How to talk about an out-group: Effects on in-group trust and out-group generosity

Jan Biermann, Hendrik Hüning and Lydia Mechtenberg[*]

October 18, 2021

Abstract

This paper examines how deliberation in an in-group on how much to share with an out-group affects in-group trust and out-group generosity. In a lab-in-the-field experiment with 13 schools, we randomly assign school minors into pairs that decide how much of a common fund to transfer to refugee minors. Treatments vary whether pairs partake in a free-form chat or write down their reasoning individually. After treatment, they vote on transfers. In our sample, communication on refugees is shaped by a political-correctness norm: it is more refugee-friendly than individual reasoning, and it increases optimism within pairs about the partner's refugee-friendliness. Subjects trust their partners the more, the more refugee-friendly they believe them to be. This is rational in our sample since more refugee-friendly subjects turn out more trustworthy. Communication also has a positive impact on willingness to collectively share funds with refugee minors. Hence, our experiment indicates that in our sample, both the in-group and the out-group profit from a political-correctness norm to speak well of the out-group.

Keywords: Communication, trust, text data, collective decisions, donation, generosity.

JEL: C93, D70, D83

[*] Department of Economics, University of Hamburg, Von-Melle-Park 5, 20146 Hamburg, Germany. E-mail addresses: Jan.Biermann@uni-hamburg.de, Hendrik.Huening@uni-hamburg.de, Lydia.Mechtenberg@uni-hamburg.de.

This paper benefited from discussions with Claudia Schwirplies. We are grateful to Sophia Schulze-Schleithoff and Jan-Patrick Mayer for excellent research assistance and the University of Hamburg WISO lab (particularly Ziad Zorgati, David Lucius, Katharina Groß, Jana Lücke and Barabros Saritos) for outstanding technical assistance. This work has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 822590. Any dissemination of results here presented reflects only the authors' view. The Agency is not responsible for any use that may be made of the information it contains. IRB approval for this research was given at the University of Hamburg on the 19th of February 2019. Declarations of interest: none.

1 Introduction

Few issues occupy center stage in social discourse within Western countries as frequently as the issue of how to deal with migrants. As main destination countries, Western societies have to decide how much resources to share. Opinions vary widely among voters.[1] Both in the EU and in the U.S., debates on how to divide resources between natives and migrants have turned out value-laden, which increased political polarization, and affects social cohesion. In general, when a group has to make a collective decision on how to treat an out-group, in-group discourse on this question may reveal individual levels of generosity as well as individual degrees of "groupiness" (Kranton et al., 2020), i.e., how much social preferences are skewed toward the in-group. As a consequence, if in-group members discuss an out-group, this may affect their mutual trust. For instance, a peer surprising another with speaking more positively about the out-group than expected may signal a generous nature, inspiring an increase in trust.[2] Alternatively, this peer's openness toward the out-group may also signal a lower-than-expected in-group identification, deterring trust as a consequence.[3]

In the experiment that we report in this paper, we let pairs of peers discuss on how much of a common fund to share with incoming refugees. Hence, transfers to the out-group are costly for the in-group. We study how changes in subjects' beliefs about their peer's attitude toward refugees affect their trust in that peer, and whether a norm exists that shields them from trust-derogating over-revelation of attitudes counting as "bad signals".

Importantly, we study in-group trust rather than trust in out-group members[4]; and we do so under a new angle: with our first research question, we focus on how deliberating on behavior toward out-group members affects in-group trust. As is well-known from the literature, trust is an important de-

[1] In-group identity can create favoritism toward the in-group and reduce the willingness to share with out-group members (Chen and Li, 2009; Lane, 2016; Grimm et al., 2017; Abbink and Harris, 2019). Perceived neediness and deservingness are important factors, too, when generosity is concerned (Cappelen et al., 2020; Engel, 2011).

[2] Trustworthiness correlates with generosity; see, e.g., Cox et al. (2016).

[3] These two opposing hypotheses can be derived from the literature. In a sample that is heterogeneous with regard to groupiness, a relatively high willingness to transfer to the out-group signals a relatively low willingness to transfer to in-group members (Lee et al., 2021). By contrast, in a sample that is heterogeneous with regard to the steepness in which preferred transfers decline in social distance, relatively high willingness to transfer to the out-group signals a relatively high willingness to transfer to in-group members (Jones and Rachlin, 2006).

[4] For research concerned with trust in out-group members see e.g. Agranov et al. (2020).

terminant of a country's economic success (Bloom et al., 2012; Butler et al., 2016; Fukuyama, 1995; Guiso et al., 2004, 2008, 2009; Knack and Keefer, 1997; Zak and Knack, 2001; La Porta et al., 1997; Leonardi et al., 2001; Algan and Cahuc, 2014). Hence, it is important to investigate whether an open hand for refugees at the cost of the non-migrant community makes non-migrant citizens less or, on the contrary, even more trustworthy in the eyes of their fellow non-migrant citizen.

If trust is affected by revealed attitudes toward refugees, and if those engaging in the discourse are aware of that risk, then intuitively one should expect them to uphold a norm protecting them from a trust-derogating communication. This leads to our second research question: Are there norms that influence the tone and content of the debate?

In the past decades, much has been done to incorporate social norms into the study of economic behavior (Fehr and Fischbacher, 2004a,b; Krupka and Weber, 2013). However, the focus has been on norms of cooperative actions, such as public-good provisions (Reuben and Riedl, 2013; Markussen et al., 2014), transfers to peers (Fehr and Fischbacher, 2004a), or donations to charities (Bartling and Özdemir, 2017). In contrast, norms of opinion expression have been largely neglected in economics.[5] However, it is not unreasonable to expect that in value-laden contexts in particular, norms of opinion expression veil the actual standpoints at least partly, thereby shielding debaters from revealing a trust-diminishing attitude. As such norms, if existent, are likely to have a great impact on deliberation, we intend this paper to contribute to introducing them as an object of interest into the economic literature.

Finally, if there are norms of opinion expression, do they affect the collective decision itself? That is, do groups with social norms that influence the expression of opinions about an out-group decide differently on how to treat that out-group, depending on whether they debated on this before? The experimental literature has shown that social discourse can foster social responsibility (Bartling et al., 2020). However, through which channels it does so - through increasing the salience of moral aspects, establishing social norms, or otherwise - is still a largely unresolved question. If this question can be answered with reference to social norms of opinion expression, the latter will become of major interest to economists. For instance, norms of

[5] One notable exception is the signaling game in Golman (2021), where opinions are informative signals about a person's values and political correctness arises when expressing an unpopular opinion is avoided because it leads to bad judgements about one's values. Other disciplines have addressed norms of opinion expressions focusing on self-censorship (Hayes et al. 2005, Steen-Johnsen and Enjolras 2016 and Chan 2018), remaining silent due to a fear of isolation (Noelle-Neumann 1974) and the dangers of norms of opinion expressions (Sunstein 2003).

political-correctness may shift beliefs about social norms regulating other types of behavior, thereby influencing behavior itself.

To address the above research questions, we conducted a lab-in-the-field experiment with 488 school minors at least 15 years old and mostly born in Germany. They were all recruited from 13 schools located in the two largest German cities, Berlin and Hamburg. We chose German school minors for our subject sample for two reasons. First, since we study trust among an in-group and their generosity towards an out-group, a sample of school minors mostly born in Germany that share a common destiny during their time at school are more suitable than the international student samples of laboratory pools. Second, we are interested in rich text data generated in a deliberation process that precedes a real decision on transfers to refugees. School minors are more likely to seriously deliberate prior to such a decision than students in the laboratory pool who in chat experiments tend to generate text that is limited in scope, both with regard to tone and content.

Overall, we contacted 214 schools, among which 13 selected themselves into our sample.[6] Hence, external validity is restricted to school minors of homogeneous nationality and 15 or more years that voluntarily participate in an anonymous online chat on refugees in a liberal democracy similar to Germany's. We discuss external validity in more detail in section 5.

In our experiment, we randomly matched the school minors into groups of two, preserving anonymity and informing them about the likely possibility of being matched with a total stranger. A random two-third of these groups had to chat on how much of a common windfall endowment destined to be equally split and added to their respective class funds they wanted to share with refugee minors. We asked them to explicitly consider arguments in favor and against sharing. The other one-third had to write down their considerations in private. Afterwards, all had to vote on how much to share with the refugee minors. Within each group, a random entry of the share was chosen and implemented. The remaining group endowment was equally split and added to the respective minors' class funds. We designed our experiment to measure the effects that pre-vote deliberation - in the form of free-form chat - would have on trust between the matched partners and on their willingness to share with refugee minors. To this purpose, we implemented six stages in both the chat and the notes treatment: an initial survey, a first trust game between the matched partners as developed by Berg et al. (1995) and Fehr et al. (1993), the deliberation or notes-taking stage as implemented

[6] Selection is based on the school headmaster's willingness to participate in a study on how school minors deliberate on sharing with refugees - an information that the school ministries in Hamburg and Berlin demanded that we would give to schools during the recruitment.

in Brandts et al. (2021), the voting stage, a follow-up trust game between matched partners, and a final survey.

The initial survey elicits the prior attitude toward refugees, the general willingness to donate to a good cause, and demographics. Moreover, within this initial survey we prime our subjects differently by randomly showing them a video with either positive, negative, or balanced content with respect to refugee minors. We intended this prime to increase heterogeneity of opinion in our sample. However, our data show no significant effect of these video primes.

The two trust games, one before and one after the chat and the voting stage, allow us to test whether communication on the decision at hand, in contrast to private reasoning, affects in-group trust among matched partners. We also twice elicit beliefs about the partner's attitude toward refugee minors, before and after the chat or notes-taking stage. This enables us to measure whether changes in subjects' assessment of the matched partner's refugee-friendliness induces changes in how much they trust this partner. Comparing voting decisions across treatments, we further test whether communication affects the willingness to share collective funds with refugee minors.

We do not find that the treatment affects trust directly. However, we find robust evidence for an indirect effect. In our sample, subjects trusted their partner more if they perceived the partner as more refugee-friendly. Reversely, a subject is less trusted if her attitude toward refugees is perceived to be negative. Hence, beliefs on how other in-group members think of an out-group are a driver of in-group trust in our sample. We find that communication affects these beliefs: After communication, many subjects ascribe a higher degree of refugee-friendliness to their matched partners than before, compared to individual reasoning. This is due to a political-correctness norm: In communication more than in individual reasoning, subjects hold back negative attitudes toward refugees. Hence, in our sample politically correct communication on transfers to refugee minors increases trust between partners indirectly, through a change in beliefs. This indicates that a political-correctness norm requiring positive talk on refugees can serve the purpose of indirectly protecting trust within an in-group.

