Dimant, Eugen

Working Paper

Hate Trumps Love: The Impact of Political Polarization on Social Preferences

CESifo Working Paper, No. 9073

Provided in Cooperation with:
Ifo Institute – Leibniz Institute for Economic Research at the University of Munich

Suggested Citation: Dimant, Eugen (2021) : Hate Trumps Love: The Impact of Political Polarization on Social Preferences, CESifo Working Paper, No. 9073, Center for Economic Studies and Ifo Institute (CESifo), Munich

This Version is available at:
https://hdl.handle.net/10419/235443

CESifo WORKING PAPERS

9073
2021

May 2021

Hate Trumps Love: The Impact of Political Polarization on Social Preferences

Eugen Dimant

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. 9073*

Hate Trumps Love: The Impact of Political Polarization on Social Preferences

Abstract

Political polarization has ruptured the fabric of U.S. society. I quantify this phenomenon through the use of 5 pre-registered studies, comprising 15 behavioral experiments and a diverse set of over 8,600 participants. The focus of this paper is to examine various behavioral-, belief-, and norm-based layers of (non-)strategic decision-making that are plausibly affected by existing polarization in the context of Donald J. Trump. I find strong heterogeneous effects: ingroup-love occurs in the *perceptional* domain (how close one feels towards others), whereas outgroup-hate occurs in the *behavioral* domain (how one helps/harms/cooperates with others). The rich setting also allows me to examine the mechanisms of observed intergroup conflict, which can be attributed to one's grim expectations regarding cooperativeness of the opposing faction, rather than one's actual unwillingness to cooperate. In a final step, I test whether popular behavioral interventions (defaults and norm-nudging) can eradicate the detrimental impact of polarization in the (non-) strategic contexts studied here. The interventions are ineffective in closing the polarization gap, suggesting that structural – on top of behavioral - changes are needed to mend existing fractions and heal the society.

JEL-Codes: C900, D010, D900.

Keywords: identity, norms, nudging, polarization, social preferences.

Eugen Dimant
Center for Social Norms and Behavioral Dynamics
University of Pennsylvania / Philadelphia / USA
edimant@sas.upenn.edu

This version: May 4, 2021
The most recent version of this working paper can be downloaded by following this link:
https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3680871
I am indebted to Simon Gächter, Daniele Nosenzo, and Silvia Sonderegger for their input on the initial project idea. I have benefited from conversations with Abraham Aldama, Loukas Balafoutas, John Beshears, Levi Boxell, Yan Chen, Danny Choi, Raymond Duch, Christina Gravert, Felix Kölle, Jeff Lees, Matthew Levendusky, Brendan Nyhan, Kyle Peyton, Sean Westwood, Rick Wilson, and Noam Yuchtman. Input received at conferences (2021 ASSA Annual Meeting, 2021 CESifo Area Conference on Public Economics, 2021 NYU CESS Experimental Political Science Conference, 2020 Global Economic Science Association (ESA) Meeting, FAIR Workshop: 'Fairness & the Moral Mind' in Bergen) and invited seminars (Harvard University, Max Planck Institute for Research on Collective Goods, University of Innsbruck, University of Michigan, University of Oxford) is greatly appreciated. I thank Riley Kennedy for excellent research assistance. The four-part data collection and analyses were pre-registered at AsPredicted.org (#42538-40, #52736-7, #52968-9, #55920-3: https://osf.io/auh4k/). I acknowledge financial support from the German Research Foundation (DFG) under Germany's Excellence Strategy - EXC 2126/1 – 390838866.

1. Introduction

Rising political polarization is often linked to fractured societies rife with racial inequality, factional conflict, and partisan animosity (Dixit and Weibull, 2007; Fiorina and Abrams, 2008; Iyengar and Westwood, 2015; Bénabou and Tirole, 2016; Reich, 2017; Autor et al., 2020; Bursztyn et al., 2020a; Graham and Svolik, 2020). At its core, polarization undermines social contracts that are necessary for a functioning society. It restrains social interactions across polarized clusters by impeding cooperativeness, trust, and altruism between political factions. It poses a credible threat to democratic values and is amplified by *false polarization*: the perception of more polarization regarding policy issues than actually exists (Levendusky and Malhotra, 2016; Moore-Berg et al., 2020).[1]

Political polarization yields direct social welfare implications in that it may affect both one's willingness to engage in altruistic behavior and the collective provision of goods within and between factions (Henrich et al., 2001; Fehr and Fischbacher, 2003; Bowles and Gintis, 2013). Arguably, not all consequences of political polarization are created equally, and thus cannot be tackled with the same policies. This warrants going beyond existing research to offer a trifecta approach – as put forward in this paper – towards examining the manifestation of polarization across beliefs, behaviors, and attitudes separately. Through the lens of one's feelings of *hate* and *love* for Donald J. Trump, I experimentally examine the mechanisms by which political intergroup conflict materializes in both strategic and non-strategic decision contexts that capture cooperativeness, altruism, and anti-social behavior.

On a theoretical level, the aggregate social consequences of polarization are unclear. Polarization (and the resulting hostile climate) can produce enough outgroup animosity to reduce individual willingness to support and cooperate with members of the opposite faction. It may also – or instead – increase intra-faction cooperation, for example, through promoting a sense of shared identity. To assess the social impact of polarization, it is thus important to compare within- and between-group behaviors in a polarized environment. Such an approach enables both strategic and non-strategic considerations. Rather than resorting to surveys (as is often done in related research, e.g. Flynn et al., 2017; Ahler and Sood, 2018; Stanley et al., 2020; Ruggeri et al., 2021), I quantify these phenomena via controlled and incentive-compatible behavioral experiments.

Across 5 preregistered studies that contain 15 incentive-compatible experiments and a diverse set of over 8,600 individuals, I approach these policy-relevant questions from several angles: non-strategic decisions, strategic decisions, nudge interventions, and norm percep-

[1] Recent examples include the divide over wearing face masks as preventative measures from COVID-19 infections: https://www.nytimes.com/interactive/2020/07/17/upshot/coronavirus-face-mask-map.html.

tions. With that, I attempt to provide a comprehensive examination of the forms in which polarization occurs, how they vary across those domains, and how to alleviate their consequences. I focus on two aspects of decision-making within and across political factions. The first is **behavioral**: How does polarization in the context of political identities affect pro-/anti-social decisions, cooperation, and social expectations in strategic and non-strategic environments? Do these take shape in the form of ingroup-love, outgroup-hate, or both?[2] The second is **perceptional**: Are these behavioral differences consistent with the observed variations in perceived interpersonal closeness and social norms and, with that, can they explain *why* we observe such affective polarization?

As such, my investigative approach is consistent with and speaks to the growing discussion on affective polarization – the animosity between and distrust towards members of the opposing faction (Druckman and Levendusky, 2019; Iyengar et al., 2019). On top of that, I am also able to quantify *behavioral* polarization. Although partisanship is found to exacerbate cross-faction discrimination, prior work has often attributed this behavior to a mix of ingroup-love and outgroup-hate by pitting one group against another.[3] The results from the experiments presented here show that these are not necessarily two sides of the same coin, but rather that ingroup-love and outgroup-hate can coexist independently from one another depending on the decision-making environment.

I introduce disparate feelings of polarization by using a participant's repugnance against (henceforth referred to as *hate*) or admiration for (henceforth referred to as *love*) the 45$^{\text{th}}$ president: Donald J. Trump. This is a particularly expedient setting since Trump's actions during his presidency has been linked to increased social divergence and hate-related consequences.[4] Donald Trump is a polarizing figure and the current symbol of the Republican party (Jacobson, 2019), which is captured by the measures that I put forward here. By comparing these differences in political identities to differences in minimal group identities, I also contribute to the *groupy* behavior (Kranton and Sanders, 2017) literature in that the impact of partisan animosity on ingroup-love and outgroup-hate can be examined separately. To tease out the mechanisms of ingroup-love/outgroup-hate, I also run the same and all exper-

[2]Humans are known to bond over common identity markers (often exemplified by one's preference *for* something or someone), the study of which has its origins in the research on ingroup-outgroup favoritism and social identity (Tajfel and Turner, 1979; Alesina et al., 1999; Akerlof and Kranton, 2000; Bernhard et al., 2006; Efferson et al., 2008; Halevy et al., 2008; Bénabou and Tirole, 2011; Chen and Li, 2009; Bénabou et al., 2018; Dimant, 2019; Charness and Chen, 2020).

[3]Throughout the paper, I will use *ingroup-love* to indicate bias in favor of one's own group, and *outgroup-hate* to indicate prejudice against the opposing group (Greene, 1999; Abramowitz and Saunders, 2006; Mason, 2015; Michelitch, 2015; Amira et al., 2019; Orr and Huber, 2020; West and Iyengar, 2020).

[4]See, e.g., Abramowitz and Webster, 2018; Mason, 2018; Klein, 2020; Müller and Schwarz, 2020. The consequences of hate are conspicuous and often erupt in form of social movements and protests (Meyer, 2004; Madestam et al., 2013; Mazumder, 2018; Cantoni et al., 2019).

imental conditions with a separate set of participants using a less polarizing minimal group identity (following Tajfel and Turner, 1979; Chen and Li, 2009), where one's preferences for Klee or Kandinsky paintings are the identity markers. From a methodological perspective, comparing these two identities[5] allows me to apply a difference-in-difference approach to the study of polarization, which is comparable to the within-subject design comparison method of Kranton et al. (2020). Doing so allows me to quantify polarization by disentangling how beliefs, preferences, and norms drive polarization both separately and collectively.

Across all experiments, the results highlight that partisan animosity evokes a state that affects all measured elements in both strategic and non-strategic settings. In particular, when comparing the results between the *Trump prime* and *Minimal Group prime* treatments I find that ingroup-love only occurs in the context of how one perceives *interpersonal closeness* to others. Conversely, outgroup-hate is manifested in one's reduced *altruism* and *cooperativeness* with the opposing faction, as well as in the form of pessimistic beliefs about the opposing faction's cooperativeness. This confirms that the results are not driven by ingroup-outgroup considerations alone, but rather that the observed disparities in perceptions, beliefs, and cooperativeness are evoked by polarization. In short, my findings in the strategic context support the conclusion that the reason people choose to cooperate less with outgroup members is because they (incorrectly) expect those people to not want to cooperate with them. With that, the contribution and main takeaway of this paper is that partisan identity not only drives costly social behavior, in part due to pessimistic beliefs, but it also comports with the perceived social norms elicited here.

Against this backdrop, while my findings indicate that the impact of polarization can be picked up across all studied (perceptional and behavioral) measures, the results emphasize the nuanced composition of polarization: the partisan rift is deep but this paper helps better understand why. In the contexts studied here, the adverse behavioral impact of intergroup conflict can be attributed to one's grim expectations about the cooperativeness of the opposing faction rather than one's categorical unwillingness to cooperate with them. Importantly, the tested behavioral interventions suggest that alleviating these negative effects is not straightforward. The examined nudges – while shifting the average level of pro-social behavior – show little success in reducing polarization as a whole, leaving the gap intact. From a policy perspective, however, it is evident that both structural and institutional changes

[5]I utilize the comparison with the minimal group identity because this is the most commonly used used identity measure across social sciences (Charness and Chen, 2020). An alternative approach is to compare the behavior in the 'hate'/'love' conditions of the Trump prime to an 'unknown' condition of the Trump prime in which the identity of the matched partner is not revealed. I have collected such data for all presented experiments and discuss their results in the Online Appendix. I remain agnostic about the *correct* comparison. Rather, I made the choice that for the purpose of this paper's story the most illuminating comparison is the one between non-minimal identities (such as Trump preferences) and minimal identities.

need to be introduced in conjunction with behavioral interventions to eliminate – or at least reduce – the detrimental impact of polarization. This demonstrates the limits of light-touch behavioral interventions, which may be insufficient in reducing the current state of affective and behavioral polarization in the United States. To the best of my knowledge, this paper is the first to target the reduction of polarization using nudges.

The experimental investigation and analyses put forth in this paper can be subdivided into several steps that logically build upon each other across the two subsequent sections. Section 3 details two studies that act as the main focal point in my examination of the research question: Study 1 (3.1), and Study 2 (3.2). I examine some additional behavioral mechanisms of these main studies in Study 3 (3.3). In Section 4, I present a series of robustness checks to bolster my findings in the core experimental investigation through Study 4 (4.1) and Study 5 (4.2). Section 5 concludes.

2. Experimental Overview

Across 15 pre-registered experiments nested within 5 distinct studies, experimental data was collected from a total of $n = 8,647$ participants between summer of 2020 and early 2021.[6] Detailed breakdowns of the data collection and analyses are indicated in Table 1.

Studies	**Conditions**	**DG** (all)	**DG** (after dropping according to pre-reg criteria)	**DG** (analyzed in the main text)	**DG** (data for additional analyses in appendix)	**PGG** (all)	**PGG** (after dropping according to pre-reg criteria)	**PGG** (analyzed in the main text)	**PGG** (data for additional analyses in appendix)
	Core Experimental Studies								
1 & 2	Trump Prime	738	588	417	171	648	517	375	142
	Minimal Group Prime	650	574	384	190	612	499	343	156
	Default Nudge	530	424	310	114	714	490	337	153
	Information Nudge	526	464	383	81	566	442	341	101
	Robustness Checks Studies								
3	Minimal Group Prime Order Change	388	350	350	-	420	336	336	-
4	Norm Elicitation	298	232	-	232	298	232	-	232
5	Biden Prime	661	458	333	125	706	406	299	107
	Sports Prime	450	405	321	84	442	361	277	84
	Sum	**4241**	**3495**	**2498**	**997**	**4406**	**3283**	**2308**	**975**

Table 1: Number of observations across all studies (the norm elicitation experiment is accounted for once because this is the only within-participant design: participants saw both the DG and PGG setting). Numbers for the DG only reflect the data collected for dictators. Data was dropped from the analyses according to the pre-registration protocols (failed attention/comprehension checks and a participant's indifference towards Donald Trump / Joe Biden / sports). Exclusion of participants is uncorrelated with the treatments and the presented results are not sensitive to the inclusion of these participants (available upon request). Each section details the exact data-handling procedure.

[6]Since MTurk is known to be liberal-leaning, I over-sampled in order to collect enough data for the Trump lovers and Biden haters, respectively. As correctly anticipated, those who indicated to love Trump appeared in the data about $\frac{1}{3}$ of the time. I calibrate the required sample size to obtain high statistical power based on a classroom pre-test that yielded an effect size of 0.54. Consequently, the power calculations yielded that 50 participants per cell are needed in order to achieve 80% at an alpha of 0.05. To ensure highest quality data collection on MTurk, I utilized a combination of CAPTCHAs and screening questions to avoid pool contamination. As per pre-registration, research assistants independently rated the response quality to open questions. I applied the following restrictions to the participant pool: U.S.-based, approval rate greater than 95%, and could participate only once in any of the three experiments presented in this paper. This corresponds to the recommended best practices to maximize data quality (Buhrmester et al., 2018). Any residual noise is accounted for by the high-powered sample and would not affect any treatment comparisons.

I begin with an examination of the affective and behavioral effects of polarization on decision-making through Study 1 (3.1) and Study 2 (3.2). By using two complementary studies, I aim to shed light on the nuances of in-group love and outgroup-hate within different contexts. With Study 1, I examine how one's identity shapes altruistic preferences in a non-strategic setting through an extended dictator game to which I refer to as the *Take-or-Give (T-o-G) Dictator Game*. In Study 2, I explore the extent to which one's identity impacts effective cooperation, beliefs, and attitudes in a non-strategic setting through an extended public goods game based on the 'Attitudes-Beliefs-Contributions (ABC) of cooperation' approach as introduced by Fischbacher et al. (2001) and Gächter et al. (2017), appropriately titled to be the *ABC of Cooperation Public Goods Game*. Note that for the purpose of brevity, "T-o-G Dictator Game" and the "ABC of Cooperation Public Goods Game" will be referred to throughout this paper as "DG" and "PGG", respectively.

In both studies, I compare the impact of the political identity markers with that of a minimal group prime to distinguish between ingroup-love and outgroup-hate. After establishing the severity of polarization, I ask and answer a crucial and policy-relevant question: can we utilize simple and cost-effective behavioral interventions, which have proven successful across various settings, to reduce the pernicious impact of political polarization in the contexts studied here? To do so, I leverage the power of nudging – particularly that of *defaults* and *norms*. I test these interventions in both the non-strategic and strategic contexts. To the best of my knowledge, my paper is the first to examine whether nudges have enough potency to reduce polarization. In a final step, I provide additional insights in the robustness and mechanisms of the studied behaviors in Study 3 (3.3).

Section 4 is devoted to examining the social dimension and robustness of the observed polarization found in Study 1 and Study 2. I assess the extent to which the observed results are consistent with social norm perceptions in Study 4 (4.1). Study 5 (4.2) tests an alternative polarizing political prime (Joseph R. Biden) and an alternative polarizing *non*-political prime (sports). Section 5 concludes with a discussion of the effects of political polarization on human behavior and suggestions for further research.

3. Core Experimental Analysis (Studies 1, 2 & 3)

In what follows, I experimentally examine the extent to which political polarization manifests itself within *non-strategic* (Study 1, Section 3.1) and *strategic* (Study 2, Section 3.2) settings. My examination of the effects of political polarization is multi-faceted in that I aim to not only measure its impact within different decision-making contexts, but also to explore opportunities to mitigate its negative effects within those contexts. Thus, the two main experiments at the core of my experimental investigation are each broken down into two parts: (1) measuring polarization and (2) reducing polarization via nudging.

Studies 1 and 2 each begin with a controlled experiment designed to provide a quantifiable measure of the impact polarization along the dimensions of human behavior and perceptions of closeness. In particular, I focus on measuring altruistic behavior with the non-strategic context, and cooperation with the strategic context. The ultimate objective of both experiments being to contribute a more nuanced understanding of the different facets of political intergroup conflict. Each study then features a unique experimental investigation into reducing the negative effects of polarization. My goal is to determine whether nudge interventions, as popularized by Thaler and Sunstein (2008), have enough potency to reduce the observed polarization and subsequently mitigate the observed detrimental impact on altruism and cooperativeness.

To this end, I examined the upper-bound of what a best-case scenario looks like and capitalize on two types of nudges that the literature has identified as the most effective widely-used behavioral interventions: a *Default Nudge* ($n = 1,244$) and a *Descriptive Norm-Nudge* ($n = 1,092$).[7] Existing experimental literature has highlighted the effectiveness of these interventions across various settings that are related to the contexts studied here (e.g., charitable giving and conformity; see Goswami and Urminsky, 2016; Benartzi et al., 2017; Altmann et al., 2019; Bicchieri and Dimant, 2019). The nudge interventions precisely follow the experimental procedures for Part 1 of each study, with the nudges introduced at the decision point in each respective study.

3.1. STUDY 1: POLARIZATION IN A NON-STRATEGIC CONTEXT

Take-or-Give Dictator Game

Study 1 examines the impact of polarization on both preferences and behavior in a *non-strategic context*. I utilize the *Take-or-Give Dictator Game* variation to measure the extent of this impact on perceptions of closeness, pro-social (giving) behavior, and anti-social (taking) behavior. By employing a context in which strategic motives are eliminated by design, the results will provide insight into how one's identity shapes altruistic preferences towards ingroups and outgroups as defined by their political preferences towards Donald J. Trump. To separate ingroup-love from outgroup-hate, the experiment was designed to contrast behavior observed in a political identity setting by running the same experimental procedure in a minimal identity setting.

[7] Hummel and Maedche (2019) document that the default nudge far outperforms alternative interventions (e.g., reminders, norm-nudges etc.), yielding an uncontested average effect size of 87% and median effect size of 50%. For norm-information nudges, these numbers are 29% and 20%, respectively. See also Jachimowicz et al., 2019; Beshears and Kosowsky, 2020.

3.1.1. Data Collection and Experimental Design

Part 1: Measuring Polarization

This experiment consisted of two stages: a belief elicitation stage (divided into two parts) followed by the T-o-G DG, with details for both stages announced sequentially to all participants. As in all experiments presented in this paper, data was collected without the use of deception for two types of primes using a between-subject design: the *Trump prime (TP)* and the *Minimal Group Paradigm prime (MGP)*.[8] Analyses will focus on comparing the differences in these identity settings. The experiment lasted 10 minutes and dictators earned an average of $4 (including a show-up payment of $0.25). This translates to an hourly wage of $24 and is well above average hourly earnings on MTurk (Hara et al., 2018). Figure 1 illustrates the experimental design.

Preference Elicitation: For the *TP treatments*, the Preference Elicitation Stage was subdivided into two elicitations: one's opinion about Trump and one's perceived closeness towards their matched partner based on their partner's opinion of Trump.

1. In the first elicitation, participants were presented with a photo of Donald J. Trump and had to rate how they personally felt about him (with a focus on the time since he became president) on a 5-point Likert scale: extreme hate, moderate hate, indifferent, moderate love, or extreme love.[9] This method was inspired by the 'feelings thermometer' in the American National Election Study (ANES).
2. In the second elicitation, participants were randomly paired with another passive participant who indicated to either hate Trump (if they indicated either extreme or moderate hate) or love Trump (if they indicated either extreme or moderate love).[10] Participants were then asked to choose one out of 7 of circles (ranging from touching to almost completely overlapping, see Figure OA.38 for illustration) on the 'Inclusion of

[8] Data for recipients was collected separately and has no bearing on the results presented here.

[9] In accordance with the pre-registration, answers for moderate and extreme hate (love) were subsumed under 'hate' ('love') and were also done so in the matching procedure during the experiment. That is, in the treatments in which the matched partner's opinion about Trump was disclosed, participants only observed whether the partner indicated to hate or love Trump, but not the strength (extreme or moderate hate/love) of their opinion. Consequently, they will be treated as a bundled characteristic throughout the experiment. For a full distribution of opinions, see Figure A.1 in the Main Appendix. Regression results as presented in Tables A.1 and A.2 are robust to using the *intensity* of one's opinion about Trump rather than the binarized measure. Results are available upon request.

[10] As per pre-registration #42538, only participants who indicated either hate or love for Trump are analyzed, whereas participants who were indifferent are not. My reasoning for this is to align the analysis with the research question and focus on the role of polarization. This renders the indifferent participants (that MTurk cannot screen out ahead of time) obsolete. As a robustness check that the identity manipulation worked, I also collected observations for conditions in which the matched partner's opinion about Trump was not disclosed to the dictator. Because the focus of this paper is to examine affective and behavioral polarization when one can identify (mis)aligned identities, I relegate those results to the Online Appendix.

Other in the Self' (IOS) scale – a standard tool in social psychology used to measure the strength of inter-personal closeness – in order to demonstrate how close they felt to their matched partner (Aron et al., 1992; Gächter et al., 2015).[11] For simplicity, this scale is converted to percentage (relative to maximum value of 7).

For the *MGP treatments*, this stage was subdivided into three elicitations: first (as before), one's opinion about Trump, then one's painting preferences, and then one's perceived closeness towards their matched partner based on their partner's painting preferences.

1. The first elicitation followed the same procedure as the TP condition. Participants were presented with the same photo of Donald J. Trump and asked to indicate their personal feelings towards him using the 5-point Likert scale: extreme hate, moderate hate, indifferent, moderate love, or extreme love.
2. In the second elicitation, participants were presented with Klee and Kandinsky paintings and were asked to choose their favorite (design following Chen and Li, 2009).
3. In the third elicitation, participants were randomly paired with a partner who said to either prefer Klee or Kandinsky.[12] Participants did not receive information about their partner's Trump preference. Participants were then asked to rate their feelings of closeness towards their matched partner using the same IOS scale as the Trump prime group. Again, in the analysis, this scale is converted to percentage (ratio of one's indicated value out of 7).

The matching procedure for the MGP treatments mirrors the procedure in the TP treatments, except that the matching and subsequent IOS closeness elicitation were done on the basis of painting preferences instead of opinions on Trump. This procedure is designed to focus on the sole effect of being matched according to one's (mis)matched Trump or painting preferences by keeping the role of the Trump prime constant across treatments. Thus, conditional on their own Trump opinion/painting preference, participants were allocated to one of the partner preferences conditions at random. Consequently, the between-design captures the following dimensions: 2 *(prime)* $\times$ 2 *(own Trump/painting preference)* $\times$ 2 *(partner's Trump/painting preference)*.

[11] In a highly-cited and highly-influential paper in the Journal of Personality and Social Psychology, Aron et al. (1992) introduced this intuitive and simple pictorial tool to measure bi-lateral relationships. Respondents are asked to assess their relationship with another individual (which, in my paper, varied based on the partner's Trump preference) by selecting one out of 7 pairs of increasingly overlapping circles. Respondents select the pair of circles that best describes their relationship with the matched partner. As later verified by Gächter et al. (2015), this scale is a "psychologically meaningful and highly reliable measure of the subjective closeness of relationships." I employ the scale in the *exact* way as in the original study.

[12] As in the Trump prime condition, I also collected observations for a treatment in which the matched partner's painting preference was not disclosed to the dictator. The results are presented in the appendix.

Take-or-Give Dictator Game: In Stage II, I aimed to capture both pro-social (giving) and anti-social (taking) behavior. To do so, I employed a variant of a dictator game inspired by existing research (List, 2007; Bardsley, 2008; Dimant, 2019). The standard dictator game involves two participants: a dictator and a recipient. The dictator receives an endowment and is tasked with deciding how to split it between herself and the recipient. Her options range from giving nothing or giving everything to the recipient, while the recipient plays a passive role. In the variant utilized by this experiment, both the dictator and the recipient began with a non-zero endowment and the dictator's action space was augmented with one additional option: the opportunity to take some or all money away from the recipient. One of the many advantages of using this modified version of the game is the ability to measure both pro-social and anti-social tendencies simultaneously (see Dimant (2019) for a discussion). Prior to making the decision, participants were told that on top of the show-up fee, half of all randomly determined dictator-recipient pairs would be paid a bonus corresponding to their in-game decisions. The remainder half only received the show-up fee. For the purpose of my study, I borrow the initial endowment structure from List (2007): the dictator starts with \$10 whereas the recipient starts with \$5.[13] The dictator makes one of the following three decisions once:

1. Take up to \$5 from the recipient's endowment and add to one's own endowment.
2. Make no change to the initial distribution of money.
3. Give out up to \$5 from one's own endowment and add to the recipient's endowment.

This variant of the dictator game allows me to take the first step towards investigating the impact of polarization on altruism, which – unlike regular dictator games – also provides me the opportunity to simultaneously study pro-social (giving) and anti-social (taking) behavior. Moreover, the contrast with the MGP prime adds an additional layer of detail in that I am able to distinguish whether the observed behavior with the political prime resembles ingroup-love, outgroup-hate, or both.

Part 2: Reducing Polarization

I follow recent evidence on the effectiveness of nudges (e.g., Hummel and Maedche, 2019; Jachimowicz et al., 2019; Beshears and Kosowsky, 2020) and test two of the most promising interventions in the toolbox of behavioral scientists: defaults and norm-nudges. Early successful research using defaults in contexts such as 401(k) savings (e.g., Madrian and Shea, 2001; Thaler and Benartzi, 2004) have helped popularize default interventions among both scientists and practitioners (Thaler and Sunstein, 2008; Benartzi et al., 2017). The concept

[13] To retain incentive-compatibility, dictators were told that their allocation decisions are paid out in 50% of the time as bonus at the end of the experiment. If not selected, they only received the show-up fee.

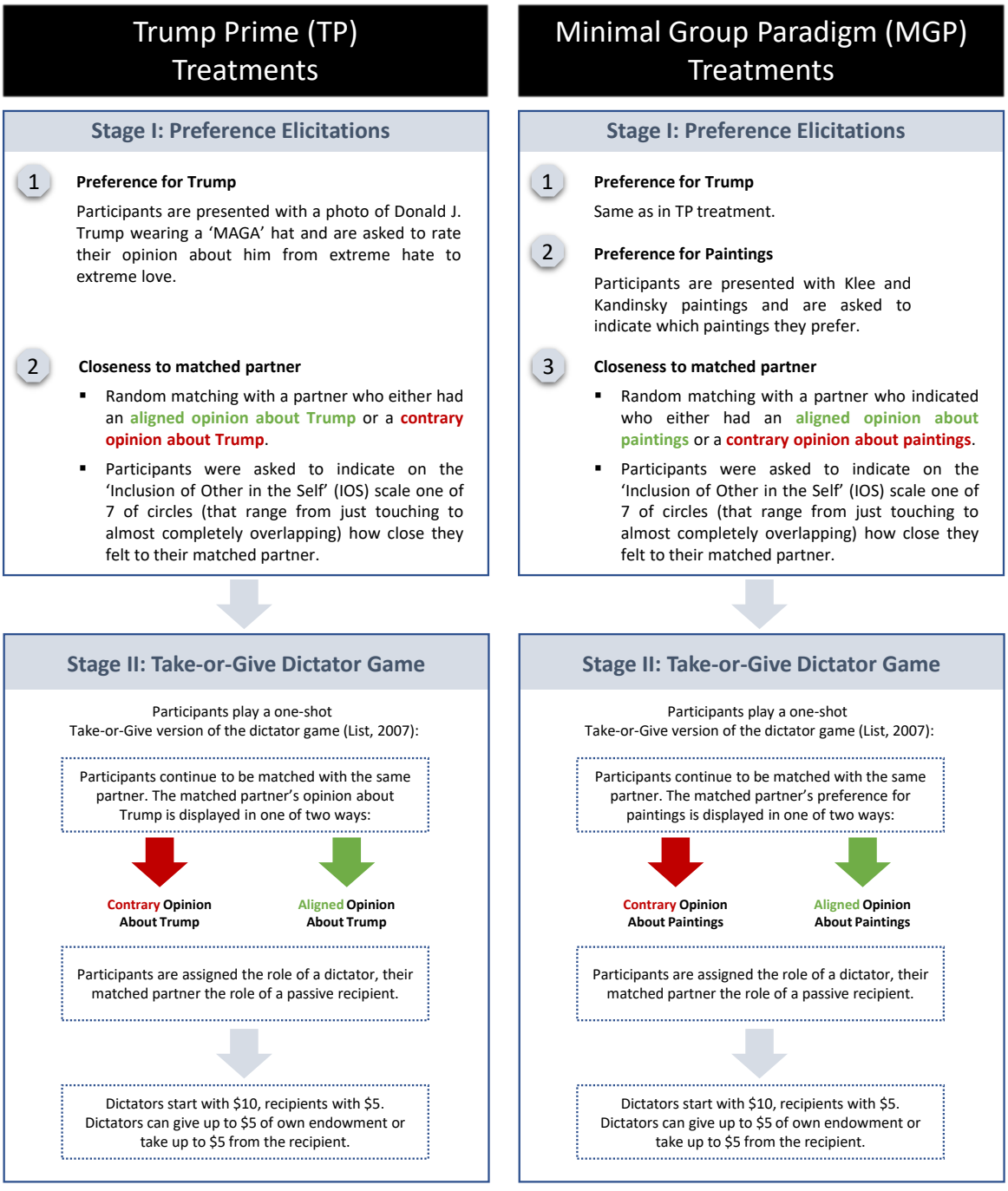

Figure 1: Experimental design of the Take-or-Give dictator game for both the Trump prime and the Minimal Group Paradigm prime conditions.

of norm-nudging has been popularized across various fields of social sciences and has been widely applied in the context of behavior change (for a theoretical conceptualization see Bicchieri and Dimant, 2019). These interventions attempt to change behavior by eliciting

and changing existing social norms through the manipulation of social expectations.

In two sets of experiments, I harness one particular aspect of norm-nudging – descriptive norms – by informing participants in this experiment truthfully of the behavior of previous participants. In particular, I emphasize the pro-social and cooperative nature of previous participants with the goal to change these participants' beliefs of what is commonly done in this situation (for applications in other related contexts see Hallsworth et al., 2017; Damgaard and Gravert, 2018; Allcott and Kessler, 2019; Bott et al., 2019; Bursztyn et al., 2020a,b; Dimant et al., 2020; Dimant and Gesche, 2020). The nudge experiments implemented here mirror the experimental procedure for the *Trump prime (TP)* in the context of the *T-o-G Dictator Game* as explained above in Part 1. I introduced both nudges, the *Default Nudge* and the *Descriptive Norm-Nudge*, at the decision stages as explained below.

Default Nudge: At the stage where participants were asked to make a give-or-take decision towards the matched partner, the option "give \$2.5" was pre-selected (in the original version of the experiment, no option was pre-selected). This pre-selected option corresponds to 50% of the amount that could be given to the partner and, most importantly, was the only choice that achieves an equal split between dictator and recipient.A participant was then given the chance to actively override the default or simply proceed to the next screen (demographic questionnaire) upon which the default was implemented.

