481.

Balafoutas, L., H. Fornwagner, and M. Sutter (2018): "Closing the gender gap in competitiveness through priming," *Nature Communications*, 9, 1–6.

Balafoutas, L. and M. Sutter (2012): "Affirmative action policies promote women and do not harm efficiency in the laboratory," *Science*, 335, 579–582.

Beblo, M., L. Görges, and E. Markowsky (2020): "Gender matters in language and economic behaviour: Can we measure a causal cognition effect of speaking?" *Labour Economics*, 65, Article 101850.

Bendel, O. (2021): "Gendersternchen," https://wirtschaftslexikon.gabler.de/definition/gendersternchen-123255/version-384841, Last accessed: 31/03/2023.

Bertrand, M. (2011): "New perspectives on gender," in *Handbook of Labor Economics*, Elsevier, vol. 4, 1543–1590.

Bilén, D., A. Dreber, and M. Johannesson (2021): "Are women more generous than men? A meta-analysis," *Journal of the Economic Science Association*, 7, 1–18.

Boggio, C., F. C. Moscarola, and A. Gallice (2020): "What is good for the goose is good for the gander? How gender-specific conceptual frames affect financial participation and decision-making," *Economics of Education Review*, 75, Article 101952.

Bolle, F. and P. Ockenfels (1990): "Prisoners' dilemma as a game with incomplete information," *Journal of Economic Psychology*, 11, 69–84.

Borghans, L., J. J. Heckman, B. H. Golsteyn, and H. Meijers (2009): "Gender differences in risk aversion and ambiguity aversion," *Journal of the European Economic Association*, 7, 649–658.

Brañas-Garza, P. (2007): "Promoting helping behavior with framing in dictator games," *Journal of Economic Psychology*, 28, 477–486.

Brañas-Garza, P., V. Capraro, and E. Rascon-Ramirez (2018): "Gender differences in altruism on Mechanical Turk: Expectations and actual behaviour," *Economics Letters*, 170, 19–23.

Brehm, J. W. (1966): *A theory of psychological reactance.*, Academic Press.

Capraro, V. (2018): "Gender differences in lying in sender-receiver games: A meta-analysis," *Judgment and Decision Making*, 13, 345–355.

Capraro, V. and A. Vanzo (2019): "The power of moral words: Loaded language generates framing effects in the extreme dictator game," *Judgment and Decision Making*, 14, 309–317.

Card, D., F. Colella, and R. Lalive (2021): "Gender preferences in job vacancies and workplace gender diversity," National Bureau of Economic Research Working Paper No. 29350.

Cartwright, E. and A. Ramalingam (2019): "Framing effects in public good games: Choices or externalities?" *Economics Letters*, 179, 42–45.

Cejka, M. A. and A. H. Eagly (1999): "Gender-stereotypic images of occupations correspond to the sex segregation of employment," *Personality and Social Psychology Bulletin*, 25, 413–423.

Chang, D., R. Chen, and E. Krupka (2019): "Rhetoric matters: A social norms explanation for the anomaly of framing," *Games and Economic Behavior*, 116, 158–178.

Chen, D. L., M. Schonger, and C. Wickens (2016): "oTree—An open-source platform for laboratory, online, and field experiments," *Journal of Behavioral and Experimental Finance*, 9, 88–97.

Chen, J. and D. Houser (2019): "When are women willing to lead? The effect of team gender composition and gendered tasks," *The Leadership Quarterly*, 30, Article 101340.

Chen, M. K. (2013): "The effect of language on economic behavior: Evidence from savings rates, health behaviors, and retirement assets," *American Economic Review*, 103, 690–731.

Conrads, J., B. Irlenbusch, R. M. Rilke, A. Schielke, and G. Walkowitz (2014): "Honesty in tournaments," *Economics Letters*, 123, 90–93.

Crawford, M. and L. English (1984): "Generic versus specific inclusion of women in language: Effects on recall," *Journal of Psycholinguistic Research*, 13, 373–381.

Croson, R. and U. Gneezy (2009): "Gender differences in preferences," *Journal of Economic Literature*, 47, 448–474.

Cubelli, R., D. Paolieri, L. Lotto, and R. Job (2011): "The effect of grammatical gender on object categorization." *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 37, 449–460.

Dato, S. and P. Nieken (2014): "Gender differences in competition and sabotage," *Journal of Economic Behavior & Organization*, 100, 64–80.

Davis, L. and M. Reynolds (2018): "Gendered language and the educational gender gap," *Economics Letters*, 168, 46–48.

Delfino, A. (2021): "Breaking gender barriers: Experimental evidence on men in pink-collar jobs," IZA Discussion Paper No. 14083.

Deutsch, M. (1960): "The effect of motivational orientation upon trust and suspicion," *Human Relations*, 13, 123–139.

Dohmen, T., A. Falk, D. Huffman, and U. Sunde (2009): "Homo reciprocans: Survey evidence on behavioural outcomes," *The Economic Journal*, 119, 592–612.

Dohmen, T., A. Falk, D. Huffman, U. Sunde, J. Schupp, and G. G. Wagner (2011): "Individual risk attitudes: Measurement, determinants, and behavioral consequences," *Journal of the European Economic Association*, 9, 522–550.

Doñate-Buendía, A., A. García-Gallego, and M. Petrović (2022): "Gender and other moderators of giving in the dictator game: A meta-analysis," *Journal of Economic Behavior & Organization*, 198, 280–301.

Dreber, A. and M. Johannesson (2008): "Gender differences in deception," *Economics Letters*, 99, 197–199.