Consistently with this political-correctness norm, subjects in the chat treatment also vote for significantly higher transfers to refugee minors than those in the notes treatment. Since the transfers that our subjects suggest at the very beginning of the chat do not differ significantly from those that subjects suggest in their private notes, we conclude that a characteristic of communication itself, compared to individual reasoning, increases benevolence in the chat treatment. We argue that the apparent norm to speak positively about refugee minors to other in-group members is actually driving the result.

Our finding that communication has a positive impact on benevolence toward third parties is in line with Ellman and Pezanis-Christou (2010) and resonates with the literature on how social information affects donations and provisions to a public good (Frey and Meier, 2004; Martin and Randal, 2008; Croson and Shang, 2008; Shang and Croson, 2009; Rotemberg, 2014).[7]

The remainder of this paper is structured as follows: Section 2 introduces our experimental design and procedures. While Section 3 presents the data, our empirical results are summarized and discussed in Section 4. Section 5 concludes.

2 Experimental Design

We conducted our experiment during regular course hours in class rooms or computer rooms at our subjects' schools. The experiment was entirely computer-based. The program for the experiment was designed using the software oTree (Chen et al. 2016). At the beginning of each experimental session, school minors were randomly matched into pairs of two, which remained fixed throughout the experiment. We matched them across different classes or schools whenever we could allocate the same time slot to two different classes or even schools (see Table 1 for details). Moreover, we told all our subjects that they might have been matched with a school minor from a different class or even from a different school.[8] Then, the six stages of the experiment were conducted as in Figure 1.

2.1 Survey I and II

Besides standard demographic questions, we asked participants how pronounced their positive and negative attitudes towards refugees are on a 5-point likert scale ranging from *very strongly* to *weakly or not at all pronounced*. Moreover, we asked participants how pronounced they think the positive (negative) attitudes towards refugees of their co-player are. We elicited these beliefs at the beginning and the end of our study and are therefore able to measure the change in beliefs. Beliefs were incentivized

[7] See also Reyniers and Bhalla (2013) who find that pairs of two subjects donate more when making the decision collectively.

[8] Our subjects did not reveal their identity or their school to each other; hence, we implemented an anonymous partner-matching protocol throughout the experiment.

such that a subject was rewarded a bonus of 0.5 Taler if she guessed correctly (See Appendix B for details).

Furthermore, each participant was randomly assigned to one of three different videos on refugee minors at school (positive, negative, or balanced) during the first survey. The purpose of including these videos was to increase heterogeneity of opinions as well as provide diverse arguments to our subjects that they could use in subsequent stages.

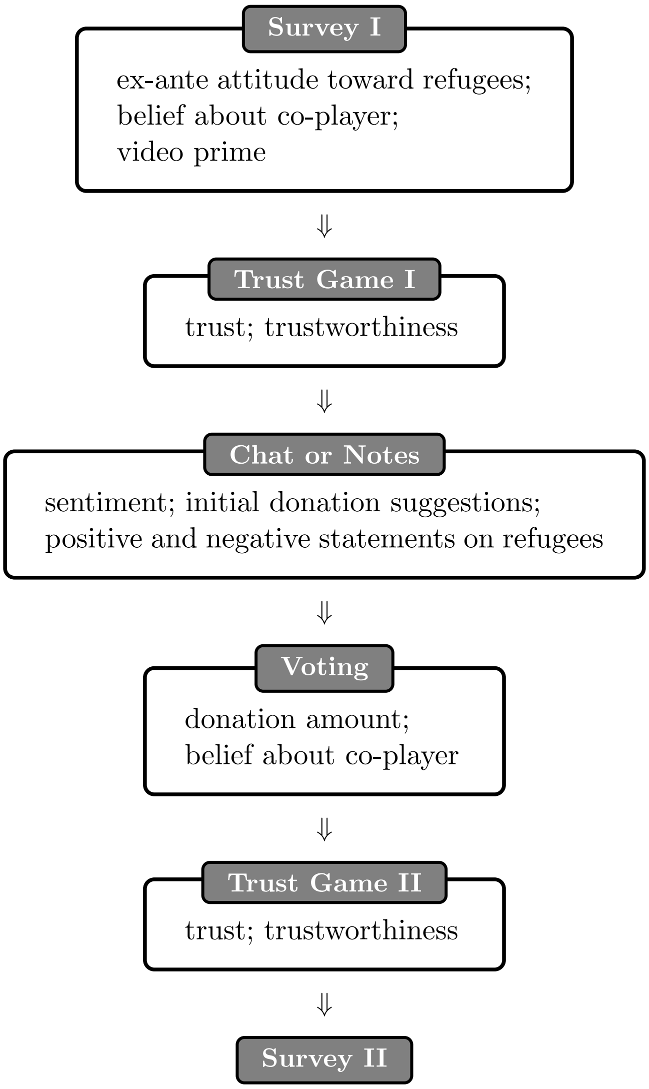


Figure 1: Stages of experiment including key variables

2.2 Trust Games

The two trust games on stages 2 and 5 are in the spirit of Berg et al. (1995): one of the two players is randomly assigned to receive an endowment of 10 units of our experimental currency, which we call "Taler", and chooses how much, if anything, of this amount to send to the other player. The amount sent is tripled. The receiving player on his part decides how much, if anything, of the amount received to send back. Every subject makes a decision as a trustor (first mover) and ten decisions as a trustee (strategy method for the second mover, one decision for every possible decision of the trustor) in every trust game stage. We hence collect information on how much the trustee wants to send back given any hypothetical amount the trustor can send (trustworthiness). Every subject is paid based either on her decisions as a trustor or based on her decisions as a trustee, which is drawn randomly. The results of all trust games are revealed only at the very end of the experiment.

2.3 Chat versus Notes

Between the two trust games, i.e. on stage 3, subject pairs enter the key stage of our experiment: they are confronted with the collective choice of a donation to refugee minors. Each pair receives a joint experimental budget of 30 Taler as a windfall gain. Next, subjects are asked: *How much, if anything, of your common budget do you want to donate to help minor unaccompanied refugees?*[9]

In the chat treatment (*Chat*), subjects have the chance to discuss this with their random partner in an online environment similar to WhatsApp; in the notes treatment (*Notes*), subjects take private notes. Both the chat and the notes taking last for seven minutes. Afterwards, both members of the matched pair cast a vote, i.e. each partner enters their preferred donation into an entry field. Voting is costless and mandatory. The decision is implemented according to random dictatorship (Gibbard 1977): The vote of one of the two members of the voting group is randomly chosen and each vote has equal probability to be chosen. The chosen amount is donated to refugee minors. The remaining share of the joint experimental budget is equally split among the two members of the voting group. By employing random dictatorship, we ensure that the subjects face incentives to reveal

[9] Our participants had the option to click on a link to get more information about the donation. The money was donated to the project “Helping Refugee Children” by the organization “Deutsches Kinderhilfswerk” which is a well-know charitable organization in Germany.

their true preferences rather than voting strategically. The information on the vote of the partner and which vote will be implemented is revealed only at the very end of the experiment together with the results of the trust games.

Pay-offs Pay-offs are determined with equal probability from the first trust game, the voting stage, or the second trust game. If the voting stage is drawn, the amount that remains after the donation is split equally between the two members of the respective pair. This pay-off structure is explained to the subjects at the beginning of the experiment, while the drawing takes place at the very end of the experiment. Participants are paid 0.6 Euro for each Taler they have earned in the experiment. Importantly, a participant's pay-off at the end of the experiment is not paid to the participants individually, but into the class fund. The class fund is a public good available to the class as a whole. Such resources are typically used to finance cultural activites of the class. Hence, in the voting stage, the participants make a decision between allocating resources to group insiders (class fund) or to group outsiders (refugee charity).

Recruitment For the experiment, we contacted 214 secondary schools from the states Saxony, Berlin and Hamburg. In total, 16 schools agreed to collaborate with us, all located in Berlin or Hamburg.[10] We conducted our experiment between December 2019 and March 2020 in thirteen of these schools. Four of those are located in Berlin and nine are located in Hamburg. In March, we had to stop our fieldwork because of school closings due to the Corona pandemic. For this reason, we could not conduct our experiment in the remaining three schools. We only considered a class if all pupils in it were at least fifteen years old; and we only accepted school minors whose parents gave informed consent, in addition to theirs.

3 Data

Overall, 501 school minors participated in our experiment in 19 different sessions.[11] Due to technical malfunction, we had to dismiss data from 13 subjects, resulting in 488 observations for our analysis. Table 1 presents some descriptive characteristics of our sample. Subjects are between 15 and 21 years old, averaging 17 years; 56% of subjects are female and one subject is

[10] Hence, we have a selected sample of schools open to collaboration with researchers from Hamburg and to the topics of deliberation and refugees. School minors with averse views on refugees are most likely under-sampled.

[11] As our power calculation in Appendix D reveals, this was the targeted quantity for our research study.

diverse. From all participating school minors, 27% went to schools in Berlin. Almost all subjects are born in Germany (96%). On average, the trustors sent 5.52 Taler in trust game 1. On average, for each Taler their trustor sent, trustees returned 1.25 Taler in trust game 1. Our measure for subjects' general willingness to donate is their individual donation to Médicins sans Frontières (MSF), which we elicited in stage 1. It averages 10.11 Taler (6.06 Euro).[12] The maximum possible donation amount to refugee minors is 30 Taler and the average donation is 20.28 Taler (see Figure A1 in Appendix A for the distribution per treatment). Two-third of school minors are randomly assigned to the chat treatment (64%) and one-third to the notes treatment. Table 1 also reveals that positive attitudes are more pronounced than negative attitudes towards refugees. This allows for considerable heterogeneity between partners in *how* positive they feel toward refugees. Negative attitudes are less prevalent, which allows for less heterogeneity in intensity of negative attitude. These ex-ante attitudes are not different across treatments (MWU-test, p-values: 0.1888 and 0.6634, respectively).

The two minors in each matched pair were either from two different schools (29% of all pairs), from two different classes (28%) or from the same class (43%). However, our instructions did not explicitly mention the three possibilities; we only informed our participants that the partner could be from a different school. Our impression from the interaction with the minors during the debriefing is that most participants believed to have played the game with a partner from a different school.