Descriptive Norm-Nudge: The protocol of this experiment was identical to that described in Part 1 with one additional piece of information. Participants were given the truthful message (as illustrated in Figure OA.6) that many participants have been benevolent towards their partners in previous sessions of this game and achieved an equal split in endowments by giving \$2.5, even if they had a contrary opinion about Trump.

3.1.2. Hypotheses

Existing literature on identity, ingroup bias, and social proximity suggests that individuals feel closer to participants who are more 'similar' to them, which culminates in stronger tendencies towards pro-sociality (Akerlof, 1997; Akerlof and Kranton, 2000; Charness et al., 2007; Fowler and Kam, 2007; Chen and Li, 2009; Christ et al., 2014; Dimant, 2019; Lees and Cikara, 2020). Compared to the minimal identity setting, greater (lower) amount of pro-sociality and closeness towards a partner with the same opinion regarding Trump will be labeled as *ingroup-love* (*outgroup-hate*).[14] Consequently, I derive the following hypothesis:

H_1: *Dictators will exhibit the largest closeness score and extent of pro-sociality towards a partner who has the same opinion of Trump: TH-TH or TL-TL. It will be lowest when the*

[14]I follow Yamagishi and Mifune (2009) and define these terms as: *Ingroup-love* (*outgroup-hate*) indicates behavior that provides ingroup (outgroup) members with preferential (spiteful) treatment.

matched partner's opinion is misaligned: TH-TL or TL-TH.

As I argued in the introduction, another scientific contribution of this paper is to examine whether 'hate' is stronger than 'love.' If so, one would expect a disproportionate effect of closeness and displayed behavior that explains a host of existing phenomena, including asymmetries between positive and negative reciprocity as well as between the contagion of pro-/anti-social behavior (Offerman, 2002; Croson and Shang, 2008; Lelkes and Westwood, 2017; Dimant, 2019; Bicchieri et al., 2020a). I thus predict:

H_2: *Dictators will exhibit disproportionately larger outgroup hate than ingroup love.*

3.1.3. Results: Measuring Polarization

Results for *Part 1: Measuring Polarization* are broken up along multiple dimensions for both TP and MGP treatments for perception of closeness (henceforth referred to as *perception*, for illustrative purposes presented as % of indicated closeness on a scale from 1 to 7) and behavior in the T-o-G dictator game (henceforth referred to as *behavior*, measured as % of dollar amount given to/taken away).[15] Figure 2 compares results from when the dictator is matched with a partner who has an aligned opinion about Trump (matching: either *Hate-Hate* or *Love-Love*) to being matched with a partner who has a contrary opinion about Trump (matching: either *Love-Hate* or *Hate-Love*). Figure 3 demonstrates the same analysis, but but broken down by a participant's opinion on Trump (hate or love).

Aligned vs. Misaligned Partner: Comparing the perceived closeness and behavior between the TP (top panel of Figure 2) and the MGP (bottom panel of Figure 2) treatments provides a clear indication of whether and where *ingroup-love*, *outgroup-hate*, or both simultaneously exist. It is evident that differences in both closeness and behavior arise only in the TP and not in the MGP, indicating that an ingroup-outgroup differentiation is evoked exclusively by the political hate-love prime.

For participants in the TP, I find highly significant differences along both dimensions: for the perception of closeness, participants feel the strongest (weakest) connection with the matched partner that has the same (opposing) view about Trump (all $p<0.001$). Interestingly, perceptions of closeness towards participants with a contrary opinion about Trump (one's outgroup, 39.8%, red bar in top-left panel) are indistinguishable from perceptions of closeness in the MGP, whether compared to someone with the same preferences (37.5%, red bar bottom-left panel, $p=0.40$) or with contrary preferences (39.1%, green bar bottom-left

[15] Consistent with the preregistration, the following statistical analyses will be performed in all three experiments: bootstrap two-sample t-test method (BSM) as proposed by Moffatt (2015) with 9999 replications. The BSM procedure retains cardinal information without distribution assumptions. Robustness checks will be performed using non-parametric Mann Whitney-U ranksum tests. Unless noted otherwise, the results can be assumed to be consistent between the two methods.

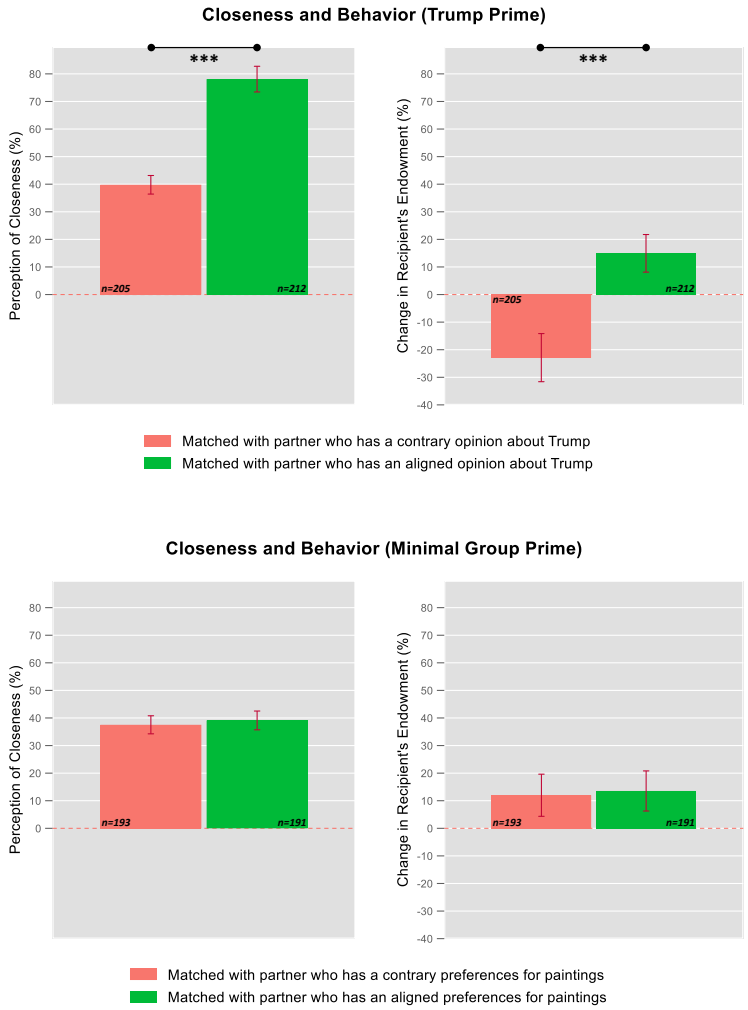

Figure 2: Closeness and behavior by being matched with a partner who has a (mis)aligned opinion about Trump for TP treatments and about painting preference for MGP treatments. Perception of closeness is converted from a 7-point scale to % for illustrative purposes. All adjacent bars (within each category) are compared. Absence of significance stars $\Rightarrow$ p-values > 0.05.

panel, p=0.87). Conversely, one can observe ingroup-love in the disparate levels of perceived closeness between the two conditions. Participants in the TP treatment demonstrated a notably higher affinity towards members of their own political faction (78.1%, green bar in top-left panel) compared to the feelings of closeness reported by participants with the same painting preferences in the MGP treatment ($\sim$39%, BSM, p<0.001). Thus, one can conclude that – rather than outgroup hate – the observed perception of closeness in the TP is the result of ingroup-love.

With respect to behavior in TP, one observes a stark difference in pro-sociality (14.9% vs. -22.9%, BSM, p<0.001). Notably, I find a marked asymmetry between ingroup-love and outgroup-hate: The absolute negative average amount in the misaligned condition is over-proportionally larger than the positive average amount in the aligned condition ($|-22.9\%|$ vs. 14.9%, BSM, p<0.01).[16] For MGP, the observed differences are trivial in size and are

[16]See also Lelkes and Westwood (2017). It is worth noting that this asymmetry is seemingly not *per se* driven by differences between Trump haters and Trump lovers. As Figure OA.1 in the Online Appendix shows, both types display the same average perception of closeness (57.5% vs. 59.0%, BSM, p=0.47). However, it is driven

insignificant for both perception of closeness (37.5% vs. 39.1%, BSM, p=0.51) and take-or-give behavior (12.0% vs. 13.5%, BSM, p=0.88). I conclude that the political identity frame evokes a state that produces traceable changes in both perceptions and behavior beyond the minimal group notion. In contrast to the results for perception of closeness, the results here display a clear indication of outgroup-hate – rather than ingroup-love – for behavior. With overlapping error bars, the pro-social behavior towards one's own political faction (14.97%, green bar in top-right panel) is indistinguishable from the behavior in MGP, regardless of whether their partners held the same (13.54%, green bar bottom-right panel, p=0.78) or different (11.99%, red bar bottom-right panel, p=0.56) painting preferences. This behavior unambiguously stands in stark contrast to the observed conduct of participants with contrary opinions about Trump. The results illustrate a clear pattern of anti-social behavior, with a significant proportion of participants exhibiting a propensity for inflicting harm (-22.9%, red bar top-right panel).

The results thus compellingly indicate that ingroup-love and outgroup-hate are context-specific. Specifically, they appear only in the political prime. In sum, ingroup-love occurs for perceived closeness and outgroup-hate occurs for actual behavior.

Heterogeneity Analysis: In Figure 3, the same two dimensions (perception and behavior) are broken down by one's own opinion of Trump. As is evident, the previous results reappear and are not driven by any particular subgroup of Trump haters or Trump lovers. Notably, I once again observe that the levels of perceived closeness in the MGP prime are indistinguishable from that towards someone with a contrary opinion on Trump, and half as much than that towards someone with an aligned opinion on Trump. As before, the degree of pro-social behavior in the MGP condition is the same as that of TP condition participants that were matched with a partner with aligned views on Trump, and this altruistic behavior occurs at a much higher level than the overall anti-social behavior towards those with contrary views on Trump.

In terms of measured behavior in the TP treatments, only those who were matched with a partner with the same opinion on Trump displayed significant pro-social behavior on average (increase in the recipient's endowment). Individuals with partners who hold differing views of Trump exhibited a significant pattern of anti-social behavior on average (decrease in the recipient's endowment). At the same time, the degree of pro-social behavior across the MGP condition is the same as that of participants in the TP condition who were matched with an aligned partner, and much higher than the overall anti-social behavior towards those with a contrary Trump opinion. [17]

by the fact that the average perceived closeness towards Trump Haters (65.7%) is significantly larger than the perceived closeness towards Trump lovers (53.8%, BSM, p<0.001).

[17] As later shown in Figure A.5, this maps well onto the different norm perceptions between Trump haters

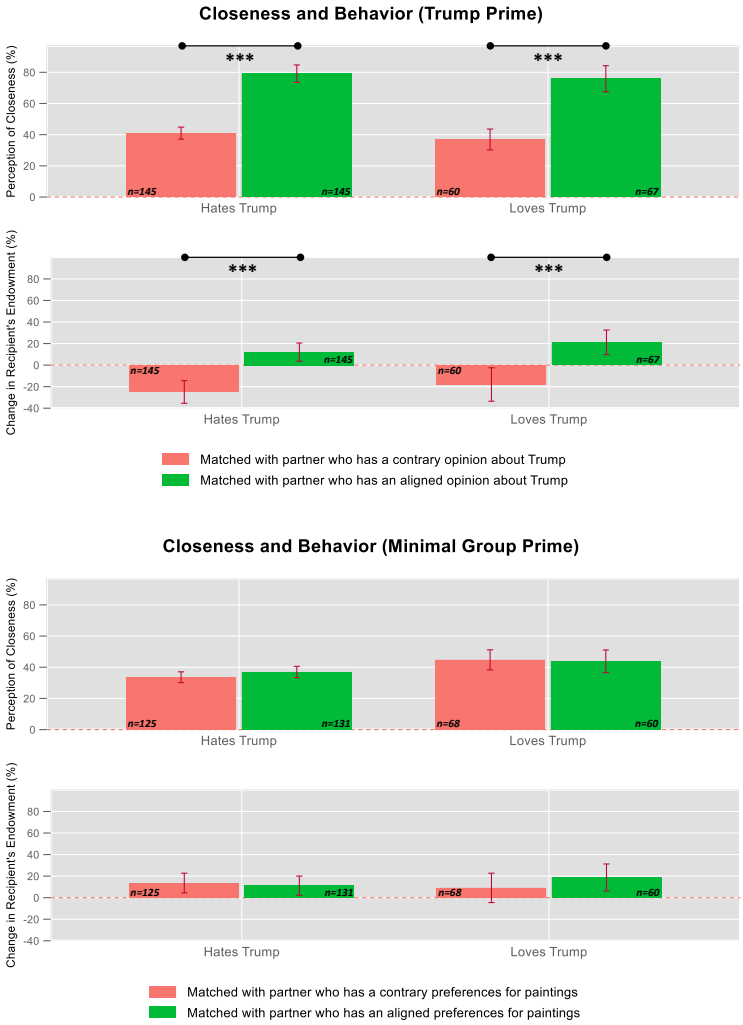

Figure 3: Closeness and behavior broken down by being matched with a partner who has a (mis)aligned opinion about Trump for both TP and MGP treatments. All adjacent bars (within each category) are compared. Absence of significance stars ⇒ p-values > 0.05.

Based on my findings in Figures 2 and 3, I conclude that the concept of ingroup-love and outgroup-hate is more nuanced than some existing research has suggested: in the context of non-strategic decisions, ingroup-love occurs with respect to perceived closeness, whereas outgroup-hate occurs with respect to altruistic behavior. The results are also evaluated in a regression framework that includes the collected controls (age, gender, level of education, political affiliation, U.S. citizenship, whether one voted in the 2016 election, and race). Without qualifications, all previously presented findings hold (Table A.1 in the Appendix).

3.1.4. Results: Reducing Polarization

Default Nudge: As evident in Figure 4, the success of the Default Nudge is nuanced. Compared to the results from Study 1 (see top-right of Figure 2 and upper panel of Figure 3), the default increases altruism at the aggregate level: Anti-social (taking) behavior towards the outgroup is reduced from about 20% to essentially 0% of the initial endowment and pro-

and lovers: the former make a clear distinction between harming their ingroup versus the outgroup, whereas the latter do not seem to make such a distinction.

social (giving) behavior to the ingroup from about 10-15% to about 25-35% of the initial endowment. These results emphasize the limits of the default nudge intervention: The ingroup-outgroup polarization remains at comparable levels in terms of both effect sizes and statistical significance and the polarization gap remains intact. One can observe the same robust result when breaking up the analysis according to the participant's opinion about Trump (see right panel of Figure 4 and Figure OA.4 in the Online Appendix).

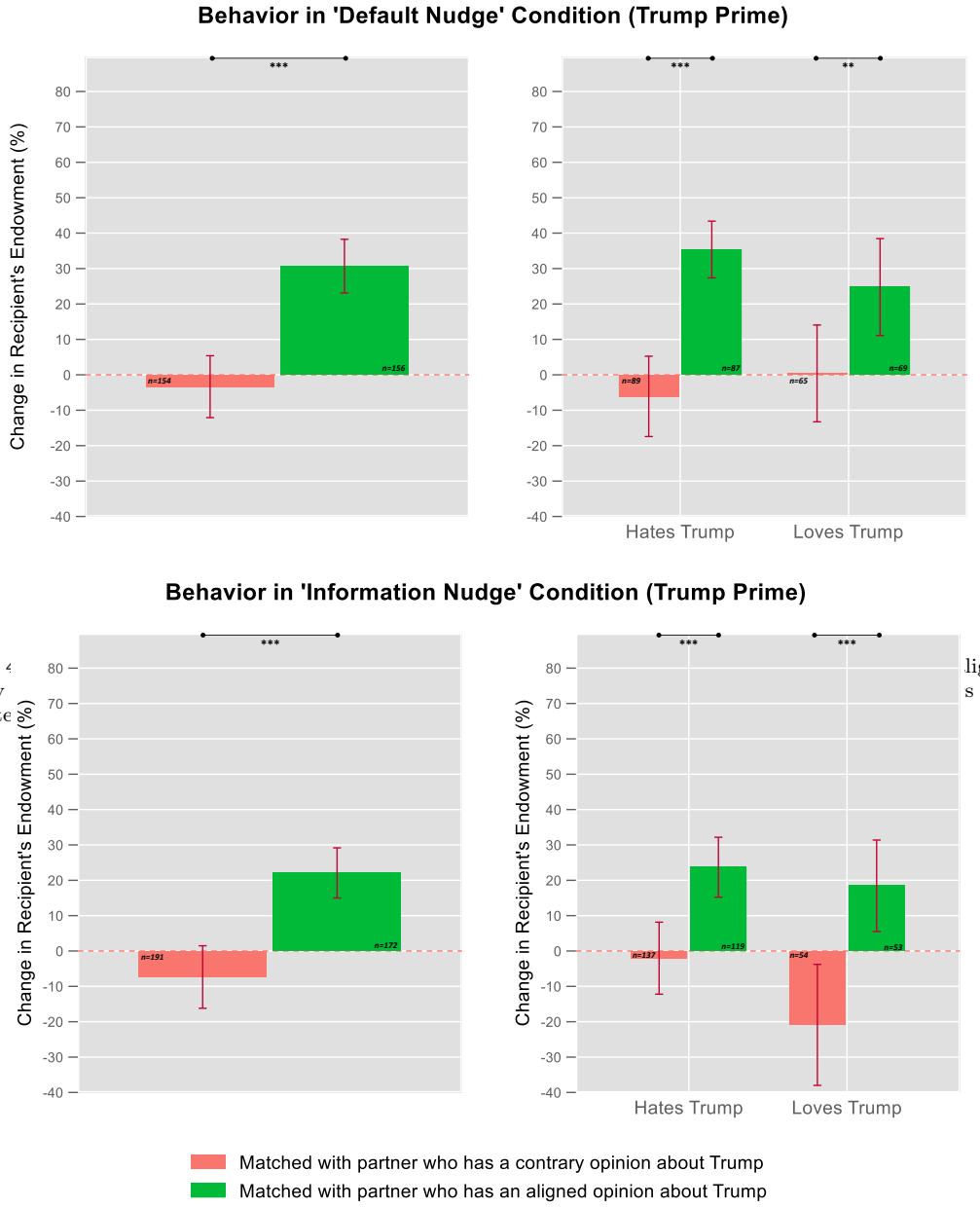

Figure 4 contrary each cate ligned or s (within

Figure 5: Left Panel: Behavior broken down by whether one is matched with a partner who either has aligned or contrary opinions. Right Panel: same but broken down by one's own opinion about Trump. All adjacent bars (within each category) are compared. Absence of significance stars ⇒ p-values > 0.05.

Descriptive Norm-Nudge: As is the case for the Default Nudge intervention discussed above, the results of the Descriptive Norm-Nudge intervention are similarly bleak: As illustrated in Figure 5 (also see Figure OA.5 in the Online Appendix), this intervention is unable to reduce the originally observed polarization in the context of the DG in Study 1. This holds true even when breaking up the results by the participant's own Trump preference (right panel). Notably, we observe some success with respect to reducing anti-social behavior towards outgroups, which is limited to those participants who hate Trump. Conversely, the outgroup animosities of those who love Trump remain unaffected, with their magnitude being comparable to that reported in the original experiment (Section 3.1). This should not distract from the ineffectiveness of this intervention in meeting its goal of reducing polarization in a non-strategic setting.

3.2. STUDY 2: POLARIZATION IN A STRATEGIC CONTEXT

ABC of Cooperation Public Goods Game

In Study 2, I examine the impact of polarization in a *strategic* context using different variants of a public goods game. The experiment is designed to build upon my findings from Study 1 about the impact of partisan animosity on ingroup-love and outgroup-hate. This approach allows me to answer important and policy-relevant questions: Does the negative impact of polarization arise because people expect individuals from the opposite faction to be less cooperative (a belief channel)? Or is it the consequence of a lower willingness to cooperate with members of the opposite faction, no matter how cooperative they are (a preference channel)? By distinguishing these mechanisms, the procedure aims to identify how deep the societal rift is and whether, in principle, cooperation might be sustained through appropriate belief management. As before, I compare the impact of the political identity markers with that of a minimal group prime to distinguish between ingroup-love and outgroup-hate.

3.2.1. Data Collection and Experimental Design

Part 1: Measuring Polarization

This experiment was carried out across the same two treatment conditions as Study 1: the *Trump prime (TP)* and the *Minimal Group Paradigm prime (MGP)*. It also consists of two stages: a preference elicitation portion followed by an extended public goods game. Stage II utilizes the 'Attitudes-Beliefs-Contributions (ABC) of cooperation' approach as introduced by Fischbacher et al. (2001) and Gächter et al. (2017) in order to measure cooperativeness and participants' beliefs and attitudes about cooperation. The average pay resulted in about \$6.15 (including a \$0.25 show-up fee). The duration of the experiment was about 15 minutes, translating to a well above-average hourly pay of ∼\$24.6 (identical to the average payoff in Study 1). The design is detailed below and illustrated in Figure 6.

Treatments: The experiment began with a Preference Elicitation Stage identical to that of Study 1 (see Section 3.1.1 for more details) with the same TP and MGP treatments. For the TP treatments, the public goods games were played either by paired subjects with the same opinion about Trump (TH-TH or TL-TL), with an opposing opinion about Trump (TH-TL or TL-TH), or by participants for which the opinion about Trump was not disclosed to the other participant (TH-TU or TL-TU). The same applied for the MGP treatments in which participants received information about their partner's painting preferences at random. As before, by directly contrasting the political identity setting to a minimal group setting, I am able to distinguish between ingroup-love and outgroup-hate. In sum, this PGG variant generates a set of conditions that allow me to answer my research questions and disentangle the mechanisms by which polarization operates.

Public Goods Game: Stage II consists of three tasks to measure participants' cooperativeness using a two-player variant of the 'ABC of cooperation' approach (Gächter et al., 2017): a one-shot sequential public goods game played with the strategy method to measure attitudes of cooperation, a belief-elicitation task to measure expectations of others' cooperation, and a one-shot simultaneous public goods game played with the direct response method to measure effective contributions. The game embodies the classic tension between private and collective interest: while fully contributing to the public goods maximizes joint payoffs, each player's self-interest is maximized by contributing nothing.

To ease participants' mental efforts, I followed the standard notion of the game and use an MPCR of 0.75 (Isaac and Walker, 1988): Each player was endowed with \$10 that she could either contribute to the public good or keep for herself. Participants were able to give any integer amount between 0 and 10, thus providing 11 options in total. Each dollar contributed to the public good was multiplied by 1.5 and then equally divided between the two participants, irrespective of each participant's individual contribution. Subjects played the tasks sequentially and in random order, but received no feedback on choices or earnings in any of the tasks until the end of the experiment. Only one of the three tasks was used to calculate earnings, and subjects were made aware of this fact at the beginning of the experiment. The task used for calculating payments was randomly selected at the end of the experiment after each subjects' choices in all tasks had been collected.

In the *first* task, I used a version of the game described above to measure players' attitudes towards cooperation. Subjects were randomly assigned to be either a first-mover or a second-mover. Participants played the game sequentially and made decisions in the role of both the first-mover and the second-mover, with only their predetermined role being used to compute their payoffs. The task was played using role-uncertainty. That is, all subjects are asked to provide decisions in both the roles of first-mover and second-mover, without knowing their role assignment in the task until *after* all decisions have been collected. I used

the Strategy Method to elicit the second-mover's choices: second-movers are asked to submit a contribution decision for each possible contribution choice made by the first-mover. This ensures that for each second-mover, one can observe a vector of contributions comprised of 11 choices. I denote subject's i contribution vector as a_{i}. I used this vector of contributions to classify subjects into "cooperation types" that reflect their willingness to cooperate as a function of their opponent's cooperativeness. Following the seminal work by Fischbacher et al. (2001), subjects were classified into four types:

1. *Free riders* if they contributed \$0, regardless of the first-mover's contribution.
2. *Conditional cooperators* if they had a vector of contributions that was either weakly monotonically increasing in relationship to the first-mover's contribution, or was not monotonically increasing but had a highly significant (at the 1% level) and positive Spearman rank correlation coefficient (between own and others' contribution).
3. *Unconditional cooperators* if they contributed a positive amount that did not vary across different first-mover's contributions.
4. *Other* if they could not be classified according to any of the previous criteria.

In the *second* task, I elicited subjects' expectations about the cooperativeness of their opponent. Subjects were asked to guess the contribution that their opponent had made in the simultaneous public goods game. Subjects were rewarded for the accuracy of their guess: If their guess was within \$2, they received a bonus of \$0.50. I denoted the subject's i belief regarding the opponent's contribution as b_{i}. In the *third* task, I elicited subjects' effective contributions using a simultaneous version of the public goods game described above. Subjects made a contribution decision in direct-response mode, without learning the contribution choice of their opponent. I denoted subject's i effective vector as c_{i}.

With that, I capitalize on a $2 \times 2 \times 2$ experimental design: *(Trump / minimal group prime) × (own opinion about Trump / preference for paintings) × (matched partner's opinion about Trump / preference for paintings).*[18]

Part 2: Reducing Polarization

In Part 2, I experimentally examine the aforementioned behavioral interventions in a strategic decision context, complementing my findings in Study 1 about their efficacy in a *non*-strategic context. Consistent with Study 1, I deploy each nudge intervention as follows:

Default Nudge: Where attitudes (a_{i}) were elicited, the default pre-selected the response of a conditional cooperator: for each question of the strategy method elicitation ("What would you contribute if your partner contributed \$0 ... \$10?") the corresponding conditionally

[18] Again, note that for the purpose of brevity, results for the conditions where one was matched with a partner for whom the opinions were undisclosed are relegated to the Online Appendix.

Figure 6: Experimental design of the public goods game for both the TP and MGP conditions.

cooperative response ($0 ... $10) was pre-selected.[19] For the elicitation of beliefs (b_i) and the effective contribution (c_i), the default pre-selected the welfare-maximizing behavior ($10). Participants had the choice to override any and all pre-selections or to continue to the next screen upon which the default nudge was implemented.

Descriptive Norm-Nudge: The protocol was identical to that described above with one additional piece of information. Participants received the truthful message (as illustrated in Figure OA.12) that many participants have been cooperative with other participants in previous sessions of this game, even if they had a contrary opinion about Trump.

3.2.2. Hypotheses

In line with the reviewed literature and the hypotheses presented in Study 1, an ingroup-love/outgroup-hate effect can be expected.[20] As such, I derive the following hypothesis:

H_3: *Participants will exhibit stronger closeness, more pronounced attitudes towards cooperation (a_i), higher beliefs about the partner's cooperativeness (b_i), and more effective cooperation (c_i) when matched with a partner who has the same opinion about Trump (TH-TH or TL-TL), whereas these numbers will be lowest when the matched partner's Trump opinion is misaligned (TH-TL or TL-TH).*

As before, one can also examine the asymmetry between hate and love, and I thus put forth a prediction of the following form:

H_4: *Participants will exhibit disproportionately larger outgroup-hate than ingroup-love.*

3.2.3. Results: Measuring Polarization

The following analysis of the results from *Part 1: Measuring Polarization* is divided along several dimensions for both the Trump prime (TP) and the Minimal Group Paradigm prime (MGP) and is visually presented in the figures below. In this discussion, I build upon my findings from Study 1 and put forth additional insights that emerge from examining polarization within this distinctive strategic decision-making context. In Figure 7, I present and compare the perceptional and behavioral findings between aligned and misaligned partners within the two treatment groups. Next, I conduct the same analysis, but I focus my comparison on participants' opinion of Trump (love vs. hatred) across both treatment groups, as demonstrated in Figure 8. Figure 9 then contains a visual representation of the distribution

[19] Recall that in the original treatments, no option was pre-selected and participants were able to choose any number between $0 and $10 for each of the possible decisions of the partner from a drop-down menu.

[20] As before, while the main focus of this experiment will focus on beliefs and behaviors towards own and opposing factions, the comparison with the participant for whom the opinion about Trump is not disclosed allows me to make to draw an inference about *whether* the observed behavior is ingroup-love or outgroup hate, or both. For brevity, I relegate these pairwise comparisons to the Online Appendix.

of the four different 'type' classifications (as outlined in the previous section) across the various treatment groups.

Aligned vs. Misaligned Partner: Figure 7 presents a comparison between the two treatment groups along measures of perceptions of closeness, beliefs (b_i) about other's contributions, and one's own effective contributions (c_i). These variables are examined with respect to partner alignment vs. misalignment, with alignment determined by individual's opinion of Trump within the TP condition and painting preferences within the MGP condition. As is evident in the graphs below, there proved to be a noteworthy disparity between the treatments along each of the the three measured dimensions.

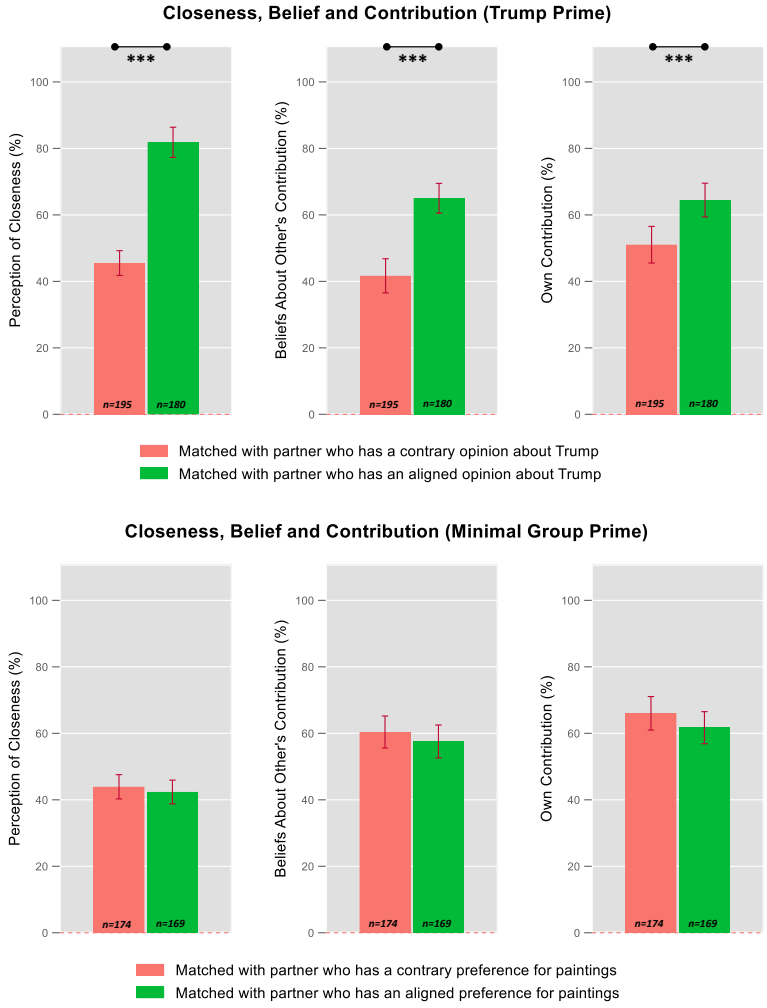

Figure 7: Closeness, belief, and behavior by being matched with a partner who has a (mis)aligned opinion about Trump (for TP treatments) and painting preference (for MGP treatments). All adjacent bars (within each category) are compared. Absence of significance stars $\Rightarrow$ p-values > 0.05.

These results reveal a pattern that is consistent with my previous findings: when matched with a partner with a concordant opinion about Trump, participants felt closer, had higher

expectations of the partner's contribution, and effectively contributed more in the PGG (Figure 7, top panel) compared to when participants were matched with a partner who had a contrary opinion about Trump (all p<0.01). In contrast, none of these differences appear in the MGP conditions (Figure 7, bottom panel).

The pattern of perceived closeness here is a clear indication of ingroup-love consistent with those previous results (red bars have about the same height, whereas the green bars are much higher in TP than in MGP). Conversely, for both beliefs about other's contributions and one's own contribution, we observe outgroup-hate rather than ingroup-love (green bars are indistinguishable, whereas the red bars are significantly lower in TP than in MGP).

Heterogeneity Analysis: Next, I continue to examine the same dimensions of closeness, beliefs, and contributions, but I shift my analysis to participants' love/hatred towards Trump across both treatments. These results are presented in Figure 8 below. For Trump haters, the results are remarkably consistent: for all three measures, the magnitude is significantly higher when matched with another Trump hater (all comparisons p<0.001). For the Trump lovers, one can observe that the discrimination between ingroup and outgroup only holds for the perceived closeness and b_i, but not for c_i.[21] There, Trump lovers contribute a statistically indistinguishable amount of about 55-62% of the maximum amount, irrespective of whether they were matched with another Trump lover or hater (p=0.253).

