Queen Bee Immigrant: The effects of status perceptions on immigration attitudes

Biljana Meiske*

October 30, 2022

Abstract

This work studies the dynamics of inter-minority relations and attempts to uncover the influence of the status position of established immigrants on their attitudes towards new waves of immigration. I hypothesize that relative status deprivation, that is, the degree to which own in-group is ranked low in the ethnic status hierarchy of the host country, has a negative impact on group members' attitudes toward an even lower-ranked status group. In an online experiment (N=1,159), participants with immigration background residing in Germany randomly receive either a positive or a negative evaluation of their own ethnic/national in-group, as evaluated by a group of ethnic German participants, while keeping constant the evaluations of other immigrant groups. The results show that participants whose in-group received a negative evaluation are systematically less willing to donate to an organization supporting refugees and express more negative views of the refugees from the Middle East. Furthermore, it is shown that participants who received a negative evaluation of their own in-group expect the native majority participants to express significantly more critical views of the refugees and other low-status out-groups, indicating that the treatment effect on behavior could be moderated by its effect on perceived norms surrounding prejudice expression. Finally, the results show that treatment affects not only the privately held attitudes but also participants' willingness to publicly express them, as participants who hold critical views of the refugees disclose them more readily when under observation of the participants from the native majority if they received a negative (rather than positive) evaluation of their in-group.

JEL Codes: C90, J15, J71
Keywords: Immigration attitudes, Discrimination, Status

Conflicting Interests: None
Ethics Clearance: Approved by Ethics Commission of the Department of Economics at Ludwig Maximilian University, Munich

*Max Planck Institute for Tax Law and Public Finance, Munich biljana.meiske@tax.mpg.de

1 Introduction

The indications that Alternative für Deutschland, an Euro-sceptic right-wing party in Germany that based the core of its platform on opposing immigration, had reached higher electoral support in the 2017 federal election among the so-called Russian-speaking German community compared to the national average (Goerres et al. (2020)), attracted a lot of media attention in Germany. Indeed, this is seemingly counter-intuitive – why would groups who themselves have a history of immigration and are also largely perceived by natives as immigrants support anti-immigration platforms? This is, however, not a sole example of such inter-minority dynamics. Cases of negative immigration attitudes expressed by the groups of immigrants were also found, for example, in Switzerland (Strijbis and Polavieja (2018)), Belgium (Meeusen et al. (2019)) and Austria (Neuhold (2020)).

This paper studies the dynamics of inter-minority relations and attempts to uncover the influence of the minority group's status position in the host country on its members' attitudes towards other minorities. I hypothesize that relative status deprivation, that is, the negative difference in status between a given ethnic/national group and that of the native majority (or other, more favorably perceived minorities), has a negative impact on this group's members' attitudes toward an even lower ranked status group.

While a considerable body of scientific literature studies the attitudes of the majority population toward immigration (for a survey of this literature, see, e.g., Hainmueller and Hopkins (2014)), less attention is paid to the immigration attitudes of established immigrants and their determinants. In principle, factors as diverse as those that have been found to impact the immigration attitudes of the majority population, including shared cultural values and perceived economic or cultural threat, could affect the positions of established immigrants as well. Furthermore, cultural and political characteristics of the sending country, as well as prevailing socio-economic conditions in a given immigrant group, might also play a role in determining the immigration attitudes of its members. Notwithstanding the potential importance of these channels, this work proposes an additional perspective and attempts to uncover the implications of

own immigration experience, encountered acceptance, and assigned status in the host society on the current immigration attitudes of established immigrants.

To investigate this idea, I run a (preregistered) survey-experiment with a sample of participants with immigration background residing in Germany and experimentally vary the status of the participants' in-group. In a separate pre-study, a smaller group of participants from the majority population, that is, those with no immigration background, is asked to evaluate different immigrant groups (structured along the region of their origin) as contributing rather positively or rather negatively to "the socio-economic and cultural life in Germany". In the second and main part of the experiment, a sample of participants with immigration background ($N = 1,159$) is presented a subset of answers elicited in the first phase. Participants are randomly chosen to be presented a subset of answers that evaluates their in-group either positively or negatively, while holding the evaluation of two other out-groups constant. The statement was designed to deliver prejudiced evaluations of the three groups and manipulate the status position of participants' in-group in the fixed ethnic hierarchy. I investigate the effect of the randomly assigned evaluation of participant's own in-group on their expressed support for refugees from the Middle East, captured by the respondents' willingness to forgo some part of their experimental earnings in order to secure a donation to the United Nations High Commissioner for Refugees (UNHCR). I additionally collect several attitudinal measures of participants' position towards refugees. Obtained results provide support for the laid out hypothesis, as the measured support for the refugees is significantly lower among participants who received a negative (rather than positive) evaluation of their in-group from the native majority.

The tendency of individuals to classify themselves and others into in- and out-groups, as well as the competition for status is a well-documented and seemingly universal characteristic of human societies. Social Identity Theory (SIT) (Tajfel et al. (1979)), starting from the assumption that individuals strive to enhance their self-esteem, offers an explanation of inter-group dynamics in the presence of a group-based identity threat. According to SIT, the lower the status assigned to a

group, less can it contribute positively to its members' social identity. In order to cope with the identity threat, the members of such a group are predicted to engage in defensive strategies, by e.g., pursuing "individual mobility" whereby individuals attempt to disassociate from the group and join a more favorably evaluated group, or by engaging in "social creativity", whereby individuals recover self-esteem by focusing on comparison with an even lower status group and emphasizing the own group's positive distinction relative to this new basis of comparison. Whereas the availability of one or another strategy depends on the contextual factors, in the inter-ethnic context observed here, both proposed strategies of coping with identity threat predict a disassociation with an overarching category of immigrants on the part of established immigrants, and, in the case of the second one, even an outright rivalry with new-coming lower-status immigrants.

The mechanism studied in this work resembles the so-called Queen-Bee phenomenon. The term, as described in Ellemers et al. (2004), should designate women occupying positions in male-dominated environments, who express a gender bias against women in evaluating their lower ranked female subordinates, sometimes even more so than their male colleagues, while at the same time distancing themselves from their own gender by expressing masculine self-descriptions. Subsequent work in this literature (for review, see, e.g. Derks et al. (2016)) has relied on both social identity theory and the system justification theory (Jost (2019), Jost et al. (2003)) to argue that rather than being a behavioral trait specific to women, the Queen-Bee behavior is in itself a response to the gender bias and identity threat in the male dominated environments. Drawing a parallel with the question considered here, one might wonder if there exists a Queen-Bee-Immigrant phenomenon. That is, do the established immigrants respond to an environment sceptical toward immigrants by distancing themselves from the immigrant status and expressing negative bias toward new immigrants. While the Queen-Bee literature considers a bias of females toward other females, that is toward own in-group, reacting by being more suspicious of the other (immigrant) out-group should arguably be even less psychologically costly, and thus more likely strategy.

The obtained experimental results support these predictions. Participants who received a negative evaluation of the own in-group donated systematically less to the UNHCR, compared to the participants who received a positive evaluation. The difference in donations amounted to 4.7 euros, representing around 13% of the average donation ($p < 0.01$). This result is not explained by participants' demographic characteristics[1], region of residence (in Germany), or region of origin of established immigrants. Additionally, the collected post-treatment measures show that the treatment effect is not propagated through its effect on participants' mood, nor through affecting participants' general generosity in an immigration-unrelated context.

In the next step, I study perceived descriptive norms surrounding expression of prejudice as a potential channel that might help to explain the observed treatment effect. Previous works on the emergence of social norms show that individuals, at least in part, infer the group's descriptive norms (what others are doing) from other individuals' behavior to which they are incidentally exposed. In particular, in situations where the behavior of interest does not produce an easily observable outcome (such as litter in public space), people combine summaries of group's behavior (e.g., election outcomes), with the direct experiences that they make to learn the descriptive norm regarding this behavior (Kwan et al. (2015), Kashima et al. (2013)). It is thus possible that the groups that were socialized in the presence of a steep ethnic hierarchy and were themselves exposed to prejudiced treatment grow to perceive inter-ethnic competition and expression of prejudice downwards (i.e. against groups ranked in the status hierarchy lower than one's own group) as pervasive, and perhaps even legitimate social dynamics in the host society, and are more likely to apply it towards the lower ranked groups once they encounter them. In particular, I hypothesize that exposing established immigrants to negative prejudice, expressed by a (high-status) majority member, updates their perceived descriptive norm such that they perceive expressing negative prejudice towards low-status groups (but not high-status ones) as more frequent among the native majority.

[1]Individual demographic controls include age, gender, equivalent household income tertile and indication of tertiary education.

In order to test this prediction, I elicit participants' empirical beliefs regarding the percentage of the pre-study participants who negatively evaluated the impact of refugees from the Middle East on the socio-economic and cultural life in Germany. To test the prediction of the hypothesis that exposure to prejudice updates the norm surrounding expression of prejudice towards low-status groups (but not towards high-status ones), I additionally measure participants' expectations of the majority participants' evaluation of one other (in Germany) salient and one non-salient low-status immigrant out-group (immigrants from Turkey, and those from Southern Africa countries), as well as one high-status out-group (immigrants from western European countries). Experimental results provide support for the prediction. Participants exposed to a lower acceptance, that is, those who received negative evaluation of their own in-group expect significantly more negative evaluations of all low-status out-groups (but not of the high-status one) on the part of the majority population participants. Whereas the treatment effect on injunctive norms (what others believe one oath to do) is not explicitly tested here, the literature on social norms provides ample evidence for the role that descriptive norms alone play in shaping intentions and behaviors (Bicchieri and Xiao (2009), Krupka and Weber (2009), Bardsley and Sausgruber (2005)) in a wide range of behavioral domains, including expression of prejudice (Álvarez-Benjumea and Winter (2020)).

I additionally explore the role of the reciprocity preferences as a potential channel for the effect that providing different evaluations of the immigrants' in-group has on their support for the refugees. The upstream indirect reciprocity designates a tendency of individuals to exhibit prosocial behavior towards others because somebody else has exhibited prosocial behavior towards them (Alexander (1987), Nowak and Sigmund (2005)). I elicit participants' preferences for upstream indirect reciprocity in an extended dictator game and provide evidence for its effect in line with the theoretical prediction. Participants with a higher preference for reciprocity donated more and were more likely to make a positive donation if they were in positive treatment, and donated less (though insignificantly) and were less likely to make a positive donation if they were in the negative treatment.

