Social Norm Perceptions in Third-Party Punishment

Katarína Čellárová* Jonathan Stäbler†

May 16, 2024

Abstract

Costly punishment by unaffected third parties can play an important role in sustaining cooperation and deterring selfish behavior. Such third-party punishment has been taken as evidence in itself that individuals care about the enforcement of social norms. In this paper, we explicitly study whether and which norm-related beliefs motivate third-party punishment. To do so, we run an experiment where we elicit punishment decisions in a modified dictator game and measure three social norm perceptions: personal norms of appropriateness, beliefs about others' appropriateness norms (normative expectations), and beliefs about typical behavior (empirical expectations). We find that higher personal norms of appropriateness and higher empirical expectations lead to an increase in punishment. Normative expectations, on the other hand, are negatively correlated with punishment when controlling for either of the other two norm perceptions. We conclude that the desire to enforce own beliefs of appropriateness or typical behavior motivates punishment decisions rather than perceived societal appropriateness views.

Keywords: Third-Party Punishment, Social Norms, Empirical Expectations, Normative Expectations, Personal Norms

JEL: C72, C91, D63, D84, D91

*Faculty of Law, Charles University, nám. Curieových 901/7, 116 40 Praha 1, Czech Republic e-mail: cellarova.katarina@gmail.com

†Department of Economics, University of Mannheim, L7 3-5, 68131 Mannheim, Germany. e-mail: jonathanstaebler@gmail.com

1 Introduction

In third-party punishment, an unaffected individual punishes another person for an act of wrongdoing. It is seen as a tool for the enforcement of social norms (Carpenter & Matthews 2012, 2009, Henrich et al. 2006, Fehr & Fischbacher 2004) and can serve to sustain cooperation by deterring selfish behavior (Lergetporer et al. 2014, Carpenter & Matthews 2012, Mathew & Boyd 2011, Charness et al. 2008, Carpenter et al. 2004, Fehr & Fischbacher 2004), to promote more egalitarian allocations (Martin et al. 2021), and more generally, to sustain different norms of behavior across societies (Kamei et al. 2023, Henrich et al. 2006). Such punishment comes at a personal cost to the punisher, which suggests that humans care about how others behave, even when they are not directly affected by the behavior.

Whatever constitutes wrongdoing, however, is subjective and not always clear, and it may depend on the punisher's perceptions of the relevant social norms. First, punishers may have their own personal beliefs about what should be done in a specific situation – *personal norms of appropriateness*[1] – and punish those who deviate from it. In this way, they could enforce their own preferences about how to behave in a specific situation. Second, because humans are a part of society, they may base their punishment decisions not only on their own appropriateness views but also on what others deem appropriate. *Normative expectations*[2] are individuals' beliefs about what others think is appropriate, and can guide punishment decisions to enforce behavior that individuals perceive to be preferred by society. Third, punishers may also rely on their *empirical expectations*[2], their beliefs about what constitutes common behavior. Empirical expectations may inform individuals' punishment decisions, as how humans typically behave may result from what they think others (and themselves) think is appropriate (Tremewan & Vostroknutov 2021).

It is often argued that the existence of third-party punishment is evidence by itself that humans care about the enforcement of social norms (Carpenter & Matthews 2012, 2009, Henrich et al. 2006, Fehr & Fischbacher 2004). However, to the best of our knowledge, no study specifically elicits personal norms, normative expectations, and empirical expectations together and addresses their

[1] We follow Bicchieri (2016) to classify the beliefs that matter for norm compliance. For a more concise text, we refer to personal norms as one of the ways in which social norms are 'perceived'.

[2] Cialdini et al. (1990) defines the injunctive norm as what people believe ought to be and the descriptive norm as what usually is (common behavior). We follow the classification of Bicchieri (2016) and it can be viewed as the individual's belief of the injunctive norm (normative expectation) and the belief of the descriptive norm (empirical expectation).

roles as the underlying motives for third-party punishment. Furthermore, no other study explicitly identifies the causal impact of the three types of social norm perceptions on third-party punishment decisions. In this paper, we close this research gap and identify whether and to what extent personal norms of appropriateness, empirical expectations, and normative expectations trigger third-party punishment and study their relative importance.