Dufwenberg, M. and G. Kirchsteiger (2000): "Reciprocity and wage undercutting," *European Economic Review*, 44, 1069–1078.

D'Acunto, F., U. Malmendier, and M. Weber (2021): "Gender roles produce divergent economic expectations," *Proceedings of the National Academy of Sciences*, 118, e2008534118.

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

——— (2001): "Chivalry and solidarity in ultimatum games," *Economic Inquiry*, 39, 171–188.

Ellingsen, T., M. Johannesson, J. Mollerstrom, and S. Munkhammar (2013): "Gender differences in social framing effects," *Economics Letters*, 118, 470–472.

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

European Parliament (2018): "Gender-neutral language in the european parliament," Guideline.

Fiedler, S. and A. Hillenbrand (2020): "Gain-loss framing in interdependent choice," *Games and Economic Behavior*, 121, 232–251.

Fornwagner, H., B. Grosskopf, A. Lauf, V. Schöller, and S. Städter (2022): "On the robustness of gender differences in economic behavior," *Scientific Reports*, 12, Article 21549.

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

Friedrich, M. C. and E. Heise (2019): "Does the use of gender-fair language influence the comprehensibility of texts?: An experiment using an authentic contract manipulating single role nouns and pronouns," *Swiss Journal of Psychology*, 78, 51–60.

Gabriel, U. and F. Mellenberger (2004): "Exchanging the generic masculine for gender-balanced forms - The impact of context valence," *Swiss Journal of Psychology*, 63, 273–278.

Galor, O., Ö. Özak, and A. Sarid (2020): "Linguistic traits and human capital formation," *AEA Papers and Proceedings*, 110, 309–313.

Gaucher, D., J. Friesen, and A. C. Kay (2011): "Evidence that gendered wording in job advertisements exists and sustains gender inequality," *Journal of Personality and Social Psychology*, 101, 109–128.

Gay, V., D. L. Hicks, E. Santacreu-Vasut, and A. Shoham (2018): "Decomposing culture: An analysis of gender, language, and labor supply in the household," *Review of Economics of the Household*, 16, 879–909.

Gerlach, P., K. Teodorescu, and R. Hertwig (2019): "The truth about lies: A meta-analysis on dishonest behavior." *Psychological Bulletin*, 145, 1–44.

Gneezy, U. (2005): "Deception: The role of consequences," *American Economic Review*, 95, 384–394.

Gneezy, U., B. Rockenbach, and M. Serra-Garcia (2013): "Measuring lying aversion," *Journal of Economic Behavior & Organization*, 93, 293–300.

Gorny, P. M., P. Nieken, and K. Ströhlein (2023): "The effects of gendered language on norm compliance," Working Paper.

Greiner, B. (2015): "Subject pool recruitment procedures: Organizing experiments with ORSEE," *Journal of the Economic Science Association*, 1, 114–125.

Grosch, K. and H. A. Rau (2017): "Gender differences in honesty: The role of social value orientation," *Journal of Economic Psychology*, 62, 258–267.

Grullón Paz, I. (2021): "F.A.A. committee recommends a pivot to gender-neutral terminology," https://www.nytimes.com/2021/06/24/us/faa-gender-neutral-drones.html?searchResultPosition=8, Last accessed: 11/07/2022.

Güth, W., R. Schmittberger, and B. Schwarze (1982): "An experimental analysis of ultimatum bargaining," *Journal of Economic Behavior & Organization*, 3, 367–388.

Gylfason, H. F., A. A. Arnardottir, and K. Kristinsson (2013): "More on gender differences in lying," *Economics Letters*, 119, 94–96.

Haerpfer, C., R. Inglehart, A. Moreno, C. Welzel, K. Kizilova, D.-M. J., M. Lagos, P. Norris, E. Ponarin, B. Puranen, et al., eds. (2020): *World Values Survey: Round Seven – Country-Pooled Datafile.*, Madrid, Spain & Vienna, Austria: JD Systems Institute & WVSA Secretariat., http://dx.doi.org/10.14281/18241.13 .

Haspelmath, M., M. S. Dryer, D. Gil, and B. Comrie (2005): *The world atlas of language structures*, OUP Oxford.

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

Hicks, D. L., E. Santacreu-Vasut, and A. Shoham (2015): "Does mother tongue make for women's work? Linguistics, household labor, and gender identity," *Journal of Economic Behavior & Organization*, 110, 19–44.

Hodel, L., M. Formanowicz, S. Sczesny, J. Valdrová, and L. von Stockhausen (2017): "Gender-fair language in job advertisements: A cross-linguistic and cross-cultural analysis," *Journal of Cross-Cultural Psychology*, 48, 384–401.

Hoffman, E., K. McCabe, K. Shachat, and V. Smith (1994): "Preferences, property rights, and anonymity in bargaining games," *Games and Economic Behavior*, 7, 346–380.

Horvath, L. K. and S. Sczesny (2016): "Reducing women's lack of fit with leadership positions? Effects of the wording of job advertisements," *European Journal of Work and Organizational Psychology*, 25, 316–328.

Houser, D., S. Vetter, and J. K. Winter (2012): "Fairness and cheating," *European Economic Review*, 56, 1645–1655.

Houston, S. H. (2019): *A survey of psycholinguistics*, De Gruyter Mouton.

Huber, C. and J. Huber (2020): "Bad bankers no more? Truth-telling and (dis) honesty in the finance industry," *Journal of Economic Behavior & Organization*, 180, 472–493.

Huber, J. and M. Kirchler (2012): "The impact of instructions and procedure on reducing confusion and bubbles in experimental asset markets," *Experimental Economics*, 15, 89–105.