Finally, experimental results show that exposure to prejudice affects not only the privately held attitudes towards refugees but also participants' willingness to publicly express them. Previous works studying how privately held opinions translate into publicly expressed attitudes and behaviors found that stigmatization and social desirability of certain beliefs play an important role in determining to which degree the discrepancy between the two emerges. In particular, individuals tend to bias their statements when publicly expressed towards positions deemed socially more appropriate (Bursztyn et al. (2018), Perez-Truglia and Cruces (2017), Enikolopov et al. (2020)), or those that are more typical of the group with which they identify (Janus (2010)). Furthermore, the work by Bursztyn et al. (2020) shows how public revelation of controversial preferences (such as xenophobic views) can impact the beliefs and behaviors of the spectators, leading them to be themselves more likely to express and less likely to condemn such attitudes. Therefore, understanding how preference falsification shapes expressed immigration attitudes among established immigrants is not only important as the observable positions might not match the privately held ones, but also because the readiness to publicly express a controversial attitude can be consequential in its own right.

I focus on one aspect of preference falsification and investigate whether established immigrants change their statements when their answers might be observed by a participant from majority population, and whether this tendency changes with the exposure to prejudice towards their in-group. Participants are asked to provide the answer to the question asking them to rate whether refugees "make Germany a better or a worse place to live" once privately, and once after being informed that a future participant, selected from a sample of majority population, might observe their answer along with the information regarding participant's region of origin. Comparing the answers provided in both settings reveals that participants indeed do answer differently when their answer is potentially observed, and the direction of misrepresentation depends largely on the initial, privately expressed preference. In particular, participants who provided a more critical assessment of the impact of refugees in Germany when

answering privately changed their answer towards expressing more supportive views in the observable setting. More interestingly, the opposite holds for the participants who privately assessed the impact of refugees highly positively, that is, they misrepresent their positions in the observable setting so as to appear more critical. Furthermore, among participants who were more critical in the private setting, those assigned to the Negative treatment misrepresent their attitudes in the observable setting (in the positive direction) systematically less than those in the Positive treatment, thus demonstrating the effect of prejudice exposure on the willingness to express a controversial position publicly.

This work contributes to the literature on the political preferences of immigrants (Dinas et al. (2021a), Strijbis and Polavieja (2018), Van der Zwan et al. (2017), Just and Anderson (2015), Dancygier and Saunders (2006)), and more specifically to the branch studying how political attitudes of the native majority shape these preferences (Dinas et al. (2021b), Fouka (2019), Kuo et al. (2017)). To the best of my knowledge, this is the first paper that provides causal evidence for the effect of status deprivation (through expressed prejudice) on immigration attitudes of the immigrant population. More generally, this work contributes to the broad literature on immigration attitudes and the drivers behind them (for survey, see Hainmueller and Hopkins (2014)). Finally, this paper also relates to the discussion on political correctness, by highlighting the negative externalities entailed by its absence in the inter-ethnical context (Braghieri (2021), Norton et al. (2006), Morris (2001)).

2 Experimental Design

The study is split into two phases, which will henceforth be referred to as the pre-study and the main experiment, both implemented as an online survey. In the following, I provide the description of both phases.

2.1 Pre-study

The pre-study was conducted with a small sample ($N = 125$) of participants residing in Germany and with no immigration background. The only purpose of the pre-study was to collect the responses from the majority population regarding their evaluations of different immigration groups that would later be used in the main experiment.

At the beginning of the survey, participants provided answers to a set of basic demographic questions, including participant's gender and age, alongside own and parental country (countries) of birth, which were used to ensure that only participants from the majority population with no immigration background, participate in the pre-study.

Thereafter, for each of the several regions/countries, participants were asked to evaluate whether people immigrating from the given region/country contribute rather positively or rather negatively to the "socio-economic and cultural life in Germany" (participants selected one of the two options as an answer). To avoid confusion in terms of which countries are encompassed by a given region, with each question participants were shown a simple political map of the relevant part of the world, with the region of interest visibly highlighted, and the text of the question explicitly listed all corresponding countries. Participants in the pre-study were paid a fixed participation fee upon completion of the survey.

2.2 Main experiment

The main part of the experiment was conducted with a sample of 1.159 participants with immigration background residing in Germany.

Demographics As in the pre-study, at the beginning of the session,

participants answered the questions regarding their basic demographic characteristics, including participants' own and parental country of birth. This information was used to match participants to one of the eleven regions or origin[2].

Treatment provision In this part of the experiment, participants are told that, in a study that took place at an earlier point of time, a group of 125 participants from Germany with no immigration background were asked to evaluate the impact of various immigrant groups on socio-economic and cultural life in Germany, and that some of the collected answers will be shown to them. Participants are then (conditional on the region that they were matched to) randomly split into two treatments. Participants in both treatments are presented with one evaluation of each of the three immigrant groups - one representing immigrants stemming from their own (parental) region of origin, and the other two representing two out-groups. In both treatments, the answers from the pre-study are selected so that one out-group (in both treatments: immigrants from western EU countries) is always evaluated positively and the other (in both treatments: immigrants from Lebanon) negatively. Here, the positive and negative evaluations refer to the group being evaluated as "contributing rather positively", and respectively as "contributing rather negatively", to the socio-economic and cultural life in Germany. The only difference between the treatments is the evaluation of the own in-group. In the **Positive treatment**, participants are shown an answer that evaluates the impact of the own in-group positively, whereas in the **Negative treatment**, participants are shown an answer that negatively evaluates the impact of the

[2]The eligible regions of origin in this study included: Countries in central-eastern European Union (Czech Republic, Slovakia, Poland, Hungary); Romania and Bulgaria; Baltic states (Estonia, Latvia, Lithuania); Countries of ex-Yugoslavia (Bosnia and Herzegovina, Croatia, Kosovo, Montenegro, North Macedonia, Serbia, Slovenia); North Africa (Morocco, Algeria, Libya, Tunisia, and Egypt); Southern European Union countries (Greece, Italy, Portugal, Spain, Cyprus, and Malta); Turkey; Southern countries of the ex-Soviet Union (Tajikistan, Turkmenistan, Georgia, Kazakhstan, Kyrgyzstan, Armenia, and Azerbaijan); Western countries of the ex-Soviet Union (Ukraine, Moldova, Belarus); Russia; and Albania. The division was made with the aim of including the regions of origin most frequently encountered among the population with immigration background in Germany. At the same time, the division attempted to achieve a trade-off between the number of regions and a sufficiently narrow definition of a region so as to allow for successful clustering.

own in-group. Figure 1, provides an example of evaluations presented to participants for both Positive and Negative treatment.

Including the two out-groups, consistently evaluated positively and negatively, ensures that provided information cannot be interpreted as a more or less positive attitude towards immigration in general, and instead ties the treatment variation to the position of the own in-group in an already set hierarchy.

Elicitation of attitudes towards refugees In this part of the study, two measures of participants' support of refugees were elicited. Following the approach of Dinas et al. (2021a), an attitudinal measure of support was constructed by collecting participants' answers to a set of seven questions. Participants provided their views (among others) on whether Germany should increase or decrease the number of people it grants asylum to, refugees' influence on the labor market, the welfare state, probability of a terrorist attack, criminality, etc. The list of all questions is provided in the Appendix A.1.

The main behavioral measure of participants' support for refugees was captured by the willingness to donate to the United Nations High Commissioner for Refugees (UNHCR). Participants were informed that, as a part of the study, a lottery would be administered whereby one randomly selected participant will be awarded 100 euros and all participants have the same chance of winning the prize. They are then asked whether they would like to donate some part of the 100 euros prize, in the case that they win the lottery, to the UNHCR, and if so, how much. Participants are informed that if they decide to dedicate some amount to refugees-support, this amount will be automatically deducted from their 100 euro prize in the case they win, and a donation in the same value will be made to an organization supporting refugees.

Mood elicitation In order to be able to control for the treatments' potential effect on participants' mood, a measure of mood is elicited via Self-Assessment Manikin questionnaire (Bradley and Lang (1994)). Three questions, intended to capture three major affective dimensions - pleasure, arousal, and dominance - asked participants to select one of the five offered manikins that they felt best describes their mood.

Empirical expectations In order to study treatment effects on participants perceived descriptive norms regarding prejudice expression, in this part of the experiment, participants were asked to guess what percentage of the 125 participants without immigration background that took part in the pre-study evaluated negatively each of several immigrant groups (categorized by their region/country of origin). Each participant was asked to guess the share of participants from the pre-study who negatively evaluated the impact of people immigrating to Germany from: participant's own (parental) region of origin, western countries of the European Union (Austria, Belgium, France, Ireland, Luxemburg, Netherlands), Lebanon, Turkey, countries of southern Africa (South African Republic, Namibia, Eswatini and Lesotho) and that of refugees immigrating from the Middle East (Syria, Iraq, Afghanistan, and Pakistan). Countries within a given region were visibly displayed to participants. Participants were informed that the answer closest to the true collected values would be rewarded by additional 25 euros.

Indirect upstream reciprocity One potential driver of treatment effects on expressed support for refugees might be the preference for upstream indirect reciprocity, that is, the tendency of individuals to exhibit prosocial (antisocial) behavior towards others because somebody else has exhibited prosocial (antisocial) behavior towards them. To facilitate studying this mechanism, in this part of the experiment, a measure of indirect upstream reciprocity was collected using an extended dictator game with three players. Each Participant is assigned one of the three roles: player A, player B, or player C. Thereby, player A is given a budget of 30 euros, out of which they can send a certain sum to another player B, who in turn can send some of the received amount to player C. The amount sent by player A is multiplied by a factor f, and the resulting amount is paid to player B. Player A and player B known that the multiplication factor can take either a high value ($f = 4$) or a low value ($f = 2$), but the realization of this value is not know to any of the players. Thus, player B observes only the resulting sum they received but is not aware whether it resulted from player A sending a higher sum that was multiplied by a low factor value, or from player A sending a lower sum

that was multiplied by a high factor value. Here, player A could select between sending 0, 8, 16, and all 30 euros. All participants assigned to role B received a total of 32 euros (corresponding to player A sending either 8 or 16 euros, and the factor being equal to either 4 or 2, respectively).

Player B is then asked to decide for both scenarios how much of the received sum they would like to send to person C. To ensure that welfare concerns do not play a role in the decision of player B, the amount sent to player C is paid to them without multiplication. Participants are informed that at the end of the study, one triplet will be selected and paid out the amounts according to the decisions they made. Most of the participants were assigned the role of player B ($n = 1164$), and the rest was distributed among the other two roles.

I take the difference in amount sent to player C in scenario where player A was more generous versus that when they were less generous as a measure of indirect upstream reciprocity of player B.

Preference falsification When individuals' are asked to state their political views while observed by the others, preference falsification might mask truly held preferences and skew them to the perceived socially appropriate positions. This part of the experiment has the aim to capture a potential difference in attitudes expressed by established immigrants when they expect these attitudes to be observed by a majority population, as compared to when this is not the case.