As before, no significant differences occur along any dimension for the MGP conditions (bottom Figure 8). This confirms that the results are not driven by ingroup-outgroup considerations alone; rather, the observed disparities in perceptions, beliefs, and cooperation largely rest on the (emotional) state evoked by polarization. I also compare both beliefs and behaviors in TP to those in MGP and reach the same conclusions for both Trump haters and Trump lovers: Ingroup-love occurs for perception of closeness, whereas outgroup-hate occurs for beliefs about other's contributions and one's own contribution.

Type-Classifications: For the final part of the investigation, I follow the tradition of Fischbacher et al. (2001) and analyze the distribution of 'types' across the various treatments. I follow the previously introduced classification and distinguish between *Conditional Cooperators* (CC), *Unconditional Cooperators* (UC), *Free Riders* (FR), and *Others* based on the participant's responses using the strategy method (a_i).

I find that essentially all type-classifications are insensitive to whom one is matched with, regardless of the treatment prime (Figure 9).[22] In combination with the previous

[21] As later shown in Figure A.7, these findings are consistent with the norm elicitation: Trump lovers do not display an ingroup/outgroup differentiation in terms of free riding or cooperation, whereas Trump haters do. The regressions in Table A.2 (see the Main Appendix) confirm these results, and it is evident that the observed behavioral differences are driven by differences in perceived closeness.

[22] Most differences do not achieve the pre-registered alpha level of 5%. For TP, the only significant dif-

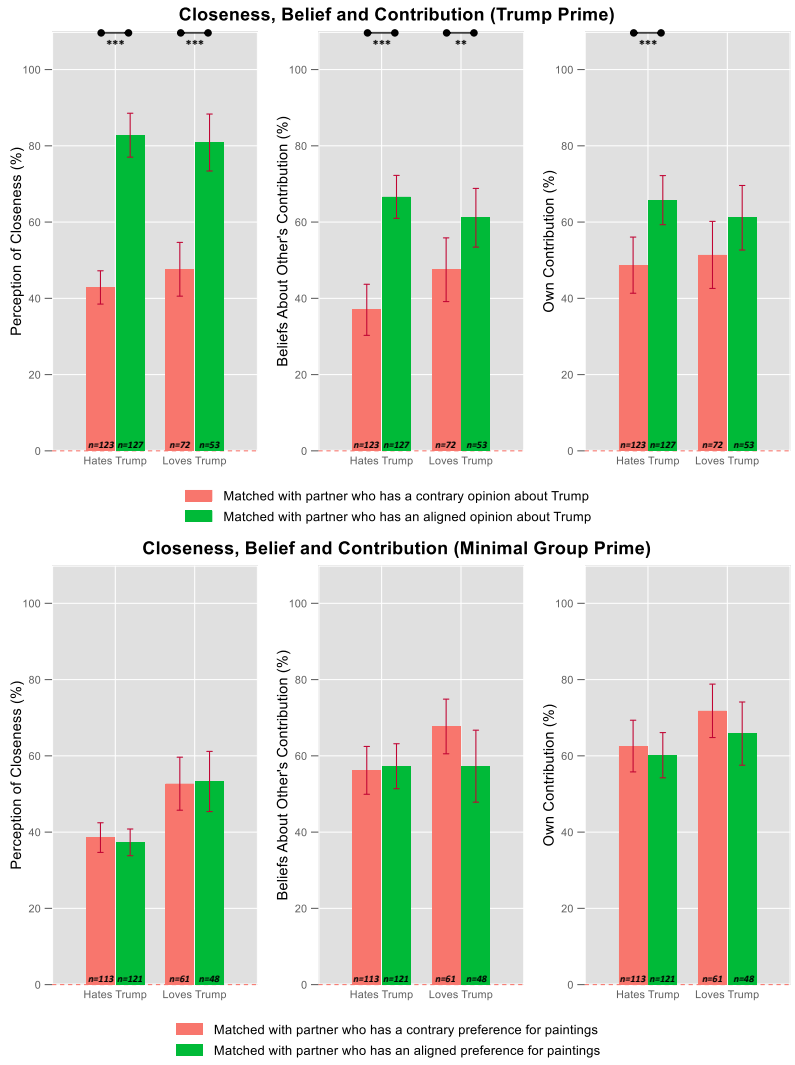

Figure 8: Closeness, belief, and behavior by being matched with a partner who has a (mis)aligned opinion about Trump for both TP and MGP treatments. All adjacent bars (within each category) are compared. Absence of significance stars ⇒ p-values > 0.05.

insights from the Trump prime conditions, this is a key result: The observed inter-faction animosity in form of ingroup/outgroup variability in contributions is a result of fallacious beliefs about the other's behavior (top panel of Figure 8) and *not* of adverse preferences per se (top panel of Figure 9). In the latter, it becomes apparent that participants are willing to cooperate with the opposing faction, regardless of one's partisanship. The implication is that polarization could potentially be counteracted by correcting the bleak expectations that the factions have about each other.

ferences is observed for the 'Others' group with p=0.011. Although visually distinct, the differences for Conditional Cooperators only achieve significance at the 10% level at p=0.06 and p=0.09 for Trump-hater and Trump-lover, respectively. For MGP, the only reliably significant difference (p<0.01) can be observed for the Unconditional Cooperators among Trump haters. For Conditional Cooperators, the differences reach p=0.11 and p=0.34 for Trump haters and Trump lovers, respectively.

Figure 9: Types (conditional cooperators, unconditional cooperators, free riders, others) by being matched with a partner who has a (mis)aligned opinion about Trump for both TP and MGP treatments. All adjacent bars (within each category) are compared. Absence of significance stars ⇒ p-values > 0.05.

3.2.4. Results: Reducing Polarization

Default Nudge: Following the original analysis presented in Figures 7 and 8, Figure 10 (also see Figure OA.8 in the Online Appendix) presents one's beliefs about the matched partner's contribution (b_i) and one's actual contribution (c_i) in the PGG in the presence of the Default Nudge (for attitudes (a_i) see Figure OA.9 in the Online Appendix). The results are consistent with those presented in Study 1 in that the introduced nudge is rather ineffective in achieving the principal goal of reducing polarization: the ingroup-outgroup differences remain similar to those observed in the original experiment where the nudge was absent, both in terms of effect sizes and statistical significance.

Descriptive Norm-Nudge: Consistent with the results above, the effectiveness of the Descriptive Norm-Nudge intervention aimed at reducing the polarization gap in the PGG is limited, too: as illustrated in Figure 11 (see also Figure OA.10 in the Online Appendix), the persistence of polarization is vivid, both for beliefs about the partner's contribution and one's own contribution. This result holds true even when breaking up the data by one's own opinion about Trump: the results are not driven by any particular subgroup (Trump haters

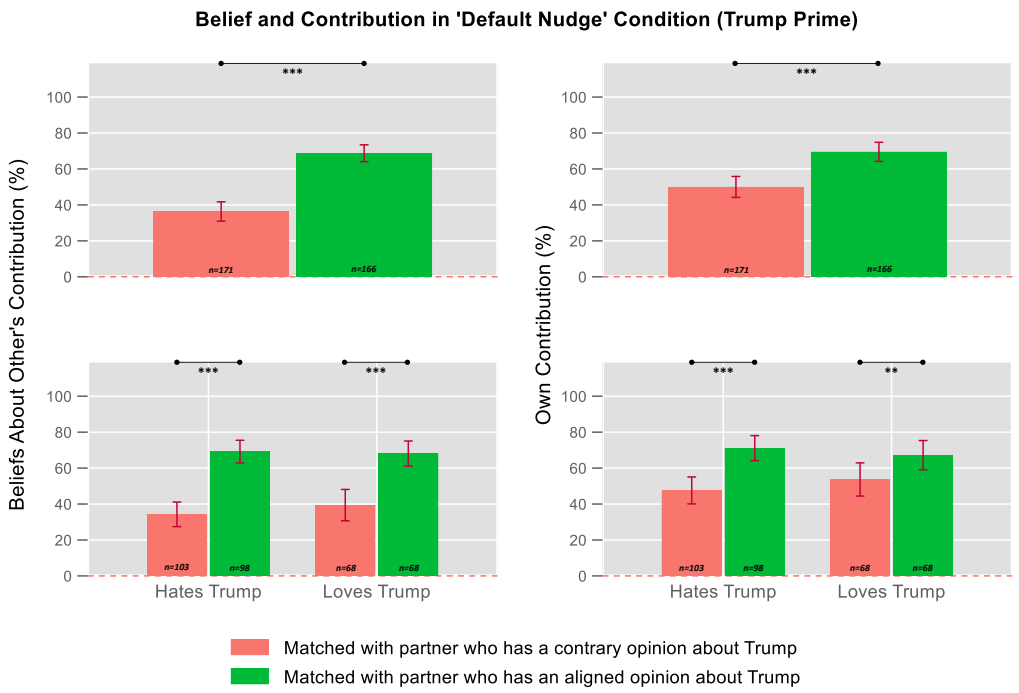

Figure 10: Left Panel: Beliefs and behavior broken down by whether one is matched with a partner who either has aligned or contrary opinions. Right Panel: same but broken down by one's own opinion about Trump. All adjacent bars (within each category) are compared. Absence of significance stars ⇒ p-values > 0.05.

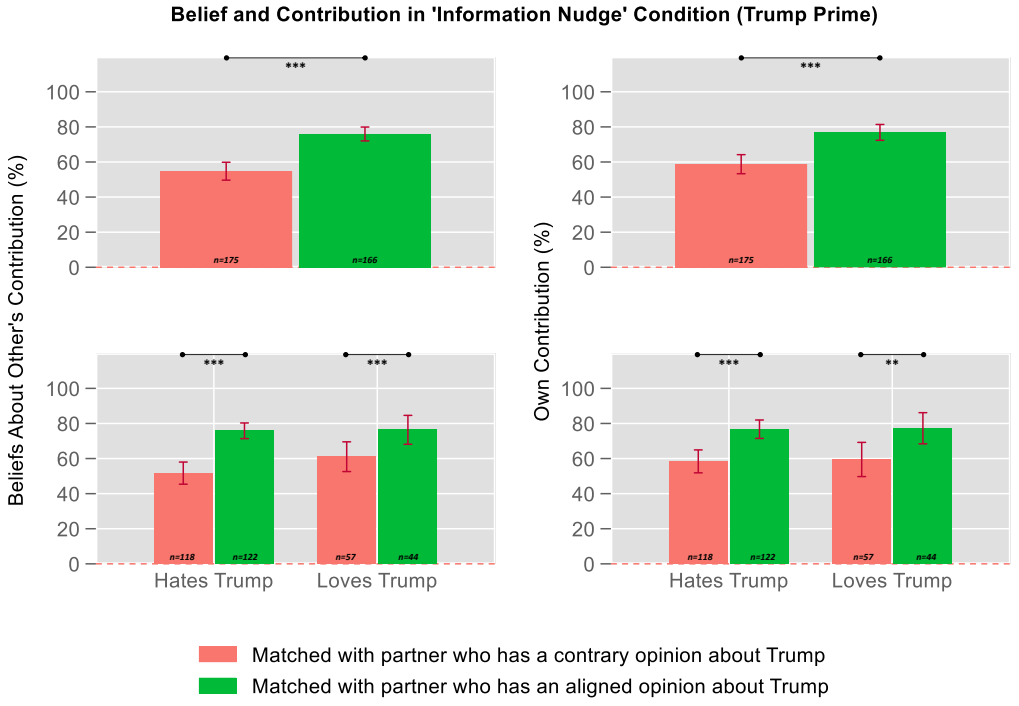

Figure 11: Left Panel: Beliefs and behavior broken down by whether one is matched with a partner who either has aligned or contrary opinions. Right Panel: Same but broken down by one's own opinion about Trump. All adjacent bars (within each category) are compared. Absence of significance stars ⇒ p-values > 0.05.

or lovers).[23] Overall, the magnitude of the persistent polarization gap following the norm-intervention is comparable to the original study (Section 3.2), absent of any intervention. The takeaway message is that attempting to nudge away polarization through means of targeting norm-relevant beliefs is insufficient, both in a strategic and non-strategic setting. While simple behavior interventions have demonstrated great success in other settings, the experiments discussed here emphasize that polarization runs deep and thus needs the additional backing of institutional change rather than behavioral interventions alone.

3.3. Addressing Alternative Explanations

The aim of this section is to provide auxiliary evidence in favor of the robustness of the null findings in the MGP treatment (as presented in Sections 3.1 and 3.2) and evidence against the presence of one plausible driver that could at least partially drive the observed affective and behavioral polarization presented in those sections.

3.3.1. STUDY 3: ON POTENTIAL ORDER EFFECTS OF TRUMP PRIME

First, it is worth stressing why beginning both Studies 1 and 2 with an elicitation of participants' opinions towards Trump was prudent, even in the treatments of the MGP that did not utilize opinions about Trump and instead varied the partner's painting preferences. Such a procedure enabled me to hold any residual effect of the thought about Trump constant across both treatments. Additionally, I was able to break down and compare the data in both treatments by one's own opinion of Trump, which is a necessary comparison when studying ingroup-love and outgroup-hate. Still, because the MGP consistently yielded null-effects in the contexts studied here, a more detailed discussion is in order.

The use of MGPs has a long tradition in social psychology and economics, with varying degrees of effectiveness (for a meta-analysis, see Lane, 2016). Although in line with existing research (e.g., Charness et al., 2007), I have remained agnostic about the presence of observed null effects in the MGP throughout this paper because my sole focus is the direct comparison with the Trump prime. There are a multitude of reasons why the prime did not yield a significant difference in the contexts studied here. Most importantly, however, a majority of the reasons (e.g., MTurkers simply care less about identity primes) would also apply to all other primes (Trump, Biden, sports) presented in this paper and thus unable to explain treatment differences. In what follows, I will highlight two candidate explanations and show that these cannot explain the observed null results: First, "*the design decision to always use the Trump prime first across all treatments (including the MGP) could have washed out any subsequent priming.*" Second, "*these are not 'true' nulls.*"

[23] As illustrated in Figure OA.11, there is some indication that those participants who indicated to hate Trump make an ingroup/outgroup differentiation with respect to the extent to which they free-ride and how conditionally cooperative they are. The same cannot be observed for those who indicated to love Trump.

With respect to the first argument, note that all experiments presented in Section 4 followed the exact MGP procedure in that – instead of using the Klee-Kandinsky paintings prime – participants saw a Biden or sports prime right *after* the Trump prime. In those conditions, both the Biden and sports prime induced stark polarization, both in terms of perception of closeness and altruistic/cooperative behavior in (non-)strategic settings. A more likely alternative explanation, which also happens to be in line with the findings by Chen and Chen (2011), is that much of the existing minimal identity literature picks up differences in domains that are less consequential than the costly and incentive-compatible behaviors studied here in which the payoff trade-offs are palpable. Much of the seminal work on MGP that has its roots in Tajfel and Turner (1979) has resorted to *other-other* rather than *self-other* (as implemented in my experimental paradigms) trade-offs (Charness and Chen, 2020). For this reason, minimal identities might not carry enough weight to induce measurable concerns in the context studied here that contain a dominant strategy and a sufficiently strong social identity changes equilibrium behavior by changing the potential function. This is also in line with a model of group-contingent social preference model à la Chen and Chen (2011) in which the relevant social preference parameter is smaller in MGP than in the non-minimal identities.[24]

To provide final and conclusive evidence for the robustness of the MGP null findings, two variants of the original MGP experiments were run for both the extended Dictator Game and the Public Goods Game (total $n = 808$). The goal is to systematically vary the order of TP and MGP and examine whether the existence of the TP can explain the consistent null findings in the MGP. These two variants are subsumed into Study 3.

The two variants followed the experimental procedure for the MGP condition in the DG and the PGG (as previously illustrated and presented in Figures 1 and 6, except that each variant systematically altered the timing at which the Trump Prime occurred, with all other aspects of the assigned MGP design followed exactly (that is, participants were matched based on their painting preferences):

1. ***Variant 1*** (*MGP First - TP End*), $n = 403$: Participants first selected their preferred painting, then they were presented with either the extended DG or PGG. Following the game's completion, participants were asked to voice their opinion about Trump. Thus, individuals were not made aware of – or had to think about – Trump until after all experimental measures were elicited.
2. ***Variant 2*** (*MGP First - TP Second*), $n = 405$: Participants were first asked to select their preferred painting, then to voice their opinion about Trump. Following these two elicitations, participants then played either the extended DG or PGG.

[24] I would like to thank Yan Chen for raising this point during a presentation of my paper.

As illustrated in Figures OA.13, OA.14, OA.15, OA.16, OA.17, OA.18, OA.19, OA.20, OA.21, and OA.22 in the Online Appendix, the results from these variants are entirely consistent with the previous results: MGP does not produce any significant perceptional or behavioral differences, irrespective of the existence and order of the Trump elicitation.

Regarding the second argument, first note that all experiments were powered according to the effect size that was expected ex ante. In many instances, I over-sampled due to the liberal-leaning nature of MTurkers. Furthermore, I provide additional evidence that those null findings are statistically reliable: I employ both a Bayesian analysis (calculation of a Bayes factor of contingency tables) and the *TOST* procedure examining the existence of the smallest effect size of interest (Lakens, 2017).[25] Both analyses point in the same direction suggesting that the presented null results are indeed reliable and 'true'.

In more detail, the results of the Bayesian analysis are clear: The Bayes factor (B_{01}) for all null-result comparisons of the MGP (perceptions of closeness as well as behavior in both the extended Dictator Game and the Public Goods Game) lies above 1,000 in every single case, yielding 'decisive evidence' in favor of the null (for which the threshold is 10; see Lee and Wagenmakers, 2014). With respect to the *TOST* procedure, first upper and lower equivalence bounds are defined for the smallest effect size of interest (SEOI). To do so, I construct the bounds by taking half of the pre-registered effect size (0.27). For the relevant comparisons discussed in Sections 3.1 and 3.2, all p-values are <0.01 (meaning that the observed effect size is significantly within the bounds), thus further strengthening reliability of the null in the minimal group prime conditions.

3.3.2. One Potential Additional Driver of Polarization in the Studied Behaviors

Studying pro-social and cooperative behavior in stylized environments comes with additional challenges. For example, although the randomization procedure implemented across all presented experiments should plausibly account for most objections, a participant's beliefs about the matched partner's socio-demographic characteristics such as wealth may affect one's social preferences.[26] I am able to provide stylized empirical evidence that this does not appear to be a source of worry in the settings presented above. In a separate incentivized survey[27], participants were asked to guess the income of a partner who either hated Trump

[25] This procedure follows the practice of equivalence testing and examines whether any null effect is close enough to zero to reject the presence of a meaningful difference compared to a smallest effect size of interest.
[26] For example, if a participant (self-servingly) believes that the matched partner is above-/below-average wealthy then this might affect one's willingness to take/give in the extended Dictator Game. However, the evidence on the role of effect sizes in (non-)strategic dilemmas is mixed (see, e.g., Larney et al., 2019).
[27] Sample was obtained on MTurk following the same quality restrictions as in all previous experiments. Total sample of n=300, with n=248 of usable observations after restriction criteria were applied and \$0.5 were paid for participation. Survey started with the obligatory question about one's own opinion of Trump and then proceeded using a within-design (all questions were presented to everyone in random order).

or loved Trump. This allows me to understand whether part of the observed altruistic behavior in the previous experiments can be explained by differential expectations about the matched partner's income (and thus need for money).

The results presented in Figure A.2 in the Appendix paint a reassuring picture: beliefs about estimated income are near-normally distributed (average of ∼$47,000) and do not appear to be mediated by the matched partner's Trump opinion.

4. Probing the Robustness of Polarization in Studies 1 & 2

The experimental conditions and analyses support a clear conclusion up until this point: polarization runs deep and impacts beliefs, behaviors, and the perception of social norms both in strategic and non-strategic settings. Two policy-relevant questions arise naturally:

1. What are the social norms in the (non-)strategic contexts studied here and are the observed behaviors consistent with them?
2. Are the results presented here specific to Donald J. Trump or are they representative of the politically polarized environment in the U.S. more generally?

Thus, in a final step, this paper sheds light on both questions using pre-registered variants of the (non-)strategic behavioral experiments as introduced in Sections 3.1 and 3.2 as well as the social norm elicitation technique by Krupka and Weber (2013).

4.1. STUDY 4: NORMS IN (NON-)STRATEGIC CONTEXTS

With Study 4, I aim to understand whether the social norm perceptions map onto heterogeneous ingroup-love and outgroup-hate by contrasting the results of a new experiment with previously discussed results in Studies 1 and 2 (Sections 3.1 and 3.2). To do so, I analyze the norm perceptions of Trump haters and Trump lovers through the incentive-compatible approach by Krupka and Weber (2013) across various contexts. For altruism, varying norm perceptions by the Left and Right can be expected (Thomsson and Vostroknutov, 2017; Chang et al., 2019) and remain an empirical question in the context of cooperativeness.

4.1.1. Data Collection and Experimental Design

To maximize statistical power and explore all variants of the behavioral experiments, the design of the norm elicitation contains both between- and within-subject variation:

1. ***Between-subject variation:*** As before, participants were presented with a picture of Donald J. Trump and were asked to indicate on a Likert scale how they feel about him. Thus, hate and love towards Trump constitute the between-subject dimension.
2. ***Within-subject variation:*** In random order, participants were informed of the structure of the DG and PGG exactly as it had been explained to participants in

Study 1 and Study 2. Subsequently, using the elicitation technique of Krupka and Weber (2013), participants were asked to rate the appropriateness of various behaviors (presented in random order) in those games.[28] Staying close to the original design, these ratings were elicited from the perspective of being matched with other participants who either had an aligned or misaligned opinion about Trump. Participants observed all variations in random order.[29] These matching variants constitute the three within-subject dimensions.

In sum, the two within- and three between-subject variations represent $2 \times 3 = 6$ dimensions, the same dimensions explored in Studies 1 and 2. To achieve proper statistical power, observations from a total of n=298 participants were collected (leaving me with n=232 participants after applying pre-registered exclusion criteria, neither of which had previously participated in Study 1 or Study 2). Out of these, 162 participants (70%) reported to hate Trump and 70 participants (30%) reported to love Trump – a split comparable to the ones obtained in the first two studies.

4.1.2. Results

For the purpose of exposition – and because the results are so clear – all illustrations and analyses are relegated to the (Online) Appendix (norms in Study 1: Figures A.4, A.5, OA.23; norms in Study 2: Figures A.6, A.7, OA.24, OA.25). These results demonstrate that the perceived social norms map convincingly onto the behaviors observed in both the T-o-G DG and ABC of Cooperation PGG. Accordingly, the elicited norms can explain the observed behavioral differences between Trump haters and Trump lovers as well as their perceptions and attitudes towards people with aligned and misaligned opinions about Donald J. Trump. This has important policy implications: understanding the extent to which behavior (mis)aligns with existing social norms enables policy makers to more effectively nudge norm enforcement (Dimant and Gesche, 2020) and, if needed, complement this with the right combination of norm-messaging and punishment (Bicchieri et al., 2021).

[28] As is customary in this norm elicitation procedure, the participants were asked to rate the appropriateness of the observed behavior along four dimensions: *Very Socially Inappropriate* (VSI), *Somewhat Socially Inappropriate* (SI), *Somewhat Socially Appropriate* (SA), and *Very Socially Appropriate* (VSA). For the dictator game, participants were asked to rate the appropriateness for three distinct behaviors: the dictator making no change to the initial endowments, the dictator taking money from the receiver, and the dictator giving money to the receiver. For the PGG, the participants rated the appropriateness for four distinct behaviors: contribute nothing, contribute nothing when others contribute something (= free-rider), contribute everything (= full cooperator), and contribute more the more the matched partner contributes (= conditional cooperator). Other results are relegated to the Online Appendix.

[29] Importantly, to ensure reliable norm-inferences, each participant's beliefs were elicited only from the perspective of one's own opinion about Trump: A participant who indicated to hate Trump would only be asked to rate the appropriateness of various behaviors based on the matching of a Trump-hater with either another Trump-hater or a Trump-lover. Similarly, a Trump-lover would only be asked to rate the appropriateness from the perspective of another Trump-lover having been matched with one of these partners.

4.2. STUDY 5: EXAMINING ALTERNATIVE POLARIZING (NON-)POLITICAL PRIMES

To further investigate the robustness of the results as presented in Studies 1 and 2 (see Sections 3.1 and 3.2 for more details), I devised four additional pre-registered experiments that examine polarization in the context of both the 46$^{\text{th}}$ president of the United States, Joseph R. Biden, and sports. Using these two new primes, I will be further examining beliefs, attitudes, and altruistic and cooperative behaviors both in *strategic* and *non-strategic* settings. The goal is to understand whether the obtained results from my main experimental analysis – stark ingroup-love/outgroup-hate differentiation along the dimensions of perceptions of closeness and behavior – are specific to Trump, or rather are representative of a societal rift in the U.S. more generally (Stewart et al., 2020). Rather than eliciting participant opinions of Trump, I will be instead utilizing participants' hate/love towards Joe Biden and sports preferences as identity markers.

The choice of Joe Biden as the first stimuli in this robustness test is straightforward since he was Donald Trump's opponent in the 2020 presidential election and was subsequently elected. Whether the polarization that Biden produces is the flip-side of Trump's polarization is an empirical question that the implemented experiments will answer. For the second identity marker, I attempted to find a setting that is polarized to a comparable degree in the U.S. but is largely unpolitical. To find such a setting, I used data from the dating app *Hater* to find the most contentious topics.[30] As illustrated in Figure A.8, among the most contentious topics are sports (such as soccer, football, mixed martial arts, and lacrosse). This is consistent with research suggesting that sports fandom is known to be a highly polarizing topic in the U.S. (Klein, 2020). I subsumed this under 'sports' and used them as a prime in the experiments, as explained in more detail below. I anticipate that the impact of these identity primes are closer to TP than to MGP because when group identities are strong and signal social preferences, stronger effects might be expected than in a minimal group setting (Fehrler and Kosfeld, 2013).[31]

4.2.1. Data Collection and Experimental Design

In designing the procedure, I followed the experimental protocol of Studies 1 and 2 (see sections 3.1.1 and 3.2.1). The structure of the design is borrowed directly from the MGP conditions of the original extended DG and PGG with the *only* difference being that –

[30] The dating app 'Hater', backed by Mark Cuban, utilizes repulsion as a social glue to facilitate love connections (with self-reported success). It matches people based on their joint hate along several dimensions, including fandom for celebrities, food (e.g., pineapple on pizza), lifestyle choices, religion, or sports.

[31] Suggestive evidence collected with a separate set of $n = 200$ participants on MTurk indicates that the polarization of primes implemented and reported in this paper are as follows: political identity > sports identity > minimal group identity. Results are available upon request.

rather than seeing and indicating their preference for Klee and Kandinsky paintings (after seeing a picture of Trump) – participants instead first saw the same picture of Trump and then either a picture of Joe Biden or a generic picture representing sports. Subsequently, just like in the MGP conditions with paintings, participants were randomly matched with other participants and received information about either their partners' *Biden* or *sports* preferences. To that end, behavior from $n = 2,259$ participants ($n = 1,367$ for the Biden prime and $n = 892$ for the sports prime) were collected between December 2020 and January 2021 (between the 2020 election and the official inauguration of President Biden).

4.2.2. Results

Biden Prime: The results indicate that the original Trump-related insights from Studies 1 and 2 replicate using a Biden prime instead of a Trump prime (for brevity, the illustrations and statistical analyses are relegated to the (Online) Appendix). This includes the striking ingroup-love/outgroup-hate differentiation in terms of:

1. Perception of closeness and the shape of altruism in a non-strategic context (Figures OA.26, OA.27, and OA.28 in the Online Appendix), as was demonstrated by the original DG conditions in Studies 1 and 2.
2. Perception of closeness and the shape of attitudes, beliefs, and cooperation in a strategic context (Figures OA.29, OA.30, and OA.31 in the Online Appendix), as shown in the original PGG conditions in Studies 1 and 2.

With that, I conclude that the previously observed results are reflective of a deeper societal rift that is not limited to Trump but rather extends to the political domain more generally.

Sports Prime: For the sports prime, the results are more nuanced:

1. For both the perception of closeness and the shape of altruism in a non-strategic setting, (Figures OA.34, OA.33, and OA.34 in the Online Appendix) the results are consistent with the previous Trump prime results: Dictators are pro-social (give ∼10% of the money) to those with an aligned opinion about sports and anti-social (take ∼10% of the money) to those with a contrary opinion about sports.
2. In the strategic context, the sports prime produces no differences with respect to attitudes, beliefs, or cooperation (Figures OA.35, OA.36, and OA.37 in the Online Appendix), with the results most closely resembling those observed in the MGP.

Taken together, the results from the sports prime suggest that substantial polarization also exists in non-political contexts.

5. Conclusion and Discussion

At its core, this paper investigates the effects of polarization and helps to understand how behavior and perceptions changes in response to it. In particular, I am is concerned with quantifying the extent to which polarization affects pro- and anti-social behavior, cooperativeness, and the perception of social norms with respect to these behaviors. I achieve this by implementing 5 pre-registered studies, comprising 15 incentive-compatible behavioral experiments and a diverse set of over 8,600 participants. To disentangle ingroup-love from outgroup-hate, I embed polarization by capitalizing on participants' negative/positive opinions about Donald J. Trump and comparing the outcomes to those observed in treatments using the minimal identity paradigm.

Along all investigated dimensions, I obtain compelling evidence for the following results: for one, polarization produces ingroup/outgroup differentiation in all studied settings, leading participants to actively harm and cooperate less with participants from the opposing faction. In addition, a lack of cooperation is not the result of a categorical unwillingness to cooperate across factions, but rather, is based on one's lack of expectation on the other participant's willingness to cooperate. Importantly, the results also cast light on the nuance with which ingroup-love and outgroup-hate – something that existing literature often takes as being two sides of the same coin – occurs. In particular, in comparing behavior between the Trump prime and Minimal Group Paradigm prime treatments, the results suggest that ingroup-love can be observed in terms of feeling close to one another, whereas outgroup-hate appears in the form of taking money away from and being less cooperative with others. The elicited norms are consistent with these observations, and also highlight that those who love Trump have a much weaker ingroup/outgroup differentiation than those who hate Trump. Through additional robustness experiments using one's opinion about Joe Biden and sports fandom as identity marker, I also find that the observed behavioral and perceptional effects caused by polarization are not limited to Trump.

Taken together, these results indicate a larger ideological rift in American society – one that I've shown to have detrimental consequences on the way people perceive one another, exhibit altruistic behavior, and show the ability to cooperate. Lastly, I attempt to reduce polarization using established behavioral interventions that utilize the concept of nudging. My findings provide new insights on the limits of light-touch solutions in changing behavior, in particular with respect to alleviating polarization.

This paper ultimately provides evidence for how exacerbated the intergroup animosities currently are in the U.S., symptomatic of a larger global societal rift (Baldassarri and Bearman, 2007; Carothers and O'Donohue, 2019). From a policy perspective, the findings in this paper contribute to affective polarization insights: alleviating the pernicious outcomes of polarization by correcting the misguided beliefs about the preferences and actions of the op-

posing faction is challenging.[32] Succeeding at this, however, is particularly important, since people are found to preferentially consume and engage with information that aligns with their prior belief. This can further aggravate partisan rifts (Dorison et al., 2019; Shi et al., 2019; Allcott et al., 2020; Bicchieri et al., 2020a,b; Dimant et al., 2020; Schwalbe et al., 2020; Dimant et al., 2021; Guess et al., 2021; Levy, 2021; Osmundsen et al., 2021). Importantly, however, the behavioral interventions tested in this paper indicates a lack of effectiveness in closing the polarization gap. This suggests that structural – on top of behavioral – changes are needed to overcome the political rift within American society.

[32] Existing research points to ways in which such social beliefs can be corrected – or at the least abate their inaccuracy – and utilized in the context of polarization (Flynn et al., 2017; Ahler and Sood, 2018; Stanley et al., 2020, for a cross-cultural perspective see Ruggeri et al., 2021). However, the set of studies presented here has one crucial advantage – and differs from those studies – in that I employ incentive-compatible behavioral experiments to measure the perceptional and behavioral impact of polarization. The literature mentioned-above does so using (often non-incentivized or not incentive-compatible) surveys. An avenue for future research is placing more weight on the role of context and methodological approaches. Further investigation should also be conducted to uncover the various forms in which polarization occurs and can be contained when decisions are incentive-compatible.

References

Abramowitz, A. I. and Saunders, K. L. (2006). Exploring the bases of partisanship in the American electorate: Social identity vs. ideology. *Political Research Quarterly*, 59(2):175–187.

Abramowitz, A. I. and Webster, S. W. (2018). Negative partisanship: Why americans dislike parties but behave like rabid partisans. *Political Psychology*, 39:119–135.