Hunt, E. and F. Agnoli (1991): "The Whorfian hypothesis: A cognitive psychology perspective." *Psychological Review*, 98, 377–389.

Jakiela, P. and O. Ozier (2021): "Gendered language," Working Paper.

Kahneman, D., J. L. Knetsch, and R. H. Thaler (1986): "Fairness and the assumptions of economics," *Journal of Business*, 59, 285–300.

Kahneman, D. and A. Tversky (2013): "Choices, values, and frames," in *Handbook of the Fundamentals of Financial Decision Making: Part I*, World Scientific, 269–278.

Kantar Public (2020): "SOEP-Core-2019: Personenfragebogen, Stichproben A-L3, M1-M2 + N-P," SOEP Survey Papers 909: Series A.

Kricheli-Katz, T. and T. Regev (2021): "The effect of language on performance: Do gendered languages fail women in maths?" *NPJ Science of Learning*, 6, Article 9.

Lankes, A. (2022): "In Argentina, one of the world's first bans on gender-neutral language," https://www.nytimes.com/2022/07/20/world/americas/argentina-gender-neutral-spanish.html?searchResultPosition=5, Last accessed: 11/07/2022.

Levin, I. P., S. L. Schneider, and G. J. Gaeth (1998): "All frames are not created equal: A typology and critical analysis of framing effects," *Organizational Behavior and Human Decision Processes*, 76, 149–188.

Liberman, V., S. M. Samuels, and L. Ross (2004): "The name of the game: Predictive power of reputations versus situational labels in determining prisoner's dilemma game moves," *Personality and Social Psychology Bulletin*, 30, 1175–1185.

Lindqvist, A., E. A. Renström, and M. Gustafsson Sendén (2019): "Reducing a male bias in language? Establishing the efficiency of three different gender-fair language strategies," *Sex Roles*, 81, 109–117.

List, J. A., A. M. Shaikh, and Y. Xu (2019): "Multiple hypothesis testing in experimental economics," *Experimental Economics*, 22, 773–793.

Lu, S. (2015): "Guidelines for gender-inclusive language in English," *Monitor on Psychology*, 46.

Majid, A., M. Bowerman, S. Kita, D. B. Haun, and S. C. Levinson (2004): "Can language restructure cognition? The case for space," *Trends in Cognitive Sciences*, 8, 108–114.

Mavisakalyan, A. (2015): "Gender in language and gender in employment," *Oxford Development Studies*, 43, 403–424.

Mavisakalyan, A. and C. Weber (2018): "Linguistic structures and economic outcomes," *Journal of Economic Surveys*, 32, 916–939.

May, T.; Ueno, H. (2020): "No more 'ladies and gentlemen' on Japan airlines," https://www.nytimes.com/2020/09/29/world/asia/japan-airlines-ladies-gentlemen.html?searchResultPosition=12, Last accessed: 11/07/2022.

Mengel, F. (2018): "Risk and temptation: A meta-study on prisoner's dilemma games," *The Economic Journal*, 128, 3182–3209.

Miettinen, T., M. Kosfeld, E. Fehr, and J. Weibull (2020): "Revealed preferences in a sequential prisoners' dilemma: A horse-race between six utility functions," *Journal of Economic Behavior & Organization*, 173, 1–25.

Muehlheusser, G., A. Roider, and N. Wallmeier (2015): "Gender differences in honesty: Groups versus individuals," *Economics Letters*, 128, 25–29.

Niederle, M., C. Segal, and L. Vesterlund (2013): "How costly is diversity? Affirmative action in light of gender differences in competitiveness," *Management Science*, 59, 1–16.

Niederle, M. and L. Vesterlund (2007): "Do women shy away from competition? Do men compete too much?" *The Quarterly Journal of Economics*, 122, 1067–1101.

Norton, E. C., H. Wang, and C. Ai (2004): "Computing interaction effects and standard errors in logit and probit models," *The Stata Journal*, 4, 154–167.

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

Perszyk, D. R. and S. R. Waxman (2018): "Linking language and cognition in infancy," *Annual Review of Psychology*, 69, 231–250.

Ploner, M. and T. Regner (2013): "Self-image and moral balancing: An experimental analysis," *Journal of Economic Behavior & Organization*, 93, 374–383.

Prewitt-Freilino, J. L., T. A. Caswell, and E. K. Laakso (2012): "The gendering of language: A comparison of gender equality in countries with gendered, natural gender, and genderless languages," *Sex Roles*, 66, 268–281.

Rains, S. A. (2013): "The nature of psychological reactance revisited: A meta-analytic review," *Human Communication Research*, 39, 47–73.

Rosenbaum, S. M., S. Billinger, and N. Stieglitz (2014): "Let's be honest: A review of experimental evidence of honesty and truth-telling," *Journal of Economic Psychology*, 45, 181–196.

Rubinstein, A. (2000): *Economics and language: Five essays*, Cambridge University Press.

Rudman, L. A. and P. Glick (2001): "Prescriptive gender stereotypes and backlash toward agentic women," *Journal of Social Issues*, 57, 743–762.

Santacreu-Vasut, E., O. Shenkar, and A. Shoham (2014): "Linguistic gender marking and its international business ramifications," *Journal of International Business Studies*, 45, 1170–1178.

Schuetze, C. F. (2020): "Can a bill have a gender? Feminine wording exposes a rift," https://www.nytimes.com/2020/10/15/world/europe/germany-gender-bill-language.html?searchResultPosition=6, Last accessed: 11/07/2022.