In this part, participants are reminded that all previously provided answers will be delivered only to the researchers in anonymized form. The participants are then informed that only in this part of the experiment they are asked to provide an answer that can be used in a potential future study to inform future participants about their views on immigration. Furthermore, the instruction clarifies that, if the future study is conducted, it will be run in Germany with a sample of German citizens and that the recipient of their answer would know their country (countries) of origin. Thereafter participants fill out the answer to the question "Is Germany made a worse or a better place to live by refugees who are granted asylum in Germany?", that was already asked as one of the attitudinal questions in the "Elicitation of attitudes" phase.

Additional demographics and debriefing At the end of the experiment, participants are shown the true percentages of participants in the pre-study who negatively evaluated each of the several groups. The session ended after collecting some additional basic demographic information.

2.3 Data and sample description

The study was conducted in the period December 2021 to January 2022. The sample for the pre-study involved 125 adult individuals with residence in Germany and with no immigration background. A participant was considered to have an immigration background if they, or at least one of their parents, was born outside of Germany. For the purposes of the main-experiment, a separate sample was recruited involving 1,175 adult individuals with residence in Germany and with an immigration background. Out of this number, 16 participants provided inconsistent answers to basic demographic questions (stated unreasonable age), and their answers were removed, resulting in a sample of 1,159 participants.

Participants with an immigration background were matched to what I will be for simplicity referring to as "region of origin", indicating one of the eleven regions encompassing their, or parental, country of birth. The regions selected to be targeted in this study encompassed all countries within Europe (except for the Western European countries), all Ex-Soviet countries, Turkey, and five northern African countries (Egypt, Tunis, Morocco, Algeria, and Libya). Table 1 in the Appendix provides an overview of all regions (and all encompassed countries), along with the share of participants matched to each region. The selection of the eligible regions attempted to match the studied sample with the groups most represented among the population with an immigration background in Germany[3], and to focus on those immigrant groups that are more likely to occupy a lower status position in German society (thus the exclusion of the Western European countries). Table 2 in the Appendix presents the descriptive statistics of the sample across both treatments. The online survey was

[3]See Statistical Office of Germany (Genesis-Online Database, code: 12211-0202)

programmed in Qualtrics and the distribution of the link to the experiment was delegated to a panel company CINT[4].

In the next section, I provide the overview of empirical results and test the following (pre-registered) hypotheses.

Hypothesis 1 Being assigned to the Negative treatment leads to a decrease in the amount donated to UNHCR and a more negative evaluation of refugees as measured by the attitudinal questions.

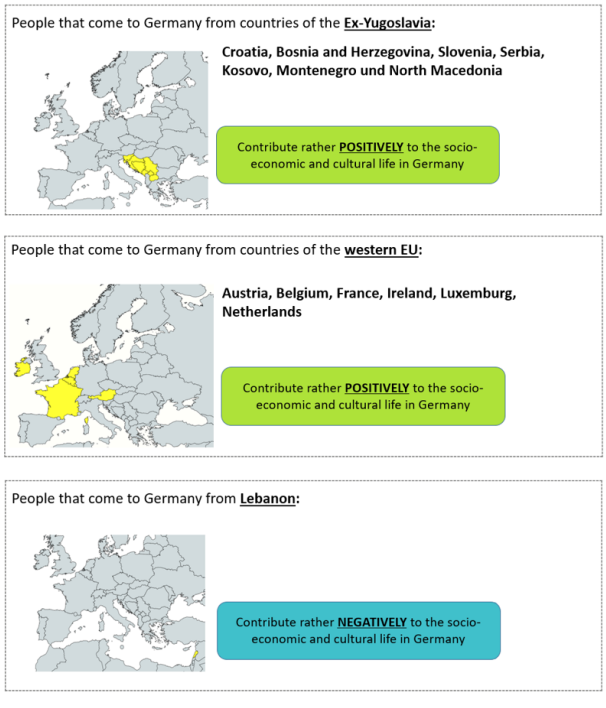
Hypothesis 2 Being assigned to the Negative treatment leads participants to expect a higher percentage of negative evaluations of refugees' impact on socio-economic and cultural life in Germany among majority participants (in the pre-study). Furthermore, assignment to the Negative status treatment leads participants to expect a higher percentage of negative evaluation of the own in-group, as well as of the other low-status groups (but not the high-status ones) among majority participants.

Hypothesis 3 Participants with higher indirect reciprocity react more strongly to treatment variation, that is, express more negative (positive) evaluations of refugees in the Negative (Positive) treatment.

Hypothesis 4 The distribution of answers provided to the question "*Do refugees who obtain asylum right in Germany make Germany a worse or a better place to live*" in "private" scenario differs from the distribution of answers provided to the same question in "observable" scenario. Furthermore, being assigned to the Negative treatment leads participants to express a less favorable opinion of refugees in "observable" scenario.

[4]https://www.cint.com/

(a) Positive treatment



(b) Negative treatment

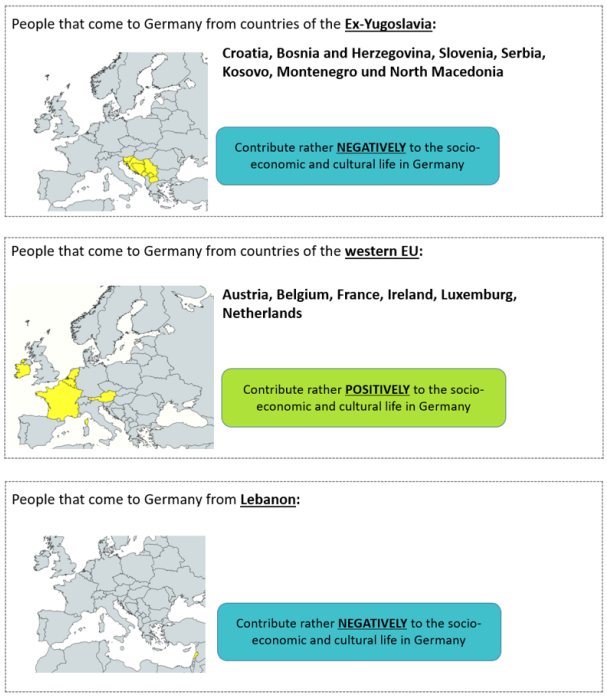


Figure 1. Treatment provision - example

The figure depicts an example of a screen that a participant, who was matched to the region of Ex-Yugoslavia, would see in the treatment provision phase if they were allocated to the Positive treatment (panel), and that if they were allocated to the Negative treatment (panel b). Participants are informed that they would see a subset of answers collected in the pre-study. Treatment variation is based on randomly matching participants to an answer from the pre-study evaluating participant's own (parental) region of origin either positively or negatively while keeping the evaluations of the other two out-groups constant.

3 Results

3.1 Pledged donation to the UNHCR

In this subsection, I present the measured effect of the treatment, that is, the effect of receiving negative status information, compared to receiving positive status information, on the behavioral measure of participants' support for refugees. The measure of support is captured by the amount that participants committed to donate to the United Nations High Commissioner for Refugees (UNHCR), from a 100-euro prize that is raffled among all participants at the end of the study. On average, participants committed to donate 36.86 euros, with individual decisions spanning across the full range of possible donations. Figure 2 provides an overview of the observed distribution of the pledged donations.

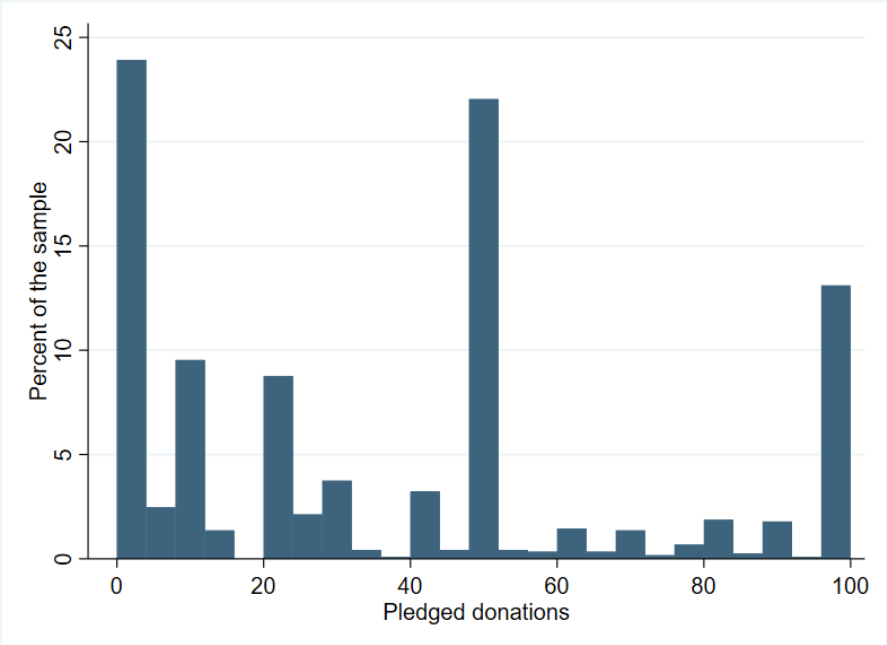


Figure 2. Distribution of pledged donations to the UNHCR

The results presented in Table 1 depict the effect of being allocated to the Negative treatment (with Positive treatment serving as a baseline) on the pledged donations. Considering that the possible value of the donation was limited at 0 from below, and at 100 from above, and that the number of participants who selected both limiting values was significant, the table presents the results of Tobit regression of the donated amount on treatment variable and

individual controls[5]. All presented regressions include fixed effects of the federal state within Germany and region of origin, and the standard errors are clustered on the level of participants' region of origin.

Table 1. Treatment effects: Pledged donation to the UNHCR

	(1)	(2)	(3)	(4)
	Pledged donation		Pr(Donation>0)	
Negative treatment	-7.049***	-6.922***	-0.189***	-0.190***
	(1.593)	(1.532)	(0.063)	(0.060)
Constant	47.418***	54.824***	1.031***	1.400***
	(3.082)	(4.608)	(0.152)	(0.212)
Marginal effects: $E(\Delta y/\Delta x)$				
Negative treatment	-4.716***	-4.630***	-0.054**	-0.054**
	(0.000)	(0.000)	(0.003)	(0.001)
Individual controls	No	Yes	No	Yes
Observations	1,159	1,159	1,159	1,159

Notes: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.
Column (1) and column (2) show Tobit regression of amount dedicated to donate to the UNHCR on treatment variable and the set of individual controls. Negative treatment indicates receiving negative status information regarding own in-group (with Positive treatment serving as a baseline). Reported marginal effects represent the average marginal effect of being allocated to Negative treatment on donated amount. Columns (3) and (4) show Probit regression of of an indicator variable for donation being larger than zero on treatment variable and the set of individual controls. All regressions include fixed effects of the federal state of residence in Germany and region of participants' (parental) origin. Individual controls (included in columns (2) and (4)) include age, gender, equivalent household income tertile and indication of tertiary education. Reported marginal effects represent the average marginal effect of being allocated to Negative treatment on probability of making a positive donation, and can be directly interpreted in terms of percentage points difference. Standard errors in parentheses are clustered on the level of region of participants' (parental) origin.