Ahler, D. J. and Sood, G. (2018). The parties in our heads: Misperceptions about party composition and their consequences. *The Journal of Politics*, 80(3):964–981.

Akerlof, G. A. (1997). Social distance and social decisions. *Econometrica: Journal of the Econometric Society*, pages 1005–1027.

Akerlof, G. A. and Kranton, R. E. (2000). Economics and identity. *The Quarterly Journal of Economics*, 115(3):715–753.

Alesina, A., Baqir, R., and Easterly, W. (1999). Public goods and ethnic divisions. *The Quarterly Journal of Economics*, 114(4):1243–1284.

Allcott, H., Braghieri, L., Eichmeyer, S., and Gentzkow, M. (2020). The welfare effects of social media. *American Economic Review*, 110(3):629–76.

Allcott, H. and Kessler, J. B. (2019). The welfare effects of nudges: A case study of energy use social comparisons. *American Economic Journal: Applied Economics*, 11(1):236–76.

Altmann, S., Falk, A., Heidhues, P., Jayaraman, R., and Teirlinck, M. (2019). Defaults and donations: Evidence from a field experiment. *Review of Economics and Statistics*, 101(5):808–826.

Amira, K., Wright, J. C., and Goya-Tocchetto, D. (2019). In-group love versus out-group hate: Which is more important to partisans and when? *Political Behavior*, pages 1–22.

Aron, A., Aron, E. N., and Smollan, D. (1992). Inclusion of other in the self scale and the structure of interpersonal closeness. *Journal of Personality and Social Psychology*, 63(4):596.

Autor, D., Dorn, D., Hanson, G., and Majlesi, K. (2020). Importing political polarization? the electoral consequences of rising trade exposure. *American Economic Review*, 110(10):3139–83.

Baldassarri, D. and Bearman, P. (2007). Dynamics of political polarization. *American Sociological Review*, 72(5):784–811.

Bardsley, N. (2008). Dictator game giving: altruism or artefact? *Experimental Economics*, 11(2):122–133.

Bénabou, R., Falk, A., and Tirole, J. (2018). Narratives, imperatives, and moral reasoning. Working Paper 24798, National Bureau of Economic Research.

Bénabou, R. and Tirole, J. (2011). Identity, morals, and taboos: Beliefs as assets. *The Quarterly Journal of Economics*, 126(2):805–855.

Bénabou, R. and Tirole, J. (2016). Mindful economics: The production, consumption, and value of beliefs. *Journal of Economic Perspectives*, 30(3):141–64.

Benartzi, S., Beshears, J., Milkman, K. L., Sunstein, C. R., Thaler, R. H., Shankar, M., Tucker-Ray, W., Congdon, W. J., and Galing, S. (2017). Should governments invest more in nudging? *Psychological Science*, 28(8):1041–1055.

Bernhard, H., Fischbacher, U., and Fehr, E. (2006). Parochial altruism in humans. *Nature*, 442(7105):912–915.

Beshears, J. and Kosowsky, H. (2020). Nudging: Progress to date and future directions. *Organizational Behavior and Human Decision Processes*, 161:3 – 19.

Bicchieri, C. (2006). *The grammar of society: The nature and dynamics of social norms*. Cambridge University Press.

Bicchieri, C. and Dimant, E. (2019). Nudging with Care: The Risks and Benefits of Social Information. *Public Choice*.

Bicchieri, C., Dimant, E., Gächter, S., and Nosenzo, D. (2020a). Social proximity and the erosion of norm compliance. Working Paper Available at SSRN: https://ssrn.com/abstract=3355028.

Bicchieri, C., Dimant, E., and Sonderegger, S. (2020b). It's not a lie if you believe the norm does not apply: Conditional norm-following with strategic beliefs. Working Paper Available at SSRN: https://dx.doi.org/10.2139/ssrn.3326146.

Bicchieri, C., Dimant, E., and Xiao, E. (2021). Deviant or wrong? The effects of norm information on the efficacy of punishment. *Journal of Economic Behavior & Organization*.

Bott, K. M., Cappelen, A. W., Sorensen, E., and Tungodden, B. (2019). You've got mail: A randomised field experiment on tax evasion. *Management Science*.

Bowles, S. and Gintis, H. (2013). *A cooperative species: Human reciprocity and its evolution*. Princeton University Press.

Buhrmester, M. D., Talaifar, S., and Gosling, S. D. (2018). An evaluation of amazon's mechanical turk, its rapid rise, and its effective use. *Perspectives on Psychological Science*, 13(2):149–154.

Bursztyn, L., Egorov, G., and Fiorin, S. (2020a). From extreme to mainstream: The erosion of social norms. *American Economic Review*, 110(11):3522–48.

Bursztyn, L., González, A. L., and Yanagizawa-Drott, D. (2020b). Misperceived social norms: Women working outside the home in saudi arabia. *American economic review*, 110(10):2997–3029.

Cantoni, D., Yang, D. Y., Yuchtman, N., and Zhang, Y. J. (2019). Protests as strategic games: experimental evidence from hong kong's antiauthoritarian movement. *The Quarterly Journal of Economics*, 134(2):1021–1077.

Carothers, T. and O'Donohue, A. (2019). *Democracies divided: The global challenge of political polarization*. Brookings Institution Press.

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

Charness, G. and Chen, Y. (2020). Social identity, group behavior, and teams. *Annual Review of Economics*, 12:691–713.

Charness, G., Rigotti, L., and Rustichini, A. (2007). Individual behavior and group membership. *American Economic Review*, 97(4):1340–1352.

Chen, R. and Chen, Y. (2011). The potential of social identity for equilibrium selection. *American Economic Review*, 101(6):2562–89.

Chen, Y. and Li, S. X. (2009). Group identity and social preferences. *American Economic Review*, 99(1):431–57.

Christ, O., Schmid, K., Lolliot, S., Swart, H., Stolle, D., Tausch, N., Al Ramiah, A., Wagner, U., Vertovec, S., and Hewstone, M. (2014). Contextual effect of positive intergroup contact on outgroup prejudice. *Proceedings of the National Academy of Sciences*, 111(11):3996–4000.

Croson, R. and Shang, J. Y. (2008). The impact of downward social information on contribution decisions. *Experimental Economics*, 11(3):221–233.

Damgaard, M. T. and Gravert, C. (2018). The hidden costs of nudging: Experimental evidence from reminders in fundraising. *Journal of Public Economics*, 157:15–26.

Dimant, E. (2019). Contagion of pro-and anti-social behavior among peers and the role of social proximity. *Journal of Economic Psychology*, 73:66–88.

Dimant, E., Galeotti, F., and Villeval, M. C. (2021). Norm-formation and the role of endogenous information acquisition. Mimeo.

Dimant, E., Gerben, A. v. K., and Shalvi, S. (2020). Requiem for a nudge: Framing effects in nudging honest. *Journal of Economic Behavior & Organization*, 172:247–266.

Dimant, E. and Gesche, T. (2020). Nudging enforcers: How norm perceptions and motives for lying shape sanctions. Working Paper Available at SSRN: https://dx.doi.org/10.2139/ssrn.3664995.

Dixit, A. K. and Weibull, J. W. (2007). Political polarization. *Proceedings of the National Academy of Sciences*, 104(18):7351–7356.

Dorison, C. A., Minson, J. A., and Rogers, T. (2019). Selective exposure partly relies on faulty affective forecasts. *Cognition*, 188:98–107.

Druckman, J. N. and Levendusky, M. S. (2019). What do we measure when we measure affective polarization? *Public Opinion Quarterly*, 83(1):114–122.

Efferson, C., Lalive, R., and Fehr, E. (2008). The coevolution of cultural groups and ingroup favoritism. *Science*, 321(5897):1844–1849.

Fehr, E. and Fischbacher, U. (2003). The nature of human altruism. *Nature*, 425(6960):785–791.

Fehrler, S. and Kosfeld, M. (2013). Can you trust the good guys? trust within and between groups with different missions. *Economics Letters*, 121(3):400–404.

Fiorina, M. P. and Abrams, S. J. (2008). Political polarization in the american public. *Annual Review of Political Science*, 11:563–588.

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

Flynn, D., Nyhan, B., and Reifler, J. (2017). The nature and origins of misperceptions: Understanding false and unsupported beliefs about politics. *Political Psychology*, 38:127–150.

Fowler, J. H. and Kam, C. D. (2007). Beyond the self: Social identity, altruism, and political participation. *The Journal of Politics*, 69(3):813–827.

Gächter, S., Kölle, F., and Quercia, S. (2017). Reciprocity and the tragedies of maintaining and providing the commons. *Nature Human Behaviour*, 1(9):650.

Gächter, S., Starmer, C., and Tufano, F. (2015). Measuring the closeness of relationships: a comprehensive evaluation of the 'inclusion of the other in the self' scale. *PloS one*, 10(6).

Gennaioli, N. and Tabellini, G. (2019). Identity, beliefs, and political conflict. *Working Paper*.

Goswami, I. and Urminsky, O. (2016). When should the ask be a nudge? the effect of default amounts on charitable donations. *Journal of Marketing Research*, 53(5):829–846.

Graham, M. H. and Svolik, M. W. (2020). Democracy in America? Partisanship, Polarization, and the Robustness of Support for Democracy in the United States. *American Political Science Review*, 114(2):392–409.

Greene, S. (1999). Understanding party identification: A social identity approach. *Political Psychology*, 20(2):393–403.

Guess, A. M., Barberá, P., Munzert, S., and Yang, J. (2021). The consequences of online partisan media. *Proceedings of the National Academy of Sciences*, 118(14).

Halevy, N., Bornstein, G., and Sagiv, L. (2008). "in-group love" and "out-group hate" as motives for individual participation in intergroup conflict: A new game paradigm. *Psychological Science*, 19(4):405–411.

Hallsworth, M., List, J. A., Metcalfe, R. D., and Vlaev, I. (2017). The behavioralist as tax collector: Using natural field experiments to enhance tax compliance. *Journal of Public Economics*, 148:14–31.

Hara, K., Adams, A., Milland, K., Savage, S., Callison-Burch, C., and Bigham, J. P. (2018). A data-driven analysis of workers' earnings on amazon mechanical turk. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, page 449. ACM.

Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., Gintis, H., and McElreath, R. (2001). In search of homo economicus: behavioral experiments in 15 small-scale societies. *American Economic Review*, 91(2):73–78.

Hummel, D. and Maedche, A. (2019). How effective is nudging? a quantitative review on the effect sizes and limits of empirical nudging studies. *Journal of Behavioral and Experimental Economics*, 80:47–58.

Isaac, R. M. and Walker, J. M. (1988). Communication and free-riding behavior: The voluntary contribution mechanism. *Economic Inquiry*, 26(4):585–608.

Iyengar, S., Lelkes, Y., Levendusky, M., Malhotra, N., and Westwood, S. J. (2019). The origins and consequences of affective polarization in the united states. *Annual Review of Political Science*, 22:129–146.

Iyengar, S. and Westwood, S. J. (2015). Fear and loathing across party lines: New evidence on group polarization. *American Journal of Political Science*, 59(3):690–707.

Jachimowicz, J. M., Duncan, S., Weber, E. U., and Johnson, E. J. (2019). When and why defaults influence decisions: A meta-analysis of default effects. *Behavioural Public Policy*, 3(2):159–186.

Jacobson, G. C. (2019). *Presidents and Parties in the Public Mind*. University of Chicago Press.

Klein, E. (2020). *Why We're Polarized*. Avid Reader Press / Simon & Schuster.

Kranton, R., Pease, M., Sanders, S., and Huettel, S. (2020). Deconstructing bias in social preferences reveals groupy and not-groupy behavior. *Proceedings of the National Academy of Sciences*, 117(35):21185–21193.

Kranton, R. E. and Sanders, S. G. (2017). Groupy versus non-groupy social preferences: Personality, region, and political party. *American Economic Review*, 107(5):65–69.

Krupka, E. L. and Weber, R. A. (2013). Identifying social norms using coordination games: Why does dictator game sharing vary? *Journal of the European Economic Association*, 11(3):495–524.

Lakens, D. (2017). Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Social Psychological and Personality Science*, 8(4):355–362.

Lane, T. (2016). Discrimination in the laboratory: A meta-analysis of economics experiments. *European Economic Review*, 90:375–402.

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

Lee, M. D. and Wagenmakers, E.-J. (2014). *Bayesian cognitive modeling: A practical course*. Cambridge university press.

Lees, J. and Cikara, M. (2020). Inaccurate group meta-perceptions drive negative out-group attributions in competitive contexts. *Nature Human Behaviour*, 4(3):279–286.

Lelkes, Y. and Westwood, S. J. (2017). The limits of partisan prejudice. *The Journal of Politics*, 79(2):485–501.

Levendusky, M. S. and Malhotra, N. (2016). (Mis) perceptions of partisan polarization in the American public. *Public Opinion Quarterly*, 80(S1):378–391.

Levy, R. (2021). Social media, news consumption, and polarization: evidence from a field experiment. *American Economic Review*, 111(3):831–70.

List, J. A. (2007). On the interpretation of giving in dictator games. *Journal of Political Economy*, 115(3):482–493.

Madestam, A., Shoag, D., Veuger, S., and Yanagizawa-Drott, D. (2013). Do political protests matter? evidence from the tea party movement. *Quarterly Journal of Economics*, 128(4):1633–1685.

Madrian, B. C. and Shea, D. F. (2001). The power of suggestion: Inertia in 401 (k) participation and savings behavior. *The Quarterly Journal of Economics*, 116(4):1149–1187.

Mason, L. (2015). "I disrespectfully agree": The differential effects of partisan sorting on social and issue polarization. *American Journal of Political Science*, 59(1):128–145.

Mason, L. (2018). *Uncivil agreement: How politics became our identity*. University of Chicago.

Mazumder, S. (2018). The persistent effect of us civil rights protests on political attitudes. *American Journal of Political Science*, 62(4):922–935.

Meyer, D. S. (2004). Protest and political opportunities. *Annu. Rev. Sociol.*, 30:125–145.

Michelitch, K. (2015). Does electoral competition exacerbate interethnic or interpartisan economic discrimination? evidence from a field experiment in market price bargaining. *The American Political Science Review*, 109(1):43.

Moffatt, P. G. (2015). *Experimetrics: Econometrics for experimental economics*. Palgrave.

Moore-Berg, S. L., Ankori-Karlinsky, L.-O., Hameiri, B., and Bruneau, E. (2020). Exaggerated meta-perceptions predict intergroup hostility between American political partisans. *Proceedings of the National Academy of Sciences*.

Müller, K. and Schwarz, C. (2020). Fanning the flames of hate: Social media and hate crime. *Journal of the European Economic Association*.

Offerman, T. (2002). Hurting hurts more than helping helps. *European Economic Review*, 46(8):1423–1437.

Orr, L. V. and Huber, G. A. (2020). The policy basis of measured partisan animosity in the United States. *American Journal of Political Science*, 64(3):569–586.

Osmundsen, M., Bor, A., Vahlstrup, P. B., Bechmann, A., and Petersen, M. B. (2021). Partisan polarization is the primary psychological motivation behind "fake news" sharing on twitter. *American Political Science Review*.

Reich, M. (2017). *Racial inequality: A political-economic analysis*. Princeton University Press.

Robbins, J. M. and Krueger, J. I. (2005). Social projection to ingroups and outgroups: A review and meta-analysis. *Personality and Social Psychology Review*, 9(1):32–47.

Ruggeri, K. et al. (2021). The general fault in our fault lines. *Nature Human Behaviour*.

Schwalbe, M. C., Cohen, G. L., and Ross, L. D. (2020). The objectivity illusion and voter polarization in the 2016 presidential election. *Proceedings of the National Academy of Sciences*.

Shi, F., Teplitskiy, M., Duede, E., and Evans, J. A. (2019). The wisdom of polarized crowds. *Nature Human Behaviour*, 3(4):329–336.

Stanley, M. L., Whitehead, P. S., Sinnott-Armstrong, W., and Seli, P. (2020). Exposure to opposing reasons reduces negative impressions of ideological opponents. *Journal of Experimental Social Psychology*, 91:104030.

Stewart, A. J., McCarty, N., and Bryson, J. J. (2020). Polarization under rising inequality and economic decline. *Science Advances*, 6(50):eabd4201.

Tajfel, H. and Turner, J. C. (1979). An integrative theory of intergroup conflict. *The Social Psychology of Intergroup Relations*, 33(47):74.

Thaler, R. and Sunstein, C. (2008). *Nudge: Improving Decisions about Health, Wealth, and Happiness*. Yale University Press.

Thaler, R. H. and Benartzi, S. (2004). Save more tomorrow$^{\text{TM}}$: Using behavioral economics to increase employee saving. *Journal of Political Economy*, 112(S1):S164–S187.

Thomsson, K. M. and Vostroknutov, A. (2017). Small-world conservatives and rigid liberals: Attitudes towards sharing in self-proclaimed left and right. *Journal of Economic Behavior & Organization*, 135:181–192.

West, E. A. and Iyengar, S. (2020). Partisanship as a social identity: Implications for polarization. *Political Behavior*, pages 1–32.

Yamagishi, T. and Mifune, N. (2009). Social exchange and solidarity: in-group love or out-group hate? *Evolution and Human Behavior*, 30(4):229–237.

Main Appendix

Appendix A. Additional Results Using Trump Primes in Take-or-Give Dictator Game(Section 3.1, Study 1) & ABC of Cooperation Public Goods Game (Section 3.2, Study 2)

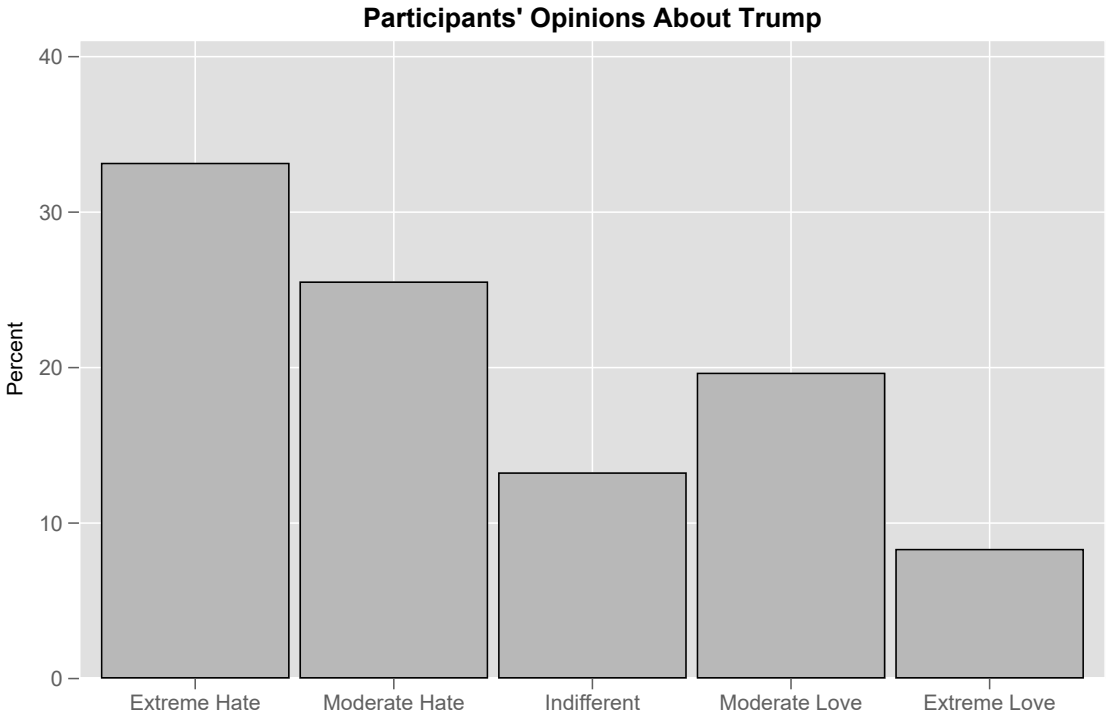

Figure A.1: Histogram of Trump opinions for both TP and MGP treatments.

Table A.1: OLS Regression Analysis of T-o-G Dictator Game Behavior

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Matched with Aligned	38.491***	31.762***	32.752***			
Trump Preference	(5.593)	(6.537)	(6.927)			
Closeness Score (%)		0.176*	0.182*		0.222***	0.244***
		(0.091)	(0.098)		(0.078)	(0.081)
Loves Trump				-32.514***	-23.298**	-17.790*
				(8.837)	(9.316)	(10.428)
Matched with Undisclosed				-15.075**	-7.334	-5.913
				(6.551)	(6.999)	(7.127)
Matched with Trump Lover				-36.979***	-28.519***	-29.041***
				(6.876)	(7.423)	(7.599)
Loves Trump ×				53.546***	44.381***	43.285***
Matched with Undisclosed				(12.054)	(12.269)	(12.441)
Loves Trump ×				78.557***	61.561***	58.618***
Matched with Trump Lover				(11.819)	(13.191)	(13.500)
Constant	-23.514***	-30.548***	-35.416	12.083***	-5.472	-31.785
	(4.398)	(5.623)	(36.886)	(4.302)	(7.286)	(30.163)
Controls	No	No	Yes	No	No	Yes
Observations	423	423	416	598	598	586

DV: Dictator game behavior (neg. = taking; pos. = giving). Standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A.2: OLS Regression Analysis of PGG Contribution Behavior

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Matched with Aligned	13.427***	4.355	5.281			
Trump Preference	(3.815)	(4.412)	(4.506)			
Closeness Score (%)		0.250***	0.196***		0.217***	0.156***
		(0.064)	(0.065)		(0.053)	(0.054)
Loves Trump				-10.873**	-3.703	-8.452
				(5.336)	(5.392)	(6.307)
Matched with Undisclosed				-7.643	-1.019	-4.873
				(4.870)	(5.081)	(5.175)
Matched with Trump Lover				-17.200***	-8.490	-10.530*
				(4.974)	(5.453)	(5.490)
Loves Trump ×				22.953***	12.896*	14.058*
Matched with Undisclosed				(7.593)	(7.751)	(7.897)
Loves Trump ×				23.904***	8.696	10.640
Matched with Trump Lover				(7.662)	(8.359)	(8.216)
Constant	51.029***	39.658***	-40.529**	65.748***	47.743***	-11.477
	(2.805)	(4.132)	(16.652)	(3.287)	(5.711)	(13.972)
Controls	No	No	Yes	No	No	Yes
Observations	388	388	373	537	537	519

Dependent variable is a participants contribution (%) to the PGG. Robust standard errors in parentheses.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Appendix B. Additional Robustness Checks: Survey Eliciting Beliefs About Matched Income & Trump Opinion (Section 3.3, Study 3)

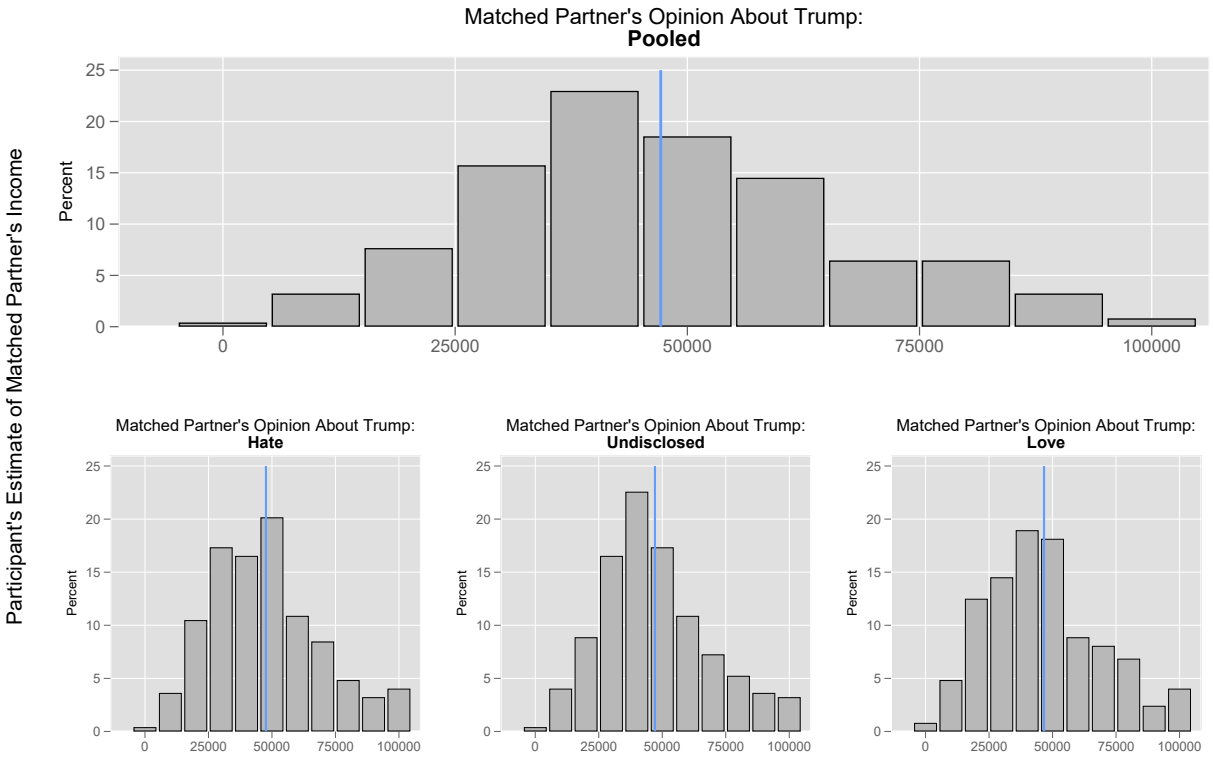

Figure A.2: A participant's beliefs of the matched partner's income level. Top panel: beliefs pooled across partner's opinion about Trump. Bottom panels: beliefs broken down by the matched partner's opinion about Trump (hate/undisclosed/love, respectively). Vertical blue lines represent the averages.

In the follow-up survey as explained in Section 3.3.2, participants were also asked to guess a matched partner's most likely opinion about Trump (that partner's Trump opinion was undisclosed to the participant). This would allow me to capture the beliefs that MTurkers had about their partner's political ideology when such information was not provided.

The results suggest that participants are more likely to assume that the matched partner's undisclosed Trump opinion is the same as their own. Such *social projection* (Robbins and Krueger, 2005; Gennaioli and Tabellini, 2019) could also explain why pro-social and cooperative behavior towards the undisclosed group has frequently shown to be similar to the behavior towards one's ingroup (see, e.g., Figures OA.2 and OA.7).

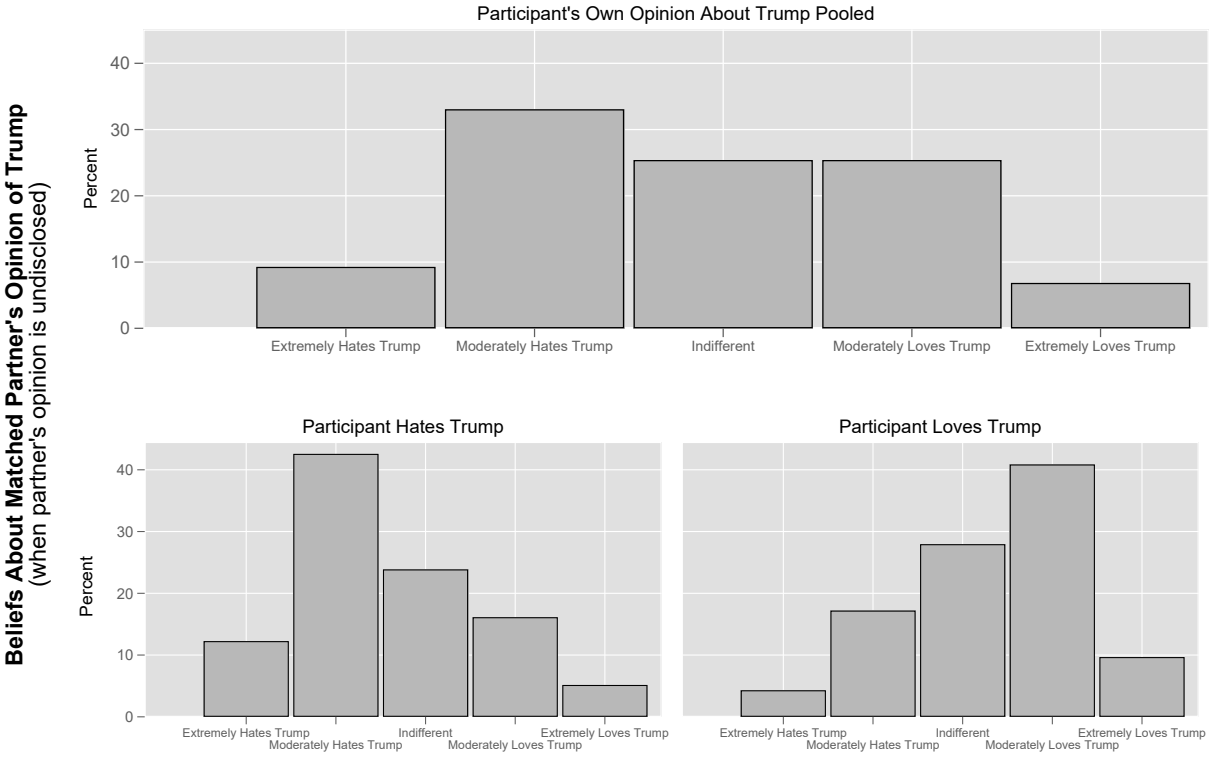

Figure A.3: A participant's beliefs of their matched partner's opinion about Trump when their true opinion was known but not disclosed in the experiment. Top panel: beliefs pooled across the participant's own opinion about Trump (pooled for everyone except those who indicated to be indifferent about Trump, since these are dropped as per pre-registration). Bottom-left and bottom-right panels: beliefs broken down by the participant's own opinion about Trump (hate and love, respectively).

Appendix C. Trump Prime Additional Results: Social Norms in (Non-)Strategic Contexts (Section 4.1, Study 4)

Appendix C.1. Results for the T-o-G Dictator Game

In this section, the results from the dictator game, as discussed in Section 3.1.3 (especially in Figures 2 and 3), will be analyzed through the lens of a norm elicitation that follows the method of Krupka and Weber (2013).[33]

The first set of results is presented in Figure A.4 and paints a picture that is extremely consistent with both the observed closeness and dictator behavior behavior, as illustrated in Figure 2: when matched with a partner who has an **aligned** opinion about Trump, taking from (giving to) that partner is perceived as more inappropriate (more appropriate) compared to when matched with a partner who has a **misaligned** opinion about Trump (all differences significant at p<0.001 using BSM tests).

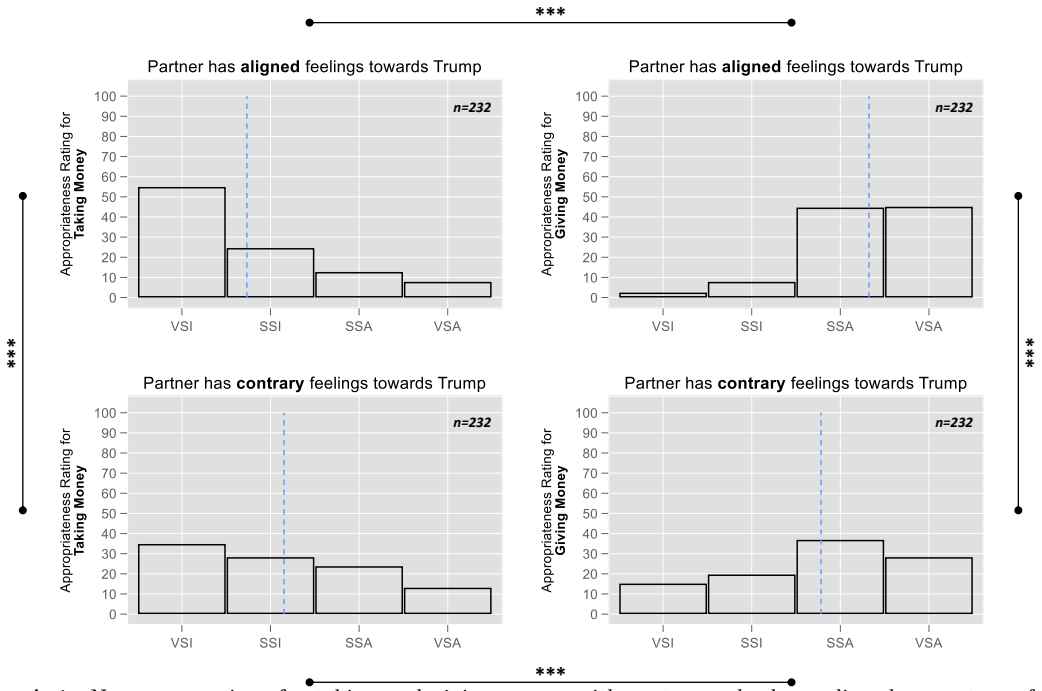

Figure A.4: Norm perceptions for taking and giving money with partners who have aligned or contrary feelings towards Trump. All adjacent quadrants are tested and statistical significance (if either ***p<0.01 or **p<0.05) is indicated where applicable. *Very Socially Inappropriate* (VSI), *Somewhat Socially Inappropriate* (SSI), *Somewhat Socially Appropriate* (SSA), and *Very Socially Appropriate* (VSA).