Previous studies indicate that social norms and individuals' beliefs about those norms matter for third-party punishment. Carpenter & Matthews (2009) test a broad set of different average behavior specifications in public goods games and find that deviations from the average contribution of the session best explain larger third-party punishment decisions.[3] Carpenter & Matthews (2012) confirm this result and further observe that deviations from the punisher's beliefs about the expected contribution, as well as from the punisher's own contribution, are associated with larger punishment decisions. Hence, in these studies, empirical expectations seem to matter for third-party punishment decisions. Other studies indicate that normative expectations matter as well. House et al. (2020) find that injunctive norm nudges in the form of messages about 'what is wrong and bad behavior' increase children's third-party punishment decisions. Zong et al. (2021) find that punishers react to information about the sender' expectations in a trust game. Dimant & Gesche (2023) observe that injunctive and descriptive norm nudges increase third-party punishment decisions and further find that both nudges increase personal appropriateness ratings of the situation in a subsequent experiment. However, they do not elicit punishment decisions in the experiment. Finally, personal norms also seem to matter. Bašić & Verrina (2023) find that both normative expectations and personal norms about third-party punishment decisions are positively correlated with own third-party punishment decisions.

Some studies investigate different types of third-party punishment norms. Kamei (2020), Fabbri & Carbonara (2017), and Lois & Wessa (2019) find that information and beliefs about others' punishment decisions influence subjects' own punishment decisions. Furthermore, Kamei (2018) finds that being observed by another punisher increases punishment, indicating that subjects care about conforming to a punishment norm. Literature on second-party punishment also identifies the importance of the descriptive and injunctive norms of cooperation for punishment (Li et al. 2021,

[3]The set of behavior specifications included the average contribution of the session, of the ingroup, or of the outgroup, own contributions, or the set of all possible (exogenously set) contributions.

Reuben & Riedl 2013), as well as the punishment norm itself (Li et al. 2021).

In summary, the literature demonstrates that various perceptions of social norms matter for third-party punishment decisions. Yet, unlike our paper, none of the above-mentioned studies explicitly elicits all three social norm perceptions – personal norms, empirical expectations, and normative expectations – with punishment decisions together. Therefore, they cannot clearly identify the channels for punishment decisions. In principle, all of the three norm perceptions inform each other and hence are correlated (Tremewan & Vostroknutov 2021). At the same time, these three norm-related perceptions can differ due to heterogeneous preferences and asymmetries in the availability and processing of information about appropriate or common behavior. It is important to consider all of these norm-related beliefs. Otherwise, the effect of one of them could be wrongly attributed to another. Furthermore, none of the papers above provides evidence for a causal effect of social norm perceptions on punishment.

To study how social norm perceptions matter, we run an online interactive experiment that consists of two phases. In the first phase – the Experience Phase – punishers go through a modified dictator game in different roles. The dictator starts with an endowment of 100 CZK and decides to transfer either 0, 10, 40, or 50 CZK to the receiver.[4] In the second phase – the Punishment Phase – subjects choose whether and how much to punish *another* dictator for their behavior in the same type of game via the strategy method. The to-be-punished dictator does not interact with the punishers in any other way. We measure personal norms, empirical expectations, normative expectations, and emotions before the Experience Phase and after the Punishment Phase.[5]

To create more pronounced heterogeneities in social norm perceptions, and to study their causal effects, we employ four treatments with an exogenous variation of the Experience Phase. Participants are randomly assigned to the role of the dictator (*Dictator treatment*), receiver (*Receiver treatment*), observer (*Observer treatment*), or to the *Baseline* treatment.[6] In the Dictator and Receiver treatments, participants played one round of the same dictator game before the Punishment

[4]We omit the choices of 20 and 30 to force more extreme transfers and hence to have a higher potential of shifting norm perceptions. 100 CZK (approx. 4 EUR) corresponded to a student wage of 50 minutes at the time of the experiment.

[5]We also control for negative emotions, as they are an important driver of third-party punishment (Jordan et al. 2016, Carpenter & Matthews 2012, Nelissen & Zeelenberg 2009).

[6]We acknowledge that the Dictator and Receiver treatments are not independent from each other. However, since the acquired experience in the Experience Phase substantially differs between the two roles, and the choices of the dictators are exogenous to the receivers, we classify them as separate treatments.

Phase. In the Observer treatment, participants observed a transfer from a dictator from an earlier session, and in the Baseline treatment, the Experience Phase was omitted.

The treatments have the potential to change social norm perceptions in the following way. First of all, the mere assignment to the roles of dictator and receiver may make subjects shift their perceptions of appropriate and common behavior in a motivated way. While dictators could tell themselves that lower transfers are more appropriate and common to justify their own low transfers, receivers may tell themselves the exact opposite. Second, subjects in the Receiver and Observer treatment receive an exogenous signal about typical behavior, which could make them update their norm perceptions. As receivers could also feel stronger emotions with the associated transfer, we explicitly control for emotions. We find that the treatment assignment leads to substantive variations in how subjects perceive social norms: all three norm perceptions are different between the treatments. Additionally, the within-subject differences between the three norm-related perceptions are also shifted by the treatments. This allows us to study their causal impact and identify their relative contributions to punishment decisions.