The results in Table 1 provide support for the Hypothesis 1. The results shown in column (1) demonstrate that participants in the Negative treatment commited to donate systematically less to the UNHCR. Participants pledged on average around 4.7 euros less to donation if they were in the Negative treatment ($p < 0.01$), which represents around 12% of the average committed sum. Furthermore, as shown in column (3), participants allocated to the Negative treatment were significantly less likely to pledge any positive donation relative to

[5]The OLS analysis produces qualitatively same results and is depicted in Table 4 in Appendix A.4.

those in the Positive treatment. In particular, reallocating a participant from Positive to Negative treatment decreased, on average, the probability of the participant pledging a positive donation by 5.4 percentage points ($p < 0.01$). The results in columns (2) and (4) show that these findings are robust to the inclusion of controls for the respondents' socio-demographic background.

The collected measure of participants' mood allows to check whether the treatment variation affected the behavior through its effect on participants' mood. However, the distribution of all three measured affective dimensions - pleasure, arousal, and dominance, elicited via Self-Assessment Manikin questionnaire (Bradley and Lang (1994)), did not differ significantly between the two treatments (Kolmogorov–Smirnov test for equality of distribution in both treatments, for each of the three affective dimensions - pleasure: $p > 0.6$; arousal: $p > 0.9$; dominance: $p > 0.9$). Furthermore, Table 3 in Appendix A.3 shows that, while valence and arousal had a positive effect on willingness to donate, controlling for these measures does not qualitatively alter the observed effect of the treatment.

Furthermore, as explained in footnote 8, treatment variation does not have any effect on participants' generosity in a context unrelated to topics of immigration (as captured by giving in a dictator game). This result provides evidence that the findings depicted in Table 1 can not be explained by changes in participants' broader generosity.

3.2 Attitudinal measures

In addition to the behavioral measure of support for refugees, a set of attitudinal measures was elicited by means of collecting answers to seven questions regarding refugees from Syria, Afghanistan, Iraq, and Pakistan who flee to Germany. The questions, among others, regraded participants' views of the influence of refugees on employment, risk of terrorism, criminality. The exact formulation of all seven questions is provided in Appendix A.1.

Compared to treatment effect on pledged donations, treatment had a smaller effect on the attitudes reported in the seven questions. The first six columns of

Table 2. Treatment effects: Attitudinal questions

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	q1	q2	q3	q4	q5	q6	q7	$\bar{q}$
Negative treatment	-0.266**	-0.082	-0.051	-0.031	0.043	-0.144	-0.049	-0.107
	(0.109)	(0.105)	(0.122)	(0.111)	(0.126)	(0.089)	(0.040)	(0.105)
Constant							1.082***	2.120***
							(0.174)	(0.178)
Individual controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	1,159	1,159	1,159	1,159	1,159	1,159	1,159	1,149

Notes: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.
Columns (1) through (6) show the results of ordered logistic regression of provided answer on treatment variable and the set of individual controls. Column (7) shows the result of Probit regression of dummy variable that takes value 1 if a participant selected "To flee war" or "Avoid political persecution" as primary reason why refugees leave their countries, and 0 otherwise. All regressions include fixed effects of the federal state of residence in Germany and region of participants' (parental) origin. Individual controls include age, gender, equivalent household income tertile and indication of tertiary education. Standard errors in parentheses are clustered on the level of region of participants' (parental) origin.

Table 2 show the results of ordered logistic regression of chosen answer for each of the (first six) questions on treatment variable and the set of socio-demographic controls. All answers are re-coded such that a higher value indicates higher support for refugees. Column (1) shows that in the case of the first question, which asked the participants' opinion on whether Germany should increase or decrease the number of people it grants the asylum to, participants were significantly more likely to provide a lower answer (decrease number of granted asylums) if they were in the Negative treatment. However, although treatment effects work in the predicted direction in most of the other questions (that is, participants in the Negative treatment provided less supportive answers), these effects are not significant.

Question q7 asked participants to provide their opinion on the primary reason why refugees abandon their countries among the following options: "To flee war", "Avoid political persecution", "Improve their economic conditions" and "Obtain access to social security payments in the destination country". I construct a dummy variable that takes value one if a participant selected one of the first two choices and show in column (7) the results of Probit regression of

this variable. Again here, being assigned to the negative treatment decreased the probability of selecting one of the two reasons that would indicate security (rather than economic) concerns as a primary reason for flight, but the effect is insignificant.

Finally, I construct an aggregate measure of participants answers to attitudinal questions by averaging for each participant seven dummy variables. The dummy variables correspond to the seven questions and each takes value one if participant selected an answer to the respective question that indicates higher support for refugees than that implied by the neutral point (selected 3 (5) on a scale 1 to 5 (0 to 10)). Column (8) shows the results of regressing this aggregate measure, denoted by $\bar{q}$, on the treatment variable and the set of individual controls.

3.3 Empirical expectations - differential evaluation based on origin

Results in the previous section showed that providing participants with negative evaluative opinion on immigrants from their own (parental) region of origin, expressed by a member of majority population, led them to significantly decrease their support for refugees. One possible explanation for this regularity might be that participants, who face differential acceptance by majority population based on their origin, might internalize this behavior as usual and perhaps legitimate in the society more generally. In other words, people from low-status regions could learn from discrimination directed towards their own in-group that discriminating downwards (i.e. against groups ranked lower than one's own group) is widespread and possibly also acceptable behavior, in the host society. As proposed by Hypothesis 2, in the context observed here this would suggest that observing lower acceptance of the own (lower-status) in-group, might update downwards participants' empirical expectations of acceptance of other lower-status groups, such as refugees, among majority population.

In order to test this prediction, I collected an incentivized measure of empirical expectations on approval of different immigrant groups by the majority population. After collecting the main outcomes of interest, participants were asked to guess the share of respondents in the pre-study (without migration background) who

evaluated *negatively* the impact of each of several immigrant groups on socio-economic and cultural life in Germany. Particularly, each participant was asked to guess the share of participants from the pre-study who negatively evaluated the impact of people immigrating to Germany from: participant's own (parental) region of origin, western countries of European Union, Lebanon, Turkey, countries of southern Africa and that of refugees immigrating from the Middle East. To avoid confusion, in cases where the evaluation regarded people immigrating from a given region, all countries within the region were listed. The exact phrasing of the question and an example screen seen by participants is provided in the instructions available in online Appendix[6].

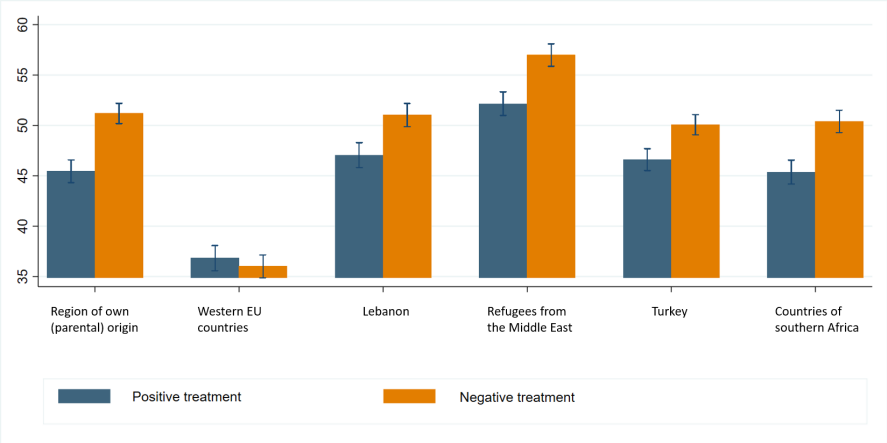


Figure 3. Treatment effect on perceived descriptive norm

The figure depicts the average elicited guesses of the share of majority population participants who evaluated negatively the influence on socio-economic and cultural life of people immigrating to Germany from countries/regions depicted on x-axis, by treatment. The vertical lines indicate the 95% confidence intervals.

Figure 3 provides an overview of measured treatment effects on collected empirical expectations. The first pair of bars on the left shows that participants who received a negative evaluation on their own in-group, on average, expected the majority population participants to be more critical towards immigrants from their region of origin. This is also intuitive, as it reflects the information that participants received in treatment provision, but is still informative as it shows that participants extrapolated from the individual evaluation that they

[6]Full survey instructions are available at http://biljanameiske.com/wp-content/uploads/2022/10/Queen-Bee-Immigrant_Instructions_ENG.pdf

received to the average opinion of the group. At the same time, it serves to confirm the successful treatment manipulation.

More interestingly, the same applies to participants' expectations of evaluations of all other low-status immigrant groups. Particularly, in accordance with Hypothesis 2, participants in the Negative treatment expected a significantly more negative evaluation of the impact of refugees from the Middle East, as well as of people immigrating from Turkey, Lebanon and from countries in the south of Africa. This is not the case for the expected evaluation of high-status immigrants, that is, those coming to Germany from the western EU countries, indicating that this is not a consequence of expecting the majority population to be more sceptical towards immigrants in general. Instead, as proposed by Hypothesis 2, it appears that receiving a negative evaluation of the own in-group led participants to expect more critical views only of those immigrant groups that were of a lower status than those who are evaluating.

I investigate these observations more formally in Table 3, which provides the results of OLS regression of the participants' estimates of the share of the majority population participants who evaluated each of the mentioned immigrant groups negatively. Confirming the indications provided by Figure 3, regression results show that participants in the Negative treatment (relative to those in the Positive treatment) expected significantly more negative evaluations of the impact of refugees, as well as of all low-status immigrant groups, but not of the high-status one. The results are significant and are not explained by participants' socio-demographic characteristics.

Whereas the main outcome of interest here was a spillover effect on the participants' empirical expectations regarding refugees from the Middle East, it is particularly interesting to note that the spillover affected not only a salient unrelated minority (Turkish), but also a very non-salient group of immigrants from countries in the south of Africa, who are effectively barely present in Germany, both as a share of population[7] and in the public discourse. The

[7]According to the data of Federal Statistical Office of Germany (Genesis-Online Database, code: 12521-0002), at the end of 2020, the number of people residing in Germany with citizenship of one of these four countries is below 8500 persons.