Sczesny, S., M. Formanowicz, and F. Moser (2016): "Can gender-fair language reduce gender stereotyping and discrimination?" *Frontiers in Psychology*, 7.

Selten, R. (1965): "Die Strategiemethode zur Erforschung des eingeschränkt rationalen Verhaltens im Rahmen eines Oligopolexperimentes," Seminar für Mathematische Wirtschaftsforschung und Ökonometrie.

Semin, G. R. (2013): "Language, culture, cognition: How do they intersect?" in *Understanding Culture*, Psychology Press, 265–276.

Stahlberg, D., F. Braun, L. Irmen, and S. Sczesny (2007): "Representation of the sexes in language," *Social Communication*, 163–187.

Stahlberg, D. and S. Sczesny (2001): "Effekte des generischen Maskulinums und alternativer Sprachformen auf den gedanklichen Einbezug von Frauen," *Psychologische Rundschau*, 52, 131–140.

Steele, C. M. (1997): "A threat in the air: How stereotypes shape intellectual identity and performance." *American Psychologist*, 52, 613–629.

Sutter, M., S. Angerer, D. Glätzle-Rützle, and P. Lergetporer (2015): "The effect of language on economic behavior: Experimental evidence from children's intertemporal choices," CESifo Working Paper.

Sutter, M. and D. Glätzle-Rützler (2015): "Gender differences in the willingness to compete emerge early in life and persist," *Management Science*, 61, 2339–2354.

Tversky, A. and D. Kahneman (1981): "The framing of decisions and the psychology of choice," *Science*, 211, 453–458.

——— (1989): "Rational choice and the framing of decisions," in *Multiple criteria decision making and risk analysis using microcomputers*, Springer, 81–126.

United Nations (2022a): "Background and purpose of gender-inclusive language in English," https://www.un.org/en/gender-inclusive-language/index.shtml, Last accessed: 10/06/2022.

——— (2022b): "Guidelines for gender-inclusive language in English," https://www.un.org/-en/genderinclusive-language/guidelines.shtml, Last accessed: 08/04/2022.

van der Velde, L., J. Tyrowicz, and J. Siwinska (2015): "Language and (the estimates of) the gender wage gap," *Economics Letters*, 136, 165–170.

Vervecken, D. and B. Hannover (2015): "Yes I can! Effects of gender fair job descriptions on children's perceptions of job status, job difficulty, and vocational self-efficacy," *Social Psychology*, 46, 76–92.

Völkening, L. (2022): "Und trotzdem denken die meisten an Männer," https://www.zeit.de/wissen-/202206/gendern-geschlechter-sprache-sprachbilder-neutralitaet?utm_referrer=https%3A%2F-%2Fwww.google.com%2F, Last accessed: 10/06/2022.

Wasserman, B. D. and A. J. Weseley (2009): "¿Qué? Quoi? Do languages with grammatical gender promote sexist attitudes?" *Sex Roles*, 61, 634–643.

Waters, M. (2021): "Where gender-neutral pronouns come from," https://www.theatlantic.com-/culture/archive/2021/06/gender-neutral-pronouns-arent-new/619092/, Last accessed: 10/06/2022.

Wu, A. H. (2018): "Gendered language on the economics job market rumors forum," *AEA Papers and Proceedings*, 108, 175–179.

A Interaction Terms and Effects

When analyzing models with binary dependent variables (*Sent any* in the dictator game and *Reciprocal* in the prisoner's dilemma) we resort to probit models. Since we use interaction terms in four of our five specifications, there is an important difference between interaction terms and effects, as pointed out by Ai and Norton (2003). To illustrate this here briefly and to explain how we report our results, let us start by considering a linear model.

$$\begin{aligned} \mathit{Reciprocal} = \beta_0 &+ \beta_1 Woman_i + \beta_2 Match_i + \beta_3 Inclusive_i \\ &+ \beta_4 Woman_i \times Match_i + \beta_5 Woman_i \times Inclusive_i \\ &+ \boldsymbol{\gamma} \boldsymbol{X}_i + \varepsilon_i \end{aligned}$$

where $\boldsymbol{X}$ is a vector of controls and γ a vector of coefficients of these controls. The interaction effect of our Woman treatment variation and our Match treatment variation in this model would be

$$\frac{\partial^2 \mathit{Reciprocal}}{\partial Woman_i \partial Match_i} = \beta_4.$$

Thus, the interaction effect would be identical to the interaction term.

This is different when our model is non-linear, like in our probit regressions.[23]

$$\begin{aligned} P(\mathit{Reciprocal} = 1) = \Phi(\beta_0 &+ \beta_1 Woman_i + \beta_2 Match_i + \beta_3 Inclusive_i \\ &+ \beta_4 Woman_i \times Match_i + \beta_5 Woman_i \times Inclusive_i \\ &+ \boldsymbol{\gamma} \boldsymbol{X}_i) \end{aligned}$$

The interaction effect is given by

$$\begin{aligned} \frac{\partial^2 P(\mathit{Reciprocal} = 1)}{\partial Woman_i \partial Match_i} = \phi'(\beta_0 &+ \beta_1 Woman_i + \beta_2 Match_i + \beta_3 Inclusive_i \\ &+ \beta_4 Woman_i \times Match_i + \beta_5 Woman_i \times Inclusive_i \\ &+ \boldsymbol{\gamma} \boldsymbol{X}_i)[\beta_1 + \beta_4 Match_i + \beta_5 Inclusive_i][\beta_2 + \beta_4 Woman_i] \\ +\phi(\beta_0 &+ \beta_1 Woman_i + \beta_2 Match_i + \beta_3 Inclusive_i \\ &+ \beta_4 Woman_i \times Match_i + \beta_5 Woman_i \times Inclusive_i \\ &+ \boldsymbol{\gamma} \boldsymbol{X}_i)\beta_4, \end{aligned} \tag{1}$$

where ϕ is the pdf associated with the cdf Φ. This expression firstly depends on participant i's characteristics. Secondly, in most cases, it will also not be equal to β_4. Thirdly, and most importantly, the estimator of this term has standard errors that differ from those of $\hat{\beta}_4$. Thus, in