Our main findings are the following: We find that an increase in all three norm perceptions leads to higher punishment decisions individually. However, when studying their joint correlations, we find that the correlation of normative expectations with punishment reverses to a significant negative correlation. The positive effect of personal norms and empirical expectations on punishment prevails. In other words, we find higher punishment by subjects who believe higher transfers are more appropriate and those who think that dictators typically transfer more. If they believe others deem higher transfers more appropriate, they punish less. We provide consistent and prevailing evidence for the positive impact of personal norms and empirical expectations on punishment. However, we find that higher normative expectations are associated with lower punishment when controlling for either one of the other norm perceptions.

To explore the negative relationship between normative expectations and punishment further, we analyze punishers' normative expectations relative to their own personal norms. We find that subjects who believe they hold higher moral standards than society punish more, whereas subjects who believe others to hold higher moral standards punish less. One explanation for this behavior is that individuals may feel a greater responsibility for punitive action when they anticipate lower moral standards among others, whereas they may not feel the necessity to enforce their own appropriateness standards if they believe society to already uphold higher moral standards.

As additional results, we find that the relative importance of the three norm-related perceptions depends on gender and the assigned role. The positive relationship between personal norms and punishment is stronger for males, whereas the positive relationship between empirical expectations and punishment is stronger for females. Moreover, receivers rely more on their empirical expectations compared to the rest of the sample. Dictators hold the lowest personal norms compared to the other treatments, which results in overall lower punishment levels.[7]

The contribution of this paper is two-fold. First, we contribute to the literature on social norms and third-party punishment (Bašić & Verrina 2023, Dimant & Gesche 2023, Zong et al. 2021, House et al. 2020, Lois & Wessa 2019, Kamei 2018, Fabbri & Carbonara 2017, Carpenter & Matthews 2012, 2009, Henrich et al. 2006, Fehr & Fischbacher 2004). We find a causal influence of social norm perceptions on punishment and thus provide evidence that third-party punishment is indeed used for the enforcement of social norms. We show that third-party punishment is used to enforce one's own view of appropriateness and typical behavior, but not society's appropriateness views. Furthermore, the importance of personal norms or empirical expectations varies depending on gender or the specific exogenously assigned role. These results provide policy implications for those aiming to increase informal sanctioning mechanisms, such as third-party punishment. Policies should focus on shifting empirical expectations (and, if possible personal norms) to influence punishment most efficiently. In addition, policymakers should evaluate who the target population is, as the relevance of the social norm perceptions for punishment varies.

Second, we also contribute to the general literature about social norms (e.g. Bicchieri, Dimant, Gelfand & Sonderegger 2023, Abeler et al. 2019, Danilov & Sliwka 2017, Kessler & Leider 2012, Andreoni & Bernheim 2009, Bénabou & Tirole 2006). Our finding that punishment is not driven by the urge to enforce societal normative views has important implications for our understanding of norm-driven economic behavior. Individuals might overall care less about societal appropriateness views and more about their own appropriateness views and typical behavior. We also show the importance of considering all three norm-related beliefs. As they are correlated but also differ from each other, the effect of one of them may be wrongly attributed to another one.

[7]This indicates that dictators could engage in motivated reasoning. In order to licence themselves to transfer less, they lower their beliefs of appropriateness in a self-serving way.

The paper is structured as follows: In Section 2, we provide a theoretical discussion, in Section 3, we describe the experimental design, and Section 4 presents the results. Lastly, Section 5 provides a discussion and concludes.

2 Theoretical Considerations

In this section, we discuss how social norm perceptions relate to each other and under which circumstances they may diverge. Based on this analysis, we derive the potential influence of social norm perceptions on third-party punishment decisions.

As Tremewan & Vostroknutov (2021) argue, individuals form social norm perceptions based on the information available to them and based on their perceptions of the availability of information to others. In the case of the personal opinion of appropriateness, individuals also take into account their own preferences about the expected outcomes that are associated with specific actions. Consequently, personal norms can differ among individuals due to differences in the information they rely on, different information processing, or heterogeneous preferences. When forming normative expectations, i.e., beliefs about others' personal views of appropriateness, individuals might compare others to themselves. If they believe others to be exactly alike, normative expectations coincide with their own personal norm of appropriateness. However, if individuals believe there are differences – either in the availability of information, information processing, or in preferences – normative expectations can differ from one's own personal norm of appropriateness.[8] Lastly, individuals' behavior may not necessarily align with their own or societal perception of appropriate behavior. Merguei et al. (2022) find that when there are several norms in a situation, subjects opportunistically follow the norm that maximizes their own payoffs – a phenomenon termed moral opportunism. Additionally, Bicchieri, Dimant & Sonderegger (2023) show that individuals motivatedly distort their social norm perceptions in a self-serving manner. They find that subjects update their empirical – but not normative – expectations about lying when presented with an upcoming opportunity to lie. Thus, individuals may choose to exploit these selfish opportunities or, crucially, expect others to do so. In addition, Kölle & Quercia (2021) show that when there is strategic uncertainty about others' behavior, normative and empirical expectations of participants

[8]See Bašić & Verrina (2023) for specific examples, where personal norms and normative expectations may differ.

differ substantially. Consequently, the perception of common behavior can differ considerably from normative views.