Table 3. Treatment effects: Empirical expectations

Elicited expectation:	What percentage of majority population participants evaluated negatively the impact of people coming to Germany from:					
	(1)	(2)	(3)	(4)	(5)	(6)
	Own (parental) region of origin	Refugees	Turkey	Lebanon	Southern Africa	Western EU countries
Negative treatment	5.863***	4.921**	3.524**	4.184*	5.112*	-1.463
	(1.118)	(2.187)	(1.161)	(2.021)	(2.444)	(1.330)
Constant	50.237***	53.229***	50.914***	38.812***	37.494***	41.056***
	(3.080)	(4.129)	(4.560)	(4.277)	(3.734)	(2.375)
Individual controls	Yes	Yes	Yes	Yes	Yes	Yes
Observations	1,159	1,159	897	1,159	1,159	1,159

Notes: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Robust standard errors in parentheses. OLS regression of the elicited guess (of participants with migration background) of the share of majority population participants who evaluated as negative the imact on socio-economic and cultural life in Germany of people immigrating to Germany from countries/regions shown in columns' headers. First column regards the people who immigrate to Germany from country/region of participant's origin (or that of their parent(s) if the participant was born in Germany). In questions that regarded immigrants from a region (rather than a country) all countries within the region were listed in the question. All regressions include fixed effects of the federal state of residence in Germany and region of participants' (parental) origin. Standard errors are clustered on the level of region of participants' (parental) origin. Individual controls include age, gender, equivalent household income tertile and indication of tertiary education.

question regarding immigrants from this region explicitly specified the countries in question (South Africa, Namibia, Eswatini, and Lesotho), thus this effect can not be the consequence of mistaking this region for other regions/countries in Africa where some percentage of the refugees came from (e.g. Eritrea). On the other hand, considering that some refugees indeed did come from some Sub-Saharan countries, it might be that the spillover effect from own in-group evaluation to the evaluation of the refugees, further spilled-over to any group that remotely resembles this group, even if the only commonality between the groups is the same continent of origin. It, therefore, illustrates how social dynamics, completely unrelated with the characteristics of the immigrant group in question, can pre-set the stage and shape attitudes towards this group, even before it is present in the host country.

While the literature on social norms provides evidence of an impact of

empirical expectations (what others are doing) on normative expectations (what others believe one ought to do) (see e.g. Bicchieri et al. (2020a), Bicchieri et al. (2020b)), the results provided here can only support the treatment effect on the former. Therefore, whether the experience of being deferentially evaluated based on the place of origin shapes as well the perceived appropriateness of such behavior remains an interesting open question. On the other hand, irrespective of their influence on normative expectations, empirical expectations have been shown to influence behavior in a wide range of domains (Bicchieri and Xiao (2009), Krupka and Weber (2009), Bardsley and Sausgruber (2005)).

Whereas the ultimate test for the behavioral effects of the empirical expectations would amount to administering a norm-manipulation experiment and is thus outside the scope of this work, these results can be taken as tentative evidence for the mediating role of expectations.

3.4 Indirect reciprocity

Another reason behind the effect that receiving evaluation on own (parental) region of origin had on support for refugees might be the upstream indirect reciprocity. Upstream indirect reciprocity designates a tendency of individuals to exhibit prosocial behaviour towards others because somebody else has exhibited prosocial behaviour towards them (Alexander (1987), Nowak and Sigmund (2005)). Previous studies have provided evidence for the upstream indirect reciprocity, both in the laboratory (Greiner and Levati (2005)) and in the field experiments (Mujcic and Leibbrandt (2018)). Exhibiting upstream indirect reciprocity in the context of inter-minorities relations would suggest that receiving a less (more) favorable evaluation from an out-group might translate into a less (more) favorable view of another out-group. Thus we would expect more reciprocal participants to react more negatively (positively) in terms of their support for refugees if they were assigned to the Negative treatment (Positive treatment).

In order to test this prediction, a measure of indirect upstream reciprocity was collected using an extended dictator game, whereby one participant (player A)

can send a certain sum to another participant (player B), who in turn can send some share of the received amount to a third participant (player C). The amount sent by participant A is multiplied by a factor, which can take either a high or a low value, but the realization of this value is not known to any of the players. Thus, player B observes only the resulting sum they received but is not aware whether it resulted from player A sending a higher sum that was multiplied by a low factor value, or from player A sending a lower sum that was multiplied by a high factor value. Player B is then asked to decide for both scenarios how much of the received sum they would like to send to person C. To ensure that welfare concerns do not play a role in the decision of player B, the amount sent to player C is paid to them without multiplication. Each participant is matched to one of the three roles, and a randomly selected triplet is paid out the amounts according to the decisions they made.

I take the difference in the amount sent to player C in the scenario where player A was more generous versus that when they were less generous as a measure of indirect upstream reciprocity of player B. In order to collect this measure for as many participants as possible, most of the participants were assigned the role of player B ($n = 1150$), and the rest was distributed among the other two roles. All participants assigned to role B received a total of 32 euros (corresponding to player A sending either 8 or 16 euros, and the factor being equal to either 4 or 2, respectively).[8] On average, participants sent 1.21 euros more to player C when player A sent them a higher amount compared to when they sent a lower amount (average amounts sent in two cases was 13.99 and 12.79 euros). This difference is significant (Wilcoxon signed-rank test: $z = 9.544$, $p < 0.001$), providing evidence for behavior consistent with indirect upstream reciprocity. Furthermore, the distribution of the measure of indirect reciprocity does not differ among treatments (Kolmogorov–Smirnov test: $p = 0.785$), supporting the view of reciprocity as a basic preference.

[8]There is no difference between treatments in sending to player C in any of the two scenarios (both-sided t-test: $p > 0.5$ in both cases). As the extended dictator game captures readiness to share income but in a context unrelated to questions of immigration, this result provides evidence that the effect of treatment variation on pledged donations is not driven by its effect on participants' generosity more generally.

Table 4. The role of upstream indirect reciprocity

	(1)	(2)
	Donation	Pr(Donation>0)
Negative treatment	-5.815***	-0.139*
	(1.557)	(0.073)
Ind. reciprocity	0.433	0.009
	(0.267)	(0.006)
Negative treatment*Ind. reciprocity	-0.751*	-0.033**
	(0.446)	(0.015)
Constant	46.482***	1.002***
	(3.089)	(0.152)
Individual controls	No	No
Observations	1,150	1,150

Notes: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Robust standard errors in parentheses. Column (1) shows Tobit regression of amount dedicated to donate to the UNHCR on the treatment variable, measure of upstream indirect reciprocity (denoted Ind. reciprocity) and their interaction. Column (2) shows the results of Probit regression of the dummy variable that takes value one if participant pledged to donate a value larger than zero on the same set of regressors. All regressions include fixed effects of the federal state of residence in Germany and region of participants' (parental) region of origin. Standard errors are clustered on the level of participants' (parental) region of origin.

Table 4 provides the results of the regression of donated amount and that of the dummy variable indicating that participant made a positive donation on the measure of treatment variable, indirect reciprocity and their interaction. The results indicate that indirect reciprocity indeed had some role in determining the decision to donate. Whereas indirect reciprocity in the positive treatment increased, albeit insignificantly, the pledged donation ($coef. = 0.433$, $p = 0.105$) and the probability to donate ($coef. = 0.009$, $p = 0.167$), it significantly reduced both values in the negative treatment. However, although providing some evidence for the role of indirect upstream reciprocity, these effects are relatively small and do not provide a systematic explanation of the found treatment effects (the treatment variable remains significant).

3.5 Preference Falsification

Previous subsections aimed at describing how exposure to expressed prejudice shapes immigration attitudes of individuals with immigration background when

these attitudes are expressed privately, that is, when they are unobservable to others (other than the experimenter). However, a broad range of political behaviors, such as protesting, signing a petition, or publically expressing political views, are per construction observable to other members of the polity, and as such are susceptible to social effects. In particular, due to perceived social pressure, individuals with counter-normative views may prefer to falsify them under observation (Kuran (1997)), such that expressed preferences might not always fully match privately held ones. Previous empirical works convincingly demonstrate that individuals care for how they are perceived by others, and that reputational concerns consequently shape observable behavior in a variety of settings, including political behavior (Valentim (2022), Bursztyn et al. (2020), Enikolopov et al. (2020), DellaVigna et al. (2016), Gerber et al. (2008)).

Provided that the final behavior exhibited by individuals is not only shaped by their individual preferences, but also by the social effects, it is worthwhile exploring the effect of the latter on the expressed immigration attitudes of the established immigrants. Understanding how perceived social pressure in the host society might impact expressed immigration attitudes of established immigrants is important not only because preference falsification might mask their genuine preferences but also in light of the findings that expressed controversial preferences, such as xenophobia, might have far-reaching spill-over effects on the beliefs and behaviors of others who observe them, and in the extreme might even lead to unraveling of norms that protected against them (Bursztyn et al. (2020), Bursztyn et al. (2018)). Thus, understanding factors that facilitate public expression of controversial preferences conditional on their existence is important in its own right.

In the context observed here, I focus on one possible channel of preference falsification and study whether established immigrants change expressed preferences towards refugees when these preferences are potentially observable by the members of the native majority. Furthermore, I analyze whether being exposed to expression of prejudice towards own in-group has an effect on the size and direction of preference falsification. Whereas most previous works provided

evidence of preference falsification in settings where individuals face strongly established norms on pro-social behavior, the context observed here is further complicated by the fact that questions of immigration and asylum policies proved to be highly polarizing among majority population in (among others) Germany. As social polarization blurs the social consensus on desirable behavior, it is not clear in which direction (if at all) established immigrants might skew their expressed preferences when observed by majority population.

To get some insight into this, one of the questions that was used in collecting attitudinal measure of support for refugees (q6) was asked again later in the survey, but participants were this time informed that their answer might (or might not) be shown to a participant in a future study. Participants knew that, if used, their response would be provided to a future participant in anonymized form, along with the indication of whether the participant has migration background, and if so from which countries, and that the person observing their answer would be a German citizen. The question asked participants to rate whether refugees who obtain asylum right in Germany make Germany a worse or a better place to live. Participants answered by selecting a number on an 11-points number line, where 0 was indicated as "worse place to live", and 10 as "better place to live". Note that participants were given the opportunity to provide a neutral answer by selecting 5 on the number line, which is exactly in the middle between the two extremes.