[23] Traditionally, we would denote the left-hand side with $P(\mathit{Reciprocal}|Z_i)$ with Z_i being the complete vector of control variables, but we suppress the conditional statement for better representation.

these models, there is a difference between the interaction term and the interaction effect and in the inference, we can make use of it.

To recognize this in our analysis we carry out the following steps. We use the *inteff* routine in Stata (Norton et al., 2004). It calculates the z-scores of the above expression for each participant in the sample and provides us with a mean z-score for the two-sided hypothesis that the interaction effect is zero. We use the square of this test statistic to run a χ^2 test. We report instances of rejections at the 10%, 5%, and 1% level, respectively, using the subscript $\star$, $\star\,\star$, and $\star\,\star\,\star$ in our regression tables on the interaction *term*. For example, if the interaction term Woman×Match was 1.5 and it was significant at the 5% level, whereas the interaction effect was only significant at the 10% level, we would denote

$$\text{Woman} \times \text{Match} \quad 1.5^{**}_{\star} \; .$$

Note that this is an abuse of notation as the subscript refers to the statistical significance of the term in (1). We attach it to the interaction term as we expect the reader to search for information on the interaction of treatment variations there.

B Tables and Figures

	M-Match	W-Mismatch	M-Inclusive	W-Inclusive	M-Mismatch	W-Match	Total
Age	26.067	25.882	25.188	24.526	24.368	23.294	24.845
	(3.634)	(6.918)	(3.331)	(2.988)	(3.483)	(2.756)	(4.091)
Semester	9.067	8.000	8.188	8.263	6.526	7.118	7.816
	(4.464)	(5.244)	(4.943)	(3.649)	(4.858)	(2.913)	(4.379)
Business and	0.333	0.294	0.438	0.526	0.474	0.294	0.398
Economics	(0.488)	(0.470)	(0.512)	(0.513)	(0.513)	(0.470)	(0.492)
Education	0.333	0.471	0.188	0.263	0.263	0.412	0.320
	(0.488)	(0.514)	(0.403)	(0.452)	(0.452)	(0.507)	(0.469)
Other majors	0.333	0.235	0.375	0.211	0.263	0.294	0.282
	(0.488)	(0.437)	(0.500)	(0.419)	(0.452)	(0.470)	(0.452)
Observations	15	17	16	19	19	17	103

Note: Standard deviations in parentheses. The p-values for Kruskal-Wallis tests for variables in the order they appear in the table are 0.301, 0.504, 0.605, 0.514, and 0.905

Table 9: Means of key demographics across treatments.

Dep. Var.: Amount sent	(1)	(2)	(3)	(4)	(5)
Woman	1.556*	0.378	0.064	-0.940	-0.288
	(0.862)	(1.288)	(1.384)	(1.644)	(1.567)
Match	-1.699*	-3.077*	-3.082*	-3.514**	-2.634*
	(0.986)	(1.551)	(1.667)	(1.688)	(1.579)
Inclusive	-1.013	-1.461	-1.254	-1.831	-1.535
	(1.057)	(1.532)	(1.573)	(1.657)	(1.630)
Woman × Match		2.724	2.831	3.302	2.378
		(1.970)	(2.131)	(2.207)	(1.976)
Woman × Inclusive		0.978	1.307	1.201	0.821
		(2.123)	(2.130)	(2.111)	(2.027)
Age			0.048	0.070	0.053
			(0.071)	(0.071)	(0.067)
Semester			-0.108	-0.110	-0.077
			(0.109)	(0.108)	(0.109)
Business and Economics			-0.791	-0.849	-0.369
			(1.142)	(1.253)	(1.244)
Education			1.248	1.192	1.662
			(1.086)	(1.210)	(1.168)
Language attitude				0.825*	0.321
				(0.475)	(0.492)
Remembered formulations				0.095	-0.512
				(0.864)	(0.942)
Language comments				3.005	2.692
				(1.953)	(1.982)
Other comments				0.684	-0.016
				(1.771)	(1.488)
Instructions clear				0.035	-0.224
				(0.345)	(0.352)
Failed attempts$_{DG}$				-0.447	0.791
				(1.738)	(1.899)
Risk aversion					-0.028
					(0.191)
First-order belief$_{DG}$					0.057**
					(0.028)
Second-order belief$_{DG}$					0.001
					(0.027)
Constant	6.654***	7.211***	6.803***	3.531	3.504
	(0.794)	(0.957)	(2.330)	(3.238)	(2.996)
R^{2}	0.053	0.069	0.116	0.187	0.290
Observations	103	103	103	103	103

Robust standard errors in parentheses, * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 10: OLS regressions with the *Amount sent* in the dictator game as the dependent variable (complete table with all coefficients).