[33]Consistent with the previous analyses, the main focus remains the behavior towards partners with the same or contrary opinion of Trump. Further below, I also present analyses that include perceptions when matched with a partner for whom the opinion about Trump remains undisclosed.

These insights complement the results from Figure 2 and suggest that the observed differences in feeling of closeness and pro-sociality towards a partner who has an aligned opinion about Trump go hand in hand with the norm perception that this is indeed the right thing to do, whereas it is perceived to be more appropriate to harm someone with a contrary opinion about Trump.

Next, following the previous analyses in Figure 3, I analyze the norm perceptions conditional on one's own opinion about Trump and present the results separately for taking behavior (top of Figure A.5) and giving behavior (bottom of Figure A.5). As before, the norm elicitations are consistent with the observed closeness and dictator game behaviors.

For taking behavior, those who identified as *Trump haters* indicate that it is more acceptable to take from a *Trump-lover* (TH-TL) than from a fellow *Trump-hater* (TH-TH), which is highly statistically significant (BSM, $p<0.001$). Conversely, I do not observe the same difference for those who identified as *Trump lovers* (BSM, $p=0.81$), which is primarily driven by the fact that those who are matched with their own kind have a substantially higher approval for taking money from their partner than Trump haters have when matched with their own kind (comparing TL-TL vs. TH-TH, BSM, $p<0.001$).

Consistent with the theme of this paper, these results indicate that hate evokes stronger norms against harming each other, and additionally, those who love Trump do not distinguish between ingroup-love and outgroup-hate with respect to harming others.

In terms of giving behavior, one can observe that being matched with a participant with the same preference for Trump leads to a significantly higher appropriateness rating compared to giving to a participant with a misaligned opinion of Trump (comparing TH-TH vs. TH-TL and TL-TL vs. TL-TH, BSM, both p-vales <0.001). In addition, consistent with the previous results, joint hate for Trump evokes a stronger bond in the form of appropriateness for giving than joint love (comparing TH-TH vs. TL-TL, BSM, $p=0.0199$).

Taken together, one can conclude that the perceived social norms map convincingly onto the observed T-o-G dictator game behavior and can explain the observed behavioral differences between Trump haters and Trump lovers as well as their perceptions and attitudes towards people with aligned and misaligned opinions about Donald J. Trump.

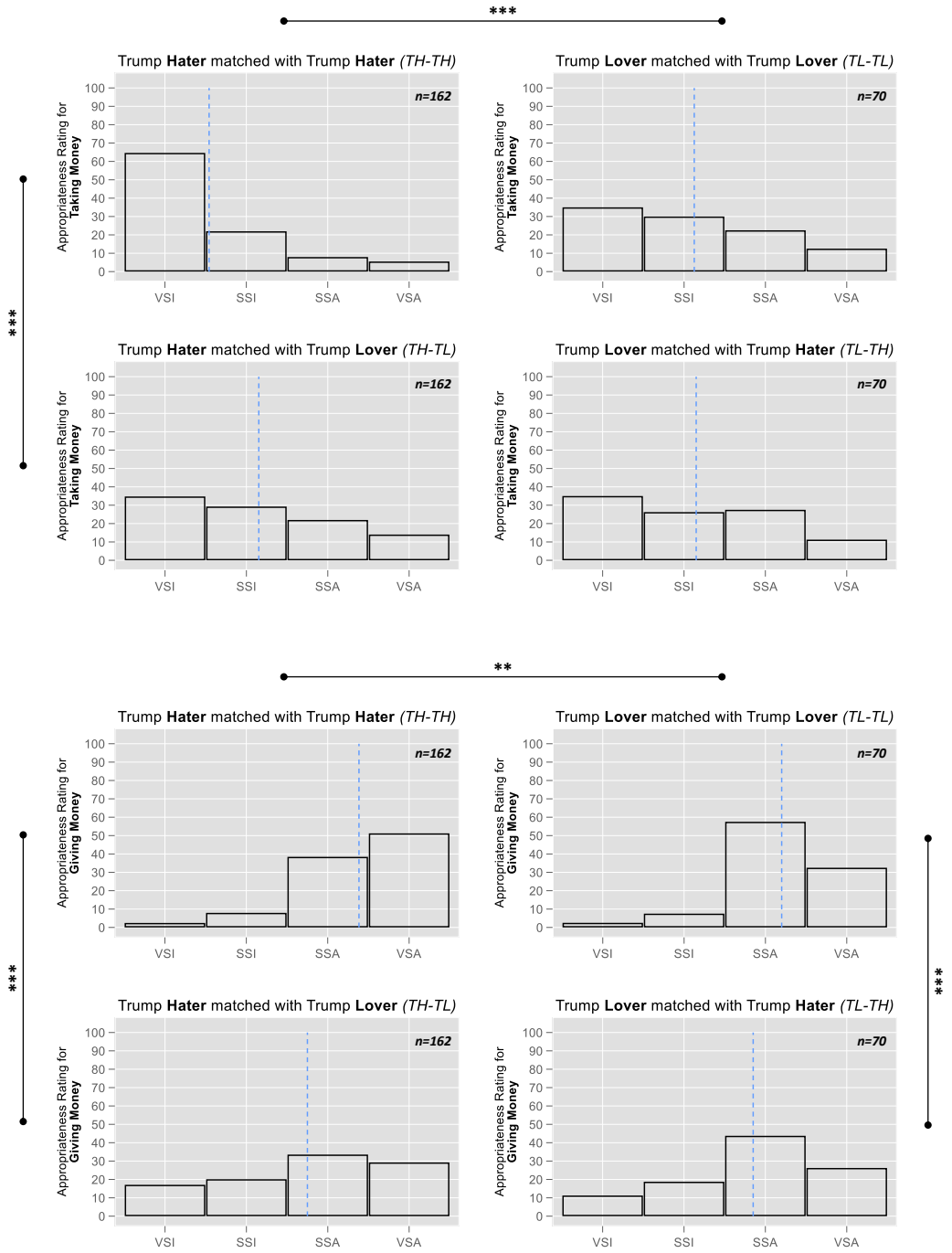

Figure A.5: Norm perceptions for taking and giving money conditional on own and matched partner's Trump opinion. All adjacent quadrants are tested and statistical significance (if either ***p<0.01 or **p<0.05) is indicated where applicable. *Very Socially Inappropriate* (VSI), *Somewhat Socially Inappropriate* (SSI), *Somewhat Socially Appropriate* (SSA), and *Very Socially Appropriate* (VSA).

Appendix C.2. Results for the Public Goods Game

Following the previous analyses, this section reports the norm perceptions across various possible behaviors in the PGG (free-riding and full cooperation) for the different treatments and reported Trump preferences.[34] These behaviors are defined as followed: **free-riding** refers to the decision to benefit from the public good by contributing nothing, even though one's matched partner contributes a non-zero amount. **Full cooperation** refers to the decision to contribute the full amount regardless of the partner's behavior. The results presented in Figure A.6 paint a clear picture: participants perceive it as *more* socially appropriate to free-ride on a partner who has a contrary opinion about Trump, but less socially appropriate to fully cooperate with the same partner (both p<0.01).

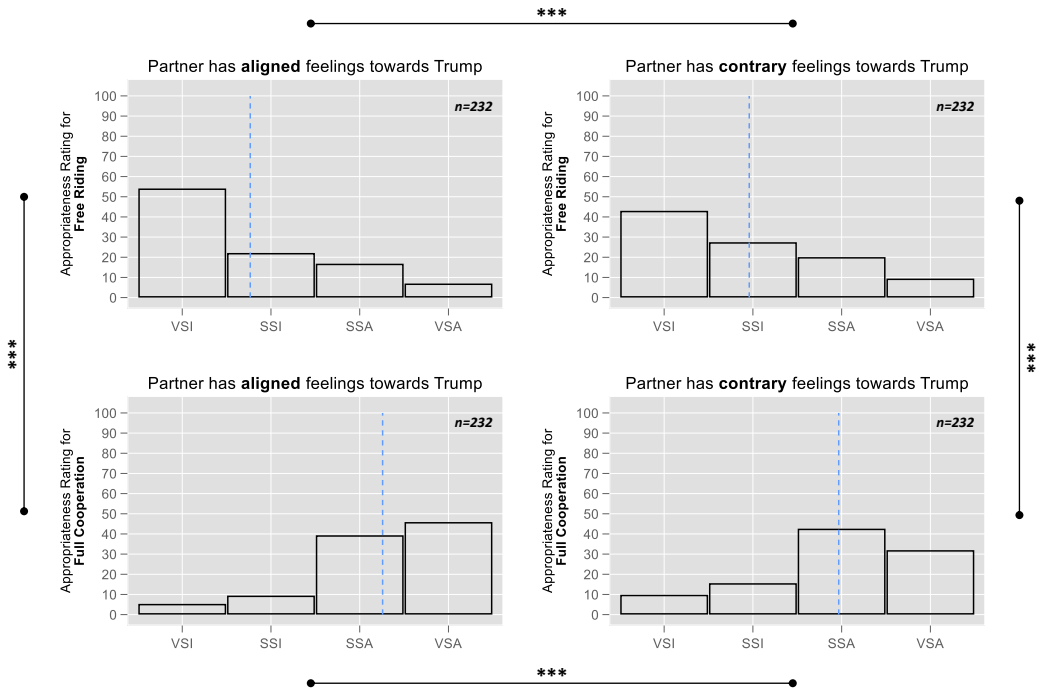

Figure A.6: Norm perceptions for free riding and full cooperation with partners who have aligned or contrary feelings towards Trump. All adjacent quadrants are tested and statistical significance (if either ***p<0.01 or **p<0.05) is indicated where applicable. *Very Socially Inappropriate* (VSI), *Somewhat Socially Inappropriate* (SSI), *Somewhat Socially Appropriate* (SSA), and *Very Socially Appropriate* (VSA).

In addition, one can observe in Figure A.7 that the previous results very much depend on one's stated preference towards Trump: the previously mentioned differential perception of appropriateness for *free-riding* is entirely driven by Trump haters (p<0.001), whereas there is no significant difference for Trump lovers (p=0.47). The same is true for full cooperation

[34] In the Online Appendix, I also present the norm elicitations for two other behaviors: *Contribute Nothing* and *Conditional Cooperator*.

(bottom of Figure A.7): Trump haters perceive it as more socially appropriate to fully cooperate with a partner who has an aligned Trump opinion ($p<0.001$). Again, Trump lovers do not make a distinction irrespective of whom they are matched with ($p=0.72$). This maps well onto the result presented in Figure 8 (top-right panel) in that only Trump haters make an ingroup-outgroup differentiation in their level of contribution.

These findings are noteworthy, as they are consistent with the results discussed in Section 3.2.3 in that Trump haters show a clear ingroup-love/outgroup-hate distinction, whereas Trump lovers do not. It is important to note that although Trump lovers do not discriminate between their matched partners, they perceive it as much more socially appropriate to free-ride on their partner than Trump haters do ($p<0.001$, not illustrated).

From a big picture perspective, the findings are in harmony with the existing social norms research and can be subsumed under the umbrella of *conditional norm followers* (Bicchieri, 2006; Bicchieri et al., 2020b): people display a preference for cooperation that is conditional on *empirical expectations* (beliefs about the matched partner's behavior, as measured in Experiment 3.2) and *normative expectations* (as measured in this experiment using the method by Krupka and Weber, 2013). Combining both types of elicitations provides a comprehensive evaluation of beliefs, preferences, and behaviors.

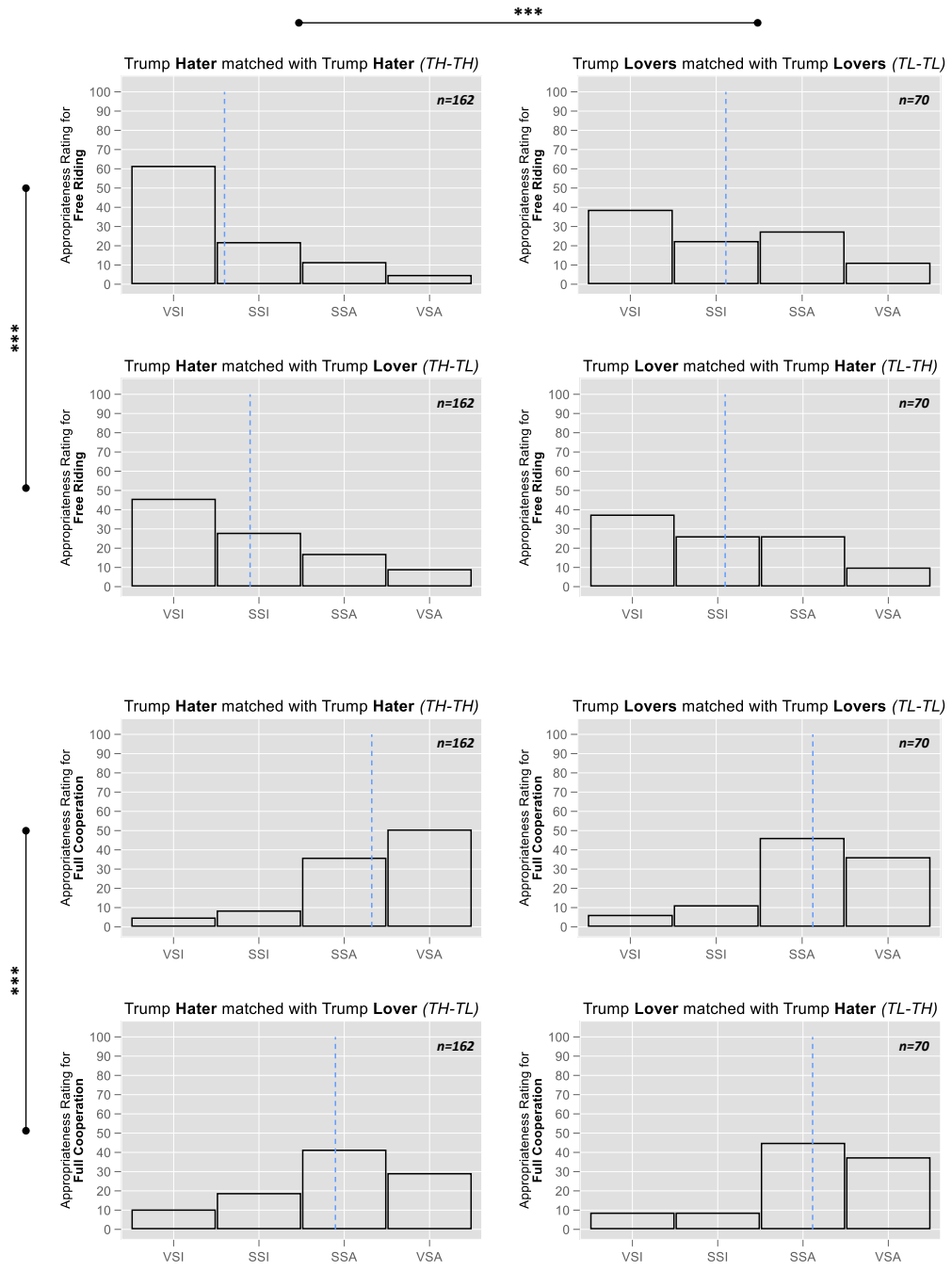

Figure A.7: Norm perceptions for free riding and full cooperation conditional on own and matched partner's Trump opinion. All adjacent quadrants are tested and statistical significance (if either ***p<0.01 or **p<0.05) is indicated where applicable. *Very Socially Inappropriate* (VSI), *Somewhat Socially Inappropriate* (SSI), *Somewhat Socially Appropriate* (SSA), and *Very Socially Appropriate* (VSA).

Appendix D. Sports Prime

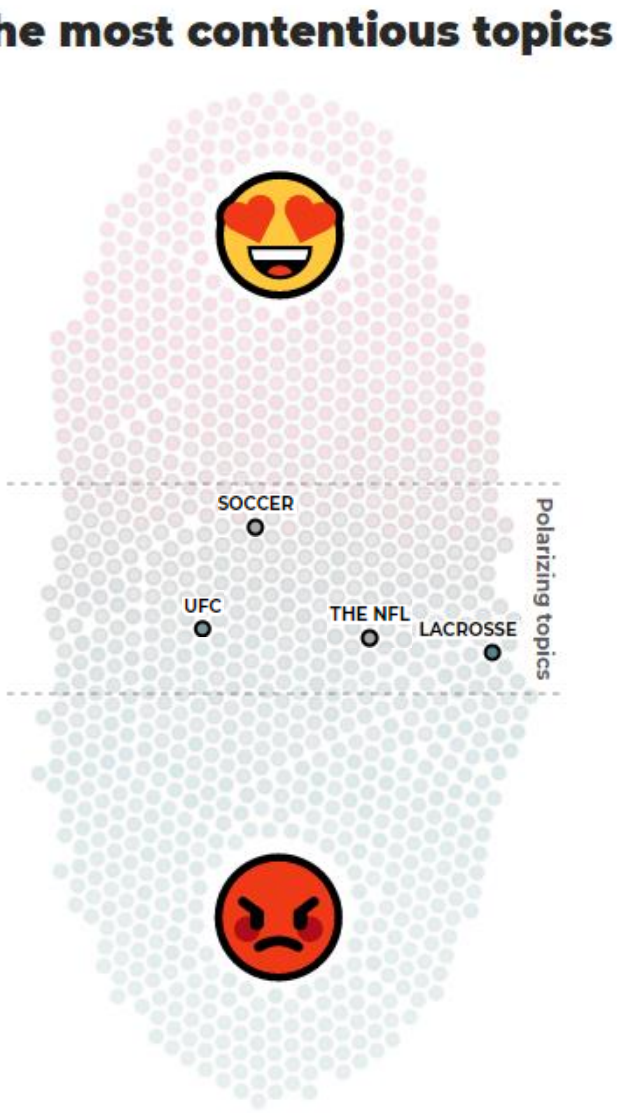

Figure A.8: Most contentious topics based on data from the dating app *Hater*.

Online Appendix to "Hate Trumps Love: The Role of Political Polarization in Social Preferences"

Eugen Dimant

Contents: Additional Results and Robustness

I. Study 1 – Additional Results and Robustness Checks

I.a. Trump Prime Conditions – Dictator Game

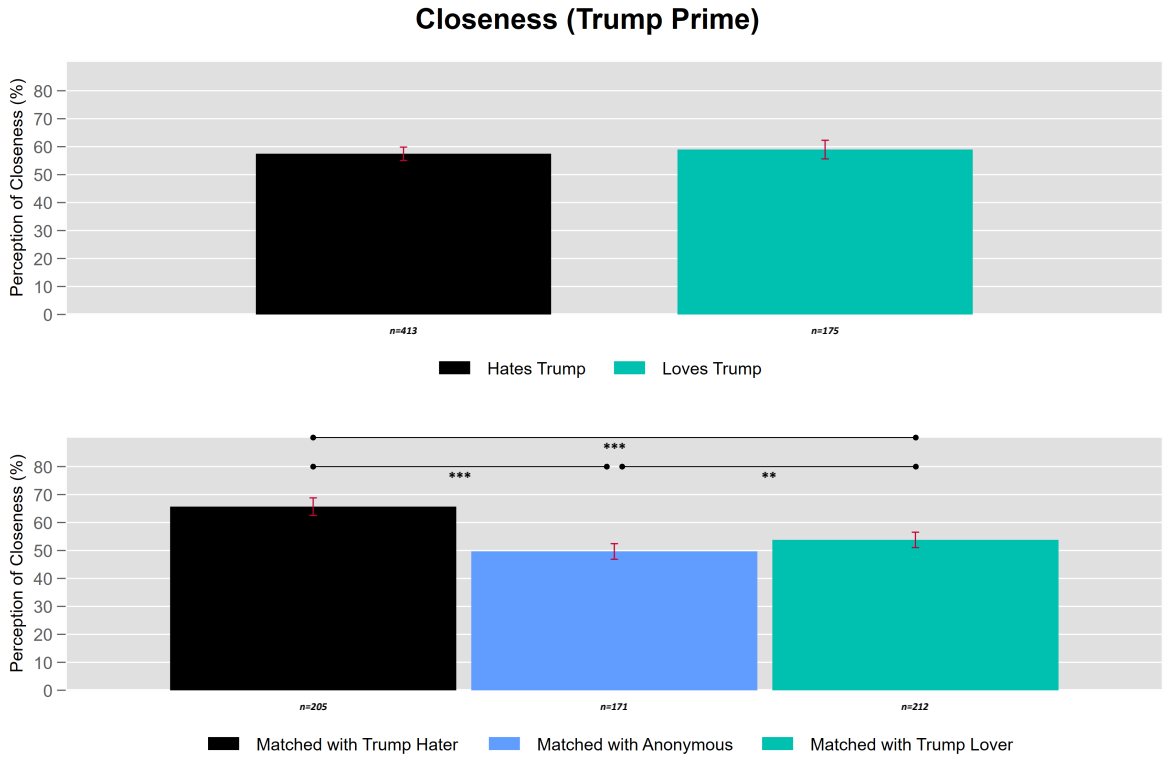

Figure OA.1: Closeness broken down by own opinion about Trump and being matched based on the partner's opinion about Trump. All adjacent bars (within each category) are compared. Absence of significance stars ⇒ p-values > 0.05.

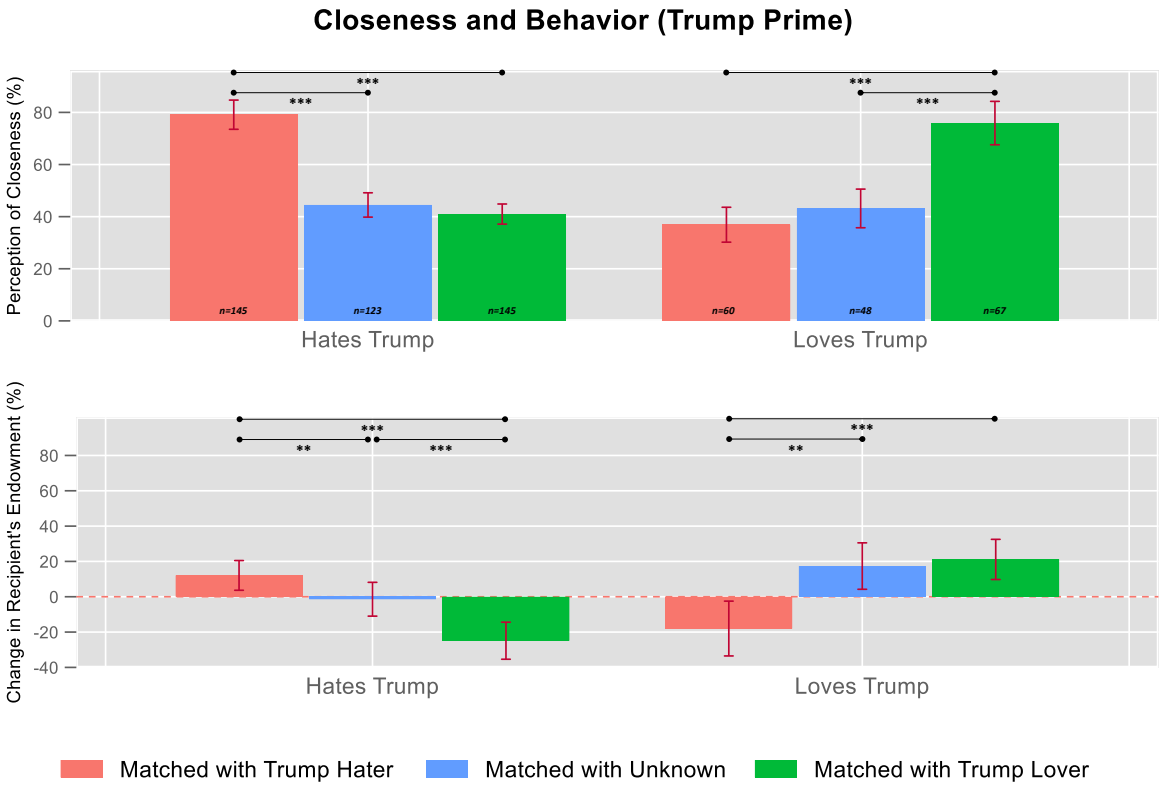

Figure OA.2: Closeness and behavior broken down by own opinion about Trump and being matched based on the partner's opinion about Trump. All adjacent bars (within each category) are compared. Absence of significance stars $\Rightarrow$ p-values > 0.05.

I.b. Minimal Group Paradigm Conditions – Dictator Game

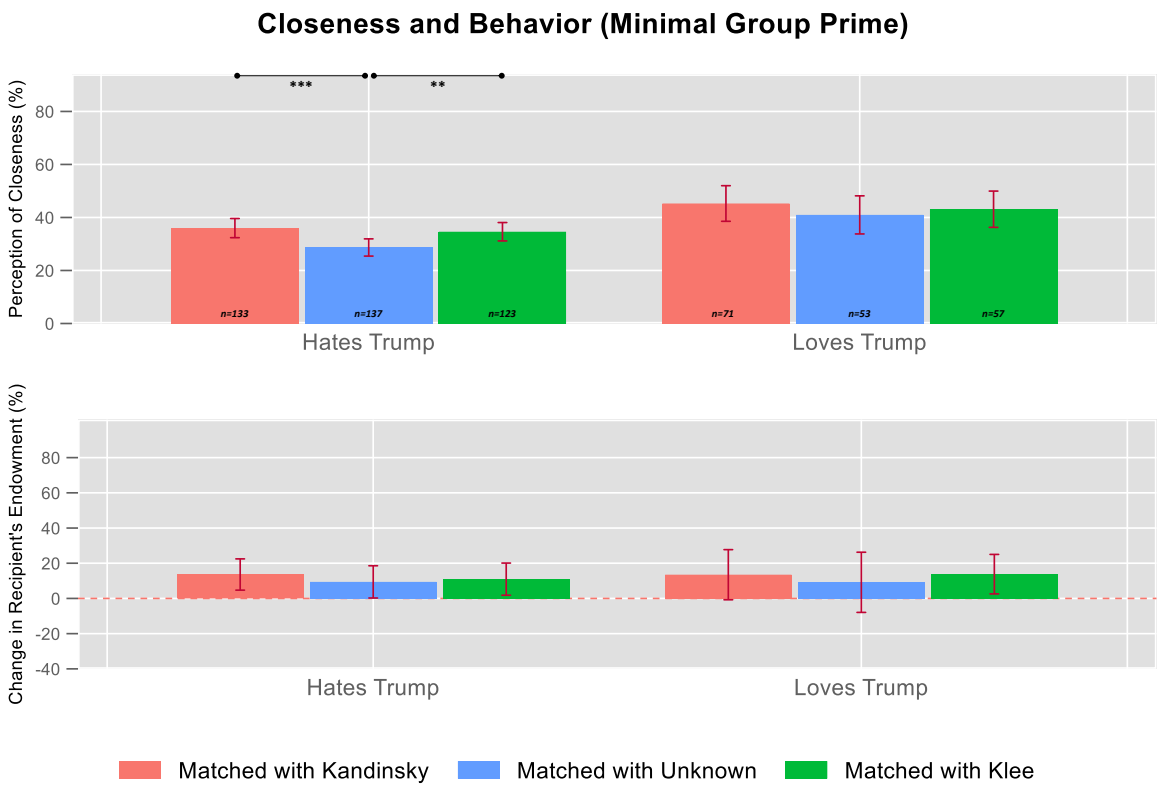

Figure OA.3: Closeness and behavior broken down by own opinion about Trump and being matched based the partner's painting preference. All adjacent bars (within each category) are compared. Absence of significance stars ⇒ p-values > 0.05.

I.c. Default Nudge – Dictator Game

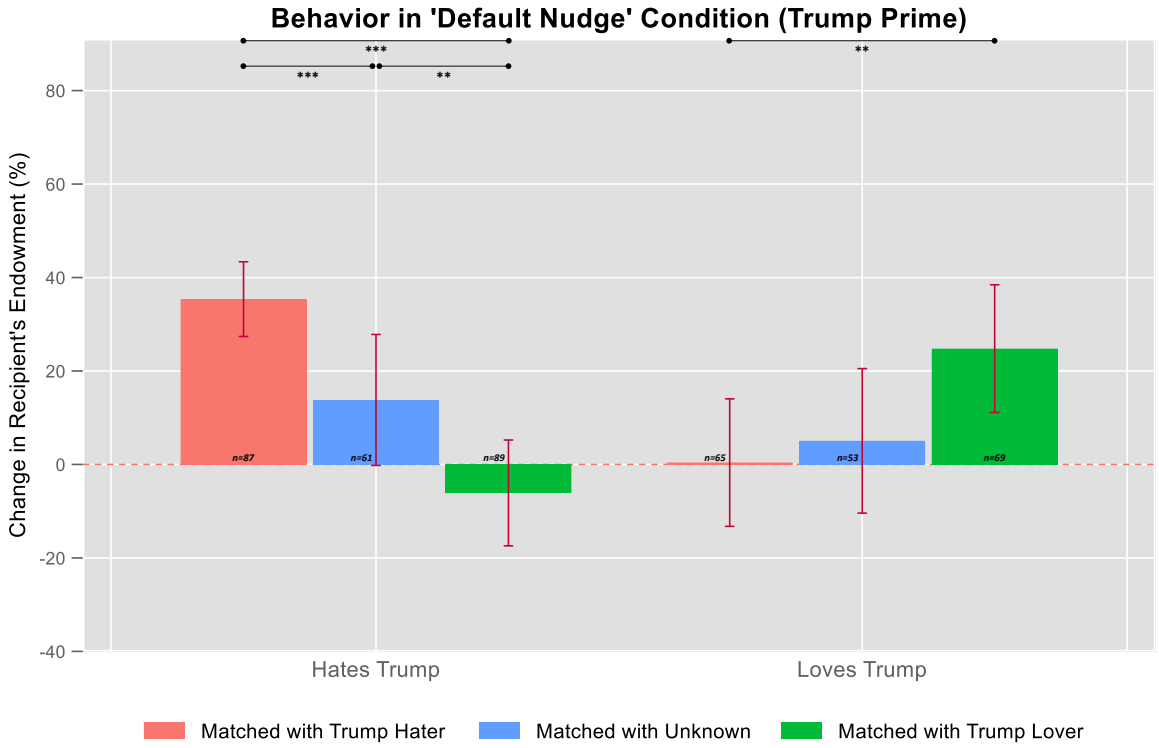

Figure OA.4: Behavior broken down by own opinion about Trump and being matched based on the partner's opinion about Trump. All adjacent bars (within each category) are compared. Absence of significance stars $\Rightarrow$ p-values $>$ 0.05.

I.d. Information Nudge – Dictator Game

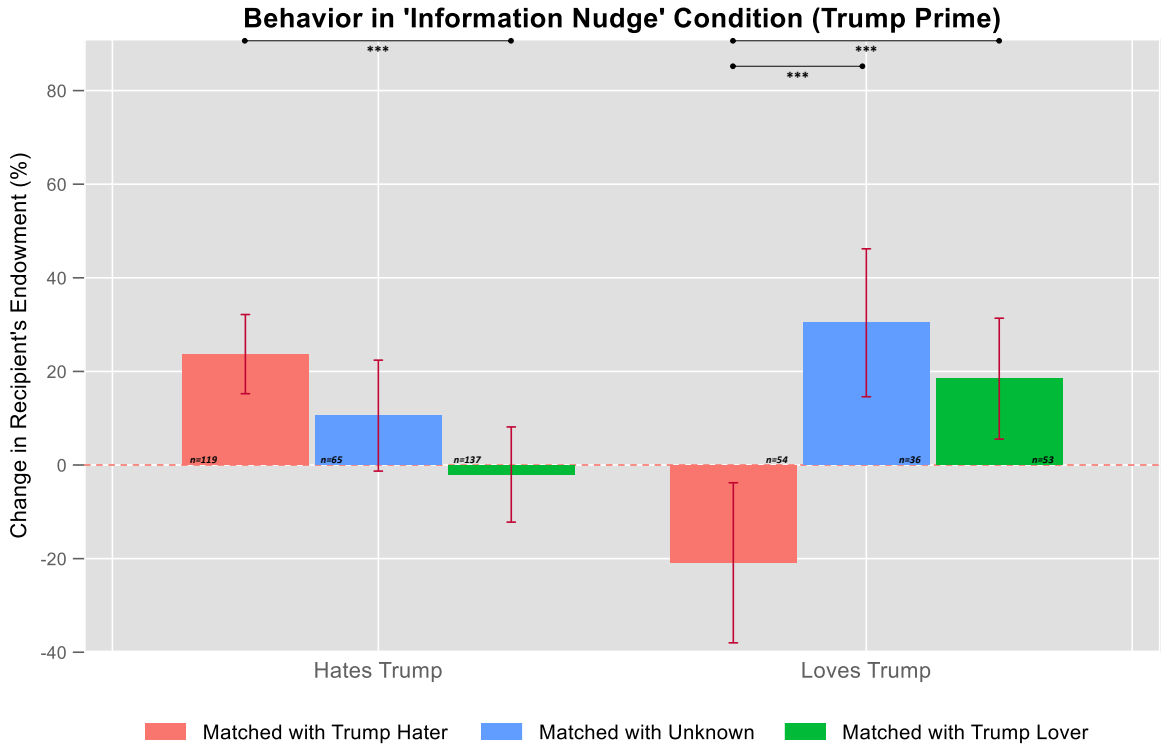

Figure OA.5: Behavior broken down by own opinion about Trump and being matched based on the partner's opinion about Trump. All adjacent bars (within each category) are compared. Absence of significance stars $\Rightarrow$ p-values $>$ 0.05.