I denote the two scenarios as "private" and "observable"[9], and the answers provided in both scenarios by a_p and a_o respectively (note that higher answer indicates a more supportive attitude towards refugees). To compare the answers provided in the two scenarios, I construct a variable $\Delta_o = a_o - a_p$, capturing the extra support that participants expressed in the observable scenario relative to that in the private scenario.[10]

Figure 4 shows the average value of Δ_o over a_p. The figure indicates that the

[9]The use of the terms "observable" and "private" here is intended only to designate and make easier the distinction between the two scenarios. The ability of the researchers to observe participants' answers renders the private setting clearly distinct from a truly private setting.

[10]I argue that calculating a difference in this case is appropriate as the question explicitly asked the participants to rate refugees' influence on a visibly enumerated line, with only end points carrying the (exactly opposite) labels. As the answer options are number values (rather than statements, as would be the case in standard Likert scale with different levels of agreement), collected answers can be considered to be interval data.

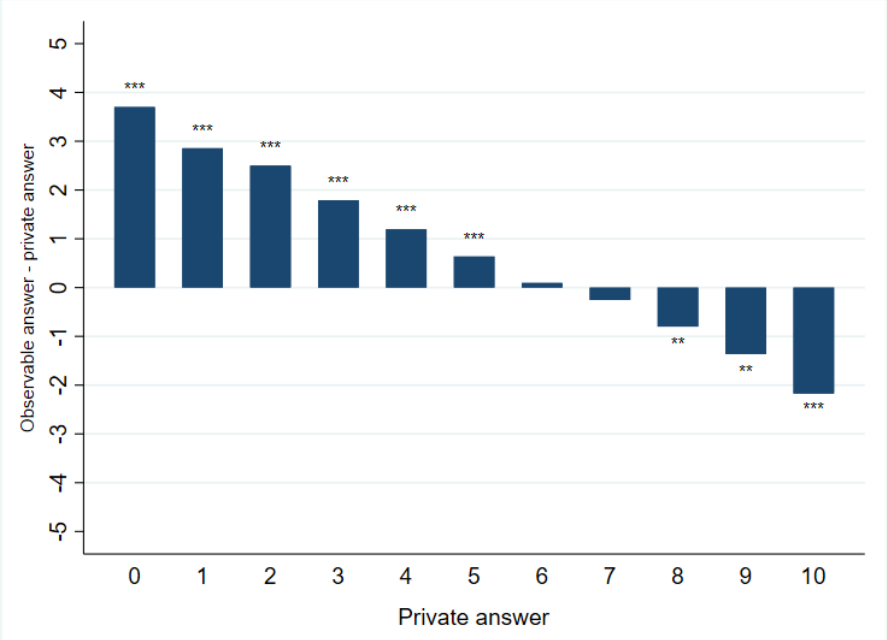


Figure 4. Difference in expressed attitudes - observable vs. private scenario

The figure depicts the average difference between the answer provided in "observable" scenario and the answer provided in "private" scenario ($\Delta_o = a_o - a_p$). The average difference between the answers is depicted per answer provided in the private scenario. A positive (negative) value indicates that on average participants provided an answer implying more (less) supportive attitude towards refugees when their answer will possibly be observed by a future participant (German citizen), than when answering privately. Note: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$ in sign test for H_0 : $median(\Delta_o) = 0$

average difference in answers strongly depends on the value of the initially provided answer in the private scenario. In particular, participants who expressed less support in the private scenario (provided any answer up to the neutral point (5)), on average, provided systematically higher answers in the observable scenario. More interestingly, participants who in the private scenario indicated highly supportive attitudes ($a_p > 7$) systematically decreased their answers in the observable scenario. This suggests that established immigrants, when given an opportunity to misrepresent their attitudes in front of the majority population, do not only use it so as to present themselves as more tolerant than they are , but also to present themselves as less tolerant than they truly are.[11]

[11] One concern here is that the presented evidence of mean reversion when comparing answers in private and observable scenarios might have also resulted if the participants randomly selected

The results depicted in Table 5, illustrate the effect of experimentally induced status on preference falsification. The table shows the results of an OLS regression of the measured preference falsification (Δ_o) on the treatment variable, while controlling for the privately expressed preference (a_p) and a set of individual characteristics. In order to account for the heterogeneous response to treatment across the distribution of the privately expressed preference, I run the regression separately for for participants expressing different levels of support in the private scenario. Particularly, columns (1), (2), and (3) include participants who, in the private scenario, chose an answer that indicates (increasingly) more critical view than the one that would be indicated by selecting a neutral point at $a_p = 5$. Accordingly, columns (4), (5), and (6) include participants who privately indicated (increasingly) more supportive attitudes.

The results show that, among participants who privately indicated more critical attitudes ($a_p < 5$, column (1)), being allocated to the Negative treatment systematically reduced preference falsification. In other words, whereas critical participants falsify their attitudes so as to appear more tolerant in both treatments, those allocated to the Negative treatment do so significantly less. The treatment effect increases in size and precision among those who expressed even more critical views privately ($a_p < 4$, column (2) and $a_p < 3$, column (3)). On the other hand, assignment to the Negative treatment (while still having a negative sign) did not significantly affect preference falsification among those who privately expressed attitudes that are more supportive than that indicated by the neutral point (i.e., for whom $a_p > 5$), neither when observed together (column (4)), nor when focusing only on those with particularly supportive views (column (5) and column (6)).

These results suggest that expressed prejudice not only negatively affects privately held attitudes towards refugees of those exposed to it, but also increases the readiness to publicly present biased views, thereby weakening the effect of the social norm against xenophobic expressions.

their answers in both cases. I discuss this possibility in Appendix A.5 and present the evidence against this case.

Table 5. Treatment effects: Preference falsification

	(1) $a_p < 5$	(2) $a_p < 4$	(3) $a_p < 3$	(4) $a_p > 5$	(5) $a_p > 6$	(6) $a_p > 7$
			$\Delta_o = a_o - a_p$			
Negative treatment	-0.435*	-0.661**	-1.154***	-0.063	-0.165	-0.157
	(0.219)	(0.234)	(0.283)	(0.232)	(0.285)	(0.431)
a_p	-0.650***	-0.667***	-0.720**	-0.594***	-0.690***	-0.826***
	(0.104)	(0.137)	(0.276)	(0.082)	(0.124)	(0.164)
Constant	4.088***	4.050***	5.098***	4.102***	4.985***	5.821***
	(0.267)	(0.403)	(1.097)	(0.913)	(0.987)	(1.487)
Individual controls	Yes	Yes	Yes	Yes	Yes	Yes
Observations	394	288	184	510	395	265

Notes: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. OLS regression of measure of preference falsification $\Delta_o = a_o - a_p$, on the privately provided answer a_p and treatment variable. Columns (1) through (3) include only those participants who in private scenario chose an answer that indicates (increasingly) more critical view than the one that would be indicated by selecting a neutral point at $a_p = 5$, as indicated in the columns' title line. Conversely, columns (4) through (6) include only those participants who in private scenario chose an answer that indicates (increasingly) more supportive view than the one that would be indicated by selecting a neutral point at $a_p = 5$. All regressions include fixed effects of the federal state of residence in Germany and region of participants' (parental) region of origin. Standard errors are clustered on the level of participants' (parental) region of origin.

4 Conclusion

While immigration attitudes received a lot of attention in both economics political science literature, previous research predominantly examined these positions from the point of view of majority populations of receiving countries. This paper studies immigration attitudes of established immigrants, that is those, who already have resided in the host countries for a longer time, toward new flows of immigration, and the drivers behind these positions. Starting from the implications of the Social Identity Theory, I hypothesize that relative status deprivation, that is, the negative difference in status between own ethnic/national group and that of the native majority (or other, more favorably perceived minorities), has a negative impact on group's members' attitudes toward an even lower ranked status group (e.g., such as refugees). I argue that low-status groups that were socialized in a steep ethnic hierarchy and were exposed to prejudiced treatment, over time come to perceive ethnic competition as usual and perhaps legitimate, and consequently

engage in it also when they are faced with an even lower-status groups.

In an online experiment a sample of participants with immigration background residing in Germany is randomly assigned to receive either a positive or a negative evaluation of the influence of their own (immigrant) in-group on "socio-economic and cultural life in Germany", as expressed by a participant from majority population (without immigration background). Participants are additionally provided with the evaluations of two other out-groups (same for all participants), which fixes the status hierarchy and only leaves the position of participant's in-group variable. Experimental results confirm the hypothesis by showing that participants who received a negative evaluation of their in-group are significantly less willing to pledge a donation to the UNHCR, and provide less supportive answers to a set of questions regarding attitudes towards refugees (albeit the latter difference is only partially significant).

Furthermore, I hypothesize that the effects of the prejudiced evaluation work through manipulating the social norms surrounding discrimination and its expressions. In particular, people from low-status regions could learn from discrimination directed towards their own in-group that discriminating downwards (i.e. against groups ranked lower than one's own group) is widespread behavior in the host society, which in turn increases the probability of them engaging in such behaviors themselves. The results show that, when asked to guess how participants from the native majority evaluated impact of other immigrant groups, participants who received a negative evaluation of their own in-group (compared to those who received a positive one) expect the evaluations to be significantly more critical. This applies to expected evaluations of all (mentioned) low-status immigrant groups, including the refugees from the Middle East, but not to evaluation of a high-status immigrant group. I provide tentative evidence for the role of perceived descriptive norm regarding acceptance of refugees in mediating treatment effect on behavior.

Lastly I show that receiving negative evaluation of the in-group increases the readiness of those participants who privately hold most negative attitudes towards refugees to publicly state their views, thus weakening the effect of the norm against

xenophobic expressions.

The findings presented in this work show how factors specific to the receiving, rather than sending country, might impact political views and behavior of immigrants. They highlight the importance of policies and public attitudes affecting perceptions of immigrant groups' status, and particularly those seeking to regulate prejudice expressions, by showing how status effects spill over into attitudes towards other (and potentially not yet present) minorities.

A Appendix

A.1 Attitudinal questions on views regarding refugees from the Middle East

Participants were asked to provide answers to the following seven questions. Other than the question number 6, all questions have been adopted from Dinas et al. (2021a). For the purposes of the analysis presented in 2 all answers were re-coded such that a higher value indicates higher support for refugees.

1. Do you think Germany should increase or decrease the number of people it grants asylum to? (1 = Greatly increase; 5 = Greatly decrease)

2. Refugees are a burden on our country because they take our jobs and social benefits.(1 = Completely agree; 5 = Completely disagree)

3. The money spent on the accommodation of refugees in our country could have been spent better to cover the needs of Germans. (1 = Completely agree; 5 = Completely disagree)

4. Refugees will increase the likelihood of a terrorist attack in our country. (1 =Completely agree; 5 = Completely disagree)

5. Refugees in our country are more to blame for crime than other groups. (1 =Completely agree; 5 = Completely disagree)

6. Is Germany made a worse or a better place to live by refugees who are granted asylum in Germany? (Respondents select their answer on a enumerated scale, where value 0 is labeled as "Worse place to live", and value 10 is labeled as "Better place to live")

7. Among the following options, which one do you think best explains why refugees from Syria and other countries leave their country? (1 = To flee war; 2 = To improve their economic conditions; 3 = To avoid political persecution; 4 = To gain access to host country's social benefits.)