4 Results

4.1 Chat stimulates donations and beliefs

Direct treatment effects We start with reporting our incentivized direct treatment effects. Results are depicted in Table 2. We see that *Chat* does not directly affect changes in trust, i.e. the difference in how much subjects trust their partner in the first and the second trust game, denoted as $\Delta\,Trust$. Thus, interaction with the co-player in the chat does not affect changes in trust *per se*. The same holds true for trustworthiness. In contrast, the willingness to donate for refugee minors is significantly higher in *Chat*. Thus, the chat interaction seems to reinforce the willingness to help refugees. Moreover,

[12] An (incentivized) question asked participants how much of 10 Euros they would donate to Médecins Sans Frontières (MSF), if they are drawn as a winner of a lottery at the end of the session. The lottery randomly picked one school minor in each session to win 10 additional Euros on top of the final pay-off.

Table 1: Summary statistics

	Overall	Notes	Chat			
Variable	Mean	Mean	Mean	St. Dev.	Min	Max
Age	17	16.99	17	0.91	15	21
Share female	0.56	0.56	0.56	0.50	0	1
Share school in Berlin	0.27	0.28	0.27	0.44	0	1
Share born in Germany	0.96	0.97	0.95	0.20	0	1
Household members	2.75	2.75	2.74	1.11	0	7
Pocket money (in Euro)	24.89	29.53	22.41	39.36	0	450
Share same class	0.43	0.39	0.46	0.50	0	1
Share same school different class	0.28	0.28	0.28	0.45	0	1
Share different school	0.29	0.33	0.27	0.45	0	1
Positive attitudes refugees	3.60	3.68	3.54	0.94	1	5
Negative attitudes refugees	2.14	2.16	2.13	0.96	1	5
Taler sent trust game 1	5.52	5.55	5.49	2.66	0	10
Taler sent trust game 2	5.75	5.83	5.71	2.93	0	10
Amount returned trust game 1	1.26	1.24	1.27	0.42	0	3
Amount returned trust game 2	1.25	1.23	1.27	0.43	0	3
Donation to MSF	10.11	10.52	9.86	6.67	0	16.67
Donation to refugees	20.28	18.64	21.19	8.86	0	30
Share positive video	0.34	0.32	0.35	0.47	0	1
Share negative video	0.33	0.32	0.32	0.47	0	1
Share balanced video	0.34	0.36	0.33	0.47	0	1
Share chat treatment	0.64	0	1	0.48	0	1
Pay-off (in Euro)	5.73	6.07	5.54	3.49	0	19

Notes: The number of observations is 488. The variables "Amount returned" correspond to the amount of Taler sent back by the trustee for each Taler received. Reported amounts are in Taler unless stated otherwise.

we find that *Chat* leads to a significantly positive update of a participant's belief about her co-player's attitudes towards refugees, denoted as $\Delta\,Opinion$ *co-player*.[13] With regard to our video primes, we find no effect of any of the videos on donations to refugees. Average donations for those that saw a positive, negative and balanced video are 20.65, 20.32 and 19.86 Taler, respectively (MWU-tests, p-values: 0.99, 0.99 and 0.99). Thus, the videos did not influence donation behavior. We correct the p-values of our direct treatment effects and video primes for multiple hypotheses testing using the Holm-method (Holm 1979).

[13] When examining only *Chat*, we see that subjects, on average, update their a priori belief about the co-player in the positive direction (Wilcoxon-signed-rank test, p-value: 0.000). In other words, on average the co-player is perceived as more refugee-friendly due to the chat interaction. In *Notes*, however, the beliefs about the co-player are not significantly updated (Wilcoxon-signed-rank test, p-value: 0.8715). For detailed information on the construction of the belief-change variable, see Appendix B.

Table 2: Summary of Treatment Effects (Notes/Chat)

Treatment Variable	Notes Mean	Chat Mean	p-value (MWU)
ΔTrust	0.25	0.22	0.999
ΔTrustworthiness	-0.01	-0.003	0.313
Donation to refugees	18.59	21.22	0.023**
ΔOpinion co-player	-0.03	0.51	0.000***

Notes: The table reports means per treatment and p-values of MWU-tests. *ΔTrust (ΔTrustworthiness)* is defined as the difference in Taler sent (Taler sent back per Taler sent) between trust game 3 and trust game 1. *Donation to refugees* is the individual donation decision ranging from zero to 30 Taler. *ΔOpinion co-player* is the belief update about the co-player's attitudes towards refugees after treatment. P-values are adjusted for multiple hypotheses testing with seven hypotheses (Chat on all four outcome variables plus videos on donations) using the Holm-method.

4.2 Trusting partners perceived as refugee-friendly

Change in trust We now have a closer look at the effect of the chat treatment on *ΔTrust*. The distribution of *ΔTrust* per treatment is depicted in Figure 2. Half of our subjects (244 subjects) exhibit a change in trust. In *Notes*, such changes cannot be rationalized with reference to new information about the matched partner. In *Chat*, by contrast, subjects have the opportunity of learning about their partner's attitude toward refugees between the two trust games - if partners are at least partially open about their attitudes. Trust in partners may, therefore, change in *Chat* if expressed attitudes toward an out-group are perceived as signals of trustworthiness toward in-group members. We hence investigate if changes in trust are affected by changes in beliefs about partners' attitudes that in turn are induced by chat interaction but not by private notes taking.

Table 3 summarizes a first pass at this analysis. It illustrates that indeed, there seems to be such an indirect effect of the chat interaction on trust.[14] First, as column 2 indicates, we find that a positive belief update indeed increases trust. Second, columns 3-5 include an interaction term between *ΔOpinion co-player* and *Chat*. Its robust significance suggests that our exogenous treatment *Chat* affects changes in trust through the mediator *ΔOpinion co-player*.[15] Interestingly, only the interaction effect but none of

[14] Table 3 also confirms our results from Table 2 in that *Chat*, on average, does not exhibit a direct effect on trust (column 1). This differs from Charness and Dufwenberg (2006); Buchan et al. (2006) and Ben-Ner et al. (2011).

[15] As Figure 2 indicates, ΔTrust is not sufficiently normally distributed to justify OLS regressions (Shapiro-Wilk test, p-value: 0.000). As a robustness check, we therefore estimate the same specifications using ordered logit. Results depicted in Table C1 indicate

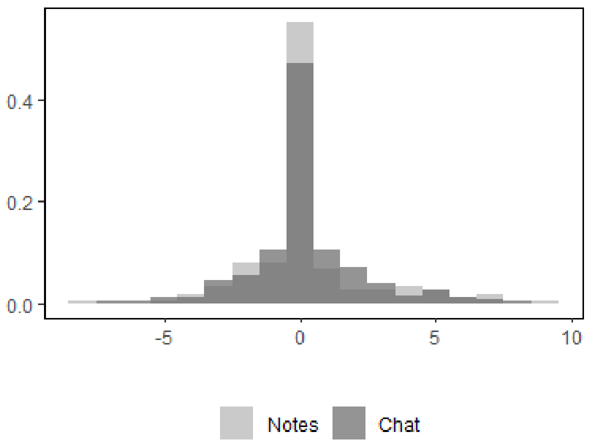


Figure 2: Distribution of ΔTrust

the main effects is significant.

Table 3: Chat interaction and trust

	(1)	(2)	(3)	(4)	(5)
Chat	−0.031	−0.117	−0.139	−0.144	−0.084
	(0.212)	(0.211)	(0.209)	(0.230)	(0.219)
ΔOpinion co-player		0.144**	−0.125	−0.138	−0.123
		(0.056)	(0.113)	(0.112)	(0.114)
Chat*ΔOpinion co-player			0.330**	0.345***	0.315**
			(0.129)	(0.130)	(0.130)
Constant	0.247	0.271	0.262	0.636	−0.024
	(0.168)	(0.169)	(0.167)	(1.944)	(2.583)
Obs.	488	486	486	481	481
Control Variables	No	No	No	Yes	Yes
School FE	No	No	No	No	Yes
R^2	0.00004	0.013	0.023	0.037	0.058
F Statistic	0.022	3.119**	3.772**	1.661*	1.215

Note: The table reports OLS regressions. Δ*trust* is the dependent variable and measures the difference between *taler sent trust game 2* and *taler sent trust game 1*. Controls: *Female* is a dummy equal to one for female subjects and zero otherwise. *Age* is numeric ranging from 15 to 21. *Born in Germany* is a dummy that is equal to one for subjects born in Germany and zero otherwise. *Pocket money* is numeric stating the money subjects receive from their parents. *Household members* indicates a subject's number of household members. *Risk aversion* is a 5-point likert scale reflecting subjects perceived risk attitudes. *Lottery bet* reflects subjects riks attitudes by measuring their willingness to pay for a lottery ticket that has a 50% chance of winning 300 Euros. Finally, *Sentiment co-player* is the number of positive words minus negative words normalized by the total amount of words written by the co-player (this variable is interacted with *Chat*). Standard errors clustered on the chat-group level are reported in parentheses. * indicates significance at the 10% level, ** at the 5% level and *** at the 1% level.

that the results are robust to the choice of the regression model.

Causal mediation analysis We now move on to performing a mediation analysis that tests for the potential causal effect of our treatment (*Chat*) through Δ*Opinion co-player* on Δ*Trust*. We hypothesize that chat interaction affects changes in trust among school minors through a change in beliefs about the co-player's attitudes towards refugees (for an illustration see Figure E1).

We identify a causal mediation effect based on the following four assumptions: There must not be confounders between treatment and outcome relationship (Assumption 1), between mediator and outcome relationship (Assumption 2), or between treatment and mediator relationship (Assumption 3). In addition, there must not be confounders affected by the treatment between mediator and outcome relationship (Assumption 4). These assumptions are also known as sequential ignorability or sequential independence assumptions (Huber 2020).