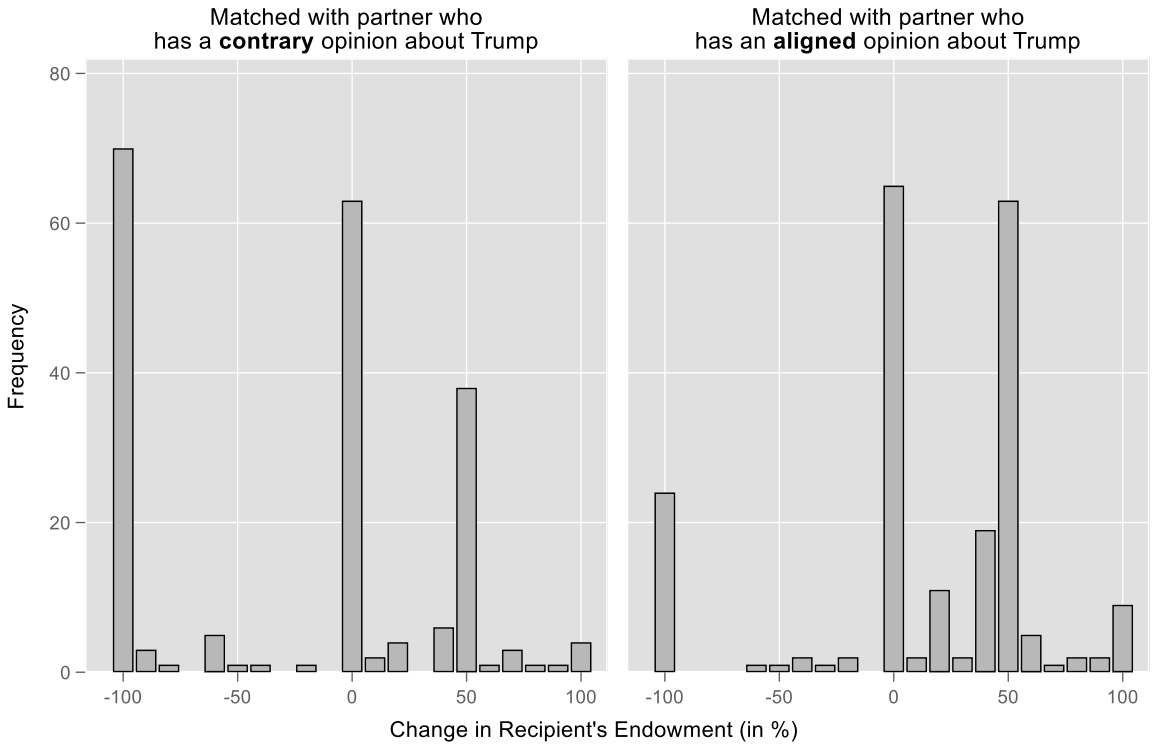

Figure OA.6: Frequency of behavior in original Trump Prime experiment (Section 3.1). Truthful information about this data was used in the Descriptive Norm-Nudge treatment of the DG.

II. Study 2 - Additional Results and Robustness Checks

II.a. Trump Prime Conditions - Public Goods Game

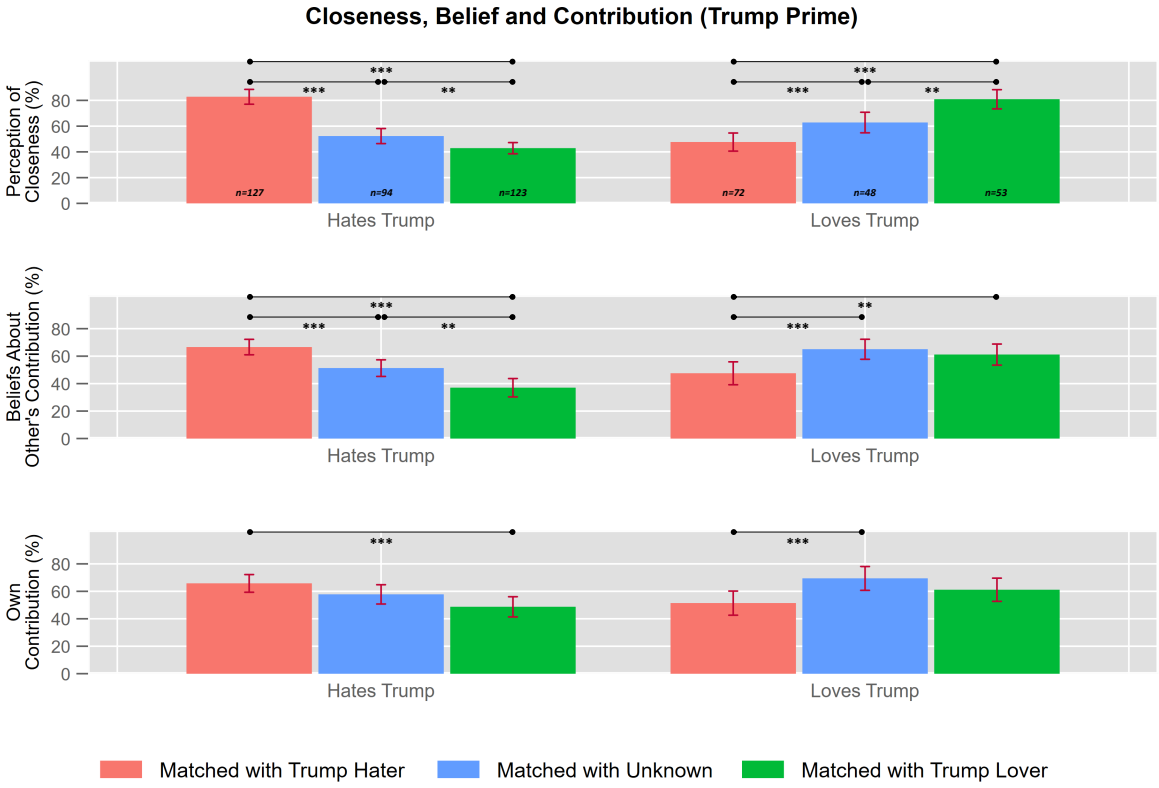

Figure OA.7: Closeness, belief, and behavior broken down by own opinion about and being matched based on the partner's opinion about Trump. All adjacent bars (within each category) are compared. Absence of significance stars $\Rightarrow$ p-values > 0.05.

II.b. Default Nudge – Public Goods Game

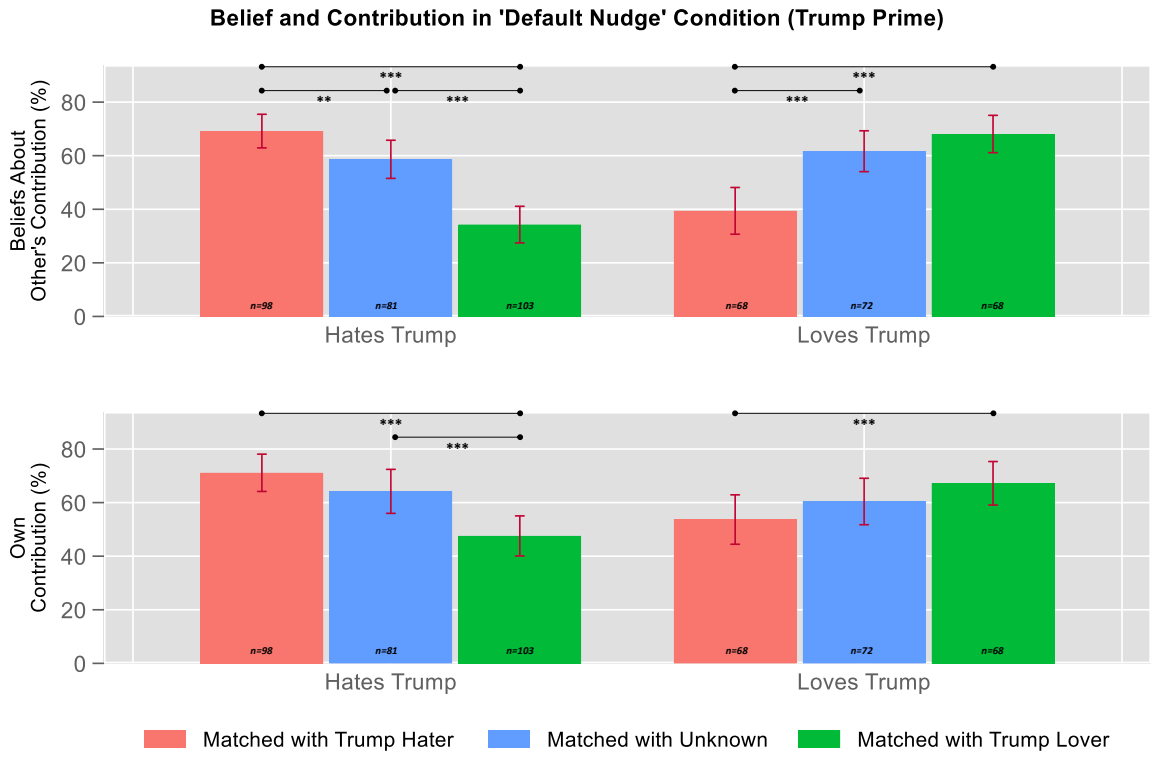

Figure OA.8: Beliefs and behavior broken down by own opinion about Trump and being matched based on the partner's opinion about Trump. All adjacent bars (within each category) are compared. Absence of significance stars $\Rightarrow$ p-values > 0.05.

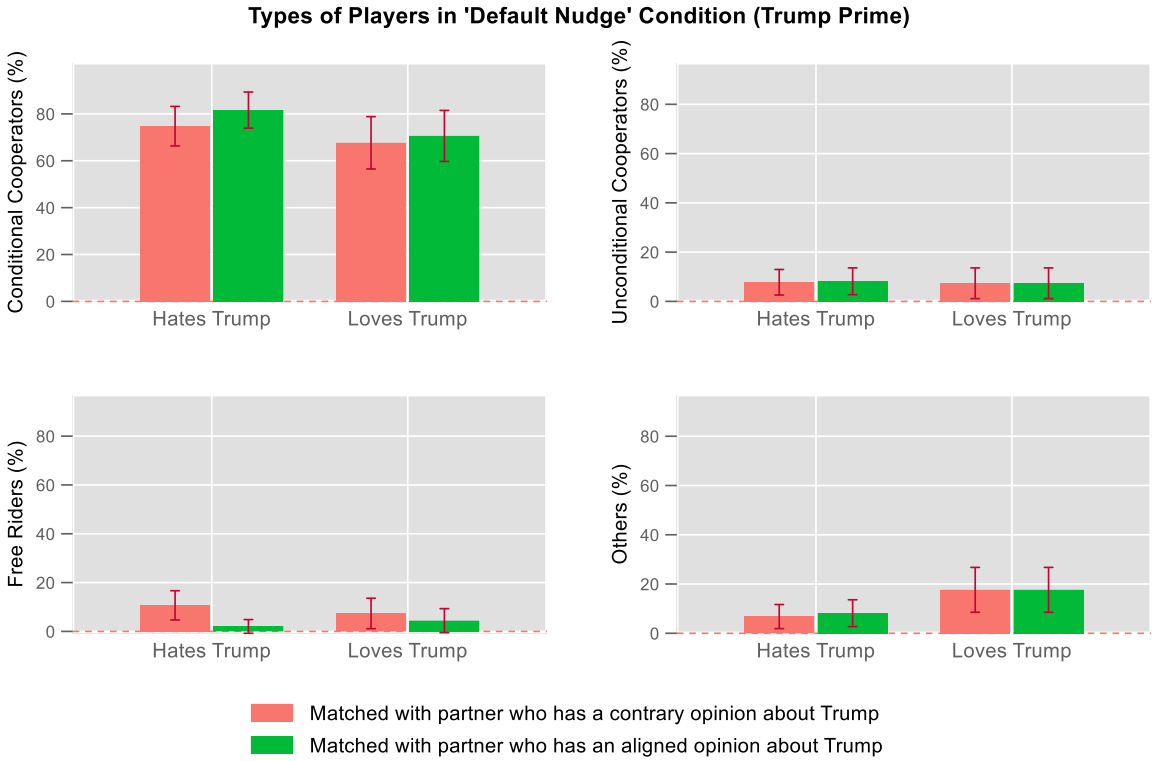

Figure OA.9: Types (conditional cooperators, unconditional cooperators, free riders, others) broken down by one's own opinion and being matched with a partner who either has aligned or contrary opinions for the TP treatment. All adjacent bars (within each category) are compared. Absence of significance stars $\Rightarrow$ p-values > 0.05.

II.c. Information Nudge – Public Goods Game

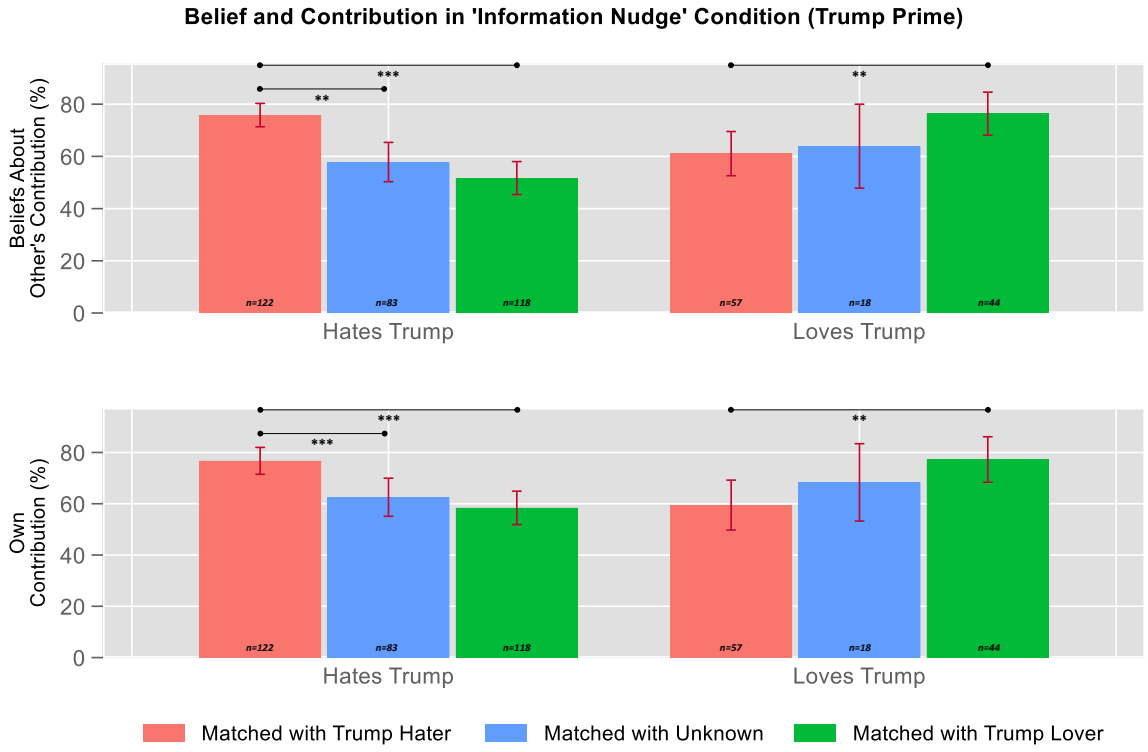

Figure OA.10: Beliefs and behavior broken down by own opinion about Trump and being matched based on the partner's opinion about Trump. All adjacent bars (within each category) are compared. Absence of significance stars $\Rightarrow$ p-values > 0.05.

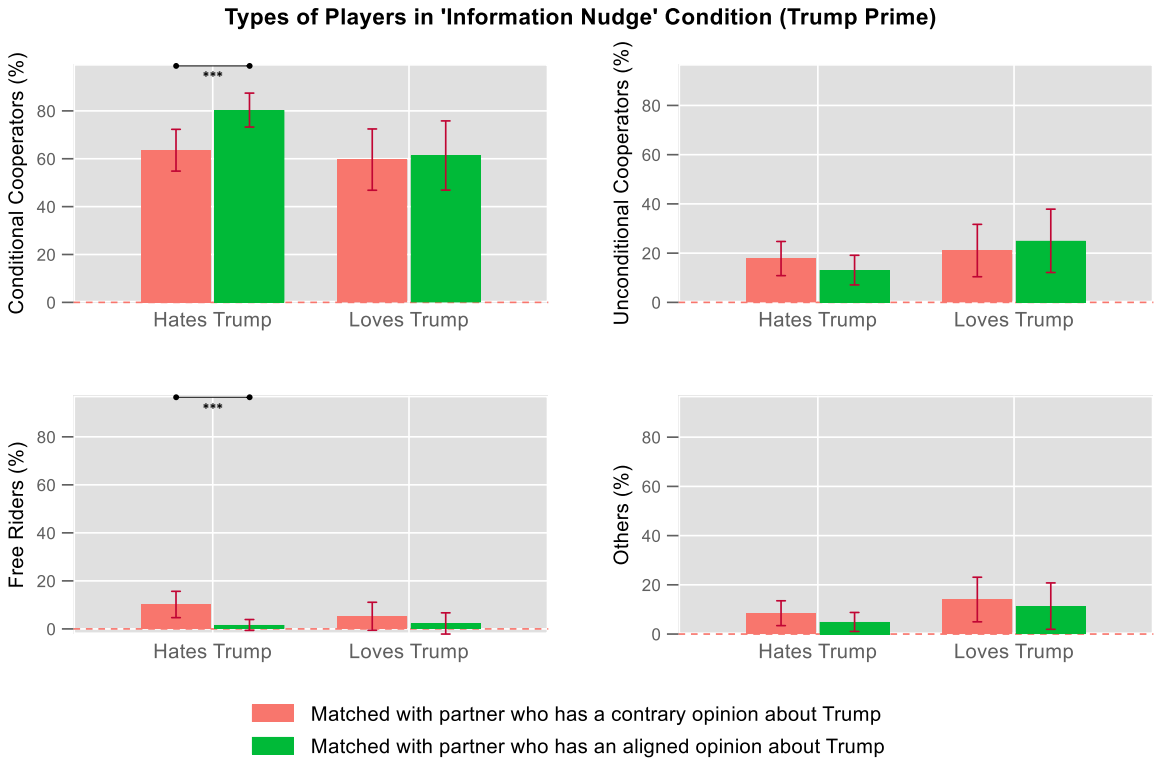

Figure OA.11: Types (conditional cooperators, unconditional cooperators, free riders, others) broken down by one's own opinion and being matched with a partner who either has aligned or contrary opinions for the TP treatment. All adjacent bars (within each category) are compared. Absence of significance stars ⇒ p-values > 0.05.

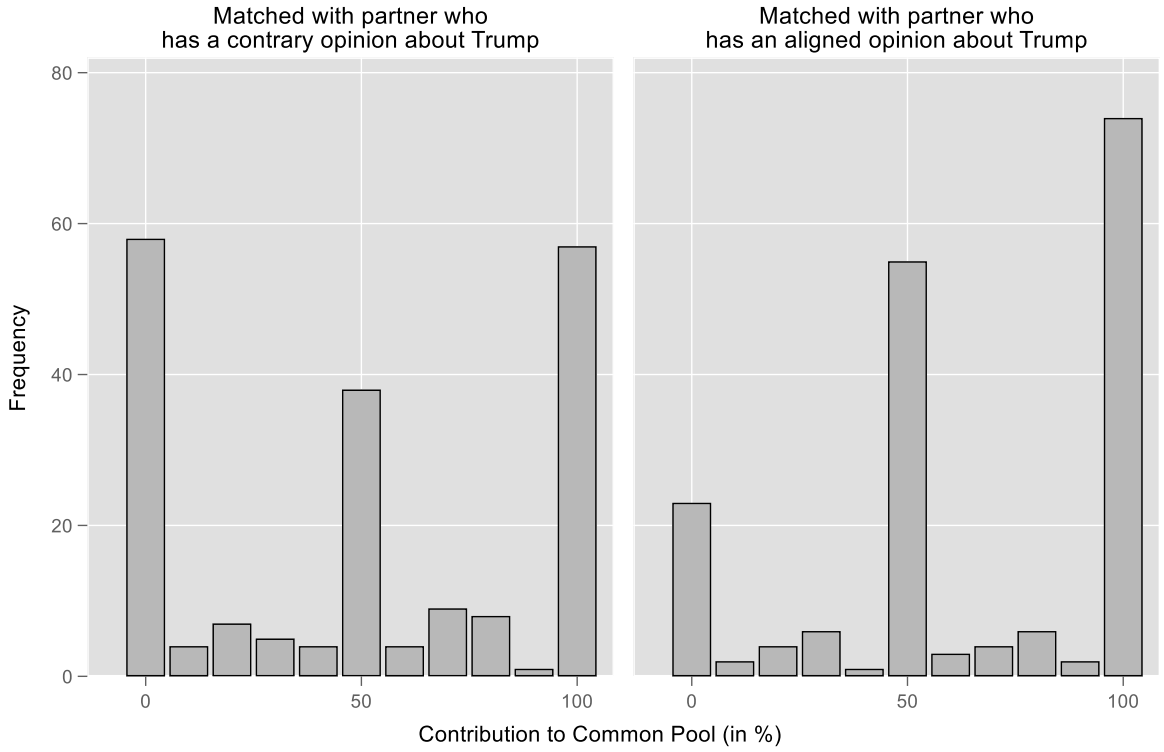

Figure OA.12: Frequency of behavior in original Trump Prime experiment (Section 3.2). Truthful information about this data was used in the descriptive norm-nudge treatment of the PGG.

III. Minimal Group Prime: Additional Results and Robustness Checks

III.a. Variant 1 (*MGP First - TP End*) – Dictator Game

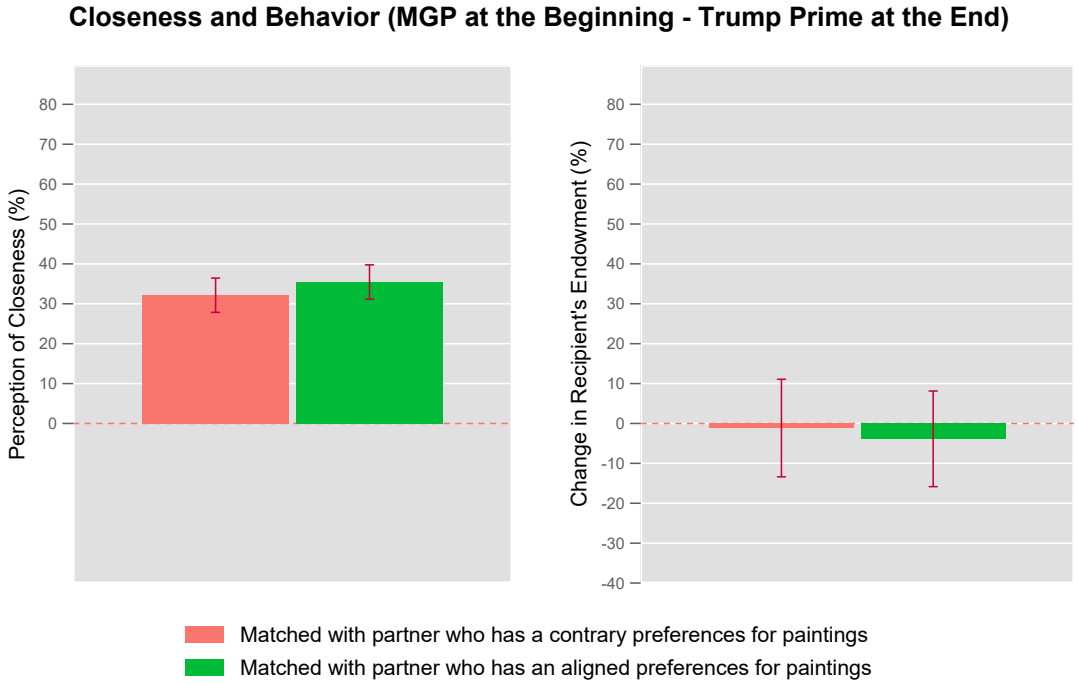

Figure OA.13: Closeness & behavior broken down by being matched with a partner who has a (mis)aligned opinion about paintings. Adjacent bars (within each category) are compared. Absence of significance stars ⇒ p-values > 0.05.

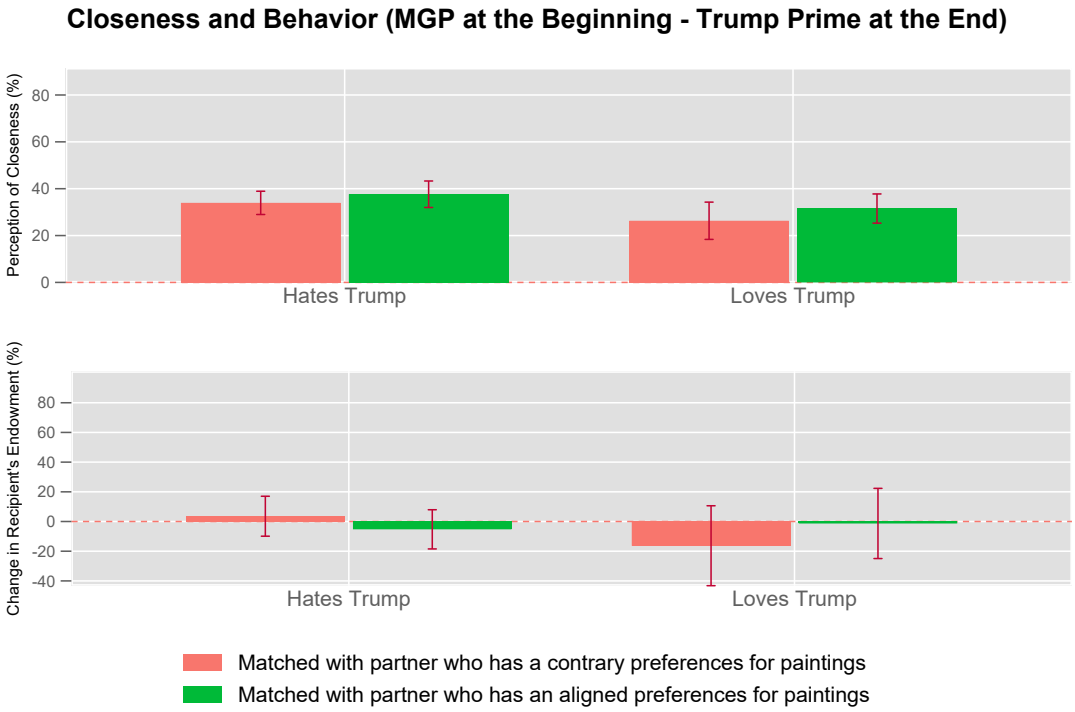

Figure OA.14: Closeness & behavior broken down by own Trump preference and being matched with a partner who has a (mis)aligned opinion about paintings. Adjacent bars (within each category) are compared. Absence of significance stars $\Rightarrow$ p-values > 0.05.

III.b. Variant 1 (*MGP First - TP End*) – Public Goods Game

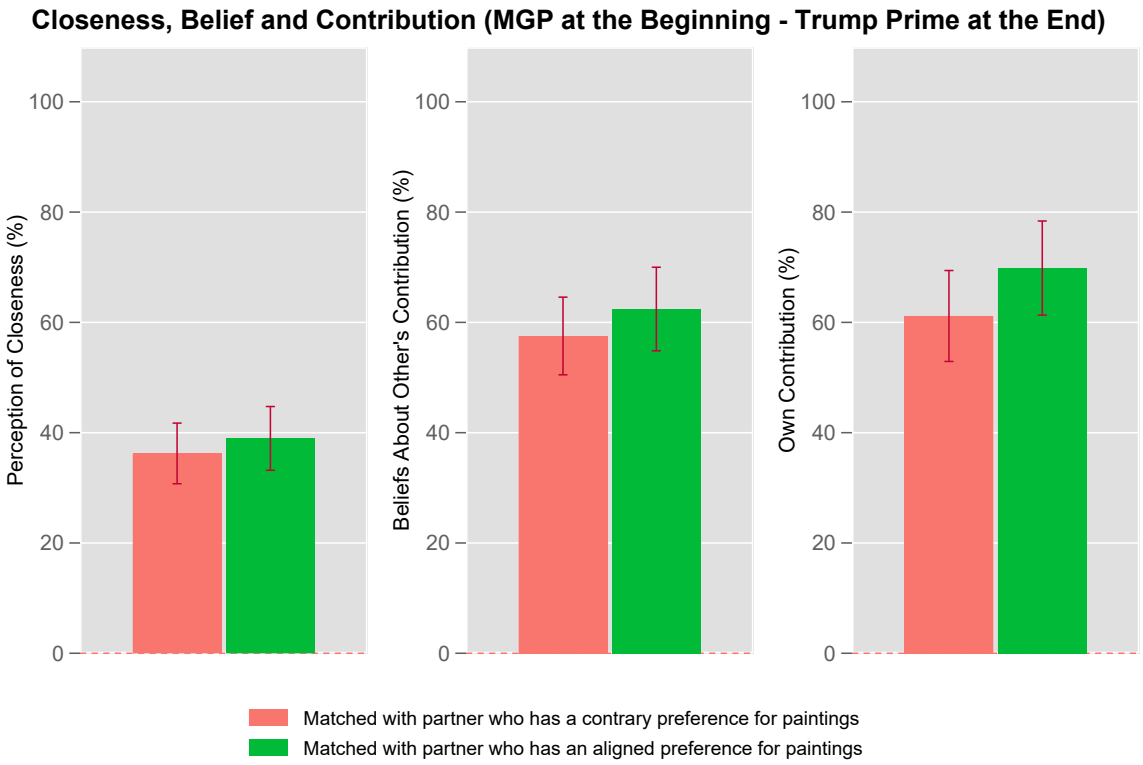

Figure OA.15: Closeness, beliefs & behavior broken down by being matched based on the partner's opinion about paintings. All adjacent bars (within each category) are compared. Absence of significance stars ⇒ p-values > 0.05.