A.2 Sample description

Tables 2 shows the basic demographic characteristics of the sample as a whole, and separately for both treatments. Table 1 shows the distribution of the sample across the targeted regions of origin.

Table 1. Regions of origin per treatment

	Share across treatments		
	Positive treatment	Negative treatment	Total
Region of (parental) origin			
Bulgaria & Romania	0.065	0.078	0.072
Central-Eastern European Union (Czech Republik, Slovakia, Poland, Hungary)	0.182	0.172	0.177
Baltic states (Estonia, Lithuania, Latvia)	0.011	0.007	0.009
Ex-Yugoslavia (Bosnia and Herzegovina, Croatia, Kosovo, Montenegro, North Macedonia, Serbia, Slovenia)	0.058	0.084	0.072
North Africa (Morocco, Algeria, Lybia, Tunesia and Egypt)	0.078	0.068	0.073
Southern European Union countries (Greece, Italy, Portugal, Spain, Cyprus and Malta)	0.106	0.145	0.127
Turkey	0.249	0.205	0.226
Southern Ex-Soviet union (Tajikistan, Turkmenistan, Georgia, Kazakhstan, Kyrgyzstan, Armenia and Azerbaijan)	0.063	0.073	0.068
Western Ex-Soviet union (Ukraine, Moldova, Belarus)	0.033	0.028	0.030
Russian federation	0.092	0.084	0.088
Albania	0.063	0.056	0.060
Observations	554	605	1,159

Notes: Regions of participants' own or parental origin across treatments.

Table 2. Sample description

	Means across treatments		
	Positive treatment	Negative treatment	Total
Age			
[18-24]	0.338	0.349	0.343
[25-34]	0.300	0.284	0.292
[35-44]	0.182	0.175	0.179
[45-54]	0.108	0.116	0.112
[55-64]	0.060	0.063	0.061
[65-74]	0.011	0.010	0.010
[75-84]	0.002	0.003	0.003
Gender			
Male	0.457	0.438	0.447
Education			
Primary or lower secondary	0.354	0.331	0.342
Secondary	0.233	0.238	0.236
Tertiary	0.413	0.431	0.423
Equivalised household income			
Tertile 1	0.372	0.367	0.369
Tertile 2	0.361	0.385	0.374
Tertile 3	0.267	0.248	0.257
Observations	554	605	1,159

Notes: Demographic characteristics of the sample per treatment.

A.3 The role of participants' mood

The following table replicates the results described in Table 1, while controlling for the three measured affective dimensions - pleasure, arousal, and dominance, elicited via Self-Assessment Manikin questionnaire.

Table 3. Treatment effects: The role of participants' mood

	(1)	(2)
	Donation	Pr(Donation>0)
Negative treatment	-6.503***	-0.170***
	(1.583)	(0.062)
Valence	1.455***	0.058***
	(0.513)	(0.011)
Arousal	0.906	0.042***
	(0.629)	(0.015)
Dominance	-0.259	-0.023
	(0.818)	(0.022)
Constant	32.758***	0.790**
	(6.289)	(0.318)
Individual controls	Yes	Yes
Observations	1,159	1,159

Notes: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.
Column (1) shows Tobit regression of amount dedicated to donate to the UNHCR on treatment variable, elicited measures of participant's mood, and the set of individual controls. Negative treatment indicates receiving negative status information regarding own in-group (with Positive treatment serving as a baseline). Columns (3) shows Probit regression of an indicator variable for donation being larger than zero on treatment variable and the set of individual controls. All regressions include fixed effects of the federal state of residence in Germany. Individual controls include age, gender, equivalent household income tertile and indication of tertiary education. Standard errors are clustered on the level of region of origin.

A.4 OLS analysis of the pledged donation amount

The following table depicts the results of the OLS regression of the amount that participants pledged to donate to the UNHCR on the treatment variable and the set of individual controls. The results corroborate the findings presented in Table 1.

Table 4. Treatment effects: Pledged donation to the UNHCR

	(1)	(2)
	Pledged donation	
Negative treatment	-4.383***	-4.316***
	(1.308)	(1.288)
Constant	45.801***	50.505***
	(1.794)	(2.716)
Individual controls	No	Yes
Observations	1,159	1,159

Notes: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.
Column (1) and column (2) show OLS regression of amount dedicated to donate to the UNHCR on treatment variable and the set of individual controls. Negative treatment indicates receiving negative status information regarding own in-group (with Positive treatment serving as a baseline). Reported marginal effects represent the average marginal effect of being allocated to Negative treatment on donated amount. All regressions include fixed effects of the federal state of residence in Germany and region of participants' (parental) origin. Individual controls (included in column (2)) include age, gender, equivalent household income tertile and indication of tertiary education. Standard errors in parentheses are clustered on the level of region of participants' (parental) origin.

A.5 Average preference falsification

One concern regarding the presented results on average preference falsification is that the presented evidence of mean reversion when comparing answers in private and observable scenarios (Figure 5) might have also resulted if the participants randomly selected their answers in both cases. However, multiple findings suggest that this is unlikely the case. Firstly, the distributions of answers in both scenarios, a_p and a_o, both significantly differ from the uniform distribution (Kolmogorov-Smirnov tests for $a_p = U(0, 10)$, and for $a_o = U(0, 10)$, both reject the null hypothesis with $p < 0.001$). Furthermore, the answer to the question in private scenario is significantly correlated both with the answer in observed scenario, as well as with the answers to all other attitudinal questions (coefficient of correlation between 0.40 and 0.47 and $p < 0.001$ in all pairwise tests) and to the donation (coef.=0.23, $p < 0.001$), suggesting that participants did not answer the question at random. Finally, as evident from Figure 5, the observed degree of preference falsification is significantly lower than the one expected if participants had answered randomly in both scenarios.

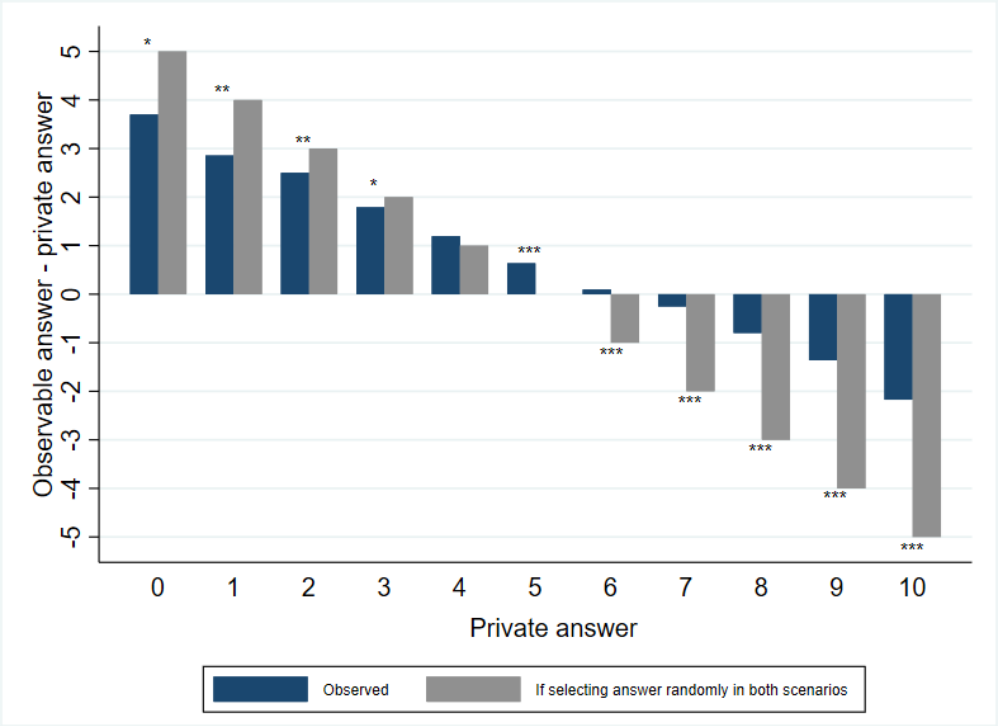


Figure 5. Observed and theoretical preference falsification

The figure depicts the average observed preference falsification ($\Delta_o = a_o - a_p$) and the preference falsification that would be expected if both a_p and a_o were selected randomly. Both values are depicted per answer provided in the private scenario. Note: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$ in sign test of equal median of observed (Δ_o) and the one that would result under random selection.

Nevertheless, this does not exclude the possibility that some share of participants randomly selected their answers, and the others tended not to falsify. However, differently than what would be expected in this case, the distance between the observed and theoretically expected falsification is not equally distributed across the whole range of a_p. Instead, the distance is significantly larger (observed falsification is lower than predicted) among those participants who privately indicated supportive attitudes ($a_p > 5$), than among those who indicated critical attitudes ($a_p < 5$). Additionally, Figure 6 depicts the share of participants who falsified upwards ($\Delta_o > 0$) in the upper panel, and the share of those who falsified downwards ($\Delta_o < 0$) in the lower panel, over a_p. As evident from the figure, the observed probability of falsification in both directions discontinuously changes around the neutral position indicated privately ($a_p = 5$), which would not be observed in case of participants randomly

selecting a_p and a_o. In particular running a probit regression of a dummy variable for observing positive (respectively negative) preference falsification, i.e., $\Delta_o > 0$ (resp. $\Delta_o < 0$) on a_p and a dummy variable ϕ that takes value one if $a_p < 5$ (resp. $a_p > 5$), yields a positive and significant coefficient for ϕ (with $p < 0.01$ and $p = 0.02$ respectively). Taken together, these findings suggest that the preference falsification was rather driven by the perceived social appropriateness of expressed views than by a random behavior.