Dep. Var.: Amount sent	(1)	(2)	(3)	(4)	(5)
Woman	1.501*	0.624	0.545	-0.431	0.052
	(0.868)	(1.058)	(1.076)	(1.300)	(1.410)
Generic male	-0.927	-2.410*	-2.493*	-2.636*	-1.875
	(0.914)	(1.426)	(1.488)	(1.483)	(1.362)
Woman × Generic male		2.831	2.586	3.022*	2.321
		(1.819)	(1.856)	(1.794)	(1.704)
Age			0.047	0.060	0.043
			(0.070)	(0.073)	(0.069)
Semester			-0.116	-0.124	-0.087
			(0.107)	(0.108)	(0.109)
Business and Economics			-0.718	-0.806	-0.324
			(1.118)	(1.254)	(1.246)
Education			1.315	1.299	1.791
			(1.055)	(1.183)	(1.106)
Language attitude				0.739*	0.238
				(0.436)	(0.445)
Remembered formulations				0.095	-0.522
				(0.864)	(0.923)
Language comments				2.931	2.669
				(1.869)	(1.916)
Other comments				0.643	-0.022
				(1.768)	(1.488)
Instructions clear				0.111	-0.165
				(0.336)	(0.349)
Failed attempts$_{DG}$				-0.204	1.078
				(1.667)	(1.748)
Risk aversion					-0.004
					(0.192)
First-order belief$_{DG}$					0.059**
					(0.027)
Second-order belief$_{DG}$					-0.000
					(0.026)
Constant	6.098***	6.543***	6.244***	2.851	2.832
	(0.691)	(0.757)	(2.198)	(2.999)	(2.711)
R^{2}	0.038	0.060	0.109	0.173	0.281
Observations	103	103	103	103	103

Robust standard errors in parentheses, * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 11: OLS regressions with the *Amount sent* in the dictator game as the dependent variable when comparing the generic male with the "non-standard" female and gender-inclusive framed instructions (complete table with all coefficients).

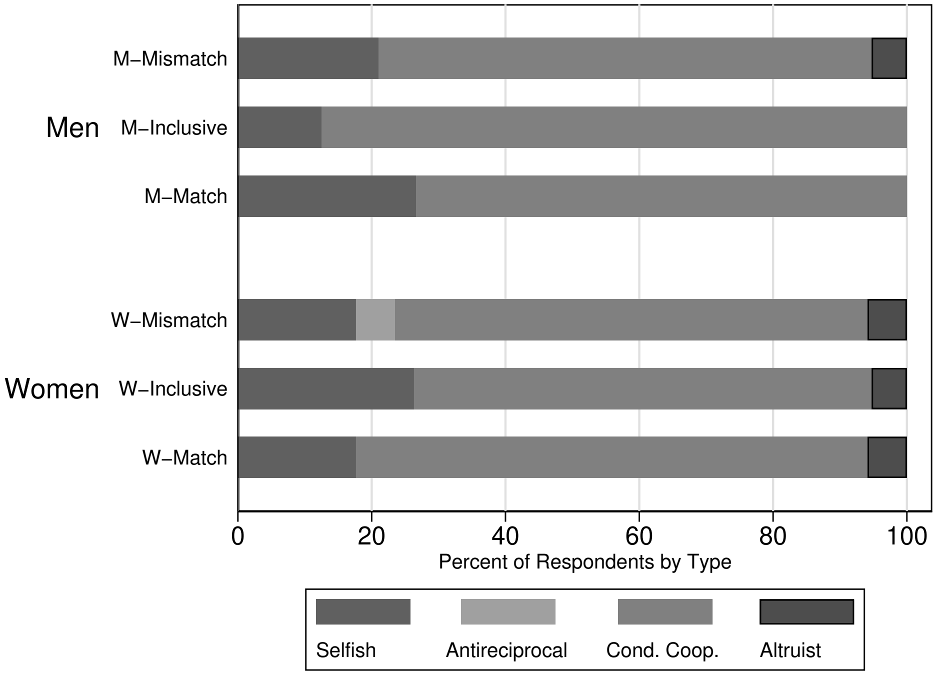
Dep. Var.: Sent any	(1)	(2)	(3)	(4)	(5)
Woman	1.011^{***}	4.614^{***}	4.679^{***}	4.505^{***}	5.281^{***}
	(0.303)	(0.329)	(0.374)	(0.509)	(0.759)
Match	-0.800^{**}	-0.721	-0.807	-0.763	-0.690
	(0.374)	(0.460)	(0.513)	(0.568)	(0.598)
Inclusive	-0.747^{**}	-0.486	-0.468	-0.698	-0.848
	(0.371)	(0.457)	(0.488)	(0.573)	(0.615)
Woman × Match		-3.511^{***}	-3.450^{***}	$-3.520^{***}_{\star\star}$	$-4.425^{***}_{\star\star\star}$
		(0.610)	(0.656)	(0.768)	(0.877)
Woman × Inclusive		$-3.930^{***}_{\star\star}$	$-3.905^{***}_{\star\star}$	$-4.123^{***}_{\star\star\star}$	$-4.876^{***}_{\star\star\star}$
		(0.577)	(0.605)	(0.664)	(0.915)
Age			0.054	0.074	0.076
			(0.053)	(0.053)	(0.049)
Semester			-0.035	-0.044	-0.028
			(0.041)	(0.044)	(0.045)
Business and Economics			-0.350	-0.463	-0.298
			(0.381)	(0.425)	(0.424)
Education			0.214	0.022	0.332
			(0.432)	(0.500)	(0.452)
Language attitude				0.399^{**}	0.248
				(0.159)	(0.186)
Remembered formulations				0.553	0.463
				(0.357)	(0.371)
Language comments				0.948	0.893
				(0.669)	(0.653)
Other comments				-0.285	-0.623
				(0.478)	(0.528)
Instructions clear				0.058	0.020
				(0.115)	(0.121)
Risk aversion					-0.093
					(0.086)
First-order belief$_{DG}$					0.010
					(0.009)
Second-order belief$_{DG}$					0.008
					(0.008)
Constant	0.921^{***}	0.805^{**}	-0.146	-2.302	-2.241
	(0.302)	(0.326)	(1.257)	(1.572)	(1.499)
Pseudo R^2	0.140	0.154	0.192	0.292	0.346
Observations	103	103	103	103	103