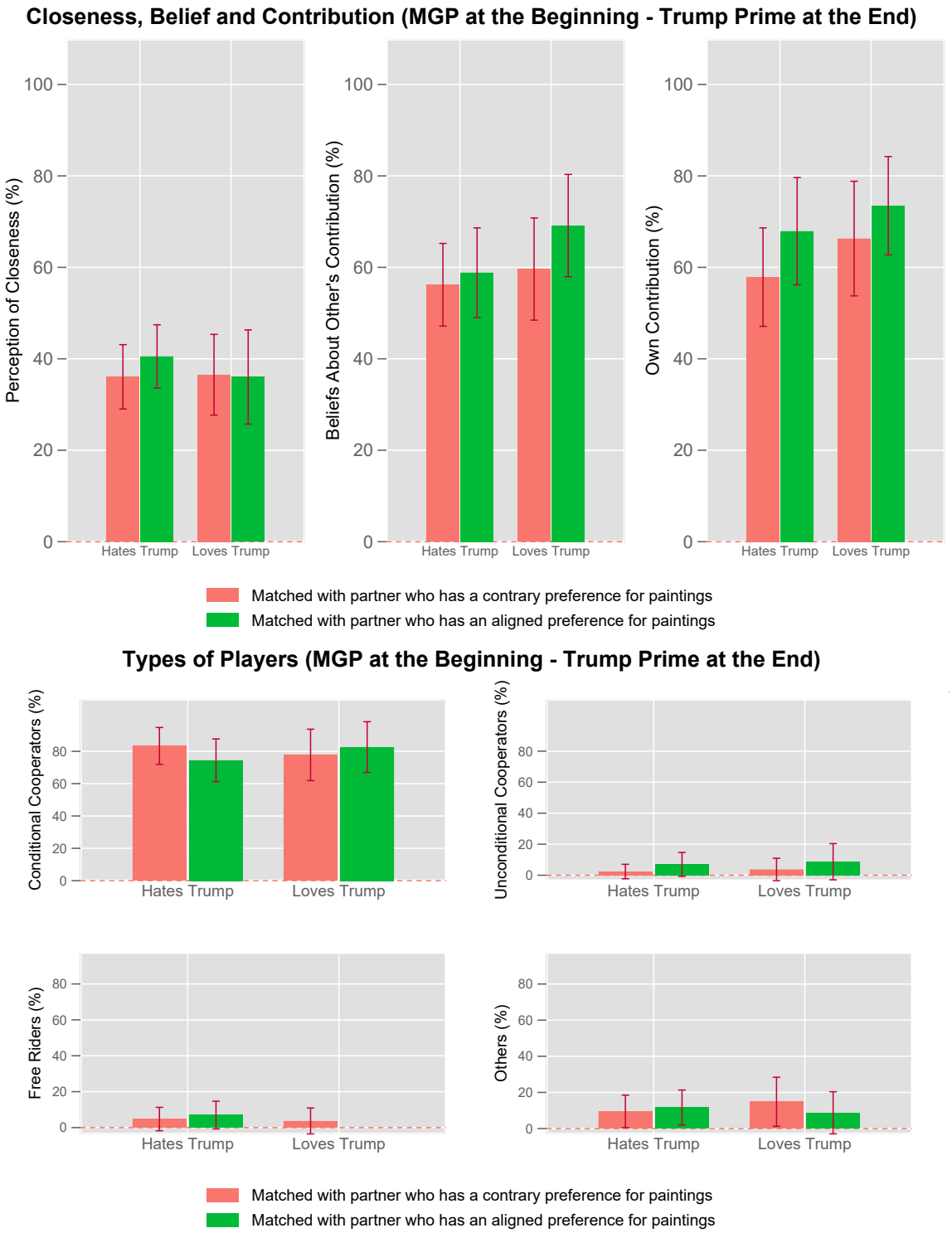

Figure OA.17: Types (conditional cooperators, unconditional cooperators, free riders, others) broken down by one's own opinion about Trump and being matched with a partner who either has aligned or contrary opinions about paintings. All adjacent bars (within each category) are compared. Absence of significance stars $\Rightarrow$ p-values > 0.05.

III.c. Variant 2 (*MGP First - TP Second*) – Dictator Game

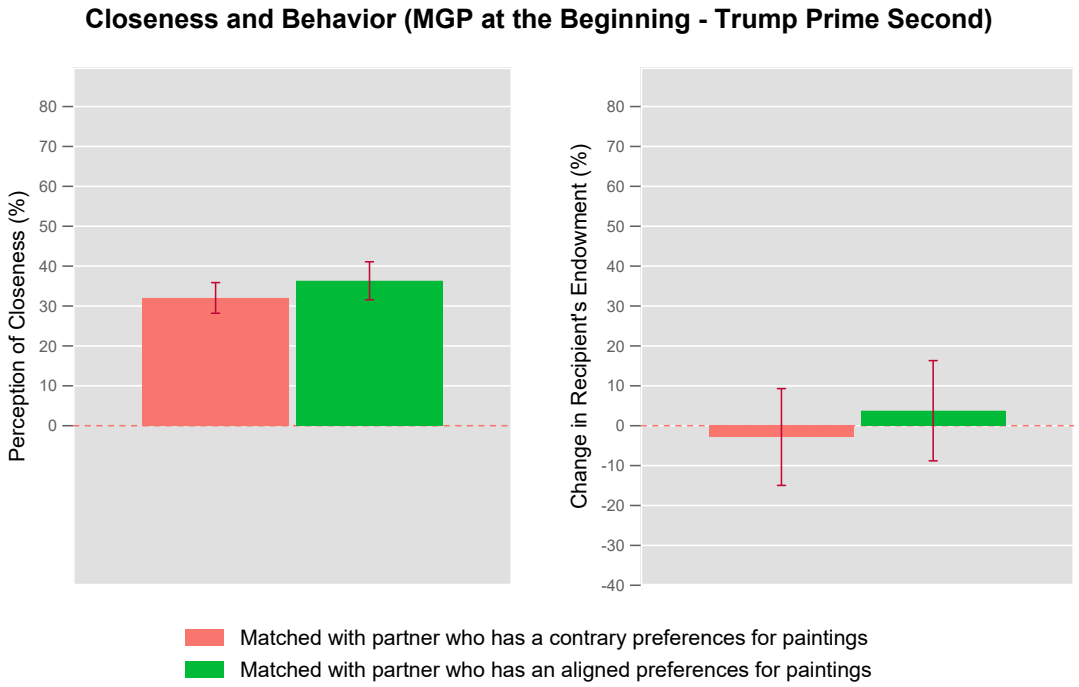

Figure OA.18: Closeness & behavior broken down by being matched with a partner who has a (mis)aligned opinion about paintings. Adjacent bars (within each category) are compared. Absence of significance stars ⇒ p-values > 0.05.

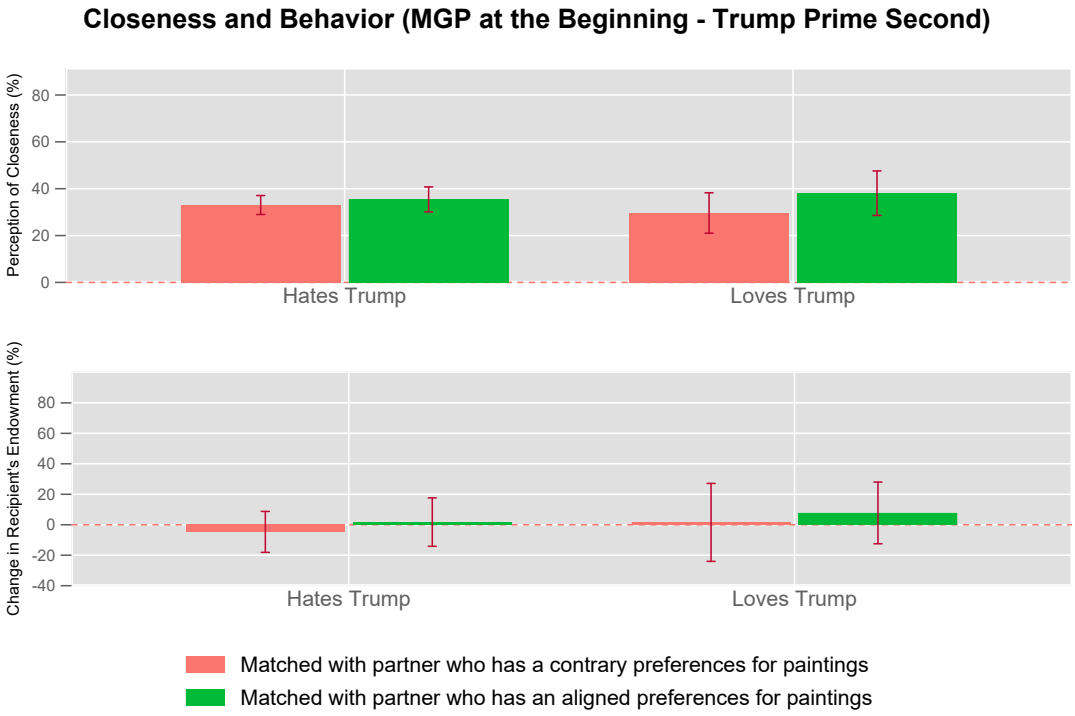

Figure OA.19: Closeness & behavior broken down by own Trump preference and being matched with a partner who has a (mis)aligned opinion about paintings. Adjacent bars (within each category) are compared. Absence of significance stars $\Rightarrow$ p-values > 0.05.

III.d. Variant 2 (*MGP First - TP Second*) – Public Goods Game

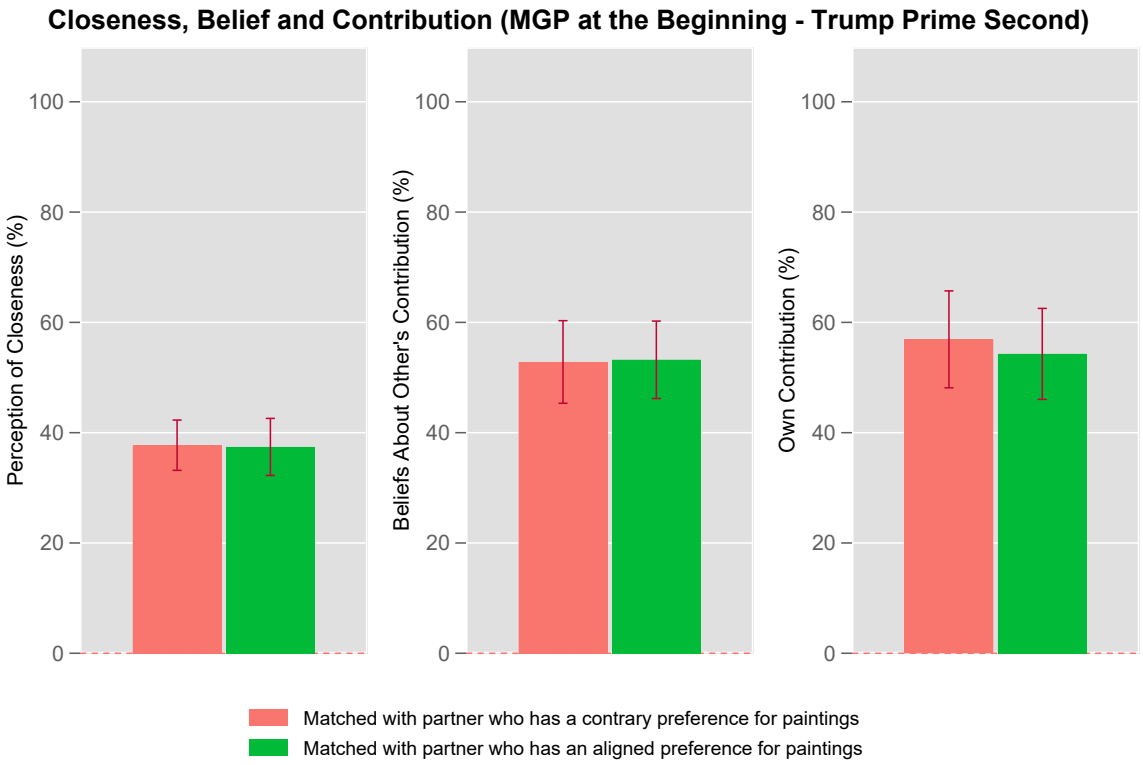

Figure OA.20: Closeness, beliefs & behavior broken down by being matched based on the partner's opinion about paintings. All adjacent bars (within each category) are compared. Absence of significance stars ⇒ p-values > 0.05.

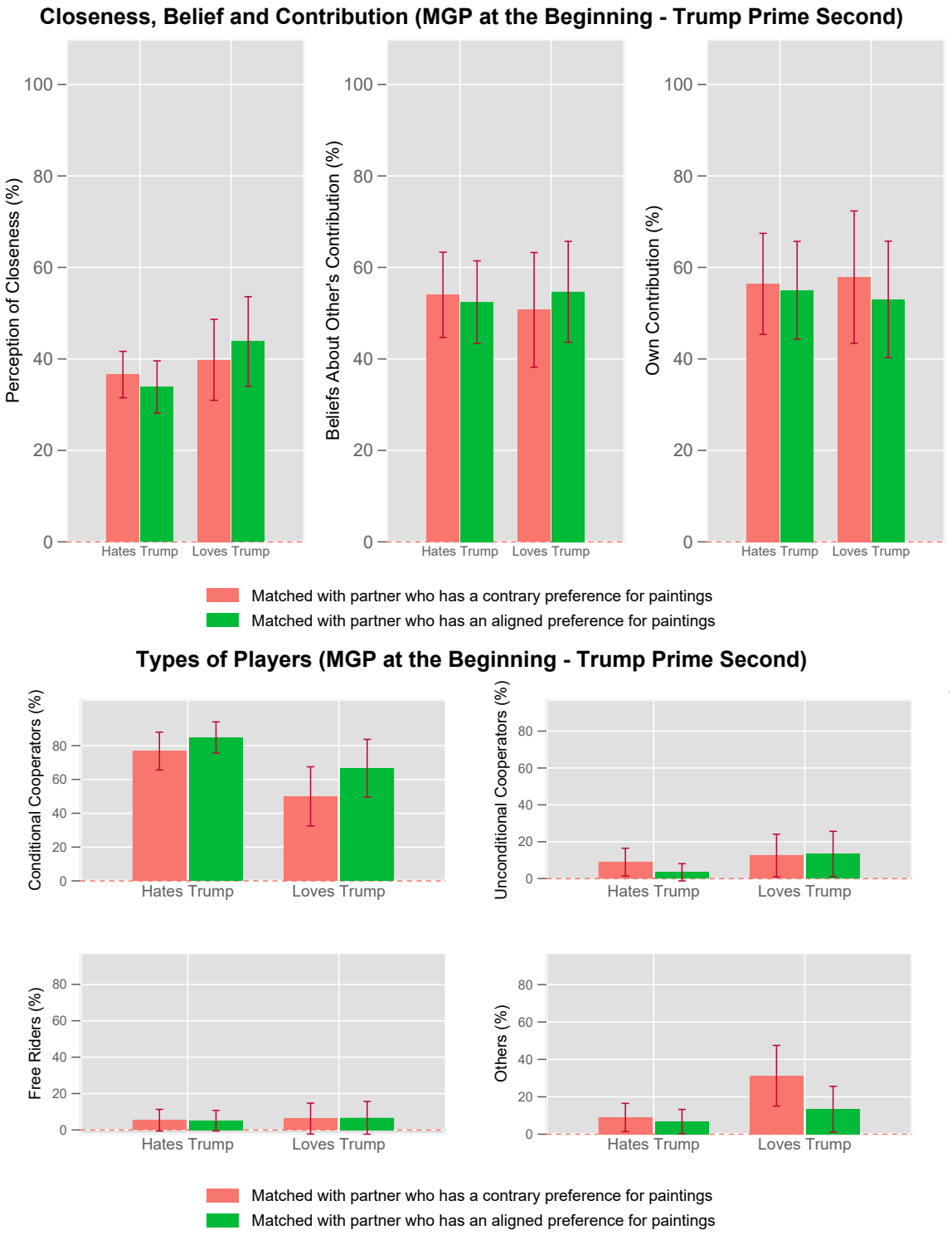

Figure OA.22: Types (conditional cooperators, unconditional cooperators, free riders, others) broken down by one's own opinion about Trump and being matched with a partner who either has aligned or contrary opinions about paintings. All adjacent bars (within each category) are compared. Absence of significance stars ⇒ p-values > 0.05.

IV. Study 4 - Additional Results and Robustness Checks

IV.a. Norm Elicitation - Dictator Game

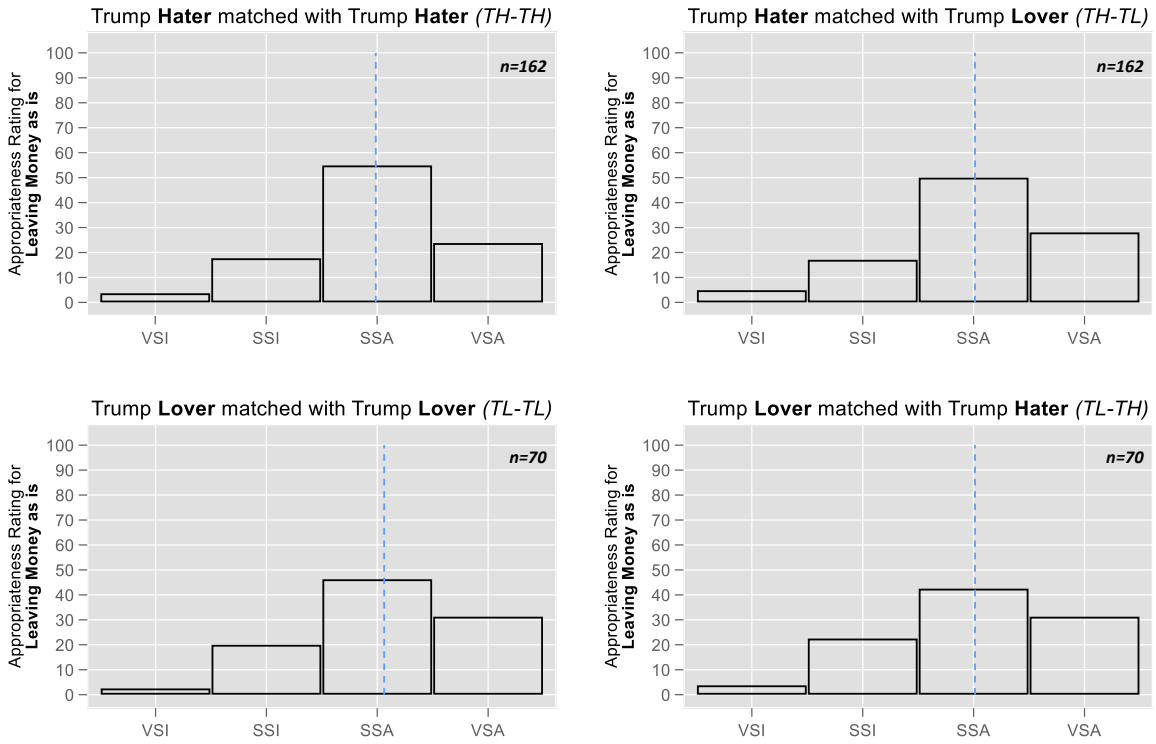

Figure OA.23: Norm perceptions for leaving initial split as is conditional on own and matched partner's Trump opinion. All adjacent quadrants are tested and statistical significance (if either ***p<0.01 or **p<0.05) is indicated where applicable. *Very Socially Inappropriate* (VSI), *Somewhat Socially Inappropriate* (SSI), *Somewhat Socially Appropriate* (SSA), and *Very Socially Appropriate* (VSA).

IV.b. Public Goods Game

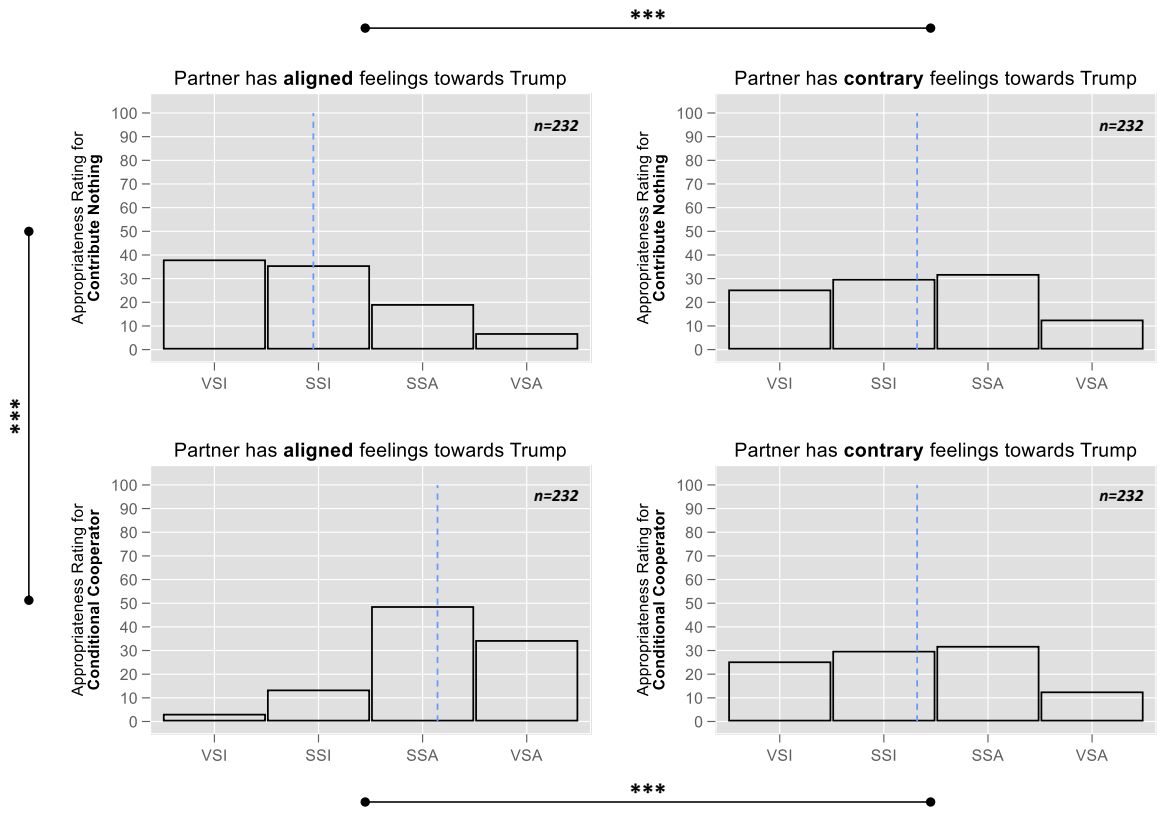

Figure OA.24: Norm perceptions for 'contribute nothing' Conditional Cooperators with partners who have aligned or contrary feelings towards Trump. All adjacent quadrants are tested and statistical significance (if either ***p<0.01 or **p<0.05) is indicated where applicable. *Very Socially Inappropriate* (VSI), *Somewhat Socially Inappropriate* (SSI), *Somewhat Socially Appropriate* (SSA), and *Very Socially Appropriate* (VSA).

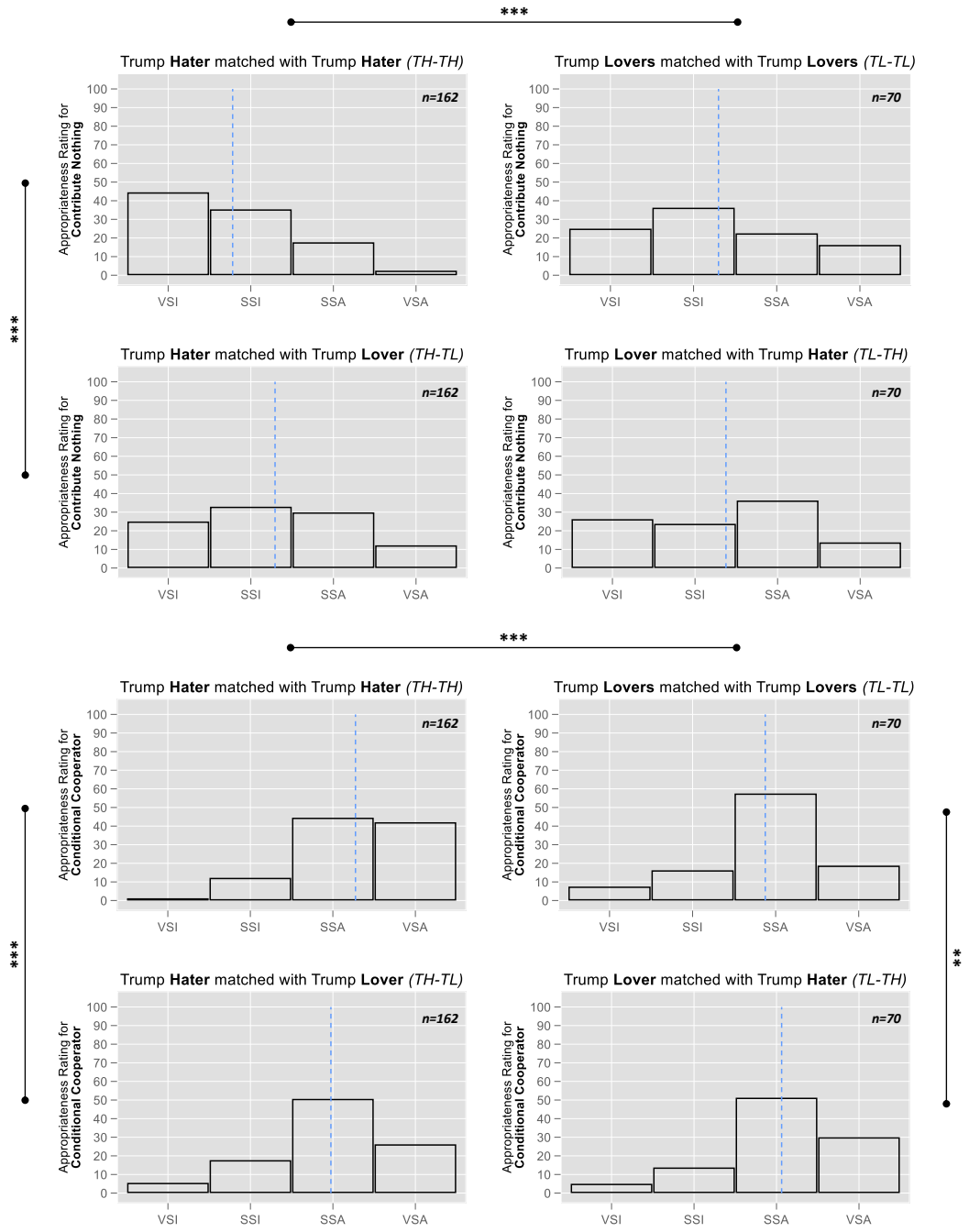

Figure OA.25: Norm perceptions for 'contribute nothing' Conditional Cooperators conditional on own and matched partner's Trump opinion. All adjacent quadrants are tested and statistical significance (if either ***p<0.01 or **p<0.05) is indicated where applicable. *Very Socially Inappropriate* (VSI), *Somewhat Socially Inappropriate* (SSI), *Somewhat Socially Appropriate* (SSA), and *Very Socially Appropriate* (VSA).

V. Study 5 - Additional Results and Robustness Checks

V.a. Biden Prime – Dictator Game

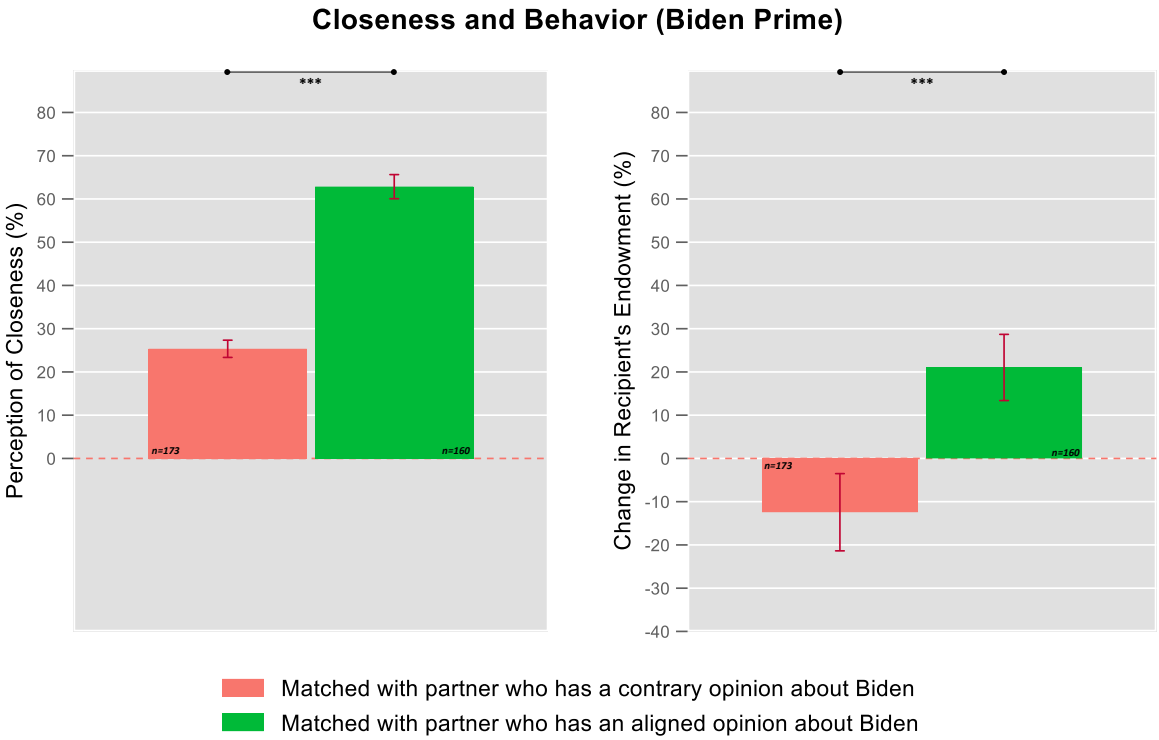

Figure OA.26: Closeness and behavior by being matched with a partner who has a (mis)aligned opinion about Biden. Perception of closeness is converted from a 7-point scale to % for illustrative purposes. All adjacent bars (within each category) are compared. Absence of significance stars ⇒ p-values > 0.05.

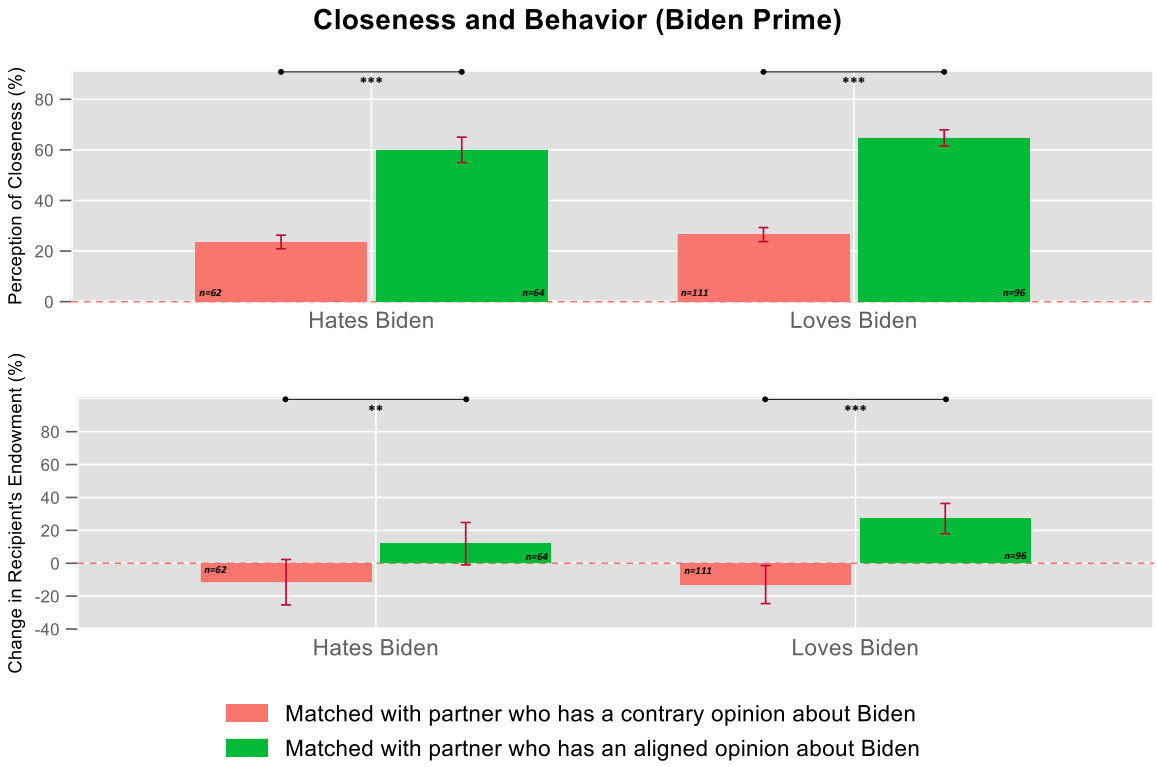

Figure OA.27: Closeness and behavior broken down by being matched with a partner who has a (mis)aligned opinion about Biden. Adjacent bars (within each category) are compared. Absence of significance stars $\Rightarrow$ p-values > 0.05.