Assumptions 1 and 3 are met because our treatment is randomized. For Assumption 2 we have to carefully think of all post-treatment potential confounders that may affect the path from the mediator to the outcome. As individuals have the chance to update their belief about their co-player's attitudes towards refugees directly after the chat and this is instantaneously followed by the second trust game, there is little reason to believe that a post-treatment confounder would dilute the path from the treatment through the mediator on changes in trust. Similarly, Assumption 4 is the more plausible, the less time is elapsed between the treatment and the mediator (VanderWeele 2016). In our case, the question about the co-player's attitudes towards refugees occurs directly after the treatment, i.e. *Chat*/*Notes*. Under these sequential ignorability assumptions, a causal mediation from the treatment to the outcome variable can be established. In the following, we empirically investigate if such a mediation effect exists, using the methods proposed in Imai et al. (2010a) and Imai et al. (2010b).[16]

As a first step, we formulate the outcome and mediator model as

$$\Delta Trust_i = \alpha + \beta Chat_i + \delta Chat_i * \Delta Opinion\ co\text{-}player_i + \gamma X_i + \epsilon_i \qquad (1)$$

$$\Delta Opinion\ co\text{-}player_i = \lambda + \theta Chat_i + \phi X_i + \eta_i, \qquad (2)$$

where X contains all covariates that are used as control variables in our trust regressions from Table 3. The regression models are subsequently used to estimate if there is an indirect causal effect from the chat interaction on the change in trust through a belief update with regard to the co-player's attitudes towards refugees.

[16] We use the *mediation* package in R (Tingley et al. 2014) that implements these methods.

Results of this mediation analysis are presented in Table 4. The ACME (Average causal mediation effect) is significant for those being treated, i.e. chat participants. This means that the chat interaction exhibits a significant indirect effect on $\Delta Trust$ via the mediator Δ*Opinion co-player*. The ADE (Average direct effect), however, is not significant, i.e. there is no direct effect of the treatment on $\Delta Trust$, confirming our previous results. Thus, the chat interaction does not per se affect changes in trust among subjects but only via the belief update about the co-player's attitudes towards refugees. In the following, we discuss the potential reason for why a change in beliefs about the partner's refugee-friendliness after the chat affects trust in that partner.

Table 4: Causal Mediation Analysis

Effect	Estimate	CI lower	CI upper	p-value
ACME (control)	-0.08	-0.23	0.04	0.206
ACME (treated)	0.11	0.04	0.20	0.002***
ADE (control)	-0.18	-0.59	0.25	0.458
ADE (treated)	0.00	-0.40	0.44	0.936
Total Effect	-0.07	-0.47	0.35	0.782
Prop. Mediated (control)	1.02	-5.14	6.37	0.780
Prop. Mediated (treated)	-1.44	-8.65	7.87	0.784
ACME (average)	0.02	-0.06	0.09	0.700
ADE (average)	-0.09	-0.49	0.35	0.756
Prop. Mediated (average)	-0.21	-2.11	3.26	0.982

Notes: Confidence intervals are obtained with nonparametric bootstrap using the percentile method. Sample size used 483. Simulations: 1000. * indicates significance at the 10% level, ** at the 5% level and *** at the 1% level.

Refugee-friendliness as a signal of generosity Refugee-friendliness might function as a credible signal of generosity, and thus trustworthiness, toward other participants. However, note that a participant's generosity toward refugees on the donation stage comes at the expense of the common class fund, and hence at the expense of the other participants. Therefore, it is not trivial if refugee-friendliness is taken as a signal of generosity toward other participants.

In fact, perceiving refugee-friendly partners as more trustworthy is rational if and only if a participant who has rather negative attitudes towards refugees is also less trustworthy, i.e. sends back smaller shares for each amount sent by the trustor, compared to a more refugee-friendly participant. The correlation between an individual's positive (negative) attitudes and her trustworthiness in the first trust game is 0.083 (-0.150) indicating that in our sample, subjects that have more negative attitudes towards refugees indeed send back smaller amounts in the trust game. We investigate this more deeply by regressing $\Delta\mathit{Trustworthiness}$ on an individual's attitudes towards refugees.

Table 5: Attitudes towards refugees and trustworthiness

	(1)	(2)	(3)	(4)	(5)	(6)
Pos. attitude	0.046**	0.054**				
	(0.023)	(0.026)				
Neg. attitude			−0.061***	−0.066***		
			(0.022)	(0.024)		
Comb. attitudes					0.032***	0.036***
					(0.012)	(0.014)
Constant	1.093***	1.153**	1.390***	1.420***	1.019***	1.034**
	(0.086)	(0.489)	(0.049)	(0.460)	(0.094)	(0.492)
Obs.	488	485	488	485	488	485
Control Variables	No	Yes	No	Yes	No	Yes
School FE	No	Yes	No	Yes	No	Yes
R^2	0.010	0.041	0.019	0.047	0.017	0.047
F Statistic	5.134**	0.994	9.580***	1.145	8.612***	1.150

Note: The table reports OLS regressions. Δ*trustworthiness* is the dependent variable and measures the difference between *amount returned trust game 2* and *amount returned trust game 1*. *Pos attitude* (*Neg attitude*) is an individual's positive (negative) attitude towards refugees stated at the beginning of the survey. The variable *Comb. attitude* combines positive and negative attitudes to one measure. Control variables are the same as in Table 3. Standard errors clustered on the chat-group level are reported in parentheses. * indicates significance at the 10% level, ** at the 5% level and *** at the 1% level.

Columns 1,3 and 5 in Table 5 show the effect of individual attitudes without control variables while columns 2,4 and 6 add controls. The findings highlight that trustworthiness increases (decreases) significantly with positive (negative) attitudes towards refugees. Hence, in our sample refugee-friendliness is indeed a credible signal of generosity toward other (non-refugee) participants in the same in-group, and the trust-enhancing effect of perceived refugee-friendliness can be rationalized.

4.3 The political-correctness norm

Given that in our sample, refugee-friendliness signals trustworthiness and perceived refugee-friendliness generates trust, which is a powerful resource, we now test whether a trust-preserving political-correctness norm exists for our sample. That is, we test whether during the chat participants with negative attitudes toward refugees refrain more strongly from fully expressing these attitudes than during the notes taking.

To this purpose, we have coders classifying each message in *Chat* and *Notes* as expressing either a positive, negative, or neutral attitude toward refugees. Based on their coding, we construct two measures that characterize how positive, respectively negative, a given participant writes about refugees.

To be precise, two coders manually and independently labeled each text message from both *Notes* and *Chat* as *Pro* or as *Contra.*[17] We only used annotated messages where both coders agreed on the labeling and discarded the rest. Krippendorff's alpha for *Pro* and *Contra* are 0.8 and 0.69, indicating substantial agreement among coders. Using these annotated text data, we construct the following variables:

$$Positivity_i = \begin{cases} \frac{Pro_i}{Pro_i+Contra_i} & if\ Pro_i + Contra_i > 0, \\ 0 & if\ Pro_i + Contra_i = 0, \end{cases} \tag{3}$$

$$Negativity_i = \begin{cases} \frac{Contra_i}{Pro_i+Contra_i} & if\ Pro_i + Contra_i > 0, \\ 0 & if\ Pro_i + Contra_i = 0, \end{cases} \tag{4}$$

where i is one chat interaction between two matched subjects or one individual note from subjects in *Notes.* These variables capture the positive and negative attitudes toward (donating to) refugee minors as expressed in *Notes* or in *Chat.*[18]

We find that overall, 610 messages in *Notes* contain expressed attitudes toward refugees; 419 (191) of those are positive (negative) attitudes. In *Chat*, 474 messages contain expressed attitudes, of which 373 (101) are positive (negative). Based on this annotation, we develop two measures that capture the attitudes toward refugees expressed by the participant: *Expressed positivity* and *Expressed negativity.* In Figure 3, we can observe that positive attitudes expressed in *Chat* do not differ systematically from those expressed in *Notes* (MWU-test, p-value: 0.288). Considering negative attitudes, however, Figure 3 illustrates that, on average, significantly fewer negative attitudes are expressed in *Chat* than in *Notes* (MWU-test, p-value: 0.000). This finding is corroborated by a simple sentiment analysis (see Appendix B). Moreover, as Figure B3 illustrates, heterogeneity in the number of positive messages is higher than those for negative messages.

We interpret these findings as follows. Since ex-ante attitudes towards refugees do not differ across treatments, our text-data analysis suggests self-

[17] We define *Pro* (*Contra*) in a broad sense. Each message expressing a positive (negative) attitude, feeling, or opinion towards refugees is labeled as *Pro* (*Contra*). Messages that argue in favor of donating more (less) are also labeled as *Pro* (*Contra*). We argue that given the nature of our setting, the expressed attitudes towards refugees and the expressed willingness to donate are inseparable.

[18] Spearman's correlation between $Positivity_i$ and $Negativity_i$ is -0.779. The two variables are not perfectly correlated because we set both variables equal to zero for subjects that do not express any positive or negative attitudes towards refugees.

censorship in the *Chat* treatment: subjects hold back at least some negative attitudes when communicating with another subject.

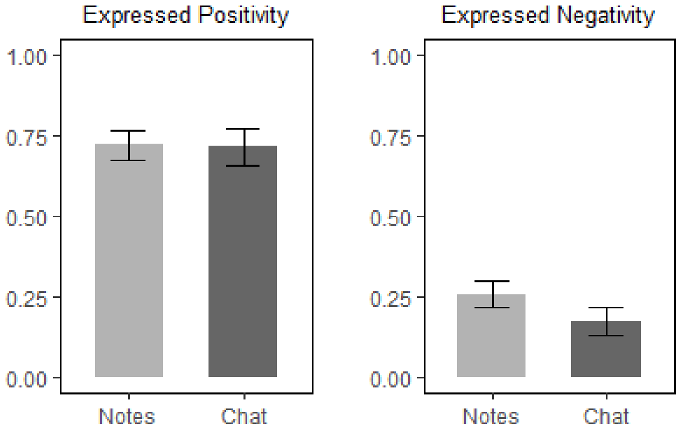


Figure 3: Expressed attitudes towards refugees

Beliefs about partners' attitudes Since subjects trust more in their matched partners if they perceive them to be refugee-friendly, it is important to know how (if at all) communication affects these beliefs. Figure 4 depicts the difference between a subject's incentivized belief about the partner's attitudes that we elicited both before and after treatment and the partner's actual attitudes ($Perceived - Actual$). In all cases, we see that belief errors are positive indicating that on average subjects believe their partners to have more positive feelings toward refugees than they actually have. Belief-errors are not significantly different across treatments, not before nor after treatment (MWU-test, p-values: 0.3392 and 0.451, respectively). Hence, subjects in *Chat* do not seem to learn significantly about the actual attitudes of their partners. In particular, they do not tune down their excess optimism about their partners' refugee-friendliness. This suggests that the political-correctness norm according to which subjects hold back their more negative attitudes toward refugees in the chat is successful in preserving this optimism which is, as shown above, helpful to sustain trust.