To sum up, the three norm-related beliefs can differ due to heterogeneities in preferences, due to differences in the availability and processing of information, and due to the anticipation of moral opportunism. Given that they differ, the question arises which among them mostly motivates individuals' third-party punishment decisions.

First, let's consider the enforcement of own personal norms. Punishing those who deviate from this view could be motivated by the desire to change future behavior. One goal could be to implement an outcome that one believes is better for everyone – or at least for individuals with the same type as themselves. Another goal could be to change future outcomes where one is directly involved. The reliance on personal norms goes in line with (Bašić & Verrina 2023), who demonstrate the importance of personal norms for economic decision-making, including third-party punishment.[9]

Second, the desire to enforce an outcome that society deems appropriate may motivate punishment for the following reasons. For instance, individuals may see it as a moral obligation to serve society and punish those that deviate from the normative views that society imposes. Another possibility is that individuals may not have very strong own appropriateness views and thus generally rely more on societal appropriate views. If normative expectations and own personal norms diverge, punishers face a trade-off between enforcing an outcome that they deem appropriate and an outcome that, in their views, society deems appropriate. Whether the one or the other dominates punishment decisions may then depend on individual factors. Bašić & Verrina (2023) find overall stronger correlations of personal norms with economic behavior compared to normative expectations. This indicates that subjects might rely more on their personal norms than normative expectations when deciding on punishment.[10]

Lastly, previous literature emphasizes that empirical expectations are more important for economic behavior than normative expectations (Bicchieri et al. 2022, Kölle & Quercia 2021, Chen

[9] Unlike this study, (Bašić & Verrina 2023) focus on appropriateness views about punishment decisions and not about behavior in the game itself.

[10] There might be cases when individuals engage in costly punitive behavior to enforce an outcome that they believe others would prefer, even if that goes against their own appropriateness views. Pluralistic ignorance – i.e., a difference between the perceived societal norm and all personal norms – is a prominent phenomenon that might be enforced in those cases (Andre et al. 2024, Bursztyn et al. 2020).

et al. 2020, Schmidt 2019, Agerström et al. 2016, Bose et al. 2023, Bicchieri & Xiao 2009) including for punishment decisions (Dimant & Gesche 2023). Beliefs about common behavior may be used for punishment, if one wants to reinforce typical behavior by punishing those who behave atypically. In this case, subjects might decide to base punishment on empirical expectations instead of normative expectations because they want to justify and reinforce deviations from higher normative standards.[11] In addition, such a reliance on empirical expectations can also be used in a self-serving way to avoid the cost of punishing others if one wants to behave opportunistically. This opportunity to decrease punishment may be especially exploited because empirical expectations are more prone to self-serving distortions compared to normative beliefs (Bicchieri, Dimant & Sonderegger 2023).

To identify the effect of each of these three norm-related beliefs, we aim to create heterogeneity in the appropriateness views and an additional mismatch between beliefs of common behavior and those appropriateness views. First, we use a (modified) dictator game, which is known to create substantial heterogeneities in behavior (Engel 2011), which might be driven by different appropriateness views. As any allocation in the dictator game is Pareto-efficient, there is not one socially optimal solution. Hence, potentially heterogeneous fairness views and social preferences shape appropriateness views. Second, we employ four treatments to shift the three norm-related beliefs to a different extent. We do that in two ways: by assigning subjects to different roles in the game (additionally to their role of punishers) and by giving a noisy signal of common behavior in the form of one transfer decisions of a dictator.

Depending on the role that subjects get assigned to (receiver, dictator, or observer), they may motivatedly distort their beliefs in self-serving ways (e.g. Bicchieri, Dimant & Sonderegger 2023, Zimmermann 2020, Epley & Gilovich 2016). For instance, dictators may tell themselves that smaller transfers are appropriate, whereas receivers may believe that higher transfers are more appropriate. Additionally, participants may update their empirical expectations downwards or upwards around the received signal of common behavior (Bicchieri et al. 2022, Hoeft et al. 2023, Gino et al. 2009, Keizer et al. 2008). At the same time, this signal may also inform normative expectations and change personal norms as they are related and inform each other.[12]

[11] This rationale also holds when own personal norms are different from normative expectations.