Robust standard errors in parentheses, * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

p-values for interaction *effects* based on Ai and Norton (2003), $_{\star}$ 0.10 $_{\star\star}$ 0.05 $_{\star\star\star}$ 0.01

Note: *Failed attempts*$_{DG}$ was excluded from the controls as it perfectly predicts the outcome.

Table 12: Probit regressions with the binary decision to send any positive amount (*Sent any*) in the dictator game as the dependent variable (complete table with all coefficients).



Note: Types are defined according to Miettinen et al. (2020).

Figure 6: Four types according to role B behavior in the prisoner's dilemma by treatment and self-reported gender.

Dep. Var.: Reciprocal	(1)	(2)	(3)	(4)	(5)
Woman	-0.072	0.077	0.152	-0.266	-1.331
	(0.284)	(0.497)	(0.539)	(0.601)	(0.824)
Match	-0.043	-0.142	-0.136	0.017	-0.565
	(0.353)	(0.481)	(0.504)	(0.547)	(0.608)
Inclusive	0.029	0.386	0.386	0.498	-0.037
	(0.345)	(0.522)	(0.530)	(0.565)	(0.634)
Woman × Match		0.187	0.023	0.060	1.084
		(0.708)	(0.770)	(0.822)	(0.945)
Woman × Inclusive		-0.638	-0.747	-0.804	-0.571
		(0.714)	(0.756)	(0.752)	(0.990)
Age			-0.050	-0.062*	-0.103**
			(0.034)	(0.037)	(0.040)
Semester			0.037	0.035	0.080*
			(0.035)	(0.036)	(0.044)
Business and Economics			-0.161	-0.261	-0.688
			(0.367)	(0.388)	(0.457)
Education			0.136	-0.127	-0.296
			(0.391)	(0.417)	(0.537)
Language attitude				0.309**	0.498**
				(0.139)	(0.194)
Remembered formulations				0.335	0.596
				(0.350)	(0.469)
Language comments				0.776	1.284
				(0.528)	(0.846)
Other comments				0.331	-0.406
				(0.570)	(0.914)
Instructions clear				0.261**	0.258**
				(0.115)	(0.127)
Failed attempts$_{PD}$				0.044	0.093***
				(0.031)	(0.034)
Risk aversion					-0.110
					(0.089)
Positive reciprocity					0.864***
					(0.326)
Negative reciprocity					-0.229*
					(0.138)
First-order belief$_{PD}$					0.030**
					(0.014)
Second-order belief$_{PD}$					0.004
					(0.014)
Constant	0.832***	0.765**	1.760*	-0.507	-5.460**
	(0.283)	(0.331)	(1.007)	(1.271)	(2.424)
Pseudo R^2	0.001	0.016	0.047	0.155	0.444
Observations	98	98	98	98	98

Robust standard errors in parentheses, * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

p-values for interaction *effects* based on Ai and Norton (2003), $_{\star}$ 0.10 $_{\star\star}$ 0.05 $_{\star\star\star}$ 0.01

Table 13: Probit regressions with *Reciprocal* as the dependent variable in the prisoner's dilemma (complete table with all coefficients).

Dep. Var.: Reciprocal	(1)	(2)	(3)	(4)	(5)
Woman	-0.072	-0.207	-0.263	-0.563	-1.270***
	(0.286)	(0.348)	(0.351)	(0.384)	(0.467)
Generic male	-0.094	-0.306	-0.305	-0.217	-0.531
	(0.307)	(0.432)	(0.443)	(0.480)	(0.532)
Woman × Generic male		0.426	0.551	0.353	0.620
		(0.617)	(0.642)	(0.713)	(0.879)
Age			-0.049	-0.060*	-0.094**
			(0.034)	(0.036)	(0.041)
Semester			0.038	0.037	0.064
			(0.035)	(0.034)	(0.042)
Business and Economics			-0.198	-0.294	-0.670
			(0.356)	(0.382)	(0.433)
Education			0.120	-0.104	-0.307
			(0.393)	(0.426)	(0.524)
Language attitude				0.285**	0.385**
				(0.133)	(0.195)
Remembered formulations				0.256	0.462
				(0.344)	(0.416)
Language comments				0.795	1.131
				(0.526)	(0.835)
Other comments				0.367	-0.204
				(0.584)	(0.827)
Instructions clear				0.258**	0.283**
				(0.112)	(0.124)
Failed attempts$_{PD}$				0.054*	0.092**
				(0.028)	(0.039)
Risk aversion					-0.131
					(0.083)
Positive reciprocity					0.767***
					(0.292)
Negative reciprocity					-0.258*
					(0.135)
First-order belief$_{PD}$					0.022*
					(0.013)
Second-order belief$_{PD}$					0.009
					(0.014)
Constant	0.858***	0.929***	1.911**	-0.220	-4.380*
	(0.222)	(0.254)	(0.969)	(1.216)	(2.291)
Pseudo R^2	0.002	0.006	0.039	0.143	0.421
Observations	98	98	98	98	98

Robust standard errors in parentheses, * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

p-values for interaction *effects* based on Ai and Norton (2003), $_{\star}$ 0.10 $_{\star\star}$ 0.05 $_{\star\star\star}$ 0.01

Table 14: Probit regressions with *Reciprocal* as the dependent variable in the prisoner's dilemma when comparing the generic male with the "non-standard" female and gender-inclusive framed instructions (complete table with all coefficients).