Effects on trust Finally, to define an upper bound on the extent to which the political-correctness norm seems to generate or to preserve trust, we now estimate how trust would have changed if subjects in *Chat* had learned the actual attitudes of their partners.

To this purpose, we first predict the change in beliefs that would have occurred under full revelation in the chat. We re-construct the variable

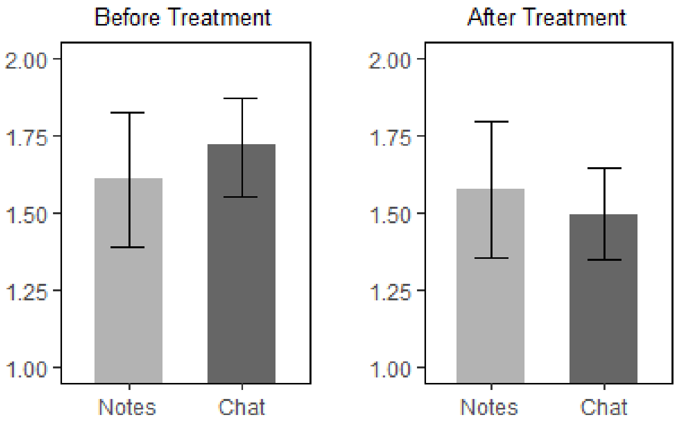


Figure 4: Error in belief about co-player's attitudes towards refugees

Δ*Opinion co-player* for the counterfactual situation in which beliefs after the chat would have been correct: in Δ*Opinion co-player* we set $BCP_{2i} = ATR_{1j}$, where ATR_{1j} are the ex-ante attitudes of subject's i's partner j. Next, we estimate how the fictitious belief change under full revelation would have affected trust. To assess the potential influence of Δ*Opinion co-player* on trust we choose the beta-parameter 0.184 from the regression of ΔTrust on Δ*Opinion co-player* and all control variables but for the sub-sample of chat participants only. Results are summarized in Figure 5.

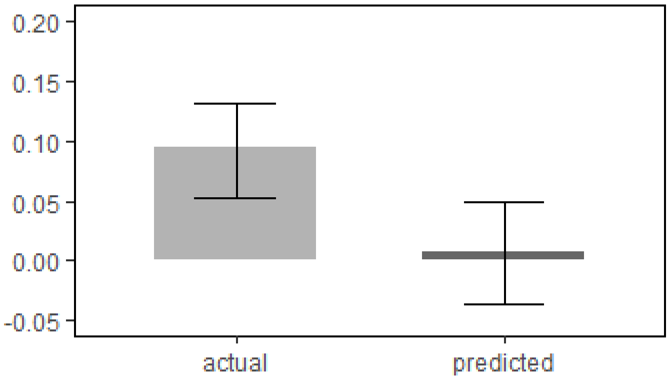


Figure 5: Predicted versus actual trust change due to Δ*Opinion co-player*

Comparing the means of the actual versus predicted influence of Δ*Opinion co-player* on ΔTrust reveals that on average post-chat trust would have been lower if beliefs were correct after the chat (MWU-test, p-value: 0.005). It is important to note, however, that trust would not have been destroyed, compared to the notes treatment. Rather, the trust premium that occurs since the chat generates overly optimistic beliefs about the partners' attitudes towards refugees would have disappeared, i.e. ΔTrust would have been closer to zero. Hence, the political-correctness norm does not only preserve trust but even turns communication about transfers to refugees into a trust-generating exercise.

4.4 The effect of *Chat* on donations

We now move on to investigating how our subjects' communication about potential transfers to refugees affects the latter, i.e. the amounts donated. As depicted in the right panel of Figure 6, we find higher donations in *Chat* than in *Notes*. This is confirmed by our regressions in Table 6. In specification (1) we examine the treatment effect on individual donation votes, not accounting for school fixed effects or control variables. The former are added in specification (2), the latter in specification (3). Our final specification (4) includes school fixed effects as well as control variables. Comparing the magnitudes of the chat coefficients reveals that the chat effect on donations to refugee minors is fairly robust across specifications. It is also statistically significant at any conventional level. Further, the chat effect is economically relevant: subjects in *Chat* donate 3.56 *Taler* more than subjects in *Notes*, which is more than 10% of the initial endowment.

In the following, we will discuss some of the possible channels through which our treatment influences the donation behavior. We particularly focus on three channels: a priori social image concerns, the political-correctness norm identified above, and social information.

Social image concerns and the political-correctness norm Social image concerns vis-à-vis the matched partner may motivate our subjects to suggest higher donations in the chat than they would have chosen privately. Two norms may give rise to such image concerns: a social norm to be generous, at least toward refugees, and the political-correctness norm to speak positively about them. The first may induce subjects to suggest higher-than-preferred donations in the chat right away, while the latter may affect donations indirectly: subjects withhold negative attitudes toward refugees from expression, speaking mainly in positive terms about them. Feeling thus committed, subjects may then follow through with their expressed generosity. To separate these two potential channels through which the chat may affect donations to refugee minors, we compare the very first suggestions of

Table 6: Chat interaction and donation

	(1)	(2)	(3)	(4)
Chat	3.232**	3.038**	3.814***	3.565***
	(1.446)	(1.345)	(1.259)	(1.189)
Attitudes Refugees			1.877***	1.539***
			(0.305)	(0.289)
Negative Video			−0.105	−0.231
			(1.312)	(1.259)
Balanced Video			−0.231	−0.137
			(1.341)	(1.233)
Female			3.126***	3.480***
			(1.175)	(1.091)
Age			1.090*	0.962
			(0.651)	(0.731)
Born in Germany			3.843	2.926
			(2.856)	(2.492)
Pocket money			−0.003	0.001
			(0.013)	(0.012)
Household members			−0.286	−0.251
			(0.448)	(0.441)
Donation to MSF			0.546***	0.564***
			(0.085)	(0.080)
December			−0.325	−3.256
			(1.691)	(3.726)
Constant	21.030***	25.632***	−21.923*	−12.615
	(1.110)	(2.703)	(11.909)	(14.757)
School FE	No	Yes	No	Yes
Clustered SE	Yes	Yes	Yes	Yes
Obs.	487	487	482	482
Wald Test	6.645***	86.780***	129.196***	202.513***

Notes: The table reports results of Tobit regressions with *Donation to Refugees* as the dependent variable. Since many subjects donate the maximum amount of 30 (see Figure A1 in Appendix A), we use Tobit regressions to estimate the effect of Chat on the latent unrestricted donation. The variable *Chat* is a dummy equal to one for subjects that chatted and zero otherwise. The variable *Attitudes Refugees* is a categorical variable indicating positive/negative attitudes towards refugees. The variable *Negative Video* (*Balanced Video*) are dummies that are equal to one for subjects that saw a negative (balanced) video about refugee minors and zero otherwise. The variable *Female* is a dummy equal to one for female subjects and zero otherwise. The variable *Age* is numeric ranging from 15 to 21. *Born in Germany* is a dummy that is equal to one for subjects born in Germany and zero otherwise. *Pocket money* is numeric stating the money subjects receive from their parents. The variable *Household members* indicates a subject's number of household members. The variable *Donation to MSF* indicates a subject's donation amount during the survey's lottery for donating to MSF. Finally, *december* is a dummy equal to one for sessions conducted in December and zero otherwise. Standard errors clustered on the chat-group level are reported in parentheses. * indicates significance at the 10% level, ** at the 5% level and *** at the 1% level.

amounts to donate (*first offer*) that subjects make in *Chat* with actual donations in *Notes*.[19] If the first donation suggestions in *Chat* exceed the actual donations in *Notes*, this indicates that the first channel is at work, i.e., that indeed a priori image concerns to appear generous drive the chat effect on donations.

Means of first offers in *Chat* and donations in *Notes* are displayed in the left panel of Figure 6. However, we find that there is no significant difference (MWU-test, p-value: 0.505). Hence, the chat effect on donations seems to be driven by differences between publicly and privately expressed attitudes rather than differences between publicly suggested and privately preferred donations. A priori social image concerns may play a role, but they do not seem due to an a priori social norm requiring generosity. Rather, discussing transfers to refugee minors under the political-correctness norm analyzed above seems to create a positive atmosphere that motivates higher donations than in *Notes*.

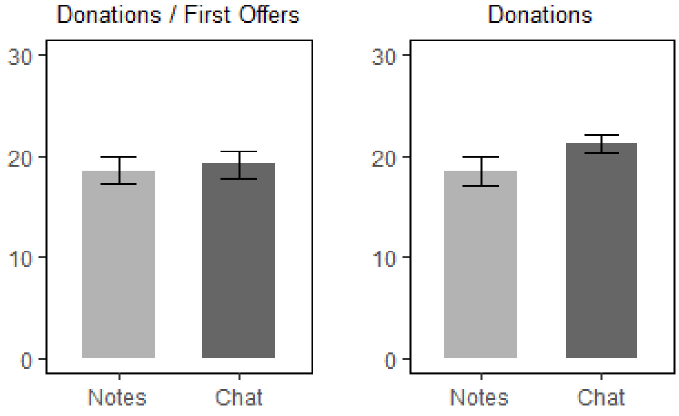


Figure 6: Analysis of first offers

Social information The third potential channel through which the chat may enhance donations is closely related to the second: As argued in section 4.2, the political-correctness norm generates (overly) optimistic beliefs about the partner's attitudes toward refugees. Since these attitudes and

[19] To obtain the *first offer* data, we manually label all messages in the chat which contain a donation amount. Based on this, we label the first of such message in every chat as *first offer*. In total, in 156 of 157 chat-groups a first suggestion of an amount to donate was made. First offers are mostly made at the very beginning of the chat interaction. The average message number in which the first offer appears is 5.6.

the willingness to donate are highly correlated (Spearman's correlation coefficient: 0.31), it is plausible to assume that the chat also creates more optimistic beliefs about the partner's willingness to donate. As the existing literature demonstrates, such beliefs may directly stimulate own generosity. For instance, Shang and Croson (2009) provide evidence that when individuals receive information about others donating large amounts, they tend to increase their own donations.

In sum, the chat seems to affect donations mainly via the political-correctness norm that shields negative attitudes toward refugees from perception, thereby creating both a donation-friendly communication dynamics and optimistic beliefs about the partners' willingness to donate. We do not find any evidence for a priori social image concerns. We confirm this interpretation of our findings by a mediation analysis in Appendix E.