[12] See Tremewan & Vostroknutov (2021) for how social norm perceptions inform each other.

3 Experiment

We ran an online experiment[13] in March and April of 2021 with subjects of the Masaryk University Experimental Economics Laboratory (MUEEL). The experiment was programmed in z-Tree (Fischbacher 2007), and we used z-Tree unleashed (Duch et al. 2020) to implement running sessions on the internet. The experiment received ethical approval from the GfeW.[14] Each session consisted of 14 participants and took an average of around 40 minutes, with average earnings of 118 CZK (approximately 4.79 EUR, which corresponded to a student wage of one hour of unqualified work). In total, 420 subjects participated in the experiment, of which 300 acted as punishers, and 120 as punishees. We analyze the punishment behavior of 296 punishers.[15] Punishees played the dictator game and were subject to potential sanctions from the punishers, and they did not interact in any other way. Hence, we ensured impartial third-party punishment decisions and removed any indirect counter-punishment considerations.

3.1 Experimental Design

The experiment consisted of two sections, Section A and Section B, where Section A was payoff-relevant with an 80% probability and Section B with 20% probability. Section A was the main part of the experiment, whereas Section B served to measure distributive preferences and demographics. Figure 1 depicts Section A of the experiment. It consisted of two phases: the Experience Phase and the Punishment Phase. In addition, we measured social norm perceptions and emotions at the beginning and at the end of Section A.[16]

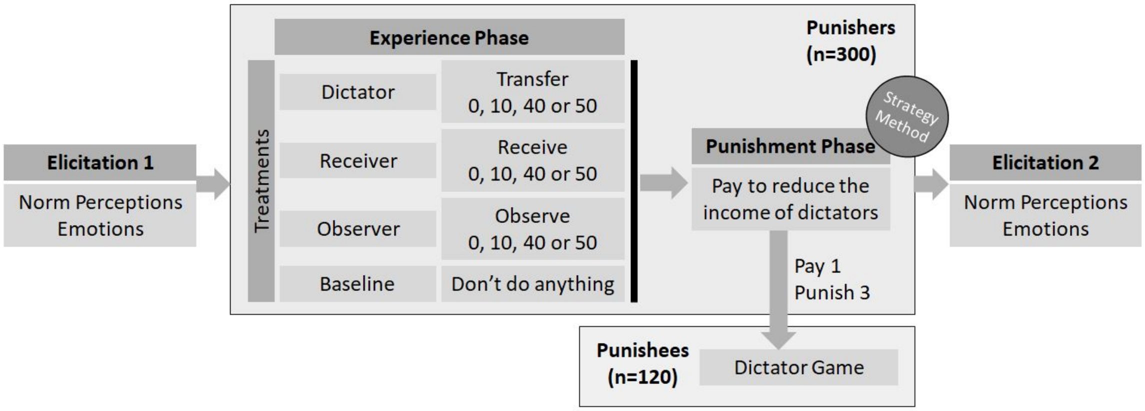
[13] Arechar et al. (2018), for example, find that an interactive public goods game with and without punishment can be conducted very reliably online and produces similar behavioral patterns as in the laboratory.

[14] German Association of Experimental Economics.

[15] We excluded four participants from the analysis because they remained inactive for several minutes and we had to forward them to the next pages.

[16] In principle, eliciting beliefs also at the beginning could overestimate the link between punishment and norm perceptions through an experimenter demand effect. However, d'Adda et al. (2016) do not find evidence that the order of norm elicitation affects behavior.

Figure 1: Experimental design Section A



In the Experience Phase, we manipulate what precedes the Punishment Phase. We employ four treatments with the goal to induce differences in the social norm perceptions. Participants were assigned to one of the following treatments: the *Dictator (N=80)* treatment, the *Receiver (N=80)* treatment, the *Observer (N=80)* treatment, and the *Baseline (N=60)* treatment.

In the Dictator and Receiver treatment, which were conducted in the same sessions, subjects were randomly assigned to either the role of dictator or the role of receiver. Dictators decided how much to transfer to a randomly matched receiver. They were endowed with 100 CZK and could transfer either 0, 10, 40, or 50 CZK. In the Observer treatment, subjects observed the transfer of a randomly chosen dictator from the Dictator treatment, which was run in a previous experimental session.[17] In the Baseline treatment, the Experience Phase was simply omitted.[18] To avoid a systematic influence of income on punishment, we equalized payoffs by different show-up fees before the start of the Experience Phase. Receiver subjects were paid a show-up fee of 50 CZK. In the treatments Observer and Baseline, everyone received an individual show-up fee that consisted of 50 CZK plus a randomly chosen payoff from the set of payoffs that Receiver subjects had obtained in earlier sessions. Each payoff from this set was used exactly once for the Observer treatment. In the Baseline treatment, the distribution of payoffs was replicated.[19] We established the same wealth level before the Punishment Phase across all treatments except in the Dictator

[17]Each dictator's decision was shown to one subject in the Observer treatment, who saw one specific decision. Dictators in the Dictator treatment did not know that their choices were shown to other players in later sessions.