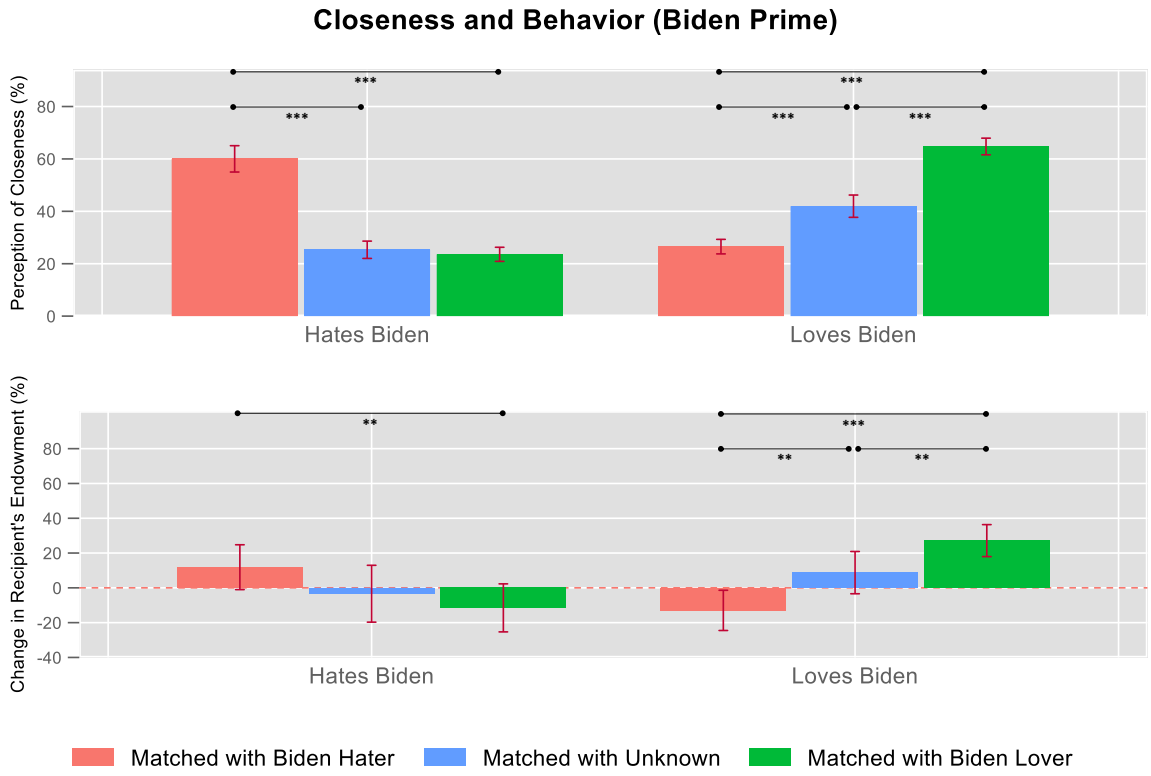

Figure OA.28: Closeness and behavior broken down by own opinion and being matched with a partner who has a (mis)aligned opinion about Biden. Adjacent bars (within each category) are compared. Absence of significance stars $\Rightarrow$ p-values > 0.05.

V.b. Biden Prime – Public Goods Game

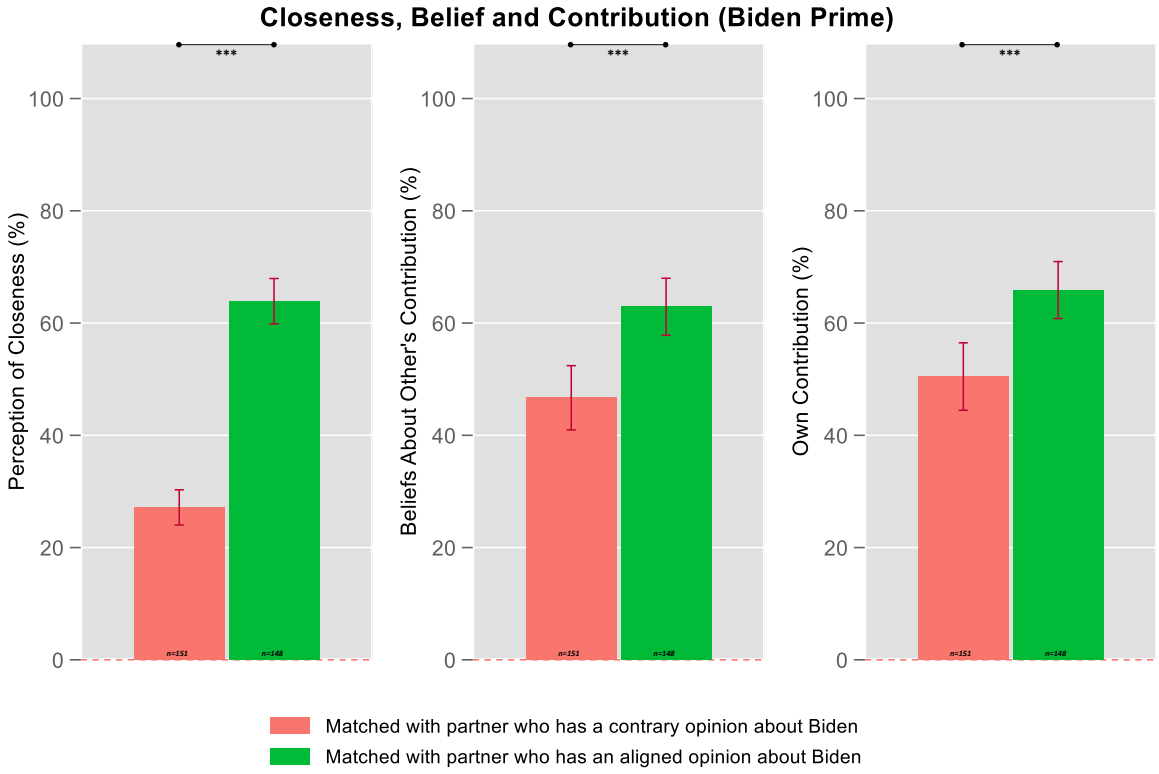

Figure OA.29: Closeness, beliefs & behavior broken down by being matched based on the partner's opinion about Biden. All adjacent bars (within each category) are compared. Absence of significance stars ⇒ p-values > 0.05.

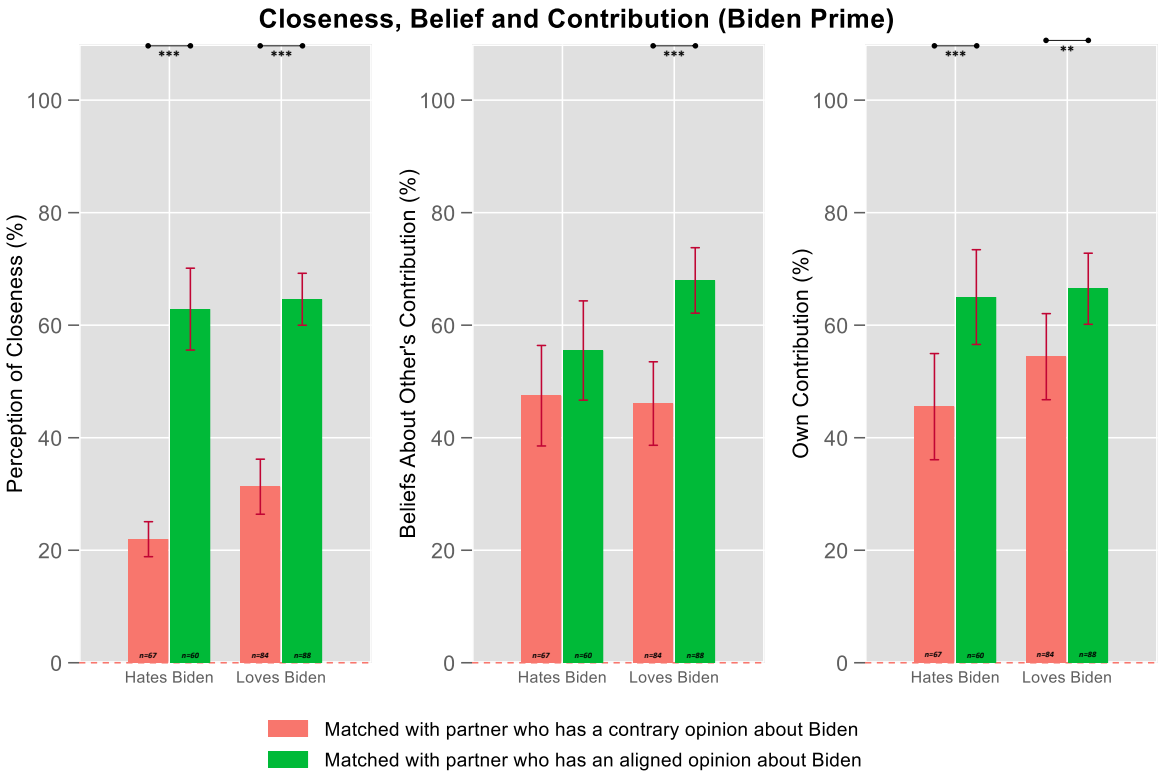

Figure OA.30: Closeness, beliefs, and behavior broken down by own opinion about Biden and being matched based on the partner's opinion about Biden. All adjacent bars (within each category) are compared. Absence of significance stars $\Rightarrow$ p-values > 0.05.

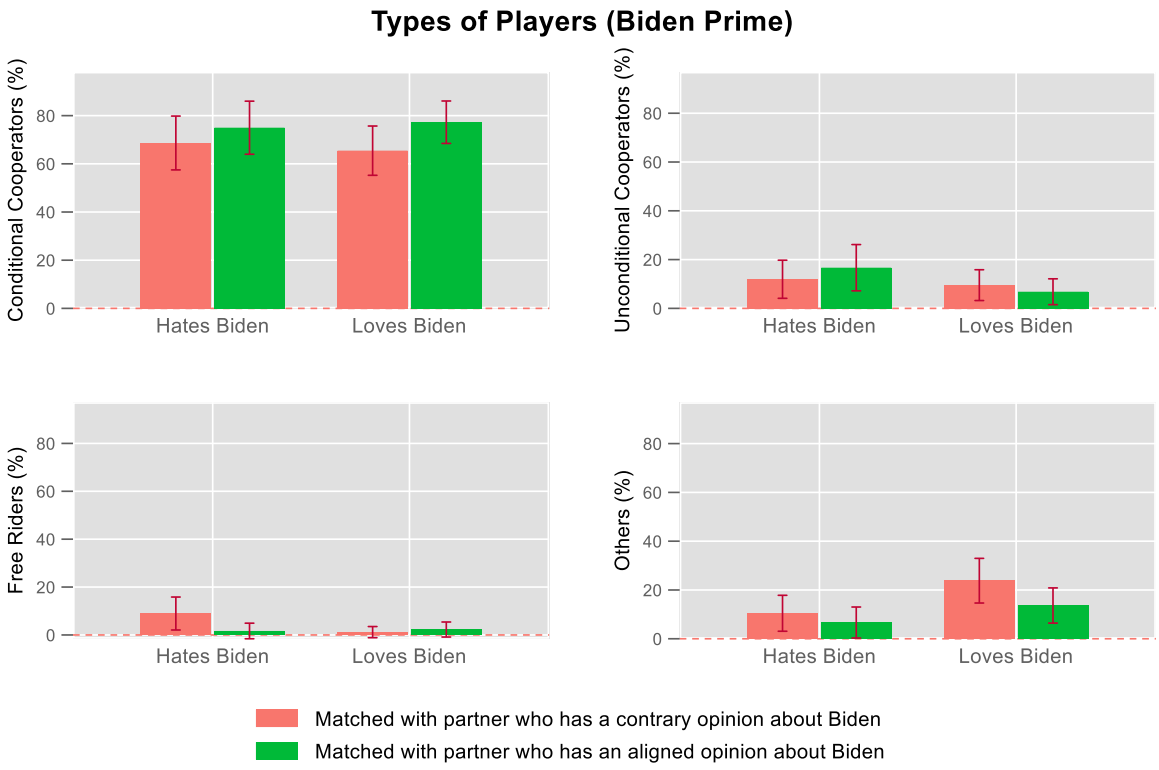

Figure OA.31: Types (conditional cooperators, unconditional cooperators, free riders, others) broken down by one's own opinion and being matched with a partner who either has aligned or contrary opinions for the Biden treatment. All adjacent bars (within each category) are compared. Absence of significance stars $\Rightarrow$ p-values > 0.05.

V.c. Sports Prime – Dictator Game

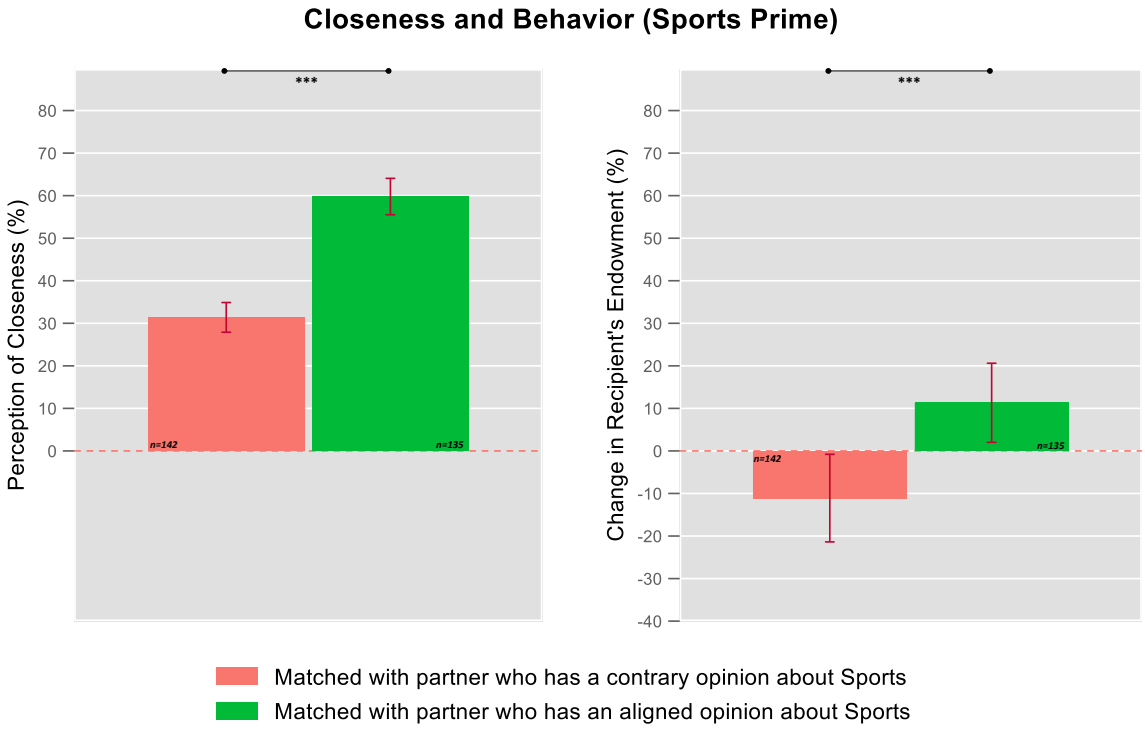

Figure OA.32: Closeness and behavior by being matched with a partner who has a (mis)aligned opinion about sports. Perception of closeness is converted from a 7-point scale to % for illustrative purposes. All adjacent bars (within each category) are compared. Absence of significance stars ⇒ p-values > 0.05.

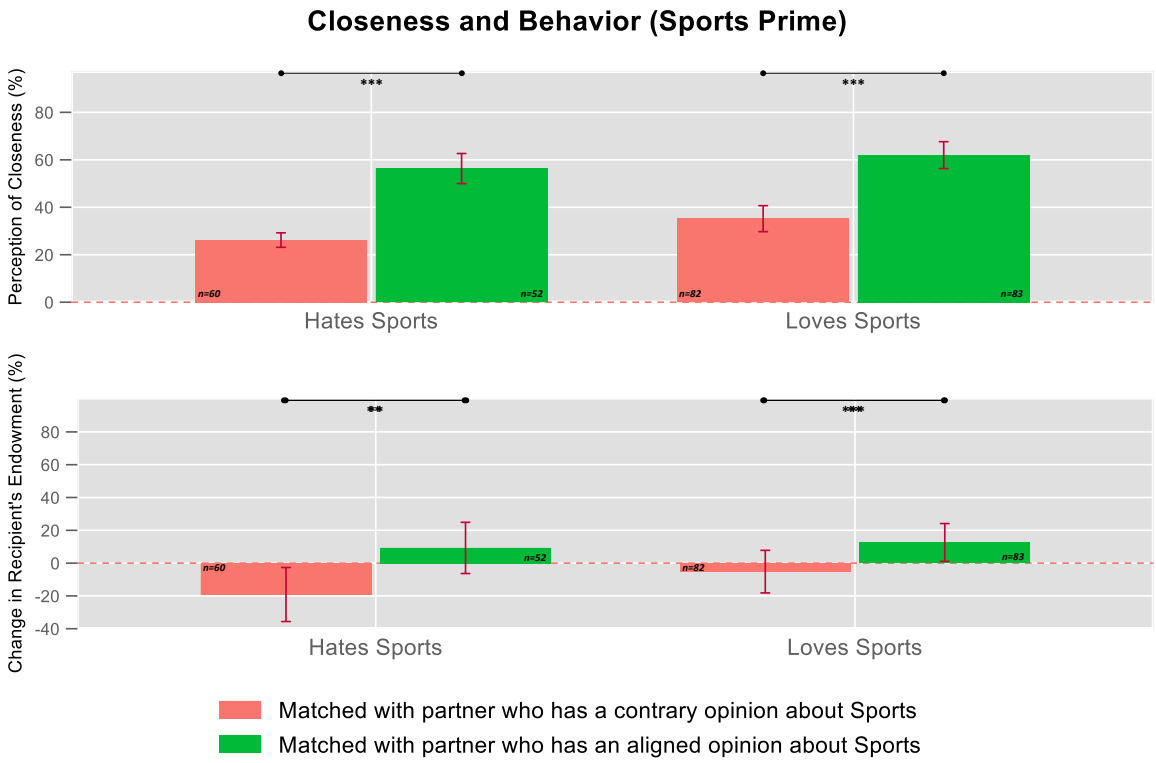

Figure OA.33: Closeness & behavior broken down by being matched with a partner who has a (mis)aligned opinion about sports. Adjacent bars (within each category) are compared. Absence of significance stars $\Rightarrow$ p-values > 0.05.

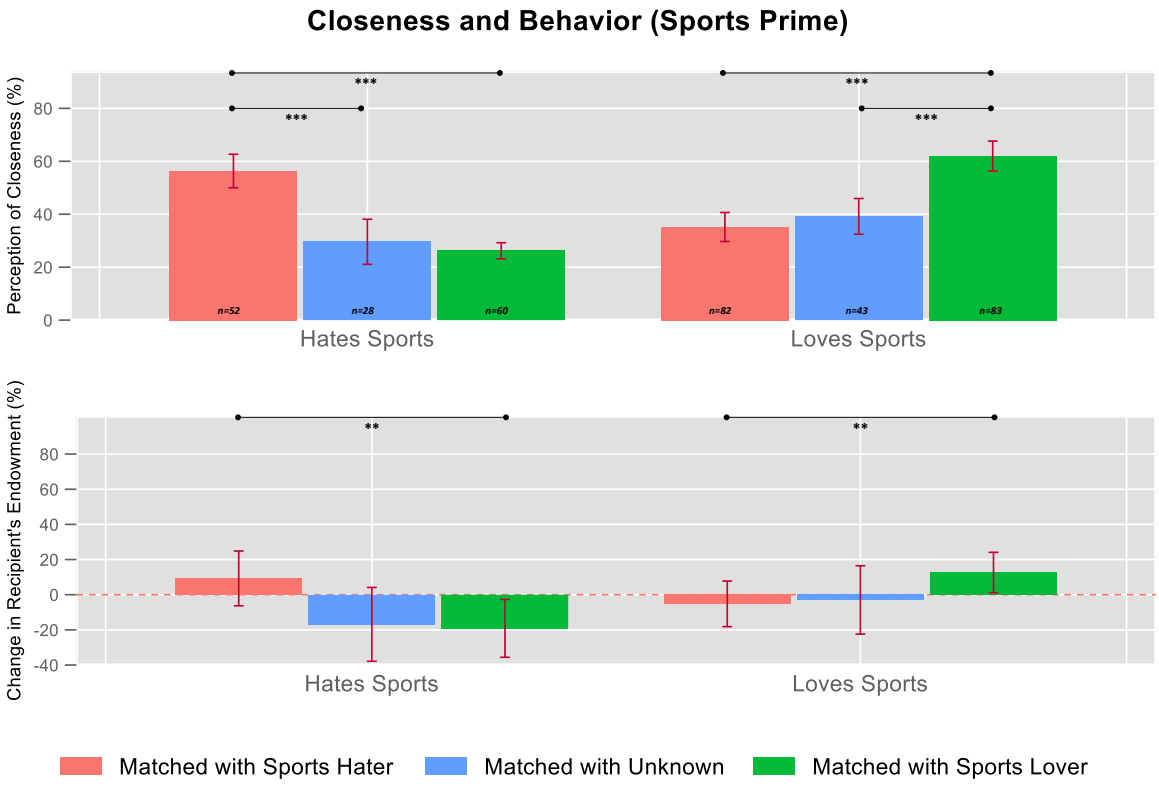

Figure OA.34: Closeness and behavior broken down by own opinion and being matched with a partner who has a (mis)aligned opinion about sports. Adjacent bars (within each category) are compared. Absence of significance stars $\Rightarrow$ p-values > 0.05.

V.d. Sports Prime – Public Goods Game

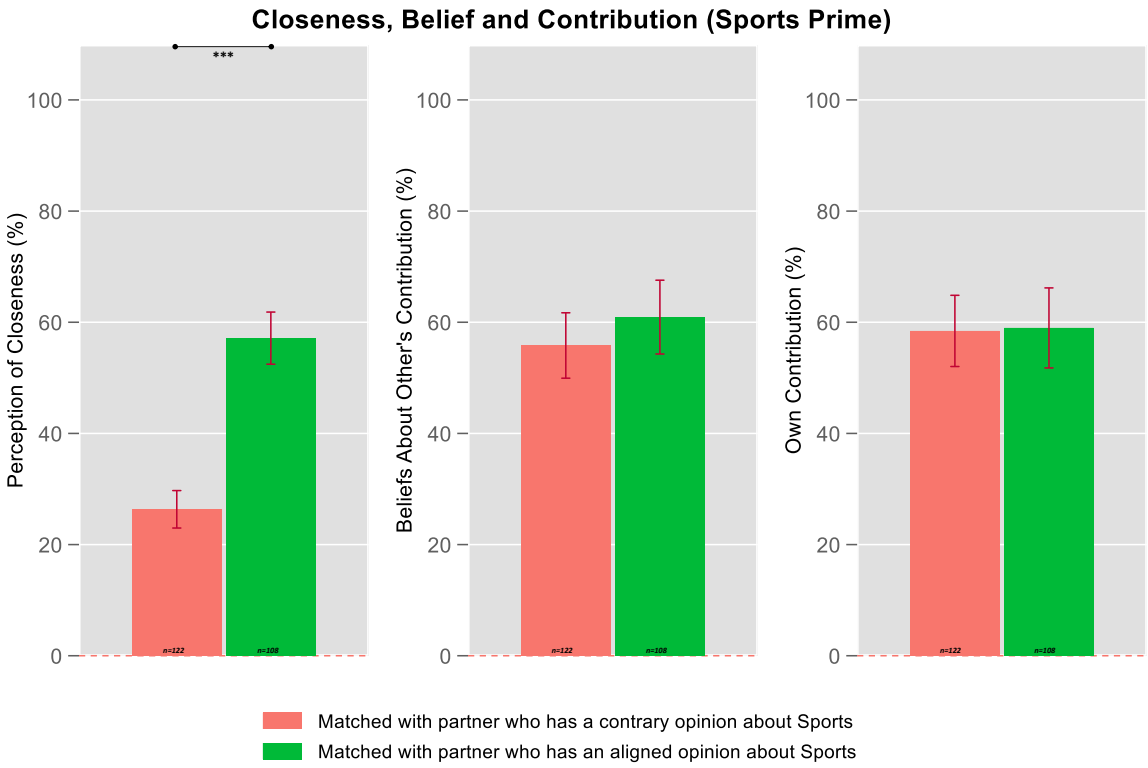

Figure OA.35: Closeness, beliefs & behavior broken down by being matched based on the partner's opinion about sports. All adjacent bars (within each category) are compared. Absence of significance stars ⇒ p-values > 0.05.

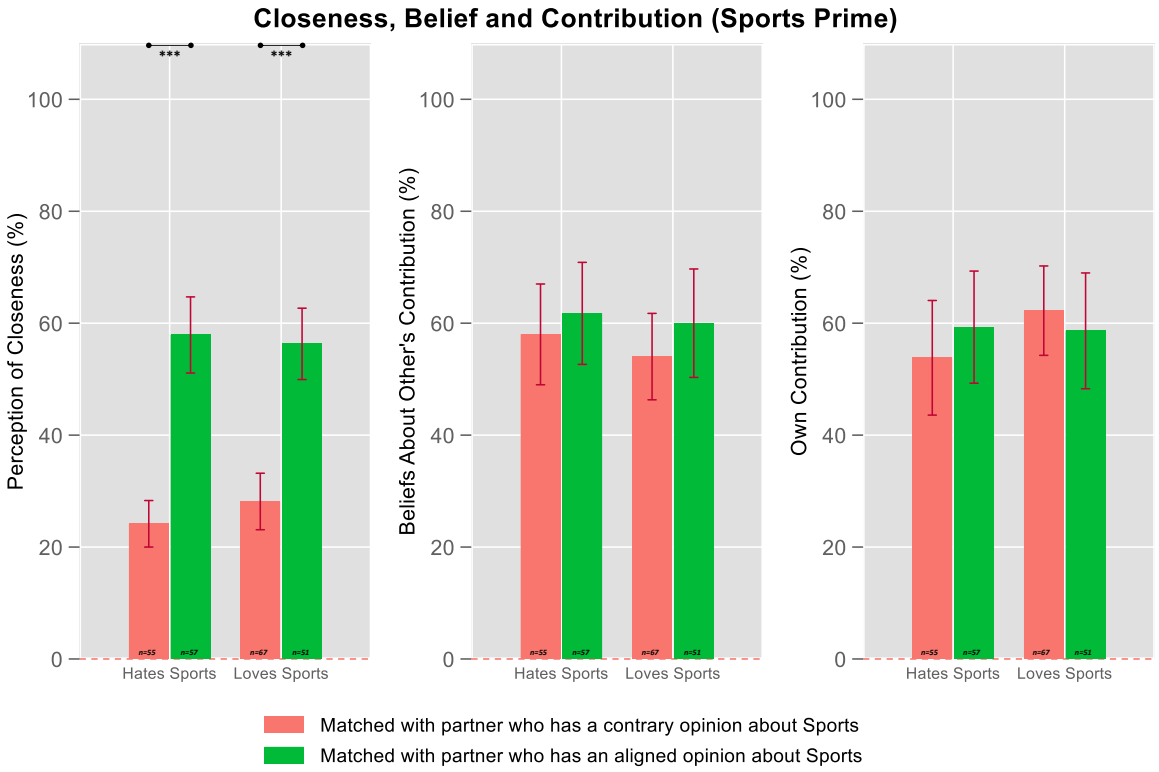

Figure OA.36: Closeness, beliefs, and behavior broken down by own opinion about Biden and being matched based on the partner's opinion about sports. All adjacent bars (within each category) are compared. Absence of significance stars $\Rightarrow$ p-values > 0.05.

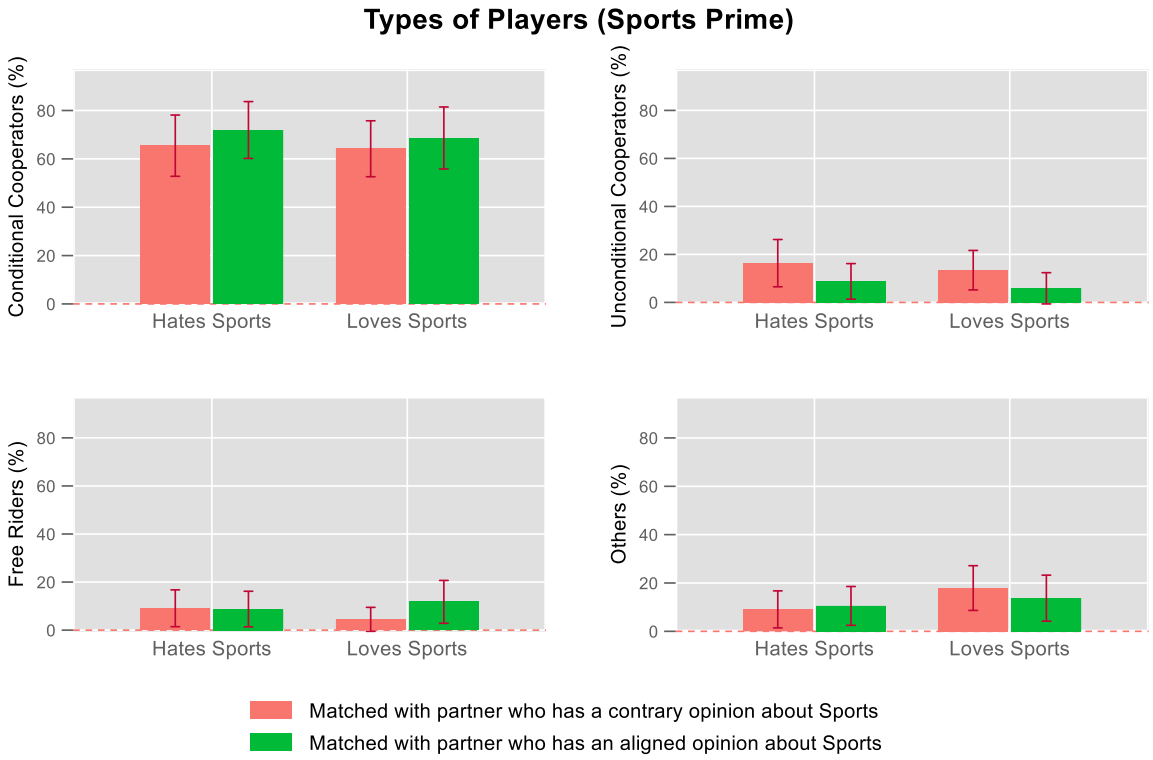

Figure OA.37: Types (conditional cooperators, unconditional cooperators, free riders, others) broken down by one's own opinion and being matched with a partner who either has aligned or contrary opinions for the sports treatment. All adjacent bars (within each category) are compared. Absence of significance stars $\Rightarrow$ p-values > 0.05.

VI. Experimental Screenshots

Below are a few selected examples that tie back to the main text. All original experimental screenshots can be downloaded from: https://osf.io/auh4k/

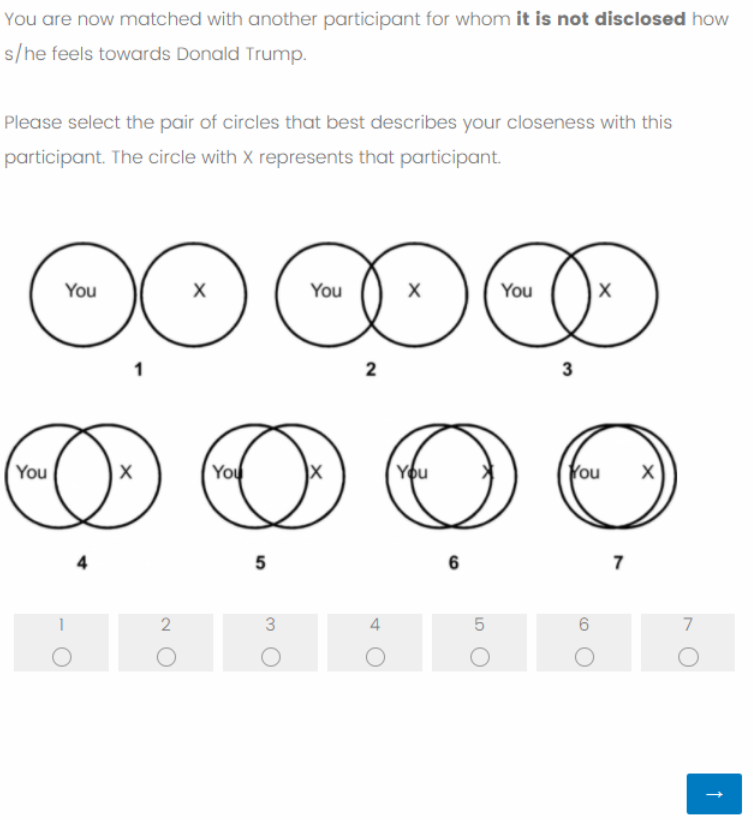

Figure OA.38: 'Inclusion of Other in the Self' (IOS) scale as used in all experiments (here exemplified for the condition in which the partner's preference remained undisclosed to the participant).