5 Conclusion

We conducted a lab-in-the-field experiment with school minors to study how discussing the highly politicized issue of refugee help affects in-group trust among discussion partners and out-group generosity towards refugees. Participants were randomly assigned into groups of two that could either partake in a free-form chat among each other or write down their reasoning with regard to refugee aid individually. Subsequently, participants decided how much of a mutual endowment to donate to refugees. Our main finding is that communication about refugee help is subject to a political-correctness norm, which indirectly increases trust through generating overly optimistic beliefs about the co-players' attitudes towards refugees that serve as credible signals of trustworthiness. This positive belief update is driven by participants withholding negative opinions about refugees more strongly in the chat than in their private notes. Moreover, communication increases school minors' willingness to share funds with refugees by more than ten percent.

As a caveat, one has to be careful in generalizing results. Our sample is self-selected and relatively homogeneous, comprising school minors from the two largest cities in Germany, mostly born in Germany, and from schools that offer "Abitur" (the exam that must be passed for university attendance). A more heterogeneous sample, comprising also school minors from rural areas and schools that do not offer "Abitur" would probably fail to coordinate on a political correctness norm, and trust may even deteriorate. However, what our experiment reveals is that homogeneous social groups discussing an out-group do coordinate on a norm of opinion expression that preserves their in-group trust.

References

Abbink, K. and D. Harris (2019). In-group favouritism and out-group discrimination in naturally occurring groups. *PloS one 14*(9), e0221616.

Agranov, M., R. Eilat, and K. Sonin (2020). A political model of trust. Working Paper 2020-50, University of Chicago.

Algan, Y. and P. Cahuc (2014). Trust, growth, and well-being: New evidence and policy implications. In *Handbook of Economic Growth*, Volume 2, pp. 49–120. Elsevier.

Bartling, B. and Y. Özdemir (2017). The limits to moral erosion in markets: Social norms and the replacement excuse. Working Paper 6696, CESifo.

Bartling, B., V. Valero, R. A. Weber, and L. Yao (2020). Public discourse and socially responsible market behavior. Working Paper 8531, CESifo.

Ben-Ner, A., L. Putterman, and T. Ren (2011). Lavish returns on cheap talk: Two-way communication in trust games. *The Journal of Socio-Economics 40*(1), 1–13.

Berg, J., J. Dickhaut, and K. McCabe (1995). Trust, reciprocity, and social history. *Games and Economic Behavior 10*(1), 122–142.

Bloom, N., R. Sadun, and J. Van Reenen (2012). The organization of firms across countries. *The Quarterly Journal of Economics 127*(4), 1663–1705.

Brandts, J., L. Gerhards, and L. Mechtenberg (2021). Deliberative structures and their impact on voting under economic conflict. Working Paper 1022, Barcelona Graduate School.

Buchan, N. R., E. J. Johnson, and R. T. Croson (2006). Let's get personal: An international examination of the influence of communication, culture and social distance on other regarding preferences. *Journal of Economic Behavior & Organization 60*(3), 373–398.

Butler, J. V., P. Giuliano, and L. Guiso (2016). The right amount of trust. *Journal of the European Economic Association 14*(5), 1155–1180.

Cappelen, A. W., S. Fest, E. Ø. Sørensen, and B. Tungodden (2020). Choice and Personal Responsibility: What Is a Morally Relevant Choice? *The Review of Economics and Statistics*, 1–35.

Chan, M. (2018). Reluctance to talk about politics in face-to-face and facebook settings: Examining the impact of fear of isolation, willingness to self-censor, and peer network characteristics. *Mass Communication and Society 21*(1), 1–23.

Charness, G. and M. Dufwenberg (2006). Promises and partnership. *Econometrica 74*(6), 1579–1601.

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

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

Cox, J. C., R. Kerschbamer, and D. Neururer (2016). What is trustworthiness and what drives it? *Games and Economic Behavior 98*, 197–218.

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

Ellman, M. and P. Pezanis-Christou (2010). Organizational structure, communication, and group ethics. *American Economic Review 100*(5), 2478–91.

Engel, C. (2011). Dictator games: A meta study. *Experimental economics 14*(4), 583–610.

Fehr, E. and U. Fischbacher (2004a). Social norms and human cooperation. *Trends in Cognitive Sciences 8*(4), 185–190.

Fehr, E. and U. Fischbacher (2004b). Third-party punishment and social norms. *Evolution and Human Behavior 25*(2), 63–87.

Fehr, E., G. Kirchsteiger, and A. Riedl (1993). Does fairness prevent market clearing? an experimental investigation. *The Quarterly Journal of Economics 108*(2), 437–459.

Frey, B. S. and S. Meier (2004). Social comparisons and pro-social behavior: Testing" conditional cooperation" in a field experiment. *American Economic Review 94*(5), 1717–1722.

Fukuyama, F. (1995). *Trust: The social virtues and the creation of prosperity*, Volume 99. Free press New York.

Gibbard, A. (1977). Manipulation of schemes that mix voting with chance. *Econometrica 45*(3), 665–681.

Golman, R. (2021). Acceptable discourse: Social norms of beliefs and opinions. Mimeo.

Grimm, V., V. Utikal, and L. Valmasoni (2017). In-group favoritism and discrimination among multiple out-groups. *Journal of Economic Behavior & Organization 143*, 254–271.

Guiso, L., P. Sapienza, and L. Zingales (2004). The role of social capital in financial development. *American Economic Review 94*(3), 526–556.

Guiso, L., P. Sapienza, and L. Zingales (2008). Trusting the stock market. *The Journal of Finance 63*(6), 2557–2600.

Guiso, L., P. Sapienza, and L. Zingales (2009). Cultural biases in economic exchange? *The Quarterly Journal of Economics 124*(3), 1095–1131.

Hayes, A. F., C. J. Glynn, and J. Shanahan (2005). Willingness to self-censor: A construct and measurement tool for public opinion research. *International Journal of Public Opinion Research 17*(3), 298–323.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics 6*(2), 65–70.

Huber, M. (2020). Mediation analysis. In *Handbook of Labor, Human Resources and Population Economics*, pp. 1–38. Springer.

Imai, K., L. Keele, and D. Tingley (2010a). A general approach to causal mediation analysis. *Psychological Methods 15*(4), 309–334.

Imai, K., L. Keele, and D. Tingley (2010b). Identification, inference, and sensitivity analysis for causal mediation effects. *Statistical Science 25*(1), 51–71.

Jones, B. and H. Rachlin (2006). Social discounting. *Psychological Science 17*(4), 283–286.

Knack, S. and P. Keefer (1997). Does social capital have an economic payoff? a cross-country investigation. *The Quarterly Journal of Economics 112*(4), 1251–1288.

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

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

La Porta, R., F. Lopez-de Silanes, A. Shleifer, and R. W. Vishny (1997). Trust in large organizations. *American Economic Review 87*(2), 333–338.

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

Lee, V. K., R. E. Kranton, P. Conzo, and S. A. Huettel (2021). The hidden cost of humanization: Individuating information reduces prosocial behavior toward in-group members. *Journal of Economic Psychology 86 (102424)*.

Leonardi, R., R. Y. Nanetti, and R. D. Putnam (2001). *Making democracy work: Civic traditions in modern Italy.* Princeton university press Princeton, NJ.

Markussen, T., L. Putterman, and J.-R. Tyran (2014). Self-organization for collective action: An experimental study of voting on sanction regimes. *Review of Economic Studies 81*(1), 301–324.

Martin, R. and J. Randal (2008). How is donation behaviour affected by the donations of others? *Journal of Economic Behavior & Organization 67*(1), 228–238.

Noelle-Neumann, E. (1974). The spiral of silence a theory of public opinion. *Journal of Communication 24*(2), 43–51.

Rauh, C. (2018). Validating a sentiment dictionary for german political language—a workbench note. *Journal of Information Technology & Politics 15*(4), 319–343.

Remus, R., U. Quasthoff, and G. Heyer (2010, May). SentiWS - a publicly available German-language resource for sentiment analysis. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Reuben, E. and A. Riedl (2013). Enforcement of contribution norms in public good games with heterogeneous populations. *Games and Economic Behavior 77*(1), 122–137.

Reyniers, D. and R. Bhalla (2013). Reluctant altruism and peer pressure in charitable giving. *Judgment and Decision Making 8*(1), 7–15.

Rotemberg, J. (2014). Charitable giving when altruism and similarity are linked. *Journal of Public Economics 114*, 36–49.

Shang, J. and R. Croson (2009). A field experiment in charitable contribution: The impact of social information on the voluntary provision of public goods. *The Economic Journal 119*(540), 1422–1439.

Steen-Johnsen, K. and B. Enjolras (2016). The fear of offending: Social norms and freedom of expression. *Society 53*, 352–362.

Sunstein, C. (2003). *Why societies need dissent.* Havard University Press, Cambridge.

Tingley, D., T. Yamamoto, K. Hirose, K. Imai, and L. Keele (2014). mediation: R package for causal mediation analysis. *Journal of Statistical Software 59*(5), 1–38.

VanderWeele, T. (2016). Mediation analysis: A practitioner's guide. *Annual Review of Public Health 37*, 17–32.

Zak, P. J. and S. Knack (2001). Trust and growth. *The Economic Journal 111*(470), 295–321.

Appendix A

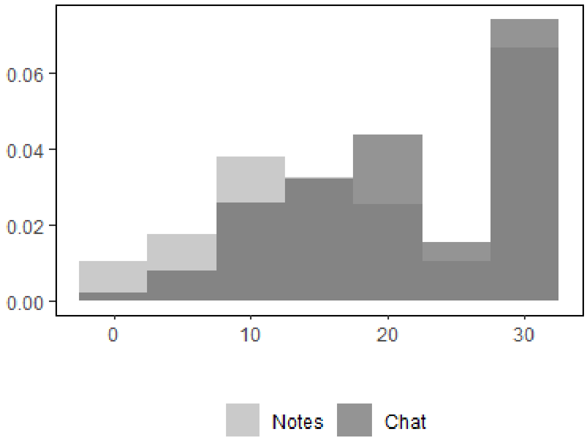


Figure A1: Distribution of *Donation to Refugees*

Appendix B

Instructions Chat and Notes We recommend the subjects to start the discussion by writing a few words about themselves (without revealing their identity). We expect this feature to reduce social distance and to make a constructive deliberation more likely. However, this personal introduction can happen in any form and is not enforced by the experimenter in any way. After this short introduction, we advise the subjects to concentrate the discussion on the donation matter and voting decision. Furthermore, we ask the subjects to provide arguments for or against donating to refugees. This allows us to analyze if expressing opinion and arguments influence the deliberation and subsequently the vote on transfers.