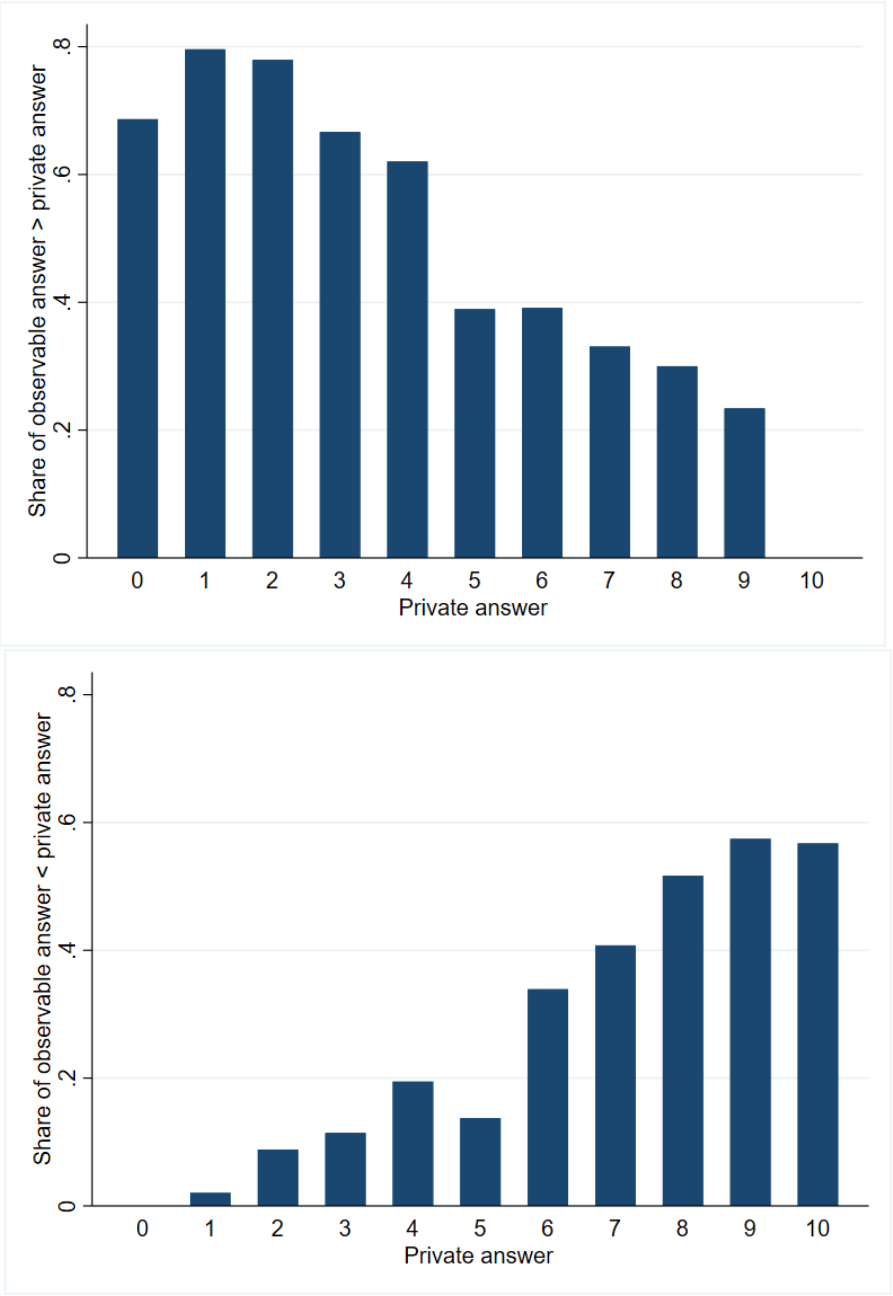


Figure 6. Share of participants with $\Delta_o > 0$ (upper panel) and participants with $\Delta_o < 0$ (lower panel)

The upper panel depicts the share of participants who provided a higher answer (a more positive view) in "observable" than in the "private" scenario ($a_o > a_p$). The lower panel depicts the share of participants who provided a lower (a more critical view) answer in "observable" than in the "private" scenario ($a_o < a_p$).

References

Alexander, R.D., 1987. The biology of moral systems. Transaction Publishers.

Álvarez-Benjumea, A., Winter, F., 2020. The breakdown of antiracist norms: A natural experiment on hate speech after terrorist attacks. Proceedings of the National Academy of Sciences 117, 22800–22804.

Bardsley, N., Sausgruber, R., 2005. Conformity and reciprocity in public good provision. Journal of Economic Psychology 26, 664–681.

Bicchieri, C., Dimant, E., Gaechter, S., Nosenzo, D., 2020a. Observability, Social Proximity, and the Erosion of Norm Compliance. Technical Report.

Bicchieri, C., Dimant, E., Sonderegger, S., 2020b. It's not a lie if you believe the norm does not apply: conditional norm-following with strategic beliefs. Available at SSRN 3326146 .

Bicchieri, C., Xiao, E., 2009. Do the right thing: but only if others do so. Journal of Behavioral Decision Making 22, 191–208.

Bradley, M.M., Lang, P.J., 1994. Measuring emotion: the self-assessment manikin and the semantic differential. Journal of behavior therapy and experimental psychiatry 25, 49–59.

Braghieri, L., 2021. Political correctness, social image, and information transmission. Technical Report. Working paper.

Bursztyn, L., Egorov, G., Fiorin, S., 2020. From extreme to mainstream: The erosion of social norms. American economic review 110, 3522–48.

Bursztyn, L., González, A.L., Yanagizawa-Drott, D., 2018. Misperceived social norms: Female labor force participation in Saudi Arabia. Technical Report. National Bureau of Economic Research.

Dancygier, R., Saunders, E.N., 2006. A new electorate? comparing preferences and partisanship between immigrants and natives. American Journal of Political Science 50, 962–981.

DellaVigna, S., List, J.A., Malmendier, U., Rao, G., 2016. Voting to tell others. The Review of Economic Studies 84, 143–181.

Derks, B., Van Laar, C., Ellemers, N., 2016. The queen bee phenomenon: Why women leaders distance themselves from junior women. The Leadership Quarterly 27, 456–469.

Dinas, E., Fouka, V., Schläpfer, A., 2021a. Family history and attitudes toward out-groups: evidence from the european refugee crisis. The journal of politics 83, 647–661.

Dinas, E., Fouka, V., Schläpfer, A., 2021b. Recognition of collective victimhood and outgroup prejudice. Public opinion quarterly 85, 517–538.

Ellemers, N., Van den Heuvel, H., De Gilder, D., Maass, A., Bonvini, A., 2004. The underrepresentation of women in science: Differential commitment or the queen bee syndrome? British Journal of Social Psychology 43, 315–338.

Enikolopov, R., Makarin, A., Petrova, M., Polishchuk, L., 2020. Social image, networks, and protest participation. Networks, and Protest Participation (April 26, 2020) .

Fouka, V., 2019. How do immigrants respond to discrimination? the case of germans in the us during world war i. American Political Science Review 113, 405–422.

Gerber, A.S., Green, D.P., Larimer, C.W., 2008. Social pressure and voter turnout: Evidence from a large-scale field experiment. American political Science review 102, 33–48.

Goerres, A., Spies, D.C., Mayer, S., 2020. Immigrant german election study (imges). GESIS Datenarchiv, Köln. ZA7495 Datenfile Version 1.0.1, https://doi.org/10.4232/1.13544. doi:10.4232/1.13544.

Greiner, B., Levati, M.V., 2005. Indirect reciprocity in cyclical networks: An experimental study. Journal of Economic Psychology 26, 711–731.

Hainmueller, J., Hopkins, D.J., 2014. Public attitudes toward immigration. Annual Review of Political Science 17, 225–249. URL: https://doi.org/10.1146/annurev-polisci-102512-194818, doi:10.1146/annurev-polisci-102512-194818, arXiv:https://doi.org/10.1146/annurev-polisci-102512-194818.

Janus, A.L., 2010. The influence of social desirability pressures on expressed immigration attitudes. Social Science Quarterly 91, 928–946.

Jost, J.T., 2019. A quarter century of system justification theory: Questions, answers, criticisms, and societal applications. British Journal of Social Psychology 58, 263–314.

Jost, J.T., Pelham, B.W., Sheldon, O., Ni Sullivan, B., 2003. Social inequality and the reduction of ideological dissonance on behalf of the system: Evidence of enhanced system justification among the disadvantaged. European journal of social psychology 33, 13–36.

Just, A., Anderson, C.J., 2015. Dual allegiances? immigrants' attitudes toward immigration. The Journal of Politics 77, 188–201.

Kashima, Y., Wilson, S., Lusher, D., Pearson, L.J., Pearson, C., 2013. The acquisition of perceived descriptive norms as social category learning in social networks. Social Networks 35, 711–719.

Krupka, E., Weber, R.A., 2009. The focusing and informational effects of norms on pro-social behavior. Journal of Economic psychology 30, 307–320.

Kuo, A., Malhotra, N., Mo, C.H., 2017. Social exclusion and political identity: The case of asian american partisanship. The Journal of Politics 79, 17–32.

Kuran, T., 1997. Private truths, public lies: The social consequences of preference falsification. Harvard University Press.

Kwan, L.Y.Y., Yap, S., Chiu, C.y., 2015. Mere exposure affects perceived descriptive norms: Implications for personal preferences and trust. Organizational Behavior and Human Decision Processes 129, 48–58.

Meeusen, C., Abts, K., Meuleman, B., 2019. Between solidarity and competitive threat?: The ambivalence of anti-immigrant attitudes among ethnic minorities. International Journal of Intercultural Relations 71, 1–13. URL: https://www.sciencedirect.com/science/article/pii/S014717671830467X, doi:https://doi.org/10.1016/j.ijintrel.2019.04.002.

Morris, S., 2001. Political correctness. Journal of political Economy 109, 231–265.

Mujcic, R., Leibbrandt, A., 2018. Indirect reciprocity and prosocial behaviour: evidence from a natural field experiment. The Economic Journal 128, 1683–1699.

Neuhold, C., 2020. Wer erbt die blauen serben? https://www.profil.at/oesterreich/wer-erbt-die-blauen-serben/400961525.

Norton, M.I., Sommers, S.R., Apfelbaum, E.P., Pura, N., Ariely, D., 2006. Color blindness and interracial interaction: Playing the political correctness game. Psychological Science 17, 949–953.

Nowak, M.A., Sigmund, K., 2005. Evolution of indirect reciprocity. Nature 437, 1291–1298.

Perez-Truglia, R., Cruces, G., 2017. Partisan interactions: Evidence from a field experiment in the united states. Journal of Political Economy 125, 1208–1243.

Strijbis, O., Polavieja, J., 2018. Immigrants against immigration: Competition, identity and immigrants' vote on free movement in switzerland. Electoral Studies 56, 150–157.

Tajfel, H., Turner, J.C., Austin, W.G., Worchel, S., 1979. An integrative theory of intergroup conflict. Organizational identity: A reader 56, 65.

Valentim, V., 2022. Social norms and preference falsification in a democracy .

Van der Zwan, R., Bles, P., Lubbers, M., 2017. Perceived migrant threat among migrants in europe. European Sociological Review 33, 518–533.