CESifo WORKING PAPERS

10459
2023

May 2023

The Effects of Gendered Language on Norm Compliance

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. 10459*

The Effects of Gendered Language on Norm Compliance

Abstract

Social norms, though often implicit, are to a great extent communicated and made salient using natural language. They carry the notions that "the participant," "the customer," or "the worker" should behave in a certain way. In English, we refer to each of these personal entity nouns using the pronouns "he," "she," or the gender-inclusive singular "they." In languages with grammatical gender, the nouns and the grammatical structure they are embedded in mark them as either male, female, or gender-inclusive. Little is known as to whether the framing of norms with respect to these grammatical genders affects norm compliance. We conducted an experiment in German with three games commonly used to study fair sharing, cooperation, and honesty. Our treatments allowed us to compare the differences in the increase of norm compliance when introducing prescriptive norms depending on the match between the participant's self-reported gender and the gender frame used in the experimental instructions. Overall, we find no 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. We observed the strongest effect for men in the sharing game, where the data suggests that a match led to a higher increase in norm compliance compared to the increase if gender-inclusive formulations were used. This line of research has important implications for the effective communication of rules and norms in organizations and administrations.

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

Keywords: norm compliance, gender in language, social identity.

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, and the audience of the ASFEE Conference 2021 and the 2022 ESA World Meeting for helpful comments 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

People often behave in ways that are not necessarily in their own best material interest (Fehr and Gächter, 1998; DellaVigna, 2009; Thaler, 2016). Donors share what they have with others, customers on online shopping platforms give positive ratings to sellers to return the favor of having received a good rating themselves, and taxpayers report income sources the state is unlikely to uncover on its own (Andreoni, 1990; Bolton et al., 2013; Mascagni, 2018). Although multiple factors are involved, social norms are crucial to explain this behavior. They carry the notion of "what ought to be done." More formally, they can be defined as the conventions and informal rules that govern behavior in groups and societies (Bicchieri et al., 2018). As such, they are ubiquitous in everyday life and particularly govern social interactions when formal laws are unavailable or cannot even be formulated (Bicchieri et al., 2022b; Fallucchi and Nosenzo, 2022).

Though often implicit, social norms are largely communicated and made salient using natural language. Prescriptive norms impose how "the participant," "the customer," or "the worker" should behave. In languages with grammatical gender, nouns are assigned a gender category, either male, female, or gender-inclusive, which also pertains to the grammatical context in which they appear. In most languages, social norms codified into official rules and laws have been prescribed for male plaintiffs and defendants using masculine pronouns (he/him/his). This is similar to more implicit norms when they surface in the shape of sayings and idioms that typically star a male or contain male pronouns.[1] The usage is supposed to be generic because all these formulations apply to every person, irrespective of sex and gender. There is yet to be conclusive evidence whether people who do not identify as men actually perceive to be addressed appropriately. Thus, little is known as to whether the framing of norms regarding these grammatical genders affects norm compliance and whether gender differences in norm compliance can be explained by how norms are formulated. As both, social norms and notions of gender, are ubiquitous in natural language, it is important to improve our understanding of how the formulation of prescriptive norms affects norm compliance. Additionally, examining whether gender differences in norm compliance can be explained by how norms are formulated can help us design more effective interventions aimed at reducing social and economic gender inequalities.

Our study aims to shed light on the following question: Are participants more likely to increase

[1] Two such examples are "A liar will not be believed even when he speaks the truth." devaluing dishonesty or "Everything comes to him who waits" valuing patience. Even "Faint heart never won fair lady," valuing courage is written from a (heterosexual) male perspective.

their norm compliance if the prescriptive norm statement is made salient using a formulation with a grammatical gender that matches their self-reported gender? We present results from a controlled experiment allowing us to make causal claims about the impact of grammatical gender on norm compliance. We made prescriptive norms ("He/She/They should") salient either before or after participants made decisions in economic games. We varied the grammatical gender in which the norm statements and experimental instructions were formulated. For these prescriptive norm statements to affect individuals, participants must perceive some sense of belongingness (Cross and Madson, 1997; Baumeister and Sommer, 1997) with the relevant social group for which the statements reflect a *social* norm. Thus, we expect that when the participant's self-reported gender matches the formulation of the prescriptive norm statements and the experimental instructions, making the norms salient before decisions are being made has a larger effect on norm compliance than when the participant's self-reported gender does not match the formulation of the prescriptive norm statements and the experimental instructions. When gender-inclusive formulations (akin to the singular they) were used for the prescriptive norm statements and the experimental instructions, participants identifying as men or women were neither explicitly excluded nor exclusively addressed. Thus, we expect making the norm salient to have a larger effect on norm compliance under the gender-inclusive formulation than when there is an explicit mismatch but a lower effect than when there is an explicit match between the grammatical gender and the self-reported gender of participants.

We implemented three games measuring pro-social behavior; in particular, participants played a dictator game (Güth et al., 1982; Kahneman et al., 1986; Forsythe et al., 1994), a sequential prisoner's dilemma (Bolle and Ockenfels, 1990; Dufwenberg and Kirchsteiger, 2000), and a deception game (Gneezy et al., 2013), commonly used to study fair sharing, cooperation, and honesty. The norms for the dictator game, the prisoner's dilemma, and the deception game were a 50-50 sharing norm, a norm to cooperate, and an honesty norm. The experiment was conducted in German, and norms, as well as the experimental instructions, were either stated describing a (generic) male participant ("der Teilnehmer"), a female participant ("die Teilnehmerin"), or the participant was described in a gender-inclusive way ("der*die Teilnehmer*in").

Overall, we find no strong evidence that a match between the participant's self-reported gender and the prescriptive norm statements and the experimental instructions led to a higher increase in norm compliance compared to the differences in a mismatch or gender-inclusive frame. We observed the strongest effect for men in the dictator game. Here, the data suggested that making

the norm salient led to an increase in norm compliance if there was a match between the self-reported gender and the prescriptive norm statements and the experimental instructions, whereas there was no such increase if gender-inclusive formulations were used.

Our study relates to the literature on norms and the interaction of norms and gender.

Norms have been extensively studied across disciplines (Sherif, 1936; Durkheim, 1950; Akerlof, 1976; Posner, 2009; Bénabou and Tirole, 2006; Lane et al., forthcoming), and there is plenty of experimental research in economics (see, e.g., Fehr et al., 2002; Kessler and Leider, 2012; Krupka and Weber, 2013; Gächter et al., 2013; Bicchieri et al., 2022a).

Whereas many studies focus on the emergence and evolution of norms (Binmore and Samuelson, 1994; Sethi and Somanathan, 1996; Ostrom, 2000), others try to disentangle how much norms contribute to moral behavior relative to other behavioral explanations, such as social preferences (Krupka and Weber, 2009; Jakiela, 2011), social identity (Benjamin et al., 2010; Akerlof and Kranton, 2010; Bénabou and Tirole, 2011), or social status (Akerlof, 1997). Another strand of literature, closer to our research question, focuses on measuring norm compliance (Spitzer et al., 2007; Bicchieri et al., 2022a) and describing environments and conditions that help enforce compliance with norms (Bernhard et al., 2006; Goette et al., 2006; Balafoutas and Nikiforakis, 2012; d'Adda et al., 2020).

Across the different lines of economic research on norms, the norms that receive the most prominent focus in the literature are social (or interpersonal) norms.[2] As such, they are only valid within the social group holding the social norm, and individuals need to know that they are part of that social group. There are studies analyzing the relationship between gender and norm compliance and the perception of norms. Friedl et al. (2020) find culture-specific gender differences in social risk-taking. Boschini et al. (2011) study the existence of a cooperation norm and find that when men interact with other men they are less likely to uphold a cooperation norm compared to women, or men in gender-mixed groups. There are documented gender differences in the ratings of social appropriateness of dictator behavior with women rating an unfair decision less acceptable than men when there is no information provided on the dictator (Krysowski and Tremewan, 2021).

Our study is also related to work describing how norms and gender correlate or interact. The results are mixed, while most studies, which we will explain in more detail in the following, do find

[2]See Bašić and Verrina (2021) for a study eliciting personal norms.

an interaction between gender, norms, and economic behavior, others do not find an influence of gender on economic behavior (Fornwagner et al., 2022). Prominent examples can be found in the labor market, where it is the norm that women negotiate less fiercely over wages and promotions (Exley et al., 2020), and men are traditionally the breadwinners in the household (Gauri et al., 2019; Bursztyn et al., 2020). One measure that can increase female labor participation and thus break such norms, particularly in typically more male-dominated domains, is the wording and naming of job advertisements in gender-neutral ways, reducing signals of male dominance and reduced belongingness for females (Gaucher et al., 2011; Horvath and Sczesny, 2016; Hodel et al., 2017). There is a large strand of literature studying the impact of gender in language (see, e.g., Crawford and English, 1984; Vervecken and Hannover, 2015; Sczesny et al., 2016), suggesting that the usage of the generic male form makes gender stereotypes more salient and can result in a male bias in readers' associations and their recall of people in texts. 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. Closest to our paper is Gorny et al. (2023) who studied the impact of gender-framed instructions on sharing, reciprocal behavior, and honest reporting.

Our study contributes to the literature on how formulations of gender in language are perceived. The more recent common practice to state preferred pronouns and internal guidelines to use gender-inclusive language sometimes create backlash (Nöstlinger, 2021; Coleman, 2022; Gonzalez Camano and Brown, 2022). The proponents of gender-inclusive language argue that such use of language is a sound strategy to empower underprivileged groups or to include minorities. To our knowledge, our study is the first empirical investigation into how gender in language affects the compliance with norms, informing these claims. This line of research, therefore, has important implications for the effective communication of rules and norms in organizations and administrations.

The paper is structured as follows. In Section 2, we describe the experimental design and derive our hypotheses, followed by our data preparation and estimation strategy in Section 3. Section 4 contains the 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

We start by describing the treatment differences and the sequence of stages in the experiment. After, we go over the procedures of how we executed the experimental sessions. Finally, we derive our hypotheses using a simple notational framework.

2.1 Treatments and Stages

To study the impact of gender in language on norm compliance, we implemented a 2×3 design.[3] First, norm salience was varied by eliciting the social appropriateness of prescriptive norm statements either *before* each of the games (Norm) or *after* all games had been played (NoNorm). This way, participants in the Norm treatments had to deliberate on the content of the prescriptive norm statements and on whether others perceived the behavior prescribed in these statements as a social norm. In contrast, the participants in the NoNorm treatments could make their decisions without such deliberation. Second, we varied whether the prescriptive norm statements and the entire experimental instructions were written using the male, female, or gender-inclusive form. Throughout the instructions and across treatments, we described the rules of the experiment, referring to "a participant." In each treatment, this generic participant was described in one of three gender frames. These gender frames either matched the participants' self-reported gender (Match), did not match their self-reported gender (Mismatch), or an inclusive form was used (Inclusive).[4] The resulting 2×3 design is summarized in Table 1.

		Gender frame		
		Match	Inclusive	Mismatch
Norm salience	NoNorm	NoNorm-Match	NoNorm-Inclusive	NoNorm-Mismatch
	Norm	Norm-Match	Norm-Inclusive	Norm-Mismatch

Table 1: Treatments in the 2×3 Design.

To induce norm salience in the Norm treatments, we elicited the participants' assessment of the social appropriateness of the prescriptive norm statements. Recently, a large part of the literature employs the method for eliciting social norms described in Krupka and Weber (2013). In a coordination task, participants have to rate the social appropriateness of behavior according to how they

[3] Our experimental design was preregistered at aspredicted.org.

[4] More precisely, in the Mismatch treatment, neither did the gender frame and the self-reported gender of the participant match *nor* was the inclusive form used.

believe all other participants rate the behavior's social appropriateness. They are incentivized to provide a rating that coincides with the modal rating of the other participants in the experiment. Given that there is no interaction between the participants, this method is incentive-compatible, as misrepresenting beliefs leads to lower expected payoffs. Other studies have shown that this method is robust to various influences such as using visual labels with different focal points, induced through varying the relative size of the visual labels and heterogenous normative expectations opposed to salient focal points such as the 50-50 sharing norm (Nosenzo and Goerges, 2020; Fallucchi and Nosenzo, 2022). We used a modified version to make prescriptive norms salient.
In our experimental design, we established three types of norms studied in the literature; the 50-50 (or fair-sharing) norm (Andreoni and Bernheim, 2009; Gächter et al., 2017), a norm for cooperation (Fehr and Rockenbach, 2004; Fehr and Fischbacher, 2004a; Goette et al., 2006), and a norm for truth-telling or honesty (Abeler et al., 2019). The statements had the following form: **A participant in the role of participant A should make a decision such that X.**[5] The participants were then asked to rate whether they personally found this statement *rather appropriate* or *rather inappropriate* and if they thought that 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 response of the other participants in the respective session. In the Norm treatments, we elicited this measure after the instructions for each game and *immediately before* participants made their decisions. In the NoNorm treatments, we elicited this information *after all three games* had been played.

The experiment proceeded in three stages. In **Stage 1**, participants received general instructions for the experiment. They were informed about their participation in three distinct two-player games. We used the strategy method (Selten, 1965) to collect data from all participants in all games. The participants knew that one game would be chosen randomly to determine the payoff. Within that randomly chosen game, the role of each participant was also selected at random. We used the perfect stranger matching protocol, ensuring that participants would not interact with another participant more than once to prevent reciprocity and reputation effects. We also informed them about the experimental currency unit (ECU) and the exchange rate of 1 ECU = €0.40. Before proceeding to the next stage, participants answered control questions to ensure their understanding of the general setup. In **Stage 2**, participants played the following games: a dictator game (Güth et al., 1982; Kahneman et al., 1986; Forsythe et al., 1994), a sequential prisoner's dilemma

[5] All translated statements in English can be found in the appendix, and the original statements in German can be found in the online appendix, together with the experimental instructions.

(Bolle and Ockenfels, 1990; Dufwenberg and Kirchsteiger, 2000), and a deception game (Gneezy et al., 2013).[6] In **Stage 3**, we elicited a range of behavioral measures and survey items, such as demographic information and attitudes toward language change.[7]

All treatments encompassed the three games mentioned earlier. We describe them in more detail in the following.

In the dictator game (Güth et al., 1982; Kahneman et al., 1986; Forsythe et al., 1994), each participant played the role of player A first. Player A had to divide 20 ECU between themselves and player B. Player A could choose any integer between 0 and 20. Player B was passive and could not make any decisions. In the dictator game, the prescriptive norm statement displayed to participants–at the end of the experiment (NoNorm treatments) or before they made their decision (Norm treatments)–in the role of player A read "A participant in the role of Participant A should make a decision on the allocation of the 20 ECU, in which both participants receive an equal share of the total 20 ECU."

In the sequential prisoner's dilemma (Bolle and Ockenfels, 1990; Dufwenberg and Kirchsteiger, 2000), participants first played the role of player A and then the role of player B. In the role of player A, they had to decide whether or not to send 8 of their 10 ECU to player B. We will refer to this as the *unconditional choice*. If player A sent the 8 ECU, the amount was doubled, thus adding 16 ECU to whatever player B kept. The game is also depicted in the game tree in Figure 1. 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. Player B could also either send 8 ECU to player A, which were doubled, or keep the endowment of 10 ECU. We will refer to this as the *conditional choice*. If this game had been selected to determine the payoff, the decision of player B was matched with the actual choice of player A to calculate the payoff for both players. Each player's role was determined using a random draw with equal probabilities.

In the prisoner's dilemma, the prescriptive norm statement, displayed to participants–at the end of the experiment (NoNorm treatments) or before they made their decision (Norm treatments)–read "A participant in the role of Participant A should make a decision in which he sends 8 of his 10 ECU to Participant B."[8]

[6] As we are interested in between-subject differences, we kept the order constant for all participants.
[7] The description of these measures can be found in Section 3.
[8] In this form ("he," "his") it was displayed to men in the Match treatments and women in the Mismatch treatments.