Sentiment analysis Chat vs. Notes To delineate potential channels through which chatting could affect donations and trust, we construct our main independent variables from the textual data obtained both in *Chat* and *Notes* of our subjects. First, we preform a sentiment analysis on the word level, measuring the general tone in which subjects deliberated or reasoned about how much to share with refugee minors. To this purpose, we use a predefined dictionary for the German language that classifies words as positive, negative or neutral (Rauh 2018).[20] We then calculate the sentiment for each chat group or a subject' notes as:

$$sentiment_i = \frac{positivewords_i - negativewords_i}{totalwords_i}, \tag{5}$$

where $positivewords_i$ ($negativewords_i$) denotes the number of all positive (negative) words and $totalwords_i$ denotes all words in a given chat group or notes.[21] Theoretically, this measure can range from -1 to +1. However, naturally occurring language often produces values close to 0.[22] For example, $sentiment_i = -0.1$ can be interpreted as a 10% overweight of negatively connoted language, suggesting a negative sentiment prevailing in chat i.
The language in *Chat* might be fundamentally different compared to *Notes* in fundamental ways, simply because the chat involves interaction with a different person. This can result in polite phrases ("nice talking to you"). In this case, any net positivity might be driven by positive words only. We find, however, that the data tells the same story when examining only positive (negative) words in the nominator of Equation 5: *Notes* contains significantly more negative words than *Chat* and the latter significantly more positive words (MWU-test, p-values: 0.000 and 0.000, respectively). Figure B1 illustrates the differences across treatments. is statistically significant.

It is important to note that this measure does not capture the context in which positive or negative expressions are used. Positive words may be used to express positive feelings (opinions, emotions, attitudes...) toward refugees (e.g. "most refugees are in general really nice people"). But they may also refer to any other topic unrelated to refugees (e.g. "it is nice talking to you"). Thus, $sentiment_i$ captures both, the general sentiment among school minors but also toward the discussed decision on how much to donate to refugee minors.

Constructing $\triangle$Opinion co-player In the questionnaire, we asked how pronounced participants think are the positive (negative) attitudes towards refugees of their co-player. The answer options to this question were *1 =*

[20] While Remus et al. (2010) developed a dictionary that set the standard for sentiment analysis for the German language, Rauh (2018) refines this measure by taking into account negation of words, that reverses the sentiment of an expression. Our results, however, do not depend on the choice of the dictionary. Results using the Remus et al. (2010) measure are available upon request.

[21] As robustness checks, we consider two measures with only positive (negative) words in the nominator of Equation 5.

[22] Rauh (2018) illustrates this by analyzing a sample of 1500 sentences from the German Parliament and showing that 31.6 % are classified as neutral.

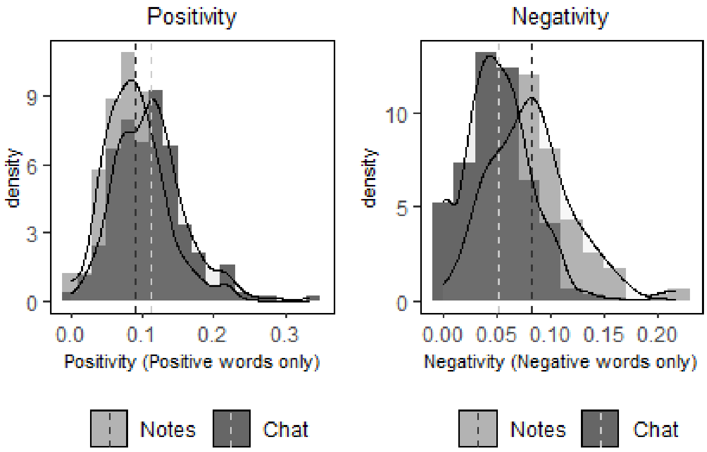


Figure B1: Sentiment - Positivity and Negativity by treatment

very strongly, *rather strongly*, *neither strongly nor weakly*, *rather weakly*, *5 = weakly or not at all pronounced*.

For our analysis, we combined the data from the two questions by first reversing the answers for the question regarding positive attitudes ('1' is recoded as '5' etc.). We then added up the answers regarding the positive attitudes (now reversed) and the answers regarding the negative attitudes. We implemented this for the answers before the chat and the answers after the chat. Finally, we subtracted the resulting value before the chat from the resulting value after the chat to obtain Δ*Opinion co-player*.

Robustness check ΔOpinion co-player Our results remain unchanged if we consider the positive attitudes instead of the combined measure: The average change is 0.058 in *Notes* and 0.323 in *Chat* (MWU-test: p-value=0.000). The same story holds for negative attitudes instead of the combined measure: The average change is -0.092 in *Notes* and 0.185 in *Chat* (MWU-test, p-value: 0.000). This impression is also confirmed when examining a Wilcoxon-signed-rank test for those who chatted in order to compare the beliefs before and after the chat. The p-value is 0.000 for the positive as well as the negative attitudes.

Differences: Actual vs. perceived attitudes towards refugees Figure B2 illustrates that in *Chat* as well as in *Notes*, the difference between the actual and perceived distance is positive indicating that individuals are underestimating the true distance to the co-player regarding attitudes towards refugees. We can also see that the perceived distance to the co-player decreases after *Chat*. This difference is, however, not statistically significant.

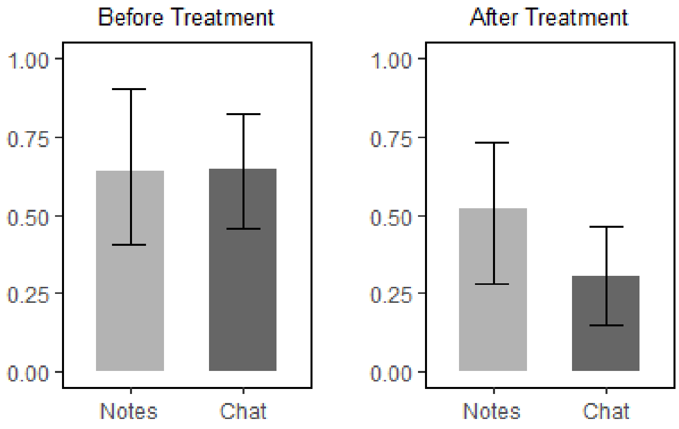


Figure B2: Error in perceived distance to co-player's attitudes

Distribution of expressed attitudes towards refugees within Chat Figure B3 illustrates the number of positive (negative) messages per participant as a share of all refugee related messages. Almost 80% of subjects did not express negative attitudes towards refugees in the chat. Overall, the heterogeneity in the number of positive messages towards refugees is a lot higher than those for negative messages.

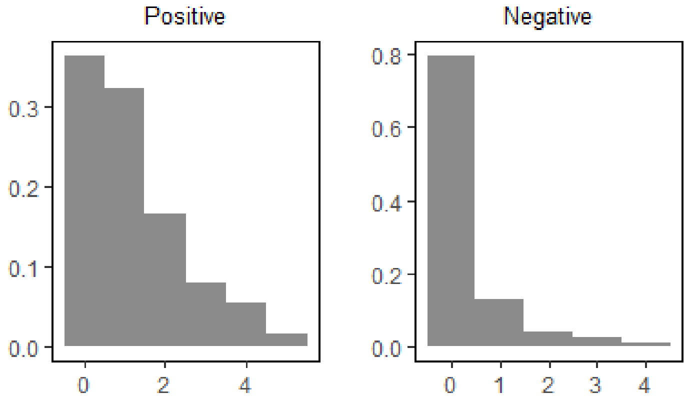


Figure B3: Number of positive and negative messages (as a share of all such messages) within *Chat*

Appendix C

As Table C1 illustrates, our results remain robust to the choice of an ordered logit regression model instead of OLS.

Table C1: Trust regressions with ordered logit

	(1)	(2)	(3)	(4)	(5)
Chat	0.085	−0.003	−0.023	−0.003	0.036
	(0.176)	(0.177)	(0.177)	(0.197)	(0.198)
∆Opinion co-player		0.165***	−0.063	−0.081	−0.077
		(0.048)	(0.103)	(0.101)	(0.105)
Chat*∆Opinion co-player			0.281**	0.301***	0.284**
			(0.116)	(0.115)	(0.119)
Obs.	488	486	486	481	481
Control Variables	No	No	No	Yes	Yes
School FE	No	No	No	No	Yes
AIC	1862.035	1839.623	1836.449	1830.558	1845.72

Note: The table reports ordered logit regressions. $\Delta trust$ is the dependent variable. Controls are as in Table 3. Log odds are reported as coefficients. Standard errors clustered on the chat-group level are reported in parentheses. * indicates significance at the 10% level, ** at the 5% level and *** at the 1% level.

As displayed in Table C2, we do not find any evidence that *Chat* affects trustworthiness. As it is the case with Δ *Trust*, we see that the effect of Δ *Opinion co-player* depends on *Chat* (interaction term in column 4). However, our data does not suggest that an individual's belief update about the co-player's attitudes towards refugees affects trustworthiness in any of the two treatment condition.

Table C2: Trustworthiness

	(1)	(2)	(3)	(4)
Chat	0.005	0.011	0.013	0.011
	(0.026)	(0.027)	(0.028)	(0.027)
ΔOpinion co-player			−0.001	−0.027
			(0.008)	(0.017)
Chat*ΔOpinion co-player				0.032*
				(0.019)
Constant	−0.008	0.110	0.126	0.145
	(0.019)	(0.259)	(0.265)	(0.263)
Obs.	488	485	483	483
Control Variables	No	Yes	Yes	Yes
School FE	No	Yes	Yes	Yes
R^2	0.0001	0.058	0.059	0.065
F Statistic	0.036	1.515*	1.453*	1.518*

Note: The table reports OLS regressions. Δ*trustworthiness* is the dependent variable and measured as the difference between *amount returned trust game 2* and *amount returned trust game 1*. Standard errors clustered on the chat-group level are reported in parentheses. * indicates significance at the 10% level, ** at the 5% level and *** at the 1% level.