[18]The Observer and Baseline treatments each took place in separate sessions.

[19]In the Baseline treatment we could not use the exact same number of payoffs because of a smaller number of subjects.

treatment, where subjects were wealthier by design.[20]

After this Experience Phase, the Punishment Phase followed. In the Punishment Phase, subjects could punish a dictator in the same version of the dictator game. This group of dictators (*punishees*, see Figure 1) was unrelated to the group of dictators in the Experience Phase and participated in the same sessions. We included punishees so that the punishment decisions of the punishers had real consequences. The Punishment Phase was the same for all treatments. Each punisher could reduce the earnings of one punishee dictator. Punishers received an endowment of 50 CZK and could use that endowment to punish. We elicited the willingness to punish via the strategy method, where the punisher assigned deduction points to the punishee for every possible transfer.[21] We applied the typically used punishment ratio of 1:3, where the punisher pays one unit of her endowment to deduct the income from the punishee by three units (e.g. Fehr & Fischbacher 2004).[22] If Section A was chosen to be payoff-relevant, the punishment decision was guaranteed to be implemented. We explain the exact matching procedure in the description of the punishees below. We made punishers aware that they themselves would not be punished at any stage of the experiment.

In the whole experiment, we used a specific modified version of the dictator game. The dictator receives an endowment of 100 CZK, while the receiver starts with zero. The dictator can choose to transfer either 0, 10, 40, or 50 CZK to the receiver. This modified version has several advantages for studying the impact of social norm perceptions on punishment. By excluding intermediate choices like 20 or 30, we enforce more extreme transfers that have a larger potential to shift social norm perceptions of receivers and observers.[23] In addition, dictators have to choose a more extreme transfer in the Experience Phase. The need to justify a transfer of 0 or 10 over 40 or 50 could induce a more pronounced shift in norm perception through motivated reasoning. Another advantage of this modified version is that participants are less familiar with what describes common behavior and what ought to be done, which could lead to more heterogeneous social norm perceptions.

[20] In principle, a higher wealth could result in higher punishment. However, we find that wealth does not play an important role for punishment decisions.

[21] Jordan et al. (2016) find that third-party punishment decisions are not influenced by the strategy method. In addition, we show in Appendix A.1.4 that this design choice does not seem to pose a threat to the validity of our results.

[22] It was possible to reduce the punishees' income by up to 150 CZK, which would result in a negative payment of the punishee. However, punishees played multiple rounds of the same dictator game. Therefore, negative payments in one round could be compensated by positive payments in another round, as well as by the elicitation stage and the show-up fee.

[23] In the pilot session, most of the initial norm perceptions were between 10 and 40 CZK. Thus, extreme transfers are further away from the initial norm perceptions and consequently have a larger potential to shift them.

Finally, as in the standard dictator game, there is no unique socially optimal allocation, and thus social norm perceptions may be more dispersed because subjects may deem different allocations as appropriate.

Before the Experience Phase and after the Punishment Phase, we elicited subjects' personal norms, normative expectations, and empirical expectations, following Bicchieri et al. (2022). First, we asked subjects what they believed *should* be transferred in the dictator game (personal norm). They could choose from the same set of transfers used throughout the whole experiment: 0, 10, 40, or 50 CZK. Second, we asked them what they thought was the *average response* to the first question by other participants of the same experimental session (normative expectation). Finally, we asked subjects *what they believed was the average choice* of the dictators in the ongoing (Dictator and Receiver treatment) or a previous (Observer treatment and Baseline) experimental session (empirical expectation). For both normative and empirical expectations, we used a continuous scale to capture small changes in individual norm perceptions.[24] The first norm elicitation took place after punishers knew their role in the Experience Phase. We incentivized the elicitation of normative and empirical expectations, paying an additional 15 CZK whenever participants were within a range of 6 CZK around the true average. In the second elicitation, which took place right after the punishment decision, subjects were shown their choices from the first elicitation. We asked them to consider whether their expectations had changed. Thus, any reported change was intentional.