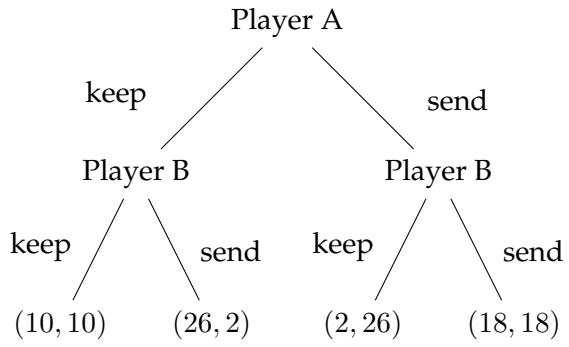


Figure 1: The sequential prisoner's dilemma.

As the third game, we implemented the deception game described in Gneezy et al. (2013). Again, each participant had to play both roles. For each possible roll of a six-sided die, player A had to decide which message $m \in \{1, 2, 3, 4, 5, 6\}$ to send to player B. The payoff for player A was given by $\pi(m) = 10 + 2m$. If the game was chosen to be payoff-relevant, player A would learn the die roll outcome together with the payoff information at the end of the experiment. Player B was asked to decide whether to follow the message or not for every possible message sent by player A. Player B's payoff was 10 ECU if player B followed the message of player A and the message was honest. If player B followed the message and the message was not honest, player B received 0 ECU. If player B decided against following the message of player A, player B received 3 ECU, irrespective of the true outcome and the message. If the game was chosen to be payoff-relevant, player B would get to know whether the message of player A for the drawn die roll outcome was honest if player B followed together with the payoff information at the end of the experiment. In the deception game, the prescriptive norm statement displayed to participants–at the end of the experiment (NoNorm treatments) or before they made their decision (Norm treatments)–read "A participant in the role of Participant A should compose a message to Participant B that contains the actually assigned number."

After all three games had been played, we elicited the perceived appropriateness of the prescriptive norm statements for participants in the NoNorm treatments and the beliefs about actual behavior in all treatments. At the end of the experiment, one of the belief elicitations and one of the prescriptive norm elicitations for the three games were chosen randomly to add to payoffs. These random draws were independent of each other and independent of the game chosen to be payoff-relevant to avoid participants balancing their expected payments across norm elicitations or between game decisions and norm elicitations. All participants had to answer a brief survey containing questions on reciprocity (Dohmen et al., 2009), risk aversion (Dohmen et al., 2011; Kan-

tar Public, 2020), moral values (Haerpfer et al., 2020), and questions regarding the comprehension of the instructions, attitude toward language. We also collected demographic information (age, gender, study degree, field of study, and past participation in experiments) and comments on the experiment. Lastly, we asked for the participant's recall of the gender frame used throughout the instructions.

2.2 Procedures

We ran the above design as a controlled online experiment on the German-speaking laboratory participant pool of a large German university.[9] Using ORSEE (Greiner, 2015), we invited the same number of female and male participants according to their selected gender upon registration in the participant pool. To assess the correct registration for the respective session and to allow participants to ask clarifying questions, the experiment was accompanied by a virtual meeting in a conferencing tool. Participants and experimenters were muted, their video feeds were disabled, and the lab rules were shown as screen-share throughout the session. Thus, communication was limited to text chat, and verbal communication was not used, unless urgently necessary, e.g., if a participant went idle for longer than five minutes. Participants received personalized links to the experimental software, which was programmed in oTree (Chen et al., 2016). A typical session lasted around 50 minutes, and participants earned €9.33 on average, including a show-up fee of €2.50.

2.3 Hypotheses

With the data from our experiment, we aim to answer the following question: Are participants more likely to increase their norm compliance if the prescriptive norm statement is made salient using a formulation with a grammatical gender that matches their self-reported gender?

As we have argued earlier, we expect participants to have different feelings of belongingness, depending on how their self-reported gender, reflecting their gender identity (Akerlof and Kranton, 2000), compares to the grammatical gender used in the different frames (Cross and Madson, 1997; Baumeister and Sommer, 1997). In the words of Akerlof and Kranton (2000), self-reported gender is a social category to which individuals belong. These social categories already encompass their

[9] The NoNorm treatments used as a baseline in this paper are the core treatments in Gorny et al. (2023).

own behavioral norms or, as Akerlof and Kranton (2000) in fact call them, "behavioral prescriptions." Whereas our prescriptive norm statements, by definition, make the behavioral norms that they prescribe salient, our variations in the gender frame potentially affect the salience of these social categories (Gorny et al., 2023). To find out what effect our norm salience treatment variation has on norm compliance, we need to compare norm compliance in the Norm treatments to a baseline that reflects the same identity prescriptions stemming from the surrounding instructions. Thus, we need a framework in which we compare the effects of our norm salience variation across the different formulations of the prescriptive norm statements and the experimental instructions.

Based on these considerations, we derive a simple and testable framework. We denote an individual's norm compliance in the norm salience condition $T \in \{NoNorm, Norm\}$ under formulation $F \in \{Match, Inclusive, Mismatch\}$ with $NC(T|F)$. We can write

$$\Delta NC_F = NC(Norm|F) - NC(NoNorm|F).$$

This difference represents the effect of our norm salience variation, holding the formulation of the prescriptive norm statements constant. We expect that when the participant's self-reported gender matches the formulation of the prescriptive norm statements, the increase of norm salience has a larger effect on norm compliance than when there is a mismatch. This translates to

$$\Delta NC_{Match} > \Delta NC_{Mismatch}.$$

When gender-inclusive formulations were used for the prescriptive norm statements, participants identifying as men or women were neither explicitly excluded nor exclusively addressed. Thus, we expect the variation of norm salience to have a larger effect on norm compliance under the gender-inclusive formulation than when there is a mismatch but a lower effect than when there is a match between the grammatical gender of the prescriptive norm statements and the self-reported gender of participants. This translates to the following three-way inequality summarizing our hypotheses.

$$\Delta NC_{Match} > \Delta NC_{Inclusive} > \Delta NC_{Mismatch} \tag{1}$$

Across games, we expect the increase in norm salience to result in the highest increase in the number of games in which participants comply with the norm in the Match frame. We expect

the increase in norm salience to result in the lowest increase in the number of games in which participants comply with the norm in the Mismatch frame. We expect the increase in norm salience to result in an increase in the number of games in which participants comply with the norm lying between these two increases in the Inclusive frame.

Hypothesis 1 (Norm compliance across games)**.**
Overall norm compliance increases the most if the participant's self-reported gender matches the formulation of the prescriptive norm statements and the least if there is a mismatch. When gender-inclusive formulations are used, the increase in norm compliance lies between these two extremes.

As we also investigate the behavior in the individual games, we also state our hypotheses in terms of these games.

In the dictator game, we expect to observe the highest increase in the share of participants complying with the 50-50 sharing norm in the Match frame, the lowest increase in compliance with the 50-50 sharing norm in the Mismatch frame, and an increase in compliance with the 50-50 sharing norm lying between these two increases in the Inclusive frame.

Hypothesis 2 (Compliance with the 50-50 sharing norm in the dictator game)**.**
Compliance with the 50-50 sharing norm in the dictator game increases the most if the participant's self-reported gender matches the formulation of the prescriptive norm statements and the least if there is a mismatch. When gender-inclusive formulations are used, the increase in norm compliance lies between these two extremes.

In the prisoner's dilemma, we expect to observe the highest increase in the share of participants complying with the cooperation norm in the Match frame, the lowest increase in compliance with the cooperation norm in the Mismatch frame, and an increase in compliance with the cooperation norm lying between these two increases in the Inclusive frame.

Hypothesis 3 (Compliance with the cooperation norm in the prisoner's dilemma)**.**
Compliance with the cooperation norm in the prisoner's dilemma increases the most if the participant's self-reported gender matches the formulation of the prescriptive norm statements and the least if there is a mismatch. When gender-inclusive formulations are used, the increase in norm compliance lies between these two extremes.

In the deception game, we expect to observe the highest increase in the share of participants complying with the honesty norm in the Match frame, the lowest increase in compliance with the

honesty norm in the Mismatch frame, and an increase in compliance with the honesty norm lying between these two increases in the Inclusive frame.

Hypothesis 4 (Compliance with the honesty norm in the deception game)**.**
Compliance with the honesty norm in the deception game increases the most if the participant's self-reported gender matches the formulation of the prescriptive norm statements and the least if there is a mismatch. When gender-inclusive formulations are used, the increase in norm compliance lies between these two extremes.

Our hypotheses compare the effects of the norm treatment variation across the three gender frames. Thus, beyond the pure comparison of treatments, we need regressions with interaction terms to test these hypotheses. We describe the relevant variables and the empirical strategy that maps to the above framework and hypotheses in the following section.

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 preregistered exclusion criteria for our sample, which we also discuss here.

3.1 Variables of Interest

The key variable of interest is the participants' norm compliance, i.e., if the participants' behavior is identical to the behavior described in the prescriptive norm statements. For each game, we study whether or not participants complied with the behavior prescribed in statements on 50-50 sharing in the dictator game (DG), cooperation in the prisoner's dilemma (PD), and honesty in the deception game (Dec). We define the dummy variable $\textit{Compliance}_G$ equal to one if the participant behaved compliant with the prescribed behavior in game $G \in \{DG, PD, Dec\}$ and zero otherwise. A participant in the role of the dictator behaved norm-compliant ($\textit{Compliance}_{DG} = 1$) if they sent 10 ECU. In the prisoner's dilemma, norm compliance ($\textit{Compliance}_{PD} = 1$) means that a participant in the role of player A chose to send 8 of their 10 ECU. In the deception game, norm compliance ($\textit{Compliance}_{Dec} = 1$) means sending truthful reports for all possible die-roll outcomes. Thus, whenever we use the term norm compliance, we refer to the actual behavior of participants in the role of player A in the games relative to the behavior described in the prescriptive norm

statements.[10] When we analyze behavior across games, we sum up these dummies to obtain $Compliance_{all}$–the number of games in which a participant behaved norm compliantly–ranging from zero to three.

Given that we are interested in studying the impact of gender in language on norm compliance, we need to take the self-reported gender of the participants into account. To control for potential interactions between the self-reported gender and the gender frame used in the instructions, we define three indicator variables relating to the participants' self-reported gender and the gender frame used in the prescriptive norm statements and throughout the experiment. 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. For the remainder of the paper, 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.[11] The variable *Match* is one if a participant's self-reported gender and the gender frame used in the instructions were identical. Thus, women in the NoNorm-Match and Norm-Match treatments saw the prescriptive norm statements in the female gender frame. In contrast, men in the NoNorm-Match and Norm-Match treatments saw the prescriptive norm statements in the male gender frame. 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 (NoNorm-Inclusive and Norm-Inclusive).

3.2 Empirical Strategy

Our 2×3 design allows us to disentangle the impact of the Norm treatments and the gender frame on norm compliance. First, in order to investigate the pure effect of the gender frame in the NoNorm and Norm treatments, we applied a conservative non-parametric approach and compared the results across treatments using two-sided Jonckheere-Terpstra tests. Given that we are particularly interested in the interaction between providing a norm statement and whether the gender frame matched the participant's self-reported gender, we need to estimate regression models, including interaction terms.

[10] These norms are highly focal and are predominant in the games we use (Krupka and Weber, 2013; Fehr and Fischbacher, 2004b; Rosenbaum et al., 2014). Thus, we refer to compliance with the behavior described in the prescriptive norm statements as norm compliance in all treatments, even though these statements were only shown to participants in the NoNorm treatments *after* the games were already played.

[11] One participant self-reported to be non-binary and was excluded from the dataset as described below.

For each regression table, we report five specifications that, in a stepwise procedure, include more variables and controls. In all regressions in our results section, *Compliance*$_{all}$ and *Compliance*$_G$ are the dependent variables. We rely on the variables *Norm*, *Match*, and *Inclusive*. In the first specification, we only include these variables. In the second step, we add the interactions between *Norm* and *Match* and between *Norm* and *Inclusive* as independent variables to test our hypotheses. For participant i and abstracting from a specific game, this model can be written as

$$\begin{aligned} Compliance_i(Norm_i, Match_i, Inclusive_i) =& \beta_0 + \beta_1 Norm_i + \beta_2 Match_i + \beta_3 Inclusive_i + \\ & \beta_4 Norm_i \times Match_i + \beta_5 Norm_i \times Inclusive_i + \varepsilon_i. \end{aligned}$$

Remember that our hypotheses can be summarized by the three-way inequality (1). The quantity ΔNC_{Match} from our conceptual framework is estimated by the difference between

$$Compliance_i(Norm = 1, Match = 1, Inclusive_i = 0) = \beta_1 + \beta_2 + \beta_4$$

and

$$Compliance_i(Norm = 0, Match = 1, Inclusive_i = 0) = \beta_2.$$

Thus, $\beta_1 + \beta_4$ provides us with an estimate of the difference between the increase in norm compliance due to making the norm salient in the Match frame. In other words, it is given by subtracting the coefficient of $Match$ (for the NoNorm-Match treatment) from the sum of the coefficients of $Norm$, $Match$, and the interaction between $Norm$ and $Match$. Making the norm salient increased norm compliance under the Match gender frame if the resulting linear term ($Norm + Norm \times Match$) is statistically significantly larger than zero.

Similarly, we can estimate $\Delta NC_{Inclusive}$ as $\beta_1 + \beta_5$ and, because the Mismatch treatments are our statistical baseline, $\Delta NC_{Mismatch}$ as β_1. The increase in norm salience due to our prescriptive norm statements increased norm compliance under the Inclusive gender frame if the resulting linear term ($Norm + Norm \times Inclusive$) is statistically significantly larger than zero. Similarly, the increase in norm salience due to our prescriptive norm statements increased norm compliance under the Mismatch gender frame if the coefficient of $Norm$ is statistically significantly larger than zero.

Since β_1 appears in all these estimates, Inequality (1) is equivalent to testing $\beta_5 > 0$, $\beta_4 > \beta_5$, and

$\beta_4 > 0$.[12] Thus, we interpret coefficients of the interaction terms that are significantly larger than zero as *direct support* for our hypotheses. We also interpret a significantly larger interaction term between *Norm* and *Match* than the interaction term between *Norm* and *Inclusive* as direct support for our hypotheses. If making the norm salient increased norm compliance under one gender frame but not under another, which is ranked lower in terms of effect sizes as per Inequality (1), we interpret this as *indirect support* for our hypotheses.

For the remaining specifications, we included control variables to determine the robustness of our estimations. In the third specification, we included demographics. Then, in specification four, we included controls for language and understanding. In the last and fifth step, we added various controls for attitudes and beliefs to show the robustness of our findings.[13]

As one of our treatment factors depends on the participants' self-reported gender and we already study interaction effects between those and our norm salience variation, we analyzed our data for men and women separately.[14]

Our dependent variable $Compliance_G$ is binary if we analyze each game separately. Thus, we applied Probit regressions. When we study the behavior across games, we used $Compliance_{all}$ which ranges from zero to three. We, thus, needed to estimate Poisson regression models when we analyzed norm compliance across the three games. As all our models are non-linear and our main interest is in the interaction terms Norm×Match and Norm×Inclusive, 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 statistical significance of the *interaction effect* as opposed to the statistical significance of the *interaction term*, which is indicated by superscript asterisks (*).[15]

3.3 Sample Selection

In total, we gathered data from 294 participants. We excluded 24 participants who failed the attention check in our post-experimental survey. A single participant self-reported to be non-

[12] The inequality for the last test is implied by the two preceding inequalities. We report the corresponding test throughout our analyses nonetheless for completeness.

[13] The description of the controls can be found in Appendix C

[14] The regressions using the full sample controlling for and interacting all treatment dummies and interactions with *Woman* can be found in the online appendix.

[15] 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).

binary and was excluded from the dataset.[16] This leaves us with a sample of 269 observations, which we refer to as the *raw sample*.