Table C3: Sample Split Trust and Trustworthiness

	Dependent variable:			
	ΔTrust	ΔTrust	ΔTrustworthiness	ΔTrustworthiness
	All	Only Chat	All	Only Chat
Chat	−0.043		0.011	
	(0.211)		(0.027)	
ΔOpinion co-player		0.183***		0.004
		(0.064)		(0.009)
Constant	0.677	0.002	0.110	−0.027
	(2.604)	(3.358)	(0.259)	(0.319)
Controls	Yes	Yes	Yes	Yes
School Fixed Effects	Yes	Yes	Yes	Yes
Observations	485	312	485	312
R^2	0.035	0.087	0.058	0.093
F Statistic	0.900	1.473*	1.515*	1.567*

*Note: Standard errors clustered on the chat-group level are reported in parentheses. * indicates significance at the 10% level, ** at the 5% level and *** at the 1% level.*

Appendix D

Ex-ante power calculation Before running our lab-in-the-field experiment in schools, we performed the following power calculation in order to get an estimate for the efficient sample size. We make the following assumptions about the minimum relevant distance (MRD) between treatments and the common standard deviation (SD) of our outcome variables across treatments. Results are summarized in Table D1.

First, we define the MRD between our treatments *Chat* and *Notes* that is economically relevant. Our two outcome variables are (a) *Donation to Refugees*, where subjects can donate between 0 and 30 ECU, and (b) $\Delta trust$. The latter is defined as the difference in ECU ("Taler") sent between the two trust games that are played. In each trust game, subjects can sent between 0 and 10 ECU to their co-player. In both cases, we want to detect a MRD of at least 10% of the initial endowment between *Chat* and *Notes*. For *Donation to Refugees*, this means we want to detect at least a mean difference of 3 ECU between treatments. For $\Delta trust$, the MRD is 1 ECU. This corresponds to a real monetary difference of 1.80 and 0.60 Euro, respectively.

Second, we need an estimate for the common standard deviation (SD) of these outcome variables across treatments. In both cases, we rely on the standard deviation of these variables from our pilot study that we conducted with seventeen participants in the WISO lab at Hamburg University. This results in estimates for the standard deviation of 8.82 for *Donation to Refugees* and 2.67 for $\Delta trust$. For comparison, the standard deviations of *Donation to Refugees* and $\Delta trust$ in our final sample are 8.86 and 2.2, indicating that our estimates from the pilot study are quite accurate already. The effect size (ES) is calculated as MRD/SD.

Table D1: Efficient sample size calculation

Outcome variable	MRD	Exp. SD	ES	Efficient sample size
Donation to Refugees	3 ECU	8.82	0.34	166
$\Delta trust$	1 ECU	2.67	0.37	141

Notes: The table displays the power calculation for our two outcome variables *donation for refugees* and $\Delta trust$. We assume a minimum relevant distance (MRD) of 10% (of the initial endowment) between treatments, i.e. 3 ECU for *donation for refugees* and 1 ECU for $\Delta trust$, respectively. The expected standard deviation (SD) of the outcome variables are taken from the pilot study. The effect size (ES) is MRD/SD. For the efficient sample size (per treatment), we correct for two hypothesis being tested, i.e. the effect of *Chat* on *donation for refugees* and $\Delta trust$.

The efficient sample size per treatment, i.e. for *Chat* and *Notes*, is 166 observations in the case of *Donation to Refugees* and 141 observations in the case of $\Delta trust$. We assume power of 0.8 (commonly used for field experiments) and an alpha of 0.05. We correct alpha for two hypotheses we want to test, i.e. the effect of *Chat* on trust and donation behavior. That means we divide alpha by two.

Overall, we conclude that for the MRD we want to detect, i.e. a 10% difference between treatments, we need at least 166 observations per treatment, i.e. *Chat* and *Notes*. Since observations in *Chat* are not independent due to interactions among two subjects, we plan twice the minimal sample size of 166 observations for *Chat*. Overall, we plan a sample size of $3 * 166 = 498$, i.e. approximately 500 observations.

Appendix E

Mediation Analysis - Trust

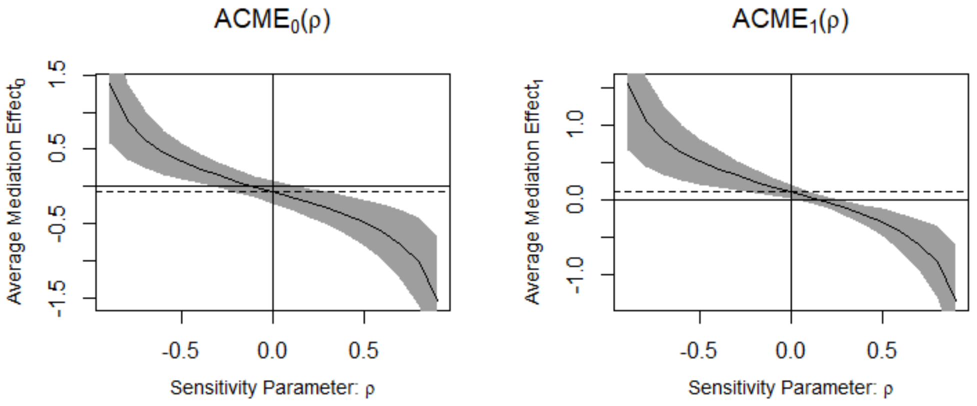
Direct and indirect path

$$\begin{tikzcd}
 & \Delta \textit{Opinion co-player} \arrow[dr, Rightarrow] & \\
\textit{Chat/Notes} \arrow[ur, Rightarrow] \arrow[rr, Rightarrow] & & \Delta \textit{Trust}
\end{tikzcd}$$

Figure E1: Potential Mediation - Trust

Sensitivity Analysis We perform a sensitivity analysis, which allows us to assess how robust our indirect effect estimates are to violations in the sequential ignorability assumptions and how substantial a violation in the assumptions would have to be in order to considerably alter our inferences about the indirect effect. The basic idea of the sensitivity analysis is to study the correlation ρ of the errors of both models (ϵ and η). Under sequential ignorability, ρ is equal to zero and thus the magnitude of this correlation coefficient represents the departure from the ignorability assumption. The correlation ρ for our outcome- and mediator model is -2.737265e-17. Assuming that our specification of both models are correct, it seems there is no evidence for a violation of the assumptions. Results for potential departures of ρ and therefore violations of sequential ignorability are summarized in Figure E2.

We see that for *Chat* the average causal mediation effect ($ACME_1(\rho)$), only departures of ρ into the higher positive domain would result in a different sign of the estimated mediation effect. Overall, ρ needs to be quite large to draw different conclusions about the mediation effect.



(a) For ACME (control)

(b) For ACME (treated)

Figure E2: Sensitivity Analysis

Mediation Analysis - Donations

$$\begin{tikzcd}
 & \Delta \textit{Opinion co-player} \arrow[dr, Rightarrow] & \\
\textit{Chat/Notes} \arrow[ur, Rightarrow] \arrow[rr, Rightarrow] & & \textit{Donations to Refugees}
\end{tikzcd}$$

Figure E3: Potential Mediation - Donations

In the following, we evaluate the potential mediation effect illustrated in Figure E3. We do so in the same way as we did for Δ *Trust*, i.e. we define the outcome and mediator model as in (1) and (2) only that this time *Donations to Refugees* is the dependent variable in (1). Again the methods proposed in Imai et al. (2010a), Imai et al. (2010b) and implemented in the R-routine by Tingley et al. (2014) are used to estimate the average direct (ADE) and average causal mediation effect (ACME). Results are depicted in Table E1.

The ADE and ACME are both significant for those being treated, i.e. chat participants, indicating that *Chat* affects willingness to donate to refugees directly but also indirectly via the belief update about the co-player's attitudes towards refugees (Δ *Opinion co-player*).

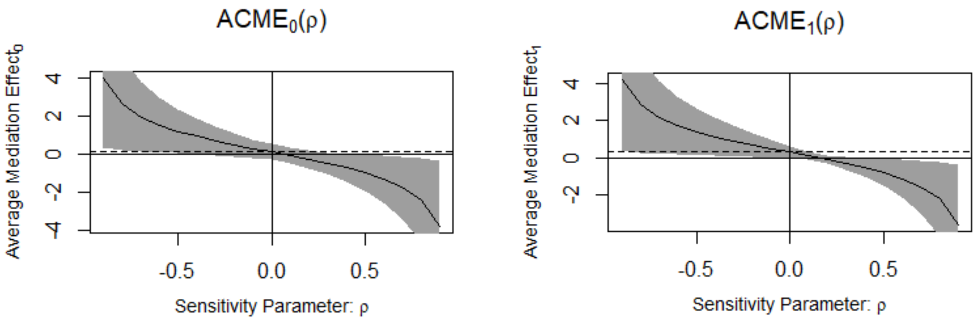
Sensitivity Analysis As for trust, we perform a sensitivity check, i.e. we evaluate how strongly the sequential ignorability assumptions have to be

Table E1: Causal Mediation Analysis - Donations

Effect	Estimate	CI lower	CI upper	p-value
ACME (control)	0.27	-0.45	1.17	0.422
ACME (treated)	0.62	0.14	1.23	0.008***
ADE (control)	2.68	0.33	5.13	0.030**
ADE (treated)	3.04	0.42	5.56	0.024**
Total Effect	3.31	0.86	5.79	0.006***
Prop. Mediated (control)	0.08	-0.18	0.58	0.428
Prop. Mediated (treated)	0.19	0.04	0.67	0.014**
ACME (average)	0.44	0.01	1.01	0.050**
ADE (average)	2.86	0.49	5.31	0.022**
Prop. Mediated (average)	0.13	-0.00	0.55	0.056*

Notes: Confidence intervals are obtained with nonparametric bootstrap using the percentile method. Sample size used 480. Simulations: 1000. * indicates significance at the 10% level, ** at the 5% level and *** at the 1% level.

violated to come to different conclusions about the causal mediation effect that we found. If important confounders are missing, the correlation ρ of the errors of the outcome and mediation model would be unequal to zero. The correlation ρ for our outcome- and mediator model is -3.738602e-17. Assuming that our specification of both models are correct, it seems there is no evidence for a violation of the assumptions. Results for potential departures of ρ and therefore violations of sequential ignorability are summarized in Figure E4.



(a) For ACME (control)

(b) For ACME (treated)

Figure E4: Sensitivity Analysis

We see that for the average causal mediation effect of *Chat* ($ACME_1(\rho)$), the range of ρ that leads to the same conclusions as those in Table E1 is quite large. The correlation ρ would need to be 0.2 to results in an mediation effect of zero.