Dep. Var.: Share honest	(1)	(2)	(3)	(4)	(5)
Woman	0.055	-6.553	-6.957	-7.114	-11.405
	(7.542)	(12.966)	(12.895)	(14.926)	(14.511)
Match	0.981	-11.520	-13.662	-11.195	-13.166
	(9.625)	(14.146)	(14.602)	(14.534)	(11.773)
Inclusive	-6.684	-4.715	-6.718	-4.368	-9.295
	(8.814)	(12.086)	(12.322)	(12.082)	(10.376)
Woman × Match		24.266	23.867	19.635	18.365
		(19.205)	(19.587)	(20.848)	(18.841)
Woman × Inclusive		-2.767	-2.330	-6.287	9.862
		(17.670)	(17.915)	(17.839)	(17.450)
Age			-1.372	-1.348	-1.156
			(0.880)	(0.982)	(1.031)
Semester			1.486	1.550*	1.047
			(0.903)	(0.869)	(0.899)
Business and Economics			-7.509	-8.924	-16.699**
			(9.312)	(9.640)	(7.464)
Education			-5.101	-8.521	-7.400
			(9.712)	(10.673)	(8.770)
Language attitude				4.608	3.375
				(3.375)	(3.217)
Remembered formulations				2.149	3.270
				(8.754)	(7.426)
Language comments				16.263	7.575
				(9.943)	(9.376)
Other comments				9.559	10.983
				(15.785)	(12.429)
Instructions clear				4.885*	2.180
				(2.925)	(3.023)
Failed attempts$_{Dec}$				6.456	2.268
				(4.948)	(5.235)
Risk aversion					1.032
					(1.663)
First-order belief$_{Dec}$					0.569***
					(0.150)
Second-order belief$_{Dec}$					0.041
					(0.152)
Constant	66.178***	69.298***	97.930***	47.509	35.030
	(7.295)	(8.788)	(25.172)	(30.591)	(31.655)
R^{2}	0.008	0.033	0.076	0.145	0.361
Observations	103	103	103	103	103

Robust standard errors in parentheses, * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 15: OLS regressions with *Share honest* as the dependent variable in the deception game (complete table with all coefficients).

Dep. Var.: Share honest	(1)	(2)	(3)	(4)	(5)
Woman	-0.067	-2.328	-4.048	-5.293	-3.410
	(7.507)	(8.778)	(8.996)	(9.459)	(9.436)
Generic male	-5.544	-9.365	-10.377	-9.370	-8.599
	(8.405)	(12.528)	(12.937)	(12.741)	(9.345)
Woman × Generic male		7.295	10.430	10.556	5.625
		(16.926)	(17.055)	(17.280)	(14.018)
Age			-1.546*	-1.539	-1.222
			(0.828)	(0.938)	(0.998)
Semester			1.349	1.438	0.971
			(0.911)	(0.879)	(0.899)
Business and Economics			-9.119	-10.376	-16.672**
			(9.125)	(9.455)	(7.652)
Education			-4.512	-8.083	-6.830
			(9.682)	(10.618)	(8.656)
Language attitude				3.410	2.855
				(3.282)	(2.920)
Remembered formulations				1.647	3.199
				(8.761)	(7.425)
Language comments				16.009	7.390
				(10.061)	(9.136)
Other comments				11.002	11.246
				(15.969)	(12.379)
Instructions clear				5.854**	2.688
				(2.782)	(2.791)
Failed attempts$_{Dec}$				6.125	2.545
				(4.904)	(4.997)
Risk aversion					1.161
					(1.674)
First-order belief$_{Dec}$					0.573***
					(0.144)
Second-order belief$_{Dec}$					0.035
					(0.153)
Constant	65.997***	67.143***	100.773***	50.225*	30.637
	(5.646)	(6.047)	(23.303)	(29.867)	(30.047)
R^2	0.005	0.007	0.051	0.125	0.355
Observations	103	103	103	103	103

Robust standard errors in parentheses, * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 16: OLS regressions with *Share honest* as the dependent variable in the deception game when comparing the generic male with the "non-standard" female and gender-inclusive framed instructions (complete table with all coefficients).

C Norm Elicitation

We first provided a brief summary of each game. We used an elicitation of social appropriateness. Specifically, we phrased our norm elicitation around fair 50-50 sharing in the dictator game (giving 10 ECU from the 20 ECU endowment), unconditional cooperation in the prisoner's dilemma, and complete honesty (i.e., a true report for each possible outcome of the die roll) in the deception game. For example, in the dictator game, we displayed the following statement.

> **A participant in the role of participant A should make a decision about the division of the 20 ECU such that both participants receive the same share of the 20 ECU.**

Participants were then asked to rate whether they personally found this statement *rather appropriate* or *rather inappropriate* and whether they think society rates this statement as *rather appropriate* or *rather inappropriate*. We incentivized the latter question with 5 ECU if the participant's answer coincided with the modal answer of the other participants in that session. We elicited this measure after all games were played, following a brief summary of each game. We did the same for the prisoners dilemma and the deception game. One of the three prescriptive norm elicitations for the three games was chosen at random to add to payoffs. The random draws for the game payoffs, the belief payoffs, and the norm payoffs were independent.