Lastly, we measured self-reported emotions following the elicitation of Bosman & Van Winden (2002) and Cubitt et al. (2011) and included both positive and negative emotions. Specifically, we asked participants about their current intensity of anger, gratitude, guilt, happiness, irritation, compassion, surprise, and envy. We measured the intensity of each emotion by self-reports on a 7-point Likert scale, from not feeling the emotion at all to feeling it very much. We elicited emotions both before the Experience Phase and after the Punishment Phase.[25] We included the elicitation of emotions as they have been shown to affect punishment (Jordan et al. 2016, Carpenter & Matthews 2012, Nelissen & Zeelenberg 2009). At the same time, we control for any emotion that the Experience Phase induces, as to not wrongly attribute an emotion effect to a norm perception

[24] The initial position of the slider was 25, which constitutes half of the highest possible transfer.

[25] At the second elicitation, choices from the first elicitation were prefilled. We asked subjects to consider whether the intensity of their emotions had changed.

effect. This may be especially relevant for subjects in the Receiver treatment, who might experience strong emotions as the received transfer directly affects their payoff.

Now we describe the procedure for punishees. Punishees started with the same elicitation of emotions and norms as the punishers, then played four rounds of the same version of the modified dictator game. In these four rounds, every punishee acted exactly twice as a dictator and twice as a receiver.[26] The exact rounds in which they acted in a specific role were randomly determined. When they were in the role of dictator, they were subject to punishment from the punishers, depending on the transfer that they chose.

In each round of the dictator game played by the punishees, we matched exactly one punisher with one round of a punishee dictator. Each session consisted of 14 participants, of which ten were punishers and four punishees. As each punishee acted exactly twice as a punishee dictator, the punishment decision of eight punishers was implemented. In this way, the punishment decision of every punisher was implemented if Section A was chosen for that punisher (since Section A was payoff-relevant with 80% probability).[27] Punishees knew that different punishers punished them in different rounds. However, we did not disclose this information to the punishers. We told punishers that punishees were in the same experimental session playing the same version of the dictator game and that punishers could reduce the income of a punishee dictator. Punishers were told about the implemented punishment decision and the choice of the punishee dictator only at the end of the experiment. Hence, the second norm elicitation of punishers remained unaffected by the transfer choice of the punishee dictator.

Finally, Section B served to measure distributional preferences. All punishers played the same dictator game as in Section A without the possibility of receiving punishment. All subjects made decisions as dictators. Afterwards, the computer decided randomly whether their role was a dictator or receiver and who they were matched with. Subjects then answered a questionnaire with demographic information, for which they received 30 CZK, in case Section B was payoff relevant (additional to a 50 CZK show-up fee). This part served as a control for differences in redistributive preferences, which we use as an additional control in a robustness check.

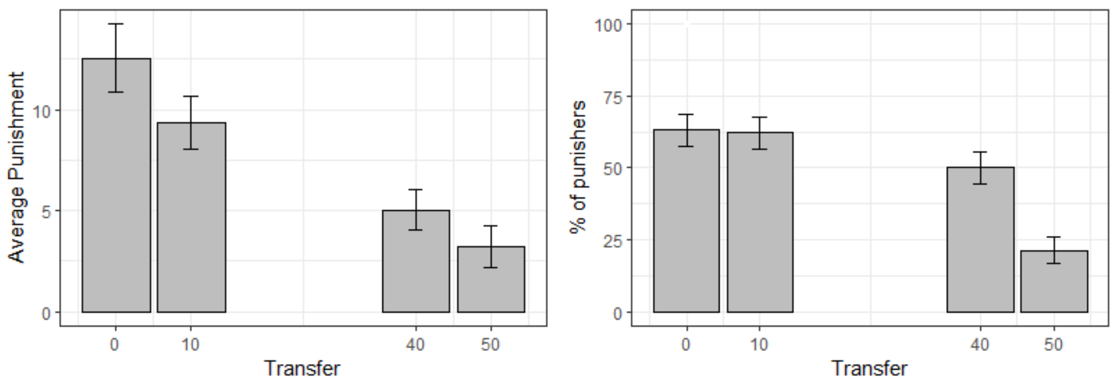
[26]We imposed the condition that each punishee would be twice in the role of dictator and twice in the role of receiver, to provide more equal payoffs for punishee participants.

[27]Section A was chosen to be payoff-relevant for exactly eight punishers, while Section B was chosen to be payoff-relevant for exactly two punishers.

4 Results

In this section, we present the results of our experiment. Figure 2 depicts the overall punishment and propensity to punish all possible transfers of the dictator game elicited via the strategy method. Subjects mainly punish dictators, who transfer 0, 10, and 40 CZK. The amount of punishment decreases with the more equal splits, indicating that subjects want to enforce more equal allocations. In our analysis, we focus on prosocial punishment and treat the transfers of 0, 10, and also of 40 as such. Approximately 60% of all subjects punished a dictator for low transfers of 0 and 10, and around 50% for a transfer of 40. Thus, we consider the punishment of 40 as prosocial as well, as any deviation from the equal split of 50 may be perceived as selfish and less social behavior.[28]

Figure 2: Average punishment decisions and frequency of punishment



Note: The left figure shows the average amount of deduction points and 95% confidence intervals for every transfer that dictators could choose. The right figure depicts how many percent of subjects decided to punish a particular transfer at all and 95% binomial confidence intervals by the normal approximation method.

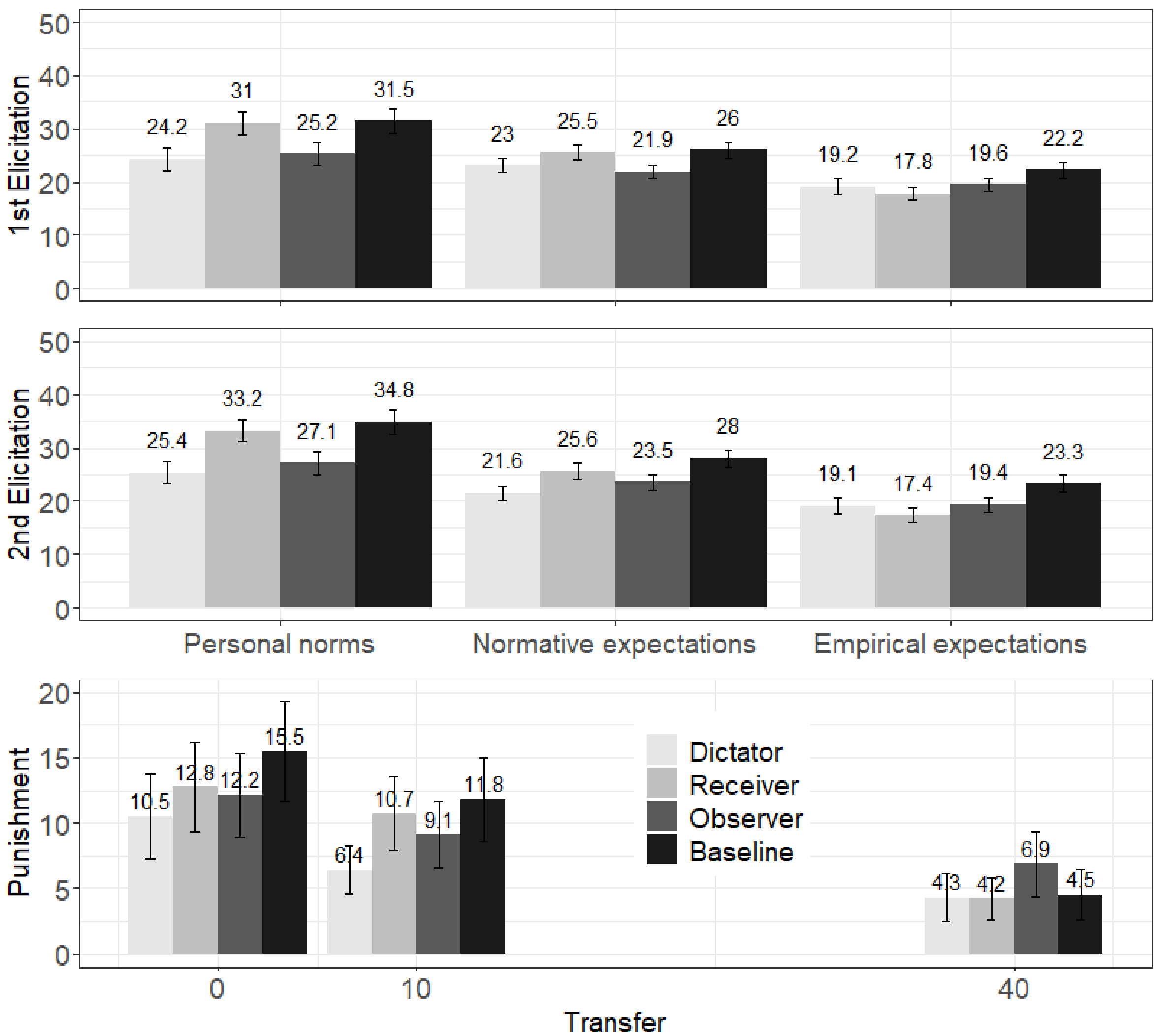
4.1 Social Norm Perceptions and Punishment

We use the treatment manipulation to increase heterogeneity among social norm perceptions by 1) assigning subjects to different roles and 2) receiving or observing different transfers. Figure 3 shows the effects of the treatments on individuals' social norm perceptions and third-party punishment levels. It reveals substantive variations in the average norm perceptions between the treatments for the first and second elicitation of norms. In addition, the treatments cause substantive variations in the distribution of the norm perceptions (see Figure 6 and Figure 7 in Appendix A.1.1). Moreover,