Given that we focus on studying norm compliance, it is important to measure the effect of our treatment manipulations if the norms were actually *social* norms to the participants. As such, for the main part of our analysis and in line with our preregistration, we only included those participants from the Norm treatments who rated the prescriptive norm statements as "rather appropriate" to society.[17] We used the ratings of the prescriptive norm statements that we elicited immediately before decisions were made in these treatments to define $\textit{Appropriateness}_G$ for each game $G \in \{DG, PD, Dec\}$. These dummy variables are one if a participant rendered the behavior described in the prescriptive norm statement relating to game G as "rather appropriate" to society and zero otherwise. $\textit{Appropriateness}_{all}$ is one if all $\textit{Appropriateness}_G$ dummies are one and zero otherwise. Note that in the Norm treatments, we elicited the ratings of the prescriptive norm statements *before* the participants made their decisions. For the Norm treatments, we excluded all participants who did not rate the respective norm as "rather appropriate" to society. In contrast, the norm rating was elicited *after* the three games in the NoNorm treatments. The answers in the NoNorm treatments might depend on the previous behavior and serve as a justification. Thus, they have to be treated with caution. We, therefore, did not exclude any participants from the NoNorm treatments leading to 103 observations for the NoNorm treatments. For the Norm treatments, the number of observations varies. Across games, we have 83, for the dictator game 139, for the prisoners dilemma 108, and for the Deception game 139 observations. We refer to our restricted sample as the *analytical sample*. The analytical sample consists of 186 observations across games, 242 in the dictator game, 211 in the prisoners dilemma, and 242 in the deception game.[18]

[16] The exact question we asked was "Which gender do you sort yourself into?" (German "Welchem Geschlecht ordnen Sie sich zu?") with the options "Männlich" ("Male") "Weiblich" ("Female") "Divers" ("Diverse," i.e. non-binary).

[17] In our preregistration we stated: "For each game, we exclude participants from the norm treatments (norm=1) from the analysis who deemed the corresponding norm inappropriate, as this means that the norm induction failed for these participants."

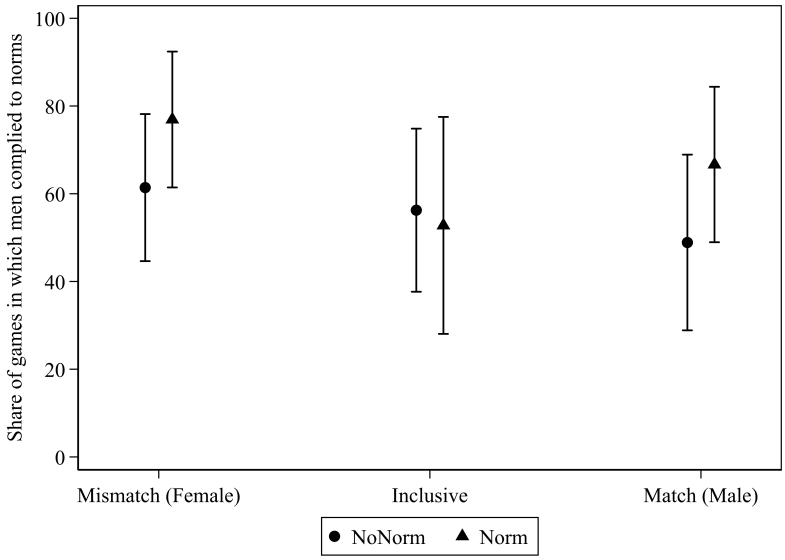
[18] Across our treatments, self-reported gender was balanced in our analytical sample. The share of women ranged from 47.22% to 60.00%, and each bilateral comparison of the shares between treatments was statistically insignificant (smallest p-value $p = 0.277$, Fisher's exact test). See Table 11 in Appendix D for more detailed summary statistics.

4 Results

4.1 Norm Compliance Across Games

To check if the introduction of prescriptive norm statements affected norm compliance, we briefly compare norm compliance between all NoNorm and all Norm treatments. In all NoNorm treatments, participants, on average, complied with the norm in 56.31% of the three games. In the Norm treatments, the participants in our analytical sample complied with the respective norms in on average 69.48% of the three games. The difference of 13.17 percentage points between these two averages is statistically significant at the 5%-level using a Mann-Whitney-U test ($p = 0.021$). Summing up, when aggregating men and women in our analytical sample, we observed significantly higher norm compliance in the Norm treatments compared to the NoNorm treatments.

Next, we study if differences in norm compliance depended on the gender frame of the experimental instructions and prescriptive norm statements. We start by analyzing men's norm compliance.



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

Figure 2: Men only–Difference in overall norm compliance across matching, inclusive, and mismatching prescriptive norm statements.

Consider Figure 2. In the NoNorm-Mismatch treatment, men, on average, complied with 61.40% of the norms, in the NoNorm-Inclusive treatment with 56.25% of the norms, and in the NoNorm-Match treatment, they complied with 48.89% of the norms. In the Norm-Mismatch treatment, men

complied with 76.92% of the norms, in the Norm-Inclusive treatment with 52.78% of the norms, and in the Norm-Match treatment, they complied with 66.67% of the norms. At first sight, the norm compliance is highest in the Mismatch and lower in the Inclusive and the Match treatment manipulations for both the Norm and the NoNorm treatments. However, we do not find statistical support for the observation that the variation of gender frames itself led to differences in overall norm compliance ($p = 0.346$ across the NoNorm treatments; $p = 0.572$ across the Norm treatments, Jonckheere-Terpstra tests).

Dep. Var.: Compliance$_{all}$	(1)	(2)	(3)	(4)	(5)
Norm	0.173	0.225	0.205	0.180	-0.272*
	(0.128)	(0.169)	(0.167)	(0.181)	(0.148)
Match	-0.171	-0.228	-0.229	-0.254	-0.484***
	(0.143)	(0.245)	(0.248)	(0.252)	(0.178)
Inclusive	-0.216	-0.088	-0.107	-0.165	-0.354**
	(0.162)	(0.213)	(0.216)	(0.221)	(0.168)
Norm × Match		0.085	0.038	0.050	0.330
		(0.295)	(0.320)	(0.318)	(0.226)
Norm × Inclusive		-0.289	-0.262	-0.222	0.210
		(0.329)	(0.327)	(0.316)	(0.228)
Constant	0.635***	0.611***	1.524***	1.045*	-0.662
	(0.112)	(0.136)	(0.558)	(0.618)	(0.875)
Pseudo R^2	0.009	0.012	0.021	0.038	0.149
Observations	92	92	92	92	92
Demographics	✗	✗	✓	✓	✓
Language & Understanding	✗	✗	✗	✓	✓
Attitudes & Beliefs	✗	✗	✗	✗	✓

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: For the complete table with all coefficients, see Tables 12a and 12b in Appendix D.

Table 2: Poisson regressions on the number of games in which men complied with the respective norm.

To study the interaction between norm salience and the gender frames, we need to look at the Poisson regressions reported in Table 2.[19] We find no support for Hypothesis 1 because the interaction terms between *Norm* and *Match* as well as *Norm* and *Inclusive* are not statistically significant. Also, their difference is not statistically significant ($p = 0.642$, Wald test).

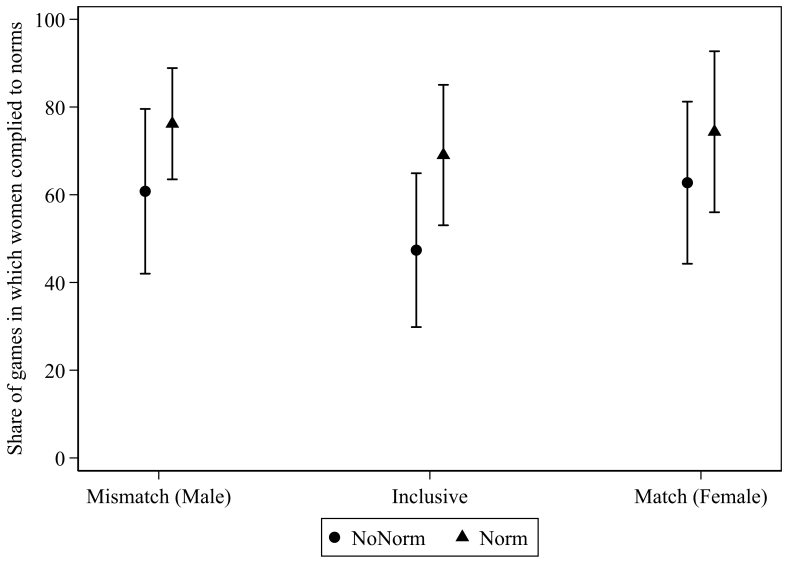
Result 1.1. (Men: Norm compliance across games)
We find no direct support for Hypothesis 1 that men's overall norm compliance increases the most if their self-reported gender matches the formulation of the prescriptive norm statements and the least if there is a mismatch. We do not find direct support that when gender-inclusive formulations

[19]The results are robust to using OLS or ordered probit regressions with robust standard errors.

are used, the increase in norm compliance lies between these two extremes.

In order to investigate if there is indirect support for Hypothesis 1, we investigate if making the norms salient increased norm compliance under one gender frame but not under another, which is ranked lower in terms of effect sizes as per Inequality (1). The coefficient of *Norm* refers to the comparison of the NoNorm-Mismatch and the Norm-Mismatch treatment. Based on the averages, one might expect a (potentially significant) difference indicating higher norm compliance in the Norm-Mismatch treatment compared to the NoNorm-Mismatch treatment. However, the coefficient is not statistically significant in the first four specifications. When controlling for beliefs and attitudes, the coefficient even turns negative and gets marginally statistically significant. Thus, we need to interpret this coefficient with caution. Comparing the NoNorm-Inclusive to the Norm-Inclusive treatment decreased the linear index of norm compliance by $|-0.272+0.210| = |-0.062| = 0.062$, but this decrease is not statistically significant ($p = 0.766$, Wald-test). When comparing the NoNorm-Match to the Norm-Match treatment the linear index of norm compliance increased by $-0.272 + 0.330 = 0.058$, but this increase is not statistically significant ($p = 0.739$, Wald-test). In addition, we observe that the coefficients for *Match* and *Inclusive* are negative and statistically significant in the last specification. Thus, controlling for the men's attitudes and beliefs, their overall norm compliance was significantly higher in the NoNorm-Mismatch treatment compared to the NoNorm-Match and the NoNorm-Inclusive treatment. Therefore, we do not find indirect support for Hypothesis 1.

Next, we analyze women's norm compliance across games. Consider Figure 3. In the NoNorm-Mismatch treatment, women complied with 60.78% of the norms. In the NoNorm-Inclusive treatment, women complied with 47.37% of the norms. In the NoNorm-Match treatment, women complied with 62.75% of the norms. In the Norm-Mismatch treatment, women complied with 76.19% of the norms. In the Norm-Inclusive treatment, women complied with 69.05% of the norms. In the Norm-Match treatment, women complied with 74.36% of the norms. The pattern looks similar to the men's behavior when comparing norm compliance in the Norm treatments. In the NoNorm treatments, norm compliance was lowest in the Inclusive frame. We do not observe a systematic variation when moving from the Mismatch over the Inclusive to the Match gender frame ($p = 0.892$ across NoNorm treatments; $p = 0.918$ across Norm treatments, Jonckheere-Terpstra tests). Consider Table 3 for the interaction terms. Again, we find no direct support for Hypothesis 1 because the interaction terms between *Norm* and *Match* as well as *Norm* and *Inclusive* are not statistically significant. Also, their difference is not statistically significant ($p = 0.622$, Wald test).



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

Figure 3: Women only–Difference in overall norm compliance across matching, inclusive, and mismatching prescriptive norm statements.

Dep. Var.: Compliance$_{all}$	(1)	(2)	(3)	(4)	(5)
Norm	0.254**	0.226	0.232	0.252	0.293*
	(0.112)	(0.174)	(0.186)	(0.184)	(0.162)
Match	0.005	0.032	0.011	0.047	0.131
	(0.130)	(0.212)	(0.227)	(0.224)	(0.192)
Inclusive	-0.173	-0.249	-0.213	-0.244	-0.024
	(0.138)	(0.240)	(0.245)	(0.240)	(0.242)
Norm × Match		-0.056	-0.009	-0.093	-0.249
		(0.258)	(0.275)	(0.277)	(0.245)
Norm × Inclusive		0.151	0.139	0.158	-0.394
		(0.279)	(0.277)	(0.279)	(0.323)
Constant	0.586***	0.601***	0.892**	0.112	-0.543
	(0.118)	(0.154)	(0.430)	(0.569)	(0.833)
Pseudo R^2	0.014	0.015	0.026	0.034	0.086
Observations	94	94	94	94	94
Demographics	✗	✗	✓	✓	✓
Language & Understanding	✗	✗	✗	✓	✓
Attitudes & Beliefs	✗	✗	✗	✗	✓

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: For the complete table with all coefficients, see Tables 13a and 13b in Appendix D.

Table 3: Poisson regressions on the number of games in which women complied with the respective norm.

Result 1.2. (Women: Norm compliance across games)
We find no direct support for Hypothesis 1 that women's overall norm compliance increases the most if their self-reported gender matches the formulation of the prescriptive norm statements and the least if there is a mismatch. We do not find direct support that when gender-inclusive formulations are used, the increase in norm compliance lies between these two extremes.

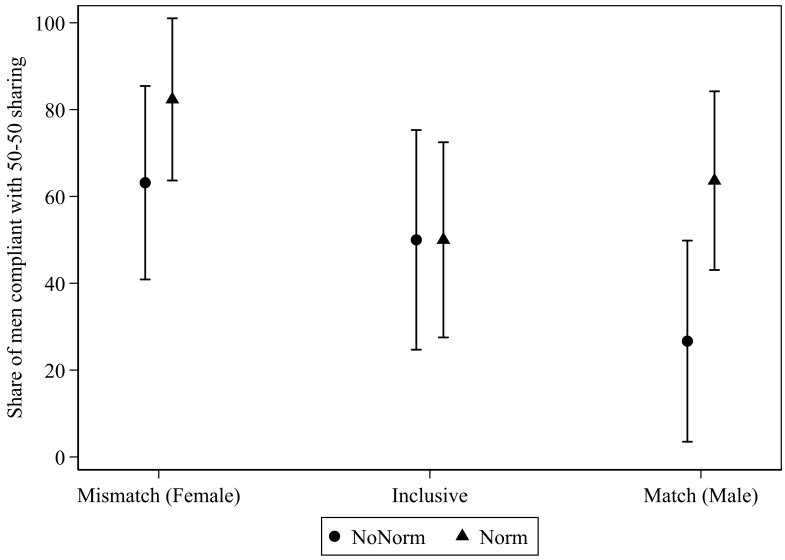
Again, we investigate whether there is indirect support for Hypothesis 1 for the women in our analytical sample. We investigate if making the norms salient increased norm compliance under one gender frame but not under another, in line with the order indicated by Inequality (1). We find mild evidence that women's norm compliance increased when comparing the NoNorm-Mismatch to the Norm-Mismatch treatment as indicated by the positive coefficient of *Norm*. The comparison of the NoNorm-Inclusive with the Norm-Inclusive treatment suggests a decrease in the linear index of norm compliance by $0.293 + (-0.394) = -0.101$, but this decrease is not statistically significant ($p = 0.687$, Wald-test). When comparing the NoNorm-Match to the Norm-Match treatment, the linear index increases by $0.293 + (-0.249) = 0.044$, but this increase is not statistically significant ($p = 0.791$, Wald-test). Therefore, we do not find indirect support for Hypothesis 1.

Given that $Compliance_{all}$ is an aggregate measure, effects and effect sizes may depend on the specific norm elicited. In the following, we focus on individual games to investigate if that was the case.

4.2 Compliance With the 50-50 Sharing Norm in the Dictator Game

We start by analyzing the men's norm compliance in the dictator game. Figure 4 depicts the compliance with the 50-50 sharing norm in the dictator game. We observe that 82.35% of the men complied with the norm in the Norm-Mismatch treatment. In the Norm-Inclusive treatment, 50.00% of the men complied with the norm, whereas the share was 63.64% in the Norm-Match treatment. In the NoNorm-Mismatch treatment, 63.16% of the men complied with the norm. In the NoNorm-Inclusive treatment, 50.00% of the men complied with the norm, whereas 26.67% of the men did so in the NoNorm-Match treatment. Considering the change in the gender frame in the NoNorm treatments, we see that norm compliance increased when moving from the NoNorm-Match, over the NoNorm-Inclusive to the NoNorm-Mismatch treatment. This increase is statistically significant ($p = 0.039$, Jonckheere-Terpstra test). When comparing norm compliance across gender

frames in the Norm treatments, we do not find a significant pattern when moving from a Match over the Inclusive gender frame to a Mismatch ($p = 0.313$, Jonckheere-Terpstra test).



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

Figure 4: Share of men who complied with the behavior described in the 50-50 sharing norm statement in the dictator game.

The regressions in Table 4 report the results from our five specifications for men only. Similar to the results across games, both interaction terms are not statistically significant. Also, their difference is not statistically significant ($p = 0.209$, Wald test). Thus, the regressions do not offer direct support for Hypothesis 2.

Result 2.1. (Men: Compliance with the 50-50 sharing norm in the dictator game)
There is no direct support for Hypothesis 2, i.e., men's compliance with the 50-50 sharing norm in the dictator game does not increase the most if their self-reported gender matches the prescriptive norm statement, and the least if there is a mismatch. We also do not find direct support for an increase in men's norm compliance that falls between these two extremes under the gender-inclusive formulation.

Let us consider if there is indirect support for Hypothesis 2, that is in line with Inequality (1). The coefficient of *Norm* is not statistically significant. Thus, we do not find significant differences when comparing the NoNorm-Mismatch to the Norm-Mismatch treatment. In addition, there was no difference between the NoNorm and Norm treatment in the Inclusive gender frame ($0.793 + (-0.515) = 0.278$, $p = 0.595$, Wald-test). In the Match gender frame, the Norm treatment

Dep. Var.: Compliance$_{DG}$	(1)	(2)	(3)	(4)	(5)
Norm	0.506**	0.593	0.603	0.519	0.793
	(0.252)	(0.464)	(0.460)	(0.497)	(0.584)
Match	-0.710**	-0.959**	-0.994**	-1.274***	-1.661**
	(0.308)	(0.457)	(0.468)	(0.485)	(0.700)
Inclusive	-0.653**	-0.336	-0.315	-0.450	-0.430
	(0.315)	(0.431)	(0.435)	(0.458)	(0.605)
Norm $\times$ Match		0.379	0.387	0.463	0.517
		(0.642)	(0.668)	(0.689)	(0.883)
Norm $\times$ Inclusive		-0.593	-0.533	-0.534	-0.515
		(0.628)	(0.634)	(0.654)	(0.821)
Constant	0.371	0.336	1.634	1.237	1.656
	(0.250)	(0.295)	(1.164)	(1.319)	(1.416)
Pseudo R^2	0.063	0.081	0.105	0.141	0.450
Observations	109	109	109	109	109
Demographics	✗	✗	✓	✓	✓
Language & Understanding	✗	✗	✗	✓	✓
Attitudes & Beliefs	✗	✗	✗	✗	✓

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

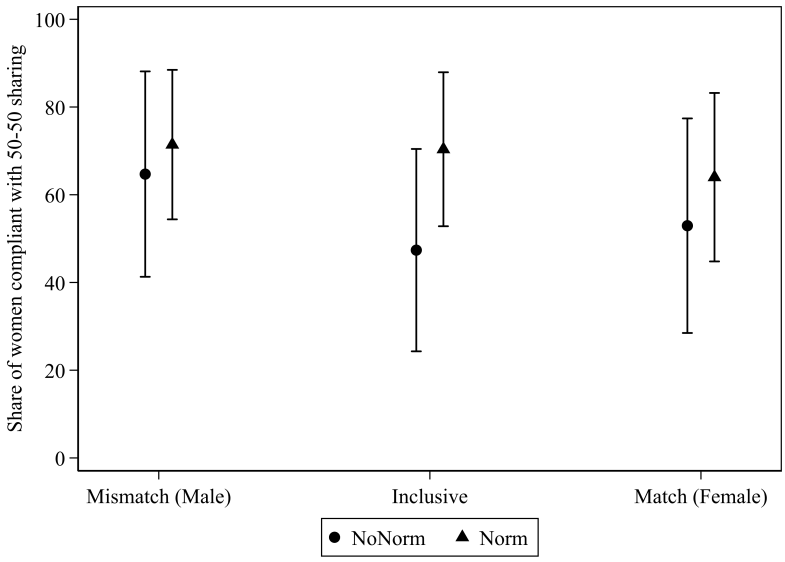
p-values for interaction *effects* based on Ai and Norton (2003), $_{\star}$ 0.10 $_{\star\star}$ 0.05 $_{\star\star\star}$ 0.01

Note: Only two men in the Norm-Match treatment failed the control question for this game. Thus, in deviation from our previous description of the specifications, we omit *Failed attempts*$_{DG}$ from the specifications reported in columns (4) and (5). For the complete table with all coefficients, see Table 14 in Appendix D.

Table 4: Probit regressions on men's compliance with the 50-50 sharing norm in the dictator game.

increased the linear index of norm compliance by $0.793 + 0.517 = 1.310$, which is statistically significant ($p = 0.042$, Wald-test). This effect is partially due to the Norm treatment increasing compliance with the 50-50 sharing norm in the Match gender frame. Mostly, however, it is due to lower norm compliance when moving from the NoNorm-Mismatch over the NoNorm-Inclusive to the NoNorm-Match treatment. Thus, making the norm salient increases norm compliance in the Match gender frame, whereas it does not in the other gender frames. Therefore, we interpret this finding as indirect support for Hypothesis 2.

We now analyze the women's norm compliance in the dictator game. As Figure 5 shows, in the NoNorm-Mismatch treatment, 64.71% of the women complied with the 50-50 sharing norm. In the NoNorm-Inclusive treatment, this share was 47.37%, whereas in the NoNorm-Match treatment, it was 52.94%. In the Norm-Mismatch treatment, 71.43% of the women complied with the norm, 70.37% of the women complied with the norm in the Norm-Inclusive treatment, and in the Norm-Match treatment, 64.00% of the women did so. There were no treatment differences across the gender frames, neither in the NoNorm treatments ($p = 0.495$, Jonckheere-Terpstra test) nor in the Norm treatments ($p = 0.571$, Jonckheere-Terpstra test).



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

Figure 5: Share of women who complied with the behavior described in the 50-50 sharing norm statement in the dictator game.

According to our regression analysis reported in Table 5, we again find no support for Hypothesis 2 because both interaction terms are not statistically significantly different from zero. However, comparing the two interaction terms, the difference between them ($0.474 - (-0.681) = 1.155$) is marginally statistically significant ($p = 0.081$, Wald-test), meaning that the (positive) difference in norm compliance between the NoNorm-Inclusive and the Norm-Inclusive treatment is greater than the (negative) difference between the NoNorm-Match and the Norm-Match treatment. This contradicts Hypothesis 2 according to which the coefficient of the interaction term between Norm and Match should be larger than the interaction term between Norm and Inclusive.

Result 2.2. (Women: Compliance with the 50-50 sharing norm in the dictator game)
We find no direct support for Hypothesis 2 that women's compliance with the 50-50 sharing norm in the dictator game increases the most if their self-reported gender matches the formulation of the prescriptive norm statements and the least if there is a mismatch. We do not find direct support that when gender-inclusive formulations are used, the increase in norm compliance lies between these two extremes.

Let us investigate if making the norm salient increased norm compliance under one gender frame but not under another. Again, we interpret this as indirect support for Hypothesis 2 if this comparison is in line with Inequality (1).

Dep. Var.: Compliance$_{DG}$	(1)	(2)	(3)	(4)	(5)
Norm	0.365	0.189	0.327	0.322	-0.030
	(0.227)	(0.402)	(0.415)	(0.437)	(0.458)
Match	-0.244	-0.304	-0.196	-0.154	0.091
	(0.280)	(0.437)	(0.455)	(0.459)	(0.442)
Inclusive	-0.204	-0.443	-0.265	-0.480	-0.665
	(0.273)	(0.426)	(0.444)	(0.466)	(0.518)
Norm $\times$ Match		0.096	-0.189	-0.247	-0.681
		(0.567)	(0.590)	(0.615)	(0.598)
Norm $\times$ Inclusive		0.413	0.258	0.366	0.474
		(0.557)	(0.572)	(0.607)	(0.692)
Constant	0.271	0.377	-0.915	-1.816	-3.904**
	(0.238)	(0.313)	(0.825)	(1.264)	(1.518)
Pseudo R^2	0.020	0.024	0.080	0.103	0.335
Observations	133	133	133	133	133
Demographics	✗	✗	✓	✓	✓
Language & Understanding	✗	✗	✗	✓	✓
Attitudes & Beliefs	✗	✗	✗	✗	✓

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

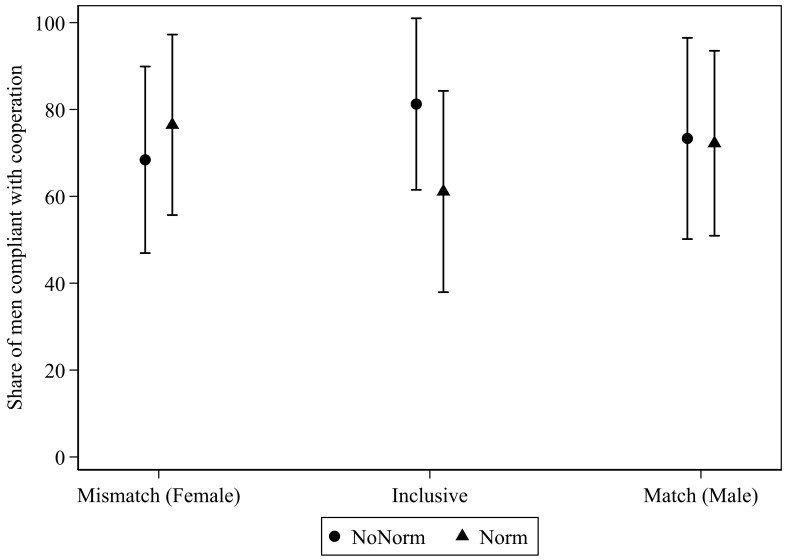
Table 5: Probit regressions on women's compliance with the 50-50 sharing norm in the dictator game.

We do not find differences between the NoNorm-Mismatch and the Norm-Mismatch treatments, as can be seen from the coefficient of *Norm* which is not statistically significant. In the Inclusive gender frame, the Norm treatment increased the linear index of norm compliance by $-0.030 + 0.474 = 0.444$. Yet, this increase is not statistically significant ($p = 0.401$, Wald-test). In the Match gender frame, the Norm treatment decreased the linear index of norm compliance by $|-0.030 + (-0.681)| = |-0.711| = 0.711$, which is marginally statistically significant ($p = 0.071$, Wald-test). Thus, the norm treatment variation had a negative impact in the Match gender frame whereas it did not have a statistically significant effect in the other gender frames. Since this treatment difference is opposite to the hypothesized comparison in Inequality (1), there is no indirect support for Hypothesis 2.

4.3 Compliance With the Cooperation Norm in the Prisoner's Dilemma

Again, we start by analyzing the men's norm compliance; see Figure 6 for a graphical overview. Of all men in the NoNorm-Mismatch treatment, 68.42% complied with the cooperation norm. In the NoNorm-Inclusive treatment, this share was 81.25%, whereas, in the NoNorm-Match treatment, it was 73.33%. In the Norm-Mismatch treatment, 76.47% of the men complied with the norm. In the

Norm-Inclusive treatment, the share was 61.11%, whereas, in the Norm-Match treatment, it was 72.22%. Considering the change in the gender frame in the NoNorm treatments, we do not find



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

Figure 6: Share of men who complied with the behavior described in the prescriptive cooperation norm statement in the prisoner's dilemma.

a significant pattern when moving from a Match over the Inclusive gender frame to a Mismatch ($p = 0.690$, Jonckheere-Terpstra test). When comparing norm compliance across gender frames in the Norm treatments, we also do not find a significant pattern when moving from a Match over the Inclusive gender frame to a Mismatch ($p = 0.819$, Jonckheere-Terpstra test).

The regressions in Table 6 report the regression results on our five specifications for men's compliance with the cooperation norm. From the shares depicted in Figure 6, one would expect a negative effect of the Norm treatment under the Inclusive gender frame. In fact, from our preferred specification in column 5, we see that the interaction term of the Norm and the Inclusive treatment variation is negative, but like the interaction term for the Norm and Match treatment variation, it is not statistically significant. Also, their difference is not statistically significant ($p = 0.524$, Wald test). Thus, the regressions do not offer direct support for Hypothesis 3.

Result 3.1. (Men: Compliance with the cooperation norm in the prisoner's dilemma)
We find no direct support for Hypothesis 3 that men's compliance with the cooperation norm in the prisoner's dilemma increases the most if their self-reported gender matches the formulation of the prescriptive norm statements and the least if there is a mismatch. We do not find direct

Dep. Var.: Compliance$_{PD}$	(1)	(2)	(3)	(4)	(5)
Norm	-0.124	0.242	0.291	0.436	-0.127
	(0.264)	(0.451)	(0.467)	(0.517)	(0.776)
Match	0.025	0.143	0.201	0.391	-0.375
	(0.325)	(0.461)	(0.511)	(0.520)	(0.808)
Inclusive	-0.037	0.408	0.435	0.489	0.278
	(0.318)	(0.473)	(0.509)	(0.559)	(0.592)
Norm × Match		-0.275	-0.538	-0.881	-0.635
		(0.653)	(0.738)	(0.735)	(1.212)
Norm × Inclusive		-0.847	-1.046	-1.189	-1.281
		(0.654)	(0.687)	(0.736)	(1.110)
Constant	0.647**	0.480	2.945**	1.952	-2.619
	(0.264)	(0.301)	(1.165)	(1.284)	(2.919)
Pseudo R^2	0.002	0.017	0.088	0.160	0.607
Observations	103	103	103	103	103
Demographics	✗	✗	✓	✓	✓
Language & Understanding	✗	✗	✗	✓	✓
Attitudes & Beliefs	✗	✗	✗	✗	✓

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

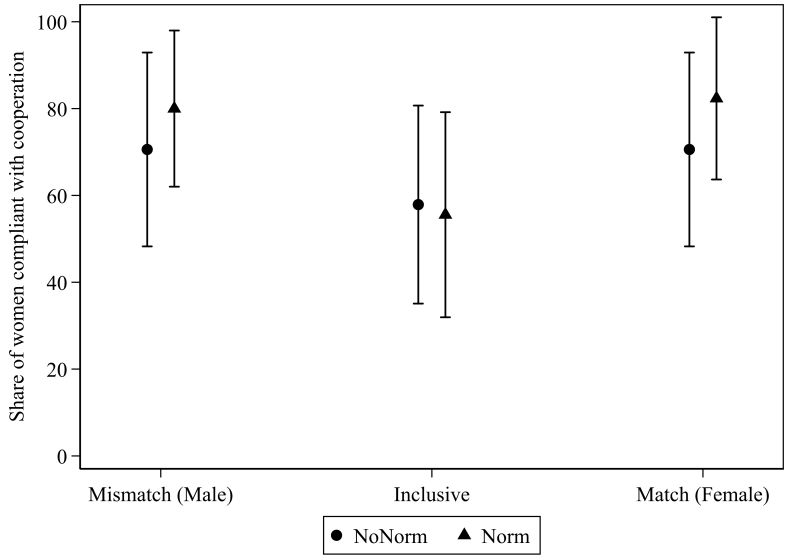
Table 6: Probit regressions on men's compliance with the cooperation norm.

support that when gender-inclusive formulations are used, the increase in norm compliance lies between these two extremes.

As in the previous game, we investigate if there were treatment effects of our norm treatment variation for each gender frame individually. If this is the case for one but not for another gender frame and that comparison is in line with Inequality (1), we interpret this as indirect support for Hypothesis 3. The Norm treatment variation did not significantly affect norm compliance in the Mismatch treatment, as indicated by the coefficient of *Norm*. However, in the Inclusive treatment variation, the Norm treatment reduced the linear index by $|-0.127 + (-1.281)| = |-1.408| = 1.408$. This effect is marginally statistically significant ($p = 0.070$, Wald-test). The Norm treatment variation did not significantly affect norm compliance in the Match treatment, as indicated by the sum of the coefficient of *Norm* and the interaction term *Norm*×*Match* ($-0.127+(-0.635) = -0.762$, $p = 0.394$, Wald-test). Thus, our norm treatment variation did reduce the men's norm compliance in the prisoner's dilemma under the Inclusive frame whereas it did not do so under the other gender frames. As we hypothesized an increase in norm compliance due to making the norm salient, we cannot interpret this finding as indirect support for Hypothesis 3.

As Figure 7 shows, in the NoNorm-Mismatch treatment, 70.59% of the women complied with the

cooperation norm. In the NoNorm-Inclusive treatment, it was 57.89% of the women, whereas in the NoNorm-Match treatment, this share was 70.59%. Of the women in the Norm-Mismatch treatment, 80.00% complied with the norm. In the Norm-Inclusive treatment, this share was 55.56%, whereas in the Norm-Match treatment, it was 82.35%.



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

Figure 7: Share of women who complied with the behavior described in the prescriptive cooperation norm statement in the prisoner's dilemma.

Considering the change in the gender frame in the NoNorm treatments, we do not find a significant pattern when moving from a Match over the Inclusive gender frame to a Mismatch ($p > 0.999$, Jonckheere-Terpstra test). When comparing norm compliance across gender frames in the Norm treatments, we also do not find a significant pattern when moving from a Match over the Inclusive gender frame to a Mismatch ($p = 0.689$, Jonckheere-Terpstra test).

The regressions in Table 7 report the regression results on our five specifications for women's compliance with the cooperation norm. Regarding Hypothesis 3, women's norm compliance was decreased in the Norm-Inclusive treatment compared to the Norm-Mismatch treatment as indicated by the negative and marginally statistically significant interaction term of *Norm* and *Inclusive* only in our preferred specification (5). In our preferred specification (5), comparing the contribution to the linear index in the Norm-Match treatment $(1.050 + 0.334 + (-0.264) = 1.120)$ with that of the Norm-Inclusive treatment $(1.050 + 0.171 + (-1.469) = -0.248)$ reveals that the Norm treatment worked significantly better under the Match gender frame than under the Inclusive gender frame

$(1.120 - (-0.248) = 1.368, p = 0.006$, Wald-test). Thus, we find that the change in women's norm compliance with the cooperation norm in the prisoner's dilemma due to making the norm salient in the Match frame was greater than the corresponding change in the Inclusive frame. Still, this is only in line with the first comparison in Inequality (1).

Result 3.2. (Women: Compliance with the cooperation norm in the prisoner's dilemma)
We find no direct support for Hypothesis 3 that for women compliance with the cooperation norm in the prisoner's dilemma increases the most if their self-reported gender matches the formulation of the prescriptive norm statements and the least if there is a mismatch. We do not find direct support that when gender-inclusive formulations are used, the increase in norm compliance lies between these two extremes.

If making the norm salient increased norm compliance under one gender frame but not under another, which is ranked lower in terms of effect sizes as per Inequality (1), we could interpret this as indirect support for Hypothesis 3 again. We find mild evidence that women's norm compliance was increased when comparing the NoNorm-Mismatch to the Norm-Mismatch treatment as indicated by the positive and statistically significant coefficient of *Norm*. Between the NoNorm-Inclusive and Norm-Inclusive treatment, we see a decrease in the linear index of norm compliance by $1.050 + (-1.469) = -0.419$, but this decrease is not statistically significant ($p = 0.378$, Wald-test). The increase between the NoNorm-Match and Norm-Match treatment is not statistically significant ($1.050 + (-0.264) = 0.786$, $p = 0.167$, Wald-test). Thus, in line with finding no direct support, we find no indirect support for Hypothesis 3.

4.4 Compliance With the Honesty Norm in the Deception Game

Again, we start by describing and analyzing the men's norm compliance in the deception game (see Figure 8). 52.63% of the men complied with the norm in the NoNorm-Mismatch treatment. In the NoNorm-Inclusive treatment, only 37.50% complied with the norm, whereas in the NoNorm-Match treatment, this share was 46.67%. In the Norm-Mismatch treatment, 66.67% of the men complied with the norm. In the Norm-Inclusive treatment, this share was 59.09%, whereas, in the Norm-Match treatment, it was 66.67%. There were neither treatment differences across the gender frames in the NoNorm treatments ($p = 0.672$, Jonckheere-Terpstra test) nor the Norm treatments ($p = 0.864$, Jonckheere-Terpstra test).

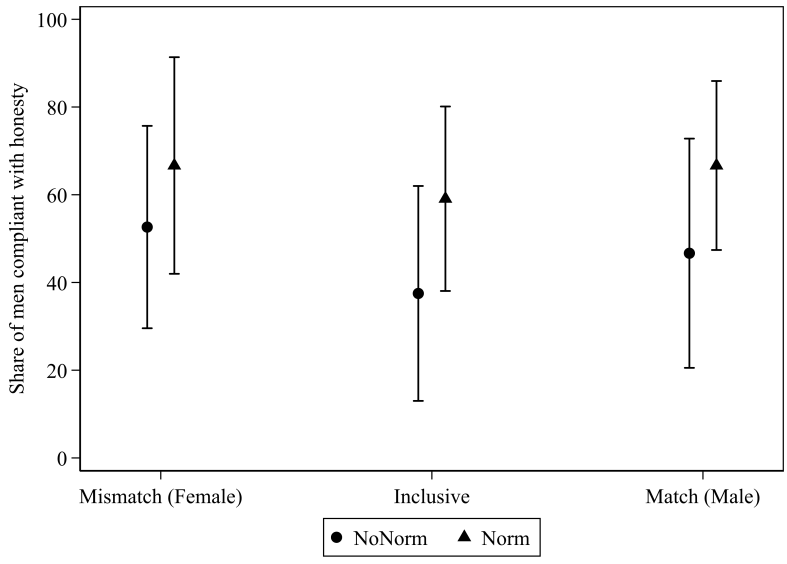
Dep. Var.: Compliance$_{PD}$	(1)	(2)	(3)	(4)	(5)
Norm	0.187	0.300	0.460	0.674	1.050*
	(0.258)	(0.455)	(0.486)	(0.537)	(0.597)
Match	0.034	0.000	0.022	0.166	0.334
	(0.328)	(0.456)	(0.499)	(0.492)	(0.556)
Inclusive	-0.519*	-0.342	-0.181	-0.065	0.171
	(0.309)	(0.434)	(0.453)	(0.470)	(0.516)
Norm × Match		0.087	0.180	-0.176	-0.264
		(0.663)	(0.705)	(0.735)	(0.766)
Norm × Inclusive		-0.360	-0.570	-0.852	-1.469*
		(0.617)	(0.638)	(0.703)	(0.779)
Constant	0.599**	0.541*	0.727	-1.051	-2.083
	(0.265)	(0.322)	(1.042)	(1.485)	(2.127)
Pseudo R^2	0.035	0.040	0.109	0.143	0.273
Observations	108	108	108	108	108
Demographics	✗	✗	✓	✓	✓
Language & Understanding	✗	✗	✗	✓	✓
Attitudes & Beliefs	✗	✗	✗	✗	✓

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: For the complete table with all coefficients, see Table 17 in Appendix D.

Table 7: Probit regressions on women's compliance with the cooperation norm.



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

Figure 8: Share of men who complied with the behavior described in the prescriptive honesty norm statement in the deception game.

The regressions in Table 8 report the probit regression results for our five specifications. We find no support for Hypothesis 4 because the interaction terms between *Norm* and *Match* as well as *Norm* and *Inclusive* are not statistically significant. When compared with the Norm-Match treatment, the index in the Norm-Inclusive treatment is larger by $1.289 - 0.176 = 1.113$, but that difference is not statistically significant ($p = 0.124$, Wald-test).

Result 4.1. (Men: Compliance with the honesty norm in the deception game)
We find no direct support for Hypothesis 4 that for men compliance with the honesty norm in the deception game increases the most if their self-reported gender matches the formulation of the prescriptive norm statements and the least if there is a mismatch. We do not find direct support that when gender-inclusive formulations are used, the increase in norm compliance lies between these two extremes.

We can again investigate if making the norm salient under one gender frame increased norm compliance whereas it did not under another. If this is the case and the differences are in line with Inequality (1), we can consider this indirect support for Hypothesis 4. As the coefficient of *Norm* is not statistically significant, the Norm treatment variation did not affect norm compliance under the Mismatch gender frame. In the Inclusive gender frame, the Norm treatment increased the linear index of norm compliance by $0.458 + 0.831 = 1.289$, and this increase is statistically significant ($p = 0.018$, Wald-test). In the Match gender frame, the Norm treatment increased the linear index of norm compliance by $0.458 + (-0.282) = 0.176$, but this increase is not statistically significant ($p = 0.751$, Wald-test).

Thus, making the norm salient increased the men's norm compliance in the Inclusive gender frame whereas it did not in the other two frames. Considering the second comparison in Inequality (1), this is partial and indirect support for Hypothesis 4.

As Figure 9 shows, in the NoNorm-Mismatch treatment, 47.06% of the women complied with the norm, whereas this share was 36.84% in the NoNorm-Inclusive, and 64.71% in the NoNorm-Match treatment. In the Norm treatment, 75.00% complied with the norm in the Norm-Mismatch treatment, 60.87% in the Norm-Inclusive treatment, and 59.26% in the Norm-Match treatment. There were no treatment differences across the gender frames neither in the NoNorm treatments ($p = 0.308$, Jonckheere-Terpstra test) nor the Norm treatments ($p = 0.364$, Jonckheere-Terpstra test).

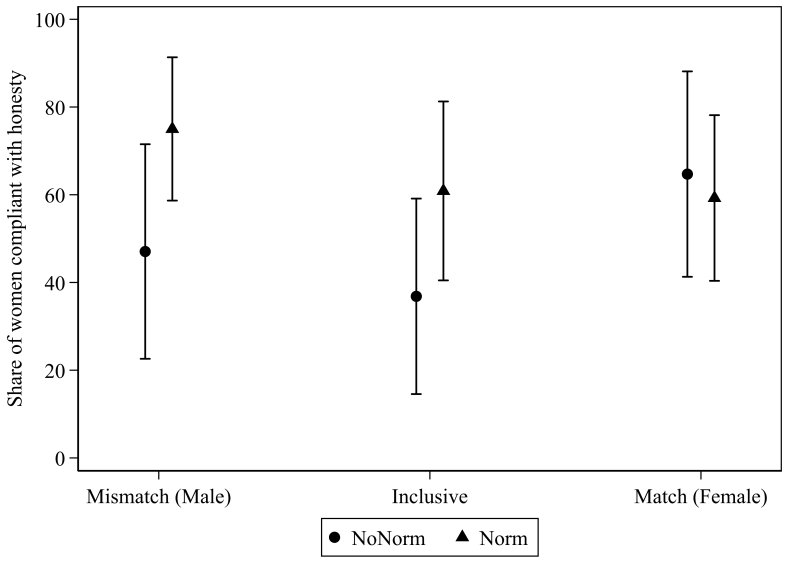
Dep. Var.: Compliance$_{Dec}$	(1)	(2)	(3)	(4)	(5)
Norm	0.480*	0.365	0.415	0.726	0.458
	(0.247)	(0.443)	(0.442)	(0.481)	(0.631)
Match	-0.080	-0.150	-0.105	0.058	-0.535
	(0.305)	(0.435)	(0.457)	(0.497)	(0.621)
Inclusive	-0.296	-0.385	-0.413	-0.406	-1.224**
	(0.305)	(0.432)	(0.445)	(0.477)	(0.553)
Norm × Match		0.150	-0.005	-0.247	-0.282
		(0.611)	(0.645)	(0.683)	(0.829)
Norm × Inclusive		0.184	0.195	0.127	0.831
		(0.611)	(0.618)	(0.653)	(0.847)
Constant	0.017	0.066	1.504	-1.180	-3.840**
	(0.244)	(0.289)	(0.993)	(1.337)	(1.619)
Pseudo R^2	0.031	0.031	0.063	0.160	0.447
Observations	111	111	111	111	111
Demographics	✗	✗	✓	✓	✓
Language & Understanding	✗	✗	✗	✓	✓
Attitudes & Beliefs	✗	✗	✗	✗	✓

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: For the complete table with all coefficients, see Table 18 in Appendix D.

Table 8: Probit regressions on men's compliance with the honesty norm.



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

Figure 9: Share of women who complied with the behavior described in the prescriptive honesty norm statement in the deception game.

Consider Table 9. In our preferred specification in column 5, we do not find direct support for Hypothesis 4, as the interaction terms are both statistically insignificant. Also, their difference is not statistically significant ($p = 0.714$, Wald test).

Result 4.2. (Women: Compliance with the honesty norm in the deception game)
We find no direct support for Hypothesis 4 that women's compliance with the honesty norm in the deception game increases the most if their self-reported gender matches the formulation of the prescriptive norm statements and the least if there is a mismatch. We do not find direct support that when gender-inclusive formulations are used, the increase in norm compliance lies between these two extremes.

Dep. Var.: Compliance$_{Dec}$	(1)	(2)	(3)	(4)	(5)
Norm	0.404*	0.748*	0.769*	0.839**	0.829*
	(0.228)	(0.400)	(0.402)	(0.413)	(0.462)
Match	-0.086	0.451	0.441	0.518	0.317
	(0.275)	(0.437)	(0.444)	(0.461)	(0.554)
Inclusive	-0.350	-0.262	-0.224	-0.358	0.069
	(0.273)	(0.424)	(0.435)	(0.461)	(0.611)
Norm × Match		-0.891	-0.844	-0.986	-0.650
		(0.564)	(0.576)	(0.600)	(0.693)
Norm × Inclusive		-0.136	-0.182	-0.203	-0.943
		(0.564)	(0.568)	(0.587)	(0.732)
Constant	0.129	-0.074	0.219	-0.105	0.103
	(0.237)	(0.305)	(0.769)	(1.056)	(1.204)
Pseudo R^2	0.029	0.046	0.057	0.071	0.436
Observations	131	131	131	131	131
Demographics	✗	✗	✓	✓	✓
Language & Understanding	✗	✗	✗	✓	✓
Attitudes & Beliefs	✗	✗	✗	✗	✓

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: For the complete table with all coefficients, see Table 19 in Appendix D.

Table 9: Probit regressions on women's compliance with the honesty norm.

One more time, we investigate if there is indirect support for Hypothesis 4 in line with Inequality (1). We find mild evidence that women's norm compliance was increased when comparing the NoNorm-Mismatch to the Norm-Mismatch treatment as indicated by the coefficient of *Norm*. The comparison of the NoNorm-Inclusive with the Norm-Inclusive treatment reveals a decrease in the linear index of norm compliance by $0.829 + (-0.943) = -0.114$, but this decrease is not statistically significant ($p = 0.840$, Wald-test). When comparing the NoNorm-Match to the Norm-Match treatment, the linear index increased by $0.829 + (-0.650) = 0.179$, but this increase is not statistically significant ($p = 0.732$, Wald-test). This means that women's norm compliance with the prescriptive honesty norm increased in the Mismatch gender frame but not in the other two frames. Thus, there is no indirect support for Hypothesis 4.

5 Discussion

Before discussing potential mechanisms and limitations, we start by briefly summarizing our results.

For men in the dictator game, we find indirect support for Hypothesis 2, as they are more likely to comply with a norm if the norm statement matches their gender. In terms of our notational framework, we find a marginally statistically greater difference in norm compliance between the NoNorm-Match and the Norm-Match treatment than between the NoNorm-Inclusive and the Norm-Inclusive treatment. We find no support for similar effects on men's norm compliance in the other two games. This is similar when considering the women's norm compliance across the three games. We will thus focus our discussion on the men's norm compliance in the dictator game. For completeness, we will also report our analysis for all considered mechanisms for men here and provide the corresponding analysis for the women in all three games in Appendix D.

With our data, we can investigate several potential mechanisms behind the result that men's norm compliance increased more in the Match gender frame than in the other two gender frames.[20] First, we look into excluded participants based on their appropriateness rating, second, the order of games, and the selected sample.

Remember that we compared all participants from the NoNorm treatments to only those participants in the Norm treatments who rated the respective prescriptive norm statement as "rather appropriate," thus rendering it a social norm for these participants. We did so to analyze whether social norms made salient in a particular gender frame affect behavior differently. Naturally, some participants rated the prescriptive norm statements as "rather inappropriate" and, in line with our preregistration, we excluded them from our analytical sample. The analysis so far reports the results in line with this (preregistered) exclusion criterion. However, this selection into the analytical sample might explain some of the observed effects. A potential mechanism that could explain some of our results is motivated reasoning (Kunda, 1990; Bénabou and Tirole, 2006; Gneezy et al., 2020). Participants' behavior and their ratings of the appropriateness of the norm statements might thus not be independent. Recall that $Appropriateness_G$ is a dummy that is one if a participant rendered the behavior described in the prescriptive norm statement relating to game G as

[20] Besides having controlled for beliefs in our regressions, we conducted Kruskal-Wallis-Tests to check for differences in beliefs in the NoNorm and Norm treatments respectively and found no statistically significant patterns (the smallest p-value was $p = 0.1327$). Thus, in line with Gorny et al. (2023), we do not find that strategic beliefs differed across our treatments.

"rather appropriate" to society and zero otherwise. *Appropriateness*$_{all}$ is one if all *Appropriateness*$_G$ dummies are one and zero otherwise. Thus, the participants we included in our analytical sample could have differed from those we excluded in a systematic way that correlates (at least partially) with our treatments. We rerun the saturated specification for norm compliance across games and for each game to investigate such a selection as a potential mechanism behind the treatment effects.

For the men in our raw sample, Table 10 reports the results from the Poisson regression for norm compliance across games and the Probit regressions for norm compliance in each game controlling for the participants' appropriateness rating. Our findings across games (column *all*) and in the deception game (column *Dec*) only change slightly when compared to the results in the analytical samples. Most importantly, in the dictator game (column *DG*), all coefficient signs are unaffected, but the difference between the two interactions Norm×Match and Norm×Inclusive is not statistically significant anymore ($p = 0.193$, Wald test).

Yet, the Norm treatment still only increases norm compliance in the Match gender frame ($0.370 + 0.810 = 1.180$, $p = 0.068$, Wald test) whereas it does not in the other two gender frames (0.370, $p = 0.480$, for the Mismatch treatment and $0.370 - 0.235 = 0.135$, $p = 0.797$, for the Inclusive treatment, Wald tests). In the deception game, the statistically significant increase in norm compliance when comparing the Norm-Inclusive to the NoNorm-Inclusive treatment remains ($0.613 + 0.671 = 1.284$, $p = 0.021$, Wald test). Only the coefficients for Appropriateness$_{PD}$ and Appropriateness$_{Dec}$ are positive and statistically significant. This indicates that for these two games, participants selected into our analytical samples were indeed more likely on average to comply with the norm. The change in the treatment coefficients suggests though, that this did not very strongly affect our results (in the case of the prisoner's dilemma it rather renders them a lower bound).

For the women in our raw sample, the regression results from including the norm ratings do not differ systematically from what we report in our results section. The coefficients are reported in Tables 21a and 21b in Appendix D.

In sum, we thus find only very mild evidence for motivated reasoning when it comes to our hypotheses. The treatment coefficients only changed marginally and all retained their sign.

However, there are further explanations and limitations connected with our experiment.

Given that we already have a 2×3 design, we chose to not randomize the order of the three games

Dep. Var.: Compliance	all	DG	PD	Dec
Appropriateness$_{all}$	-0.076			
	(0.092)			
Appropriateness$_{DG}$		-0.289		
		(0.398)		
Appropriateness$_{PD}$			0.831**	
			(0.418)	
Appropriateness$_{Dec}$				0.994*
				(0.526)
Norm	-0.128	0.370	-0.134	0.613
	(0.148)	(0.524)	(0.701)	(0.605)
Match	-0.388**	-1.364**	-0.487	-0.618
	(0.184)	(0.654)	(0.693)	(0.640)
Inclusive	-0.361**	-0.454	0.069	-1.398**
	(0.182)	(0.617)	(0.546)	(0.578)
Norm × Match	0.208	0.810	-0.391	-0.544
	(0.217)	(0.833)	(1.038)	(0.814)
Norm × Inclusive	0.077	-0.235	-1.615*	0.671
	(0.217)	(0.766)	(0.868)	(0.798)
Constant	-0.267	2.070	-3.049	-5.389***
	(0.659)	(1.426)	(2.322)	(2.040)
Pseudo R^2	0.141	0.447	0.536	0.481
Observations	122	122	122	122
Demographics	✓	✓	✓	✓
Language & Understanding	✓	✓	✓	✓
Attitudes & Beliefs	✓	✓	✓	✓

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: For the complete table with all coefficients, see Tables 20a and 20b in Appendix D.

Table 10: Poisson regressions on how many norms men complied with across games controlling for whether they rated all norms rather appropriate or not (column *all*) and probit regressions on men's norm compliance in the individual games controlling for whether they rated the respective norm rather appropriate or not (columns *DG* through *Dec*) using the saturated specification.

played. Randomization would add another layer of complexity to the analysis. However, this choice comes at the cost that it remains a question for further research if the battery of games influences the salience of the gender frames and the effect on the norms.

We ran our experiment with a rather small student sample. Students are relatively homogenous in terms of education and age. Our results could therefore possibly be an upper bound of the effects of gendered language on norm compliance since older people grew up before gender-inclusive forms were introduced. In addition, students might be heavily exposed to gender-inclusive language and the related discussion whereas, for the general population, this topic might be less salient. Furthermore, due to our small sample size, effects need to be rather large to be picked up by statistical tests.

Overall, our data suggest that selection into the analytical sample is the strongest driver of our

results. Depending on self-reported gender and treatment, the participants' rating of the prescriptive social norm statements varies and reflects their norm compliance in the respective ensuing game.

6 Conclusion

We report results from a controlled online experiment in which we made prescriptive norms salient and systematically varied the grammatical gender used in the formulation of these prescriptive norms and the experimental instructions. We hypothesized that a match between the self-reported gender of participants and the gender used in the norm statement increases the participants' norm compliance.

In the dictator game, we find mild support for our hypothesis that men are more likely to comply with a norm if the norm statement matches their gender. We find no support for similar effects on norm compliance in the other two games. For women, we did not find evidence in favor of our hypotheses.

We initially excluded participants from our analysis who did not consider our prescriptive norm statements a social norm. Including them in our analysis and controlling for the participant's appropriateness ratings for the norm statements only slightly affected our result. There was still some support for men's compliance to be higher when the prescriptive norm statement was framed with a matching, i.e. male, frame.

Due to the limitations we discussed, there is a need for further research on how to effectively communicate prescriptive norms. For gendered language in specific, we provide a first empirical basis for an otherwise heated debate. How these effects vary over time, with the pool of participants, and with the language spoken by them, are questions left for further research.

References

ABELER, J., D. NOSENZO, AND C. RAYMOND (2019): "Preferences for truth-telling," *Econometrica*, 87, 1115–1153.

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

AKERLOF, G. (1976): "The economics of caste and of the rat race and other woeful tales," *The Quarterly Journal of Economics*, 90, 599–617.

AKERLOF, G. A. (1997): "Social distance and social decisions," *Econometrica*, 65, 1005–1027.

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

——— (2010): *Identity economics*, Princeton University Press.

ANDREONI, J. (1990): "Impure altruism and donations to public goods: A theory of warm-glow giving," *The Economic Journal*, 100, 464–477.

ANDREONI, J. AND B. D. BERNHEIM (2009): "Social image and the 50–50 norm: A theoretical and experimental analysis of audience effects," *Econometrica*, 77, 1607–1636.

BALAFOUTAS, L., H. FORNWAGNER, E. HAUSER, AND O. HAUSER (2023): "Gender-Inclusive Language and Economic Decision-Making," SSRN Working Paper 4411481.

BALAFOUTAS, L. AND N. NIKIFORAKIS (2012): "Norm enforcement in the city: A natural field experiment," *European Economic Review*, 56, 1773–1785.

BAŠIĆ, Z. AND E. VERRINA (2021): "Personal norms—and not only social norms—shape economic behavior," Discussion Papers of the Max Planck Institute for Research on Collective Goods 2020/25.

BAUMEISTER, R. F. AND K. L. SOMMER (1997): "What do men want? Gender differences and two spheres of belongingness: Comment on Cross and Madson (1997)," *Psychological Bulletin*, 122, 38–44.

BÉNABOU, R. AND J. TIROLE (2006): "Incentives and prosocial behavior," *American Economic Review*, 96, 1652–1678.

——— (2011): "Identity, morals, and taboos: Beliefs as assets," *The Quarterly Journal of Economics*, 126, 805–855.

BENJAMIN, D. J., J. J. CHOI, AND A. J. STRICKLAND (2010): "Social identity and preferences," *American Economic Review*, 100, 1913–1928.

BERNHARD, H., E. FEHR, AND U. FISCHBACHER (2006): "Group affiliation and altruistic norm enforcement," *American Economic Review*, 96, 217–221.

BICCHIERI, C., E. DIMANT, S. GÄCHTER, AND D. NOSENZO (2022a): "Social proximity and the erosion of norm compliance," *Games and Economic Behavior*, 132, 59–72.

BICCHIERI, C., E. DIMANT, M. GELFAND, AND S. SONDEREGGER (2022b): "Social norms and behavior change: The interdisciplinary research frontier," *Journal of Economic Behavior & Organization*, 205, A4–A7.

BICCHIERI, C., R. MULDOON, AND A. SONTUOSO (2018): "Social Norms," in *The Stanford Encyclopedia of Philosophy*, ed. by E. N. Zalta, Metaphysics Research Lab, Stanford University, Winter 2018 ed.

BINMORE, K. AND L. SAMUELSON (1994): "An economist's perspective on the evolution of norms," *Journal of Institutional and Theoretical Economics (JITE)/Zeitschrift für die gesamte Staatswissenschaft*, 150, 45–63.

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

BOLTON, G., B. GREINER, AND A. OCKENFELS (2013): "Engineering trust: Reciprocity in the production of reputation information," *Management Science*, 59, 265–285.

BOSCHINI, A., A. MUREN, AND M. PERSSON (2011): "Men among men do not take norm enforcement seriously," *The Journal of Socio-Economics*, 40, 523–529.

BURSZTYN, L., A. L. GONZÁLEZ, AND D. YANAGIZAWA-DROTT (2020): "Misperceived social norms: Women working outside the home in Saudi Arabia," *American Economic Review*, 110, 2997–3029.

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.

COLEMAN, L. (2022): "Cops being told to use gender-neutral terms instead of calling people sir or ma'am," https://www.thesun.co.uk/news/18292950/cops-use-gender-neutral-terms-instead-sir-maam/.

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

CROSS, S. E. AND L. MADSON (1997): "Models of the self: Self-construals and gender," *Psychological Bulletin*, 122, 5–37.

D'ADDA, G., M. DUFWENBERG, F. PASSARELLI, AND G. TABELLINI (2020): "Social norms with private values: Theory and experiments," *Games and Economic Behavior*, 124, 288–304.

DELLAVIGNA, S. (2009): "Psychology and economics: Evidence from the field," *Journal of Economic Literature*, 47, 315–372.

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.

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

DURKHEIM, E. (1950): "The rules of sociological method," *Chicago University*.

EXLEY, C. L., M. NIEDERLE, AND L. VESTERLUND (2020): "Knowing when to ask: The cost of leaning in," *Journal of Political Economy*, 128, 816–854.

FALLUCCHI, F. AND D. NOSENZO (2022): "The coordinating power of social norms," *Experimental Economics*, 25, 1–25.

FEHR, E. AND U. FISCHBACHER (2004a): "Social norms and human cooperation," *Trends in Cognitive Sciences*, 8, 185–190.

——— (2004b): "Third-party punishment and social norms," *Evolution and Human Behavior*, 25, 63–87.

FEHR, E., U. FISCHBACHER, AND S. GÄCHTER (2002): "Strong reciprocity, human cooperation, and the enforcement of social norms," *Human Nature*, 13, 1–25.

FEHR, E. AND S. GÄCHTER (1998): "Reciprocity and economics: The economic implications of homo reciprocans," *European Economic Review*, 42, 845–859.

FEHR, E. AND B. ROCKENBACH (2004): "Human altruism: Economic, neural, and evolutionary perspectives," *Current Opinion in Neurobiology*, 14, 784–790.

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.

FRIEDL, A., A. PONDORFER, AND U. SCHMIDT (2020): "Gender differences in social risk taking," *Journal of Economic Psychology*, 77, Article 102182.

GÄCHTER, S., L. GERHARDS, AND D. NOSENZO (2017): "The importance of peers for compliance with norms of fair sharing," *European Economic Review*, 97, 72–86.

GÄCHTER, S., D. NOSENZO, AND M. SEFTON (2013): "Peer effects in pro-social behavior: Social norms or social preferences?" *Journal of the European Economic Association*, 11, 548–573.

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.

GAURI, V., T. RAHMAN, AND I. K. SEN (2019): "Measuring social norms about female labor force participation in Jordan," World Bank Policy Research Working Paper.

GNEEZY, U., B. ROCKENBACH, AND M. SERRA-GARCIA (2013): "Measuring lying aversion," *Journal of Economic Behavior & Organization*, 93, 293–300.

GNEEZY, U., S. SACCARDO, M. SERRA-GARCIA, AND R. VAN VELDHUIZEN (2020): "Bribing the self," *Games and Economic Behavior*, 120, 311–324.

GOETTE, L., D. HUFFMAN, AND S. MEIER (2006): "The impact of group membership on cooperation and norm enforcement: Evidence using random assignment to real social groups," *American Economic Review*, 96, 212–216.

GONZALEZ CAMANO, E. AND R. BROWN (2022): ""¡Hola a Todes!" Language becomes a political battleground in Latin America," https://www.americasquarterly.org/article/hola-a-todes-language-becomes-a-political-battleground-in-latin-america/.

GORNY, P. M., P. NIEKEN, AND K. STRÖHLEIN (2023): "He, she, they? The impact of gendered language on economic behavior." Working Paper.

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

GÜTH, W., R. SCHMITTBERGER, AND B. SCHWARZE (1982): "An experimental analysis of ultimatum bargaining," *Journal of Economic Behavior & Organization*, 3, 367–388.

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 .

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.

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.

JAKIELA, P. (2011): "Social preferences and fairness norms as informal institutions: Experimental evidence," *American Economic Review*, 101, 509–513.

KAHNEMAN, D., J. L. KNETSCH, AND R. H. THALER (1986): "Fairness and the assumptions of economics," *Journal of Business*, 59, 285–300.

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

KESSLER, J. B. AND S. LEIDER (2012): "Norms and contracting," *Management Science*, 58, 62–77.

KRUPKA, E. AND R. A. WEBER (2009): "The focusing and informational effects of norms on pro-social behavior," *Journal of Economic Psychology*, 30, 307–320.

——— (2013): "Identifying social norms using coordination games: Why does dictator game sharing vary?" *Journal of the European Economic Association*, 11, 495–524.

KRYSOWSKI, E. AND J. TREMEWAN (2021): "Why does anonymity make us misbehave: Different norms or less compliance?" *Economic Inquiry*, 59, 776–789.

KUNDA, Z. (1990): "The case for motivated reasoning." *Psychological Bulletin*, 108, 480–498.

LANE, T., D. NOSENZO, AND S. SONDEREGGER (forthcoming): "Law and norms: Empirical evidence," *American Economic Review*.

MASCAGNI, G. (2018): "From the lab to the field: A review of tax experiments," *Journal of Economic Surveys*, 32, 273–301.

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.

NOSENZO, D. AND L. GOERGES (2020): "Measuring social norms in economics: Why it is important and how it is done," *Analyse & Kritik*, 42, 285–312.

NÖSTLINGER, N. (2021): "Debate over gender-neutral language divides Germany," https://www.politico.eu/article/debate-over-gender-inclusive-neutral-language-divides-germany/.

OSTROM, E. (2000): "Collective action and the evolution of social norms," *Journal of Economic Perspectives*, 14, 137–158.

POSNER, E. A. (2009): *Law and social norms*, Harvard University Press.

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.

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.

SETHI, R. AND E. SOMANATHAN (1996): "The evolution of social norms in common property resource use," *American Economic Review*, 86, 766–788.

SHERIF, M. (1936): *The psychology of social norms*, Harper.

SPITZER, M., U. FISCHBACHER, B. HERRNBERGER, G. GRÖN, AND E. FEHR (2007): "The neural signature of social norm compliance," *Neuron*, 56, 185–196.

THALER, R. H. (2016): "Behavioral economics: Past, present, and future," *American Economic Review*, 106, 1577–1600.

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.

A Interaction Terms and Effects

When analyzing the behavior within the three games played by the participants, we resort to probit models. This is due to the binary nature of our dependent variables, i.e., they are one if a participant complied to the respective prescriptive norm statement and zero otherwise. 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.

$$\begin{aligned} Compliance_{G,i} = \beta_0 &+ \beta_1 Norm_i + \beta_2 Match_i + \beta_3 Inclusive_i \\ &+ \beta_4 Norm_i \times Match_i + \beta_5 Norm_i \times Inclusive_i \\ &+ \boldsymbol{\gamma} \boldsymbol{X}_i + \varepsilon_i \end{aligned}$$

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

$$\frac{\partial^2 Compliance_{G,i}}{\partial Norm_i \partial Match_i} = \beta_4.$$

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.[21]

$$\begin{aligned} P(Compliance_{G,i} = 1) = \Phi(&\beta_0 + \beta_1 Norm_i + \beta_2 Match_i + \beta_3 Inclusive_i \\ &+ \beta_4 Norm_i \times Match_i + \beta_5 Norm_i \times Inclusive_i \\ &+ \boldsymbol{\gamma} \boldsymbol{X}_i) \end{aligned}$$

The interaction effect is given by

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

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 impor-

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

tantly, the estimator of this term has standard errors that differ from those of $\widehat{\beta}_4$. Thus, in 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 Norm$\times$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

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

Note that this is an abuse of notation as the subscript refers to the statistical significance of the term in (2). We attach it to the interaction term as we expect the reader to search for information on the interaction of treatment variations there.

As our interaction variables are dummies, we could alternatively compare the probabilities of observing compliance between treatments. Consider the NoNorm-Match and Norm-Match treatment. Abstracting from $\boldsymbol{\gamma X}_i$, the linear index for the NoNorm-Match treatment is given by β_2, whereas the linear index for the Norm-Match treatment is given by $\beta_1 + \beta_2 + \beta_4$. Their difference is given by $\beta_1 + \beta_4$, which we can test to be significantly different from zero. We do so throughout our analysis to investigate whether the Norm treatment variation systematically affected compliance in some gender frames, but not in others.

This procedure can also be applied to the Poisson regressions we ran for analyzing norm compliance across games by replacing $P(Compliance_{G,i} = 1)$ with

$$\begin{aligned} E(Compliance_{all,i}) = \exp\big(&\beta_0 + \beta_1 Norm_i + \beta_2 Match_i + \beta_3 Inclusive_i \\ &+ \beta_4 Norm_i \times Match_i + \beta_5 Norm_i \times Inclusive_i \\ &+ \boldsymbol{\gamma X}_i + \varepsilon_i\big) \, . \end{aligned}$$

B Prescriptive Norm Statements

The prescriptive norm statement used in the dictator game was the following.

> **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.**

The prescriptive norm statement used in the prisoner's dilemma was the following.

> **A participant in the role of participant A should make a decision in which *she* sends 8 ECUs of *her* 10 ECUs to participant B.**[22]

The prescriptive norm statement used in the deception game was the following.

> **A participant in the role of participant A should compose a message to participant B, which contains the actually assigned number.**

C Controls

We now introduce all the additional controls we used throughout our analysis. *Age* measures the participants' age in years. The variable *Undergraduate* is one if the participant was currently enrolled for a bachelor's degree and zero otherwise. We asked participants for the current *Semester* they are in, including bachelor semesters if the participant was in their masters. We asked participants about the subjects in which they major. We grouped those in majors related to *Business and Economics*, *Education*, and *Other*, with the latter category serving as a baseline unless mentioned otherwise. 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. The variable *Language comments* is one whenever a free-text comment referred to the instructions or norm gender frames and zero otherwise. We also asked participants to rate the clarity of instructions on a 7-point Likert scale. We refer to the resulting variable as *Instructions clear*. Before the beginning of Stage 2, participants had to pass a short survey on the general understanding of the experiment. Before each individual game, we also conducted control questions on the understanding of 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* 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 and when pooling all games because participants can only reciprocate in the prisoner's dilemma. To elicit the variables *First-order belief*$_G$ and *Second-order*

[22] Emphasis is added to indicate that this statement is a translation from the female treatment.

belief$_G$, we first provided a brief summary of each game $G \in \{DG, PD, Dec\}$. Subsequently, we elicited beliefs relative to the prescriptive norm statements in the respective game. Specifically, we phrased our belief elicitation around 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. [23]

D Tables

	NoNorm			Norm		
	Male	Inclusive	Female	Male	Inclusive	Female
Age in years	26.000	24.829	23.861	23.861	23.552	24.920
	(5.512)	(3.120)	(3.164)	(3.315)	(3.474)	(7.315)
Woman	0.536	0.543	0.472	0.569	0.527	0.600
	(0.508)	(0.505)	(0.506)	(0.500)	(0.504)	(0.495)
Semester	8.406	8.229	6.806	7.034	7.309	7.320
	(4.924)	(4.222)	(4.013)	(4.357)	(3.237)	(4.177)
Undergraduate	0.469	0.371	0.556	0.690	0.691	0.700
	(0.507)	(0.490)	(0.504)	(0.467)	(0.466)	(0.463)
Observations	32	35	36	58	55	50

Note: Standard deviations in parentheses; for the Norm treatments, all observations which are in at least one of the analytical samples are included.

Table 11: Descriptive statistics across treatments.

[23] A complete list of descriptions and summary statistics of the variables is available from the authors upon request.

Dep. Var.: Compliance$_{all}$	(1)	(2)	(3)	(4)	(5)
Norm	0.173	0.225	0.205	0.180	-0.272*
	(0.128)	(0.169)	(0.167)	(0.181)	(0.148)
Match	-0.171	-0.228	-0.229	-0.254	-0.484***
	(0.143)	(0.245)	(0.248)	(0.252)	(0.178)
Inclusive	-0.216	-0.088	-0.107	-0.165	-0.354**
	(0.162)	(0.213)	(0.216)	(0.221)	(0.168)
Norm $\times$ Match		0.085	0.038	0.050	0.330
		(0.295)	(0.320)	(0.318)	(0.226)
Norm $\times$ Inclusive		-0.289	-0.262	-0.222	0.210
		(0.329)	(0.327)	(0.316)	(0.228)
Age			-0.036	-0.043*	-0.018
			(0.024)	(0.023)	(0.019)
Semester			0.016	0.024	0.030**
			(0.018)	(0.018)	(0.014)
Business and Economics			-0.209	-0.170	-0.148
			(0.143)	(0.146)	(0.135)
Education			-0.147	-0.147	0.144
			(0.184)	(0.183)	(0.166)
Language attitude				0.099**	-0.002
				(0.050)	(0.045)
Remembered formulations				0.002	-0.193
				(0.155)	(0.123)
Language comments				0.247	0.207
				(0.181)	(0.178)
Instructions clear				0.041	-0.050
				(0.040)	(0.035)
Failed attempts$_{all}$				0.015	0.026**
				(0.010)	(0.011)
Pseudo R^2	0.009	0.012	0.021	0.038	0.149
Observations	92	92	92	92	92

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 12a: Poisson regressions on the number of games in which men complied with the respective norm (complete table with all coefficients, see Table 12b for the remaining coefficients).

Dep. Var.: Compliance$_{all}$	(1)	(2)	(3)	(4)	(5)
Risk aversion					-0.025
					(0.025)
Positive reciprocity					0.158*
					(0.096)
Negative reciprocity					-0.061*
					(0.036)
First-order belief$_{DG}$					0.010***
					(0.003)
Second-order belief$_{DG}$					-0.004
					(0.004)
First-order belief$_{PD}$					0.010*
					(0.006)
Second-order belief$_{PD}$					-0.003
					(0.006)
First-order belief$_{Dec}$					0.005
					(0.006)
Second-order belief$_{Dec}$					0.004
					(0.007)
Constant	0.635***	0.611***	1.524***	1.045*	-0.662
	(0.112)	(0.136)	(0.558)	(0.618)	(0.875)
Pseudo R^2	0.009	0.012	0.021	0.038	0.149
Observations	92	92	92	92	92

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 12b: Poisson regressions on the number of games in which men complied with the respective norm (complete table with all coefficients, continued from Table 12a).

Dep. Var.: Compliance$_{all}$	(1)	(2)	(3)	(4)	(5)
Norm	0.254**	0.226	0.232	0.252	0.293*
	(0.112)	(0.174)	(0.186)	(0.184)	(0.162)
Match	0.005	0.032	0.011	0.047	0.131
	(0.130)	(0.212)	(0.227)	(0.224)	(0.192)
Inclusive	-0.173	-0.249	-0.213	-0.244	-0.024
	(0.138)	(0.240)	(0.245)	(0.240)	(0.242)
Norm × Match		-0.056	-0.009	-0.093	-0.249
		(0.258)	(0.275)	(0.277)	(0.245)
Norm × Inclusive		0.151	0.139	0.158	-0.394
		(0.279)	(0.277)	(0.279)	(0.323)
Age			-0.010	-0.003	-0.005
			(0.012)	(0.011)	(0.009)
Semester			0.012	0.012	0.001
			(0.017)	(0.017)	(0.015)
Business and Economics			-0.317**	-0.298*	-0.197
			(0.161)	(0.160)	(0.168)
Education			-0.094	-0.086	-0.037
			(0.134)	(0.133)	(0.142)
Language attitude				0.100**	0.054
				(0.045)	(0.046)
Remembered formulations				0.030	0.110
				(0.131)	(0.114)
Language comments				0.075	-0.022
				(0.179)	(0.165)
Instructions clear				0.025	-0.023
				(0.043)	(0.044)
Failed attempts$_{all}$				0.018	-0.039
				(0.044)	(0.044)
Pseudo R^2	0.014	0.015	0.026	0.034	0.086
Observations	94	94	94	94	94

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 13a: Poisson regressions on the number of games in which women complied with the respective norm (complete table with all coefficients, see Table 13b for the remaining coefficients).

Dep. Var.: Compliance$_{all}$	(1)	(2)	(3)	(4)	(5)
Risk aversion					0.007
					(0.025)
Positive reciprocity					0.147*
					(0.088)
Negative reciprocity					-0.118***
					(0.046)
First-order belief$_{DG}$					0.011***
					(0.004)
Second-order belief$_{DG}$					-0.005
					(0.004)
First-order belief$_{PD}$					0.004
					(0.006)
Second-order belief$_{PD}$					-0.005
					(0.006)
First-order belief$_{Dec}$					0.003
					(0.003)
Second-order belief$_{Dec}$					0.001
					(0.004)
Constant	0.586***	0.601***	0.892**	0.112	-0.543
	(0.118)	(0.154)	(0.430)	(0.569)	(0.833)
Pseudo R^2	0.014	0.015	0.026	0.034	0.086
Observations	94	94	94	94	94

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 13b: Poisson regressions on the number of games in which women complied with the respective norm (complete table with all coefficients, continued from Table 13a).

Dep. Var.: Compliance$_{DG}$	(1)	(2)	(3)	(4)	(5)
Norm	0.506**	0.593	0.603	0.519	0.793
	(0.252)	(0.464)	(0.460)	(0.497)	(0.584)
Match	-0.710**	-0.959**	-0.994**	-1.274***	-1.661**
	(0.308)	(0.457)	(0.468)	(0.485)	(0.700)
Inclusive	-0.653**	-0.336	-0.315	-0.450	-0.430
	(0.315)	(0.431)	(0.435)	(0.458)	(0.605)
Norm × Match		0.379	0.387	0.463	0.517
		(0.642)	(0.668)	(0.689)	(0.883)
Norm × Inclusive		-0.593	-0.533	-0.534	-0.515
		(0.628)	(0.634)	(0.654)	(0.821)
Age			-0.047	-0.054	-0.132**
			(0.048)	(0.050)	(0.057)
Semester			0.003	0.005	0.082*
			(0.035)	(0.035)	(0.048)
Business and Economics			-0.412	-0.461	-0.777*
			(0.312)	(0.331)	(0.415)
Education			0.138	0.039	1.055**
			(0.377)	(0.410)	(0.477)
Language attitude				0.247**	0.043
				(0.116)	(0.156)
Remembered formulations				-0.260	-0.699*
				(0.308)	(0.371)
Language comments				0.535	0.362
				(0.508)	(0.639)
Instructions clear				0.020	-0.112
				(0.084)	(0.108)
Risk aversion					-0.068
					(0.076)
First-order belief$_{DG}$					0.044***
					(0.012)
Second-order belief$_{DG}$					0.005
					(0.011)
Constant	0.371	0.336	1.634	1.237	1.656
	(0.250)	(0.295)	(1.164)	(1.319)	(1.416)
Pseudo R^2	0.063	0.081	0.105	0.141	0.450
Observations	109	109	109	109	109

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: Only two men in the Norm-Match treatment failed the control question for this game. Thus, in deviation to our previous description of the specifications we omit *Failed attempts*$_{DG}$ from the specifications reported in columns (4) and (5).

Table 14: Probit regressions on men's compliance with the 50-50 sharing norm (complete table with all coefficients).

Dep. Var.: Compliance$_{DG}$	(1)	(2)	(3)	(4)	(5)
Norm	0.365	0.189	0.327	0.322	-0.030
	(0.227)	(0.402)	(0.415)	(0.437)	(0.458)
Match	-0.244	-0.304	-0.196	-0.154	0.091
	(0.280)	(0.437)	(0.455)	(0.459)	(0.442)
Inclusive	-0.204	-0.443	-0.265	-0.480	-0.665
	(0.273)	(0.426)	(0.444)	(0.466)	(0.518)
Norm $\times$ Match		0.096	-0.189	-0.247	-0.681
		(0.567)	(0.590)	(0.615)	(0.598)
Norm $\times$ Inclusive		0.413	0.258	0.366	0.474
		(0.557)	(0.572)	(0.607)	(0.692)
Age			0.054*	0.081**	0.107**
			(0.028)	(0.039)	(0.050)
Semester			-0.042	-0.050	-0.071*
			(0.032)	(0.037)	(0.042)
Business and Economics			-0.017	-0.025	-0.013
			(0.326)	(0.324)	(0.350)
Education			0.586*	0.558*	0.589*
			(0.328)	(0.328)	(0.332)
Language attitude				0.137	0.231*
				(0.112)	(0.123)
Remembered formulations				0.038	0.107
				(0.259)	(0.269)
Language comments				0.435	0.275
				(0.489)	(0.548)
Instructions clear				-0.037	-0.134
				(0.088)	(0.107)
Failed attempts$_{DG}$				-1.088	-1.662**
				(0.729)	(0.826)
Risk aversion					0.082
					(0.060)
First-order belief$_{DG}$					0.044***
					(0.011)
Second-order belief$_{DG}$					-0.018
					(0.011)
Constant	0.271	0.377	-0.915	-1.816	-3.904**
	(0.238)	(0.313)	(0.825)	(1.264)	(1.518)
Pseudo R^2	0.020	0.024	0.080	0.103	0.335
Observations	133	133	133	133	133

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 15: Probit regressions on women's compliance with the 50-50 sharing norm (complete table with all coefficients).

Dep. Var.: Compliance$_{PD}$	(1)	(2)	(3)	(4)	(5)
Norm	-0.124	0.242	0.291	0.436	-0.127
	(0.264)	(0.451)	(0.467)	(0.517)	(0.776)
Match	0.025	0.143	0.201	0.391	-0.375
	(0.325)	(0.461)	(0.511)	(0.520)	(0.808)
Inclusive	-0.037	0.408	0.435	0.489	0.278
	(0.318)	(0.473)	(0.509)	(0.559)	(0.592)
Norm × Match		-0.275	-0.538	-0.881	-0.635
		(0.653)	(0.738)	(0.735)	(1.212)
Norm × Inclusive		-0.847	-1.046	-1.189	-1.281
		(0.654)	(0.687)	(0.736)	(1.110)
Age			-0.098**	-0.131***	-0.235***
			(0.047)	(0.049)	(0.067)
Semester			0.051	0.056	0.185***
			(0.038)	(0.042)	(0.052)
Business and Economics			-0.465	-0.513	-1.462***
			(0.351)	(0.381)	(0.476)
Education			-0.705*	-0.955**	-1.412**
			(0.387)	(0.441)	(0.559)
Language attitude				0.235*	0.667**
				(0.136)	(0.276)
Remembered formulations				0.258	0.006
				(0.320)	(0.447)
Language comments				0.004	1.807**
				(0.496)	(0.843)
Instructions clear				0.174**	0.095
				(0.089)	(0.133)
Failed attempts$_{PD}$				0.057	0.081*
				(0.065)	(0.043)
Risk aversion					0.015
					(0.113)
Positive reciprocity					0.550
					(0.355)
Negative reciprocity					-0.264
					(0.166)
First-order belief$_{PD}$					0.059***
					(0.021)
Second-order belief$_{PD}$					0.012
					(0.022)
Constant	0.647**	0.480	2.945**	1.952	-2.619
	(0.264)	(0.301)	(1.165)	(1.284)	(2.919)
Pseudo R^2	0.002	0.017	0.088	0.160	0.607
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

Table 16: Probit regressions on men's compliance with the cooperation norm (complete table with all coefficients).

Dep. Var.: Compliance$_{PD}$	(1)	(2)	(3)	(4)	(5)
Norm	0.187	0.300	0.460	0.674	1.050*
	(0.258)	(0.455)	(0.486)	(0.537)	(0.597)
Match	0.034	0.000	0.022	0.166	0.334
	(0.328)	(0.456)	(0.499)	(0.492)	(0.556)
Inclusive	-0.519*	-0.342	-0.181	-0.065	0.171
	(0.309)	(0.434)	(0.453)	(0.470)	(0.516)
Norm × Match		0.087	0.180	-0.176	-0.264
		(0.663)	(0.705)	(0.735)	(0.766)
Norm × Inclusive		-0.360	-0.570	-0.852	-1.469*
		(0.617)	(0.638)	(0.703)	(0.779)
Age			0.005	0.020	0.033
			(0.031)	(0.035)	(0.045)
Semester			0.032	0.030	0.029
			(0.037)	(0.039)	(0.040)
Business and Economics			-1.060***	-1.096***	-1.067**
			(0.396)	(0.387)	(0.423)
Education			-0.546	-0.633	-0.729
			(0.393)	(0.391)	(0.446)
Language attitude				0.111	0.087
				(0.117)	(0.132)
Remembered formulations				0.371	0.728**
				(0.295)	(0.316)
Language comments				-0.173	-0.633
				(0.418)	(0.464)
Instructions clear				0.125	0.126
				(0.102)	(0.114)
Failed attempts$_{PD}$				0.095	0.115
				(0.168)	(0.219)
Risk aversion					0.025
					(0.067)
Positive reciprocity					0.141
					(0.232)
Negative reciprocity					-0.423***
					(0.119)
First-order belief$_{PD}$					0.009
					(0.011)
Second-order belief$_{PD}$					0.002
					(0.011)
Constant	0.599**	0.541*	0.727	-1.051	-2.083
	(0.265)	(0.322)	(1.042)	(1.485)	(2.127)
Pseudo R^2	0.035	0.040	0.109	0.143	0.273
Observations	108	108	108	108	108

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 17: Probit regressions on women's compliance with the cooperation norm (complete table with all coefficients).

Dep. Var.: Compliance$_{Dec}$	(1)	(2)	(3)	(4)	(5)
Norm	0.480*	0.365	0.415	0.726	0.458
	(0.247)	(0.443)	(0.442)	(0.481)	(0.631)
Match	-0.080	-0.150	-0.105	0.058	-0.535
	(0.305)	(0.435)	(0.457)	(0.497)	(0.621)
Inclusive	-0.296	-0.385	-0.413	-0.406	-1.224**
	(0.305)	(0.432)	(0.445)	(0.477)	(0.553)
Norm × Match		0.150	-0.005	-0.247	-0.282
		(0.611)	(0.645)	(0.683)	(0.829)
Norm × Inclusive		0.184	0.195	0.127	0.831
		(0.611)	(0.618)	(0.653)	(0.847)
Age			-0.039	-0.035	0.013
			(0.040)	(0.043)	(0.048)
Semester			-0.010	0.012	0.008
			(0.033)	(0.035)	(0.044)
Business and Economics			-0.565*	-0.399	-0.382
			(0.302)	(0.333)	(0.437)
Education			-0.584	-0.669*	-0.218
			(0.362)	(0.386)	(0.427)
Language attitude				0.206*	0.272*
				(0.118)	(0.144)
Remembered formulations				0.241	0.081
				(0.313)	(0.389)
Language comments				1.602***	2.597**
				(0.605)	(1.040)
Instructions clear				0.220**	0.163
				(0.090)	(0.114)
Failed attempts$_{Dec}$				0.279*	0.204
				(0.164)	(0.214)
Risk aversion					0.005
					(0.072)
First-order belief$_{Dec}$					0.041***
					(0.013)
Second-order belief$_{Dec}$					-0.005
					(0.014)
Constant	0.017	0.066	1.504	-1.180	-3.840**
	(0.244)	(0.289)	(0.993)	(1.337)	(1.619)
Pseudo R^2	0.031	0.031	0.063	0.160	0.447
Observations	111	111	111	111	111

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 18: Probit regressions on men's compliance with the honesty norm (complete table with all coefficients).

Dep. Var.: Compliance$_{Dec}$	(1)	(2)	(3)	(4)	(5)
Norm	0.404*	0.748*	0.769*	0.839**	0.829*
	(0.228)	(0.400)	(0.402)	(0.413)	(0.462)
Match	-0.086	0.451	0.441	0.518	0.317
	(0.275)	(0.437)	(0.444)	(0.461)	(0.554)
Inclusive	-0.350	-0.262	-0.224	-0.358	0.069
	(0.273)	(0.424)	(0.435)	(0.461)	(0.611)
Norm × Match		-0.891	-0.844	-0.986	-0.650
		(0.564)	(0.576)	(0.600)	(0.693)
Norm × Inclusive		-0.136	-0.182	-0.203	-0.943
		(0.564)	(0.568)	(0.587)	(0.732)
Age			-0.013	-0.014	-0.057***
			(0.023)	(0.022)	(0.022)
Semester			0.018	0.028	-0.004
			(0.029)	(0.029)	(0.041)
Business and Economics			-0.305	-0.355	-0.828*
			(0.352)	(0.369)	(0.502)
Education			-0.046	-0.090	-0.500
			(0.337)	(0.347)	(0.480)
Language attitude				0.120	0.057
				(0.102)	(0.131)
Remembered formulations				0.172	0.479
				(0.261)	(0.312)
Language comments				-0.130	0.008
				(0.459)	(0.400)
Instructions clear				-0.055	-0.126
				(0.084)	(0.106)
Failed attempts$_{Dec}$				0.105	-0.077
				(0.169)	(0.198)
Risk aversion					-0.080
					(0.065)
First-order belief$_{Dec}$					0.038***
					(0.008)
Second-order belief$_{Dec}$					0.003
					(0.009)
Constant	0.129	-0.074	0.219	-0.105	0.103
	(0.237)	(0.305)	(0.769)	(1.056)	(1.204)
Pseudo R^2	0.029	0.046	0.057	0.071	0.436
Observations	131	131	131	131	131

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 19: Probit regressions on women's compliance with the honesty norm (complete table with all coefficients).

Dep. Var.: Compliance	all	DG	PD	Dec
Appropriateness$_{all}$	-0.076			
	(0.092)			
Appropriateness$_{DG}$		-0.289		
		(0.398)		
Appropriateness$_{PD}$			0.831**	
			(0.418)	
Appropriateness$_{Dec}$				0.994*
				(0.526)
Norm	-0.128	0.370	-0.134	0.613
	(0.148)	(0.524)	(0.701)	(0.605)
Match	-0.388**	-1.364**	-0.487	-0.618
	(0.184)	(0.654)	(0.693)	(0.640)
Inclusive	-0.361**	-0.454	0.069	-1.398**
	(0.182)	(0.617)	(0.546)	(0.578)
Norm $\times$ Match	0.208	0.810	-0.391	-0.544
	(0.217)	(0.833)	(1.038)	(0.814)
Norm $\times$ Inclusive	0.077	-0.235	-1.615*	0.671
	(0.217)	(0.766)	(0.868)	(0.798)
Age	-0.031*	-0.112**	-0.169***	0.028
	(0.016)	(0.048)	(0.053)	(0.049)
Semester	0.022*	0.042	0.152***	0.000
	(0.013)	(0.043)	(0.047)	(0.045)
Business and Economics	-0.191*	-0.759*	-0.751	-0.435
	(0.112)	(0.404)	(0.510)	(0.437)
Education	-0.022	0.506	-1.170**	-0.092
	(0.141)	(0.437)	(0.583)	(0.426)
Language attitude	0.039	0.005	0.448**	0.336**
	(0.039)	(0.146)	(0.199)	(0.146)
Remembered formulations	-0.033	-0.531	0.155	0.159
	(0.099)	(0.341)	(0.392)	(0.393)
Language comments	0.043	0.119	0.307	2.409**
	(0.157)	(0.626)	(0.624)	(0.947)
Instructions clear	-0.025	-0.125	0.050	0.141
	(0.032)	(0.100)	(0.102)	(0.114)
Pseudo R^2	0.141	0.447	0.536	0.481
Observations	122	122	122	122

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: Since there were only few participants who failed the control question for the dictator game, we omit *Failed attempts*$_{DG}$ from the specification reported in column DG.

Table 20a: Poisson regressions on how many norms men complied to across games controlling for whether they rated all norms rather appropriate or not (column 1) and probit regressions on men's norm compliance in the individual games controlling for whether they rated the respective norm rather appropriate or not (columns 2 to 4) using the saturated specification (complete table with all coefficients, see Table 20b for the remaining coefficients).

Dep. Var.: Compliance	all	DG	PD	Dec
Failed attempts$_{all}$	0.034***			
	(0.011)			
Failed attempts$_{PD}$			0.169	
			(0.291)	
Failed attempts$_{Dec}$				0.295
				(0.195)
Risk Aversion	-0.025	-0.116	-0.062	0.008
	(0.020)	(0.071)	(0.084)	(0.071)
Positive reciprocity	0.110*		0.674**	
	(0.066)		(0.275)	
Negative reciprocity	-0.029		-0.360**	
	(0.031)		(0.170)	
First-order belief$_{DG}$	0.006**	0.041***		
	(0.003)	(0.012)		
Second-order belief$_{DG}$	-0.000	0.009		
	(0.004)	(0.011)		
First-order belief$_{PD}$	0.013***		0.035***	
	(0.005)		(0.013)	
Second-order belief$_{PD}$	-0.005		0.015	
	(0.005)		(0.017)	
First-order belief$_{Dec}$	0.006			0.044***
	(0.005)			(0.014)
Second-order belief$_{Dec}$	0.001			-0.005
	(0.006)			(0.014)
Constant	-0.267	2.070	-3.049	-5.389***
	(0.659)	(1.426)	(2.322)	(2.040)
Pseudo R^2	0.141	0.447	0.536	0.481
Observations	122	122	122	122

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: Since there were only few participants who failed the control question for the dictator game, we omit *Failed attempts*$_{DG}$ from the specification reported in column DG.

Table 20b: Poisson regressions on how many norms men complied to across games controlling for whether they rated all norms rather appropriate or not (column 1) and probit regressions on men's norm compliance in the individual games controlling for whether they rated the respective norm rather appropriate or not (columns 2 to 4) using the saturated specification (complete table with all coefficients, continued from Table 20a).

Dep. Var.: Compliance	all	DG	PD	Dec
Appropriateness$_{all}$	0.127			
	(0.078)			
Appropriateness$_{DG}$		-0.224		
		(0.347)		
Appropriateness$_{PD}$			0.669**	
			(0.265)	
Appropriateness$_{Dec}$				-0.466
				(0.354)
Norm	0.209	0.053	0.688	0.754*
	(0.152)	(0.426)	(0.473)	(0.452)
Match	0.136	0.014	0.705	0.315
	(0.181)	(0.434)	(0.580)	(0.554)
Inclusive	0.040	-0.617	0.116	-0.016
	(0.237)	(0.487)	(0.490)	(0.615)
Norm $\times$ Match	-0.242	-0.745	-0.668	-0.673
	(0.202)	(0.565)	(0.666)	(0.682)
Norm $\times$ Inclusive	-0.370	0.336	-0.891	-0.557
	(0.276)	(0.644)	(0.643)	(0.713)
Age	0.005	0.074*	-0.001	-0.058***
	(0.007)	(0.040)	(0.025)	(0.022)
Semester	-0.003	-0.057	0.036	-0.012
	(0.011)	(0.037)	(0.031)	(0.039)
Business and Economics	-0.142	0.001	-0.562*	-0.591
	(0.135)	(0.339)	(0.331)	(0.419)
Education	-0.021	0.518	-0.196	-0.309
	(0.106)	(0.323)	(0.341)	(0.401)
Language attitude	0.027	0.249**	-0.024	0.040
	(0.037)	(0.119)	(0.110)	(0.127)
Remembered formulations	0.171*	0.159	0.554**	0.346
	(0.088)	(0.260)	(0.254)	(0.295)
Language comments	0.065	0.343	-0.142	0.219
	(0.167)	(0.507)	(0.461)	(0.431)
Instructions clear	-0.032	-0.097	0.068	-0.129
	(0.034)	(0.095)	(0.091)	(0.104)
Pseudo R^2	0.094	0.324	0.236	0.448
Observations	147	147	147	147

vce(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: Since there were only few participants who failed the control question for the dictator game, we omit *Failed attempts*$_{DG}$ from the specification reported in column DG.

Table 21a: Poisson regressions on how many norms women complied to across games controlling for whether they rated all norms rather appropriate or not (column 1) and probit regressions on women's norm compliance in the individual games controlling for whether they rated the respective norm rather appropriate or not (columns 2 to 4) using the saturated specification (complete table with all coefficients, see Table 21b for the remaining coefficients).

Dep. Var.: Compliance	all	DG	PD	Dec
Failed attempts$_{all}$	-0.059***			
	(0.020)			
Failed attempts$_{PD}$			-0.173**	
			(0.082)	
Failed attempts$_{Dec}$				-0.103
				(0.197)
Risk Aversion	0.008	0.056	0.028	-0.099
	(0.017)	(0.057)	(0.053)	(0.063)
Positive reciprocity	0.080		0.048	
	(0.082)		(0.204)	
Negative reciprocity	-0.121***		-0.367***	
	(0.036)		(0.091)	
First-order belief$_{DG}$	0.012***	0.043***		
	(0.003)	(0.010)		
Second-order belief$_{DG}$	-0.008***	-0.015		
	(0.003)	(0.010)		
First-order belief$_{PD}$	0.005		0.016*	
	(0.004)		(0.010)	
Second-order belief$_{PD}$	-0.006		-0.005	
	(0.004)		(0.010)	
First-order belief$_{Dec}$	0.002			0.040***
	(0.003)			(0.008)
Second-order belief$_{Dec}$	0.005			0.005
	(0.003)			(0.009)
Constant	-0.424	-3.395**	-0.917	0.443
	(0.740)	(1.531)	(1.670)	(1.098)
Pseudo R^2	0.094	0.324	0.236	0.448
Observations	147	147	147	147

vce(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: Since there were only few participants who failed the control question for the dictator game, we omit *Failed attempts*$_{DG}$ from the specification reported in column DG.

Table 21b: Poisson regressions on how many norms women complied to across games controlling for whether they rated all norms rather appropriate or not (column 1) and probit regressions on women's norm compliance in the individual games controlling for whether they rated the respective norm rather appropriate or not (columns 2 to 4) using the saturated specification (complete table with all coefficients, continued from Table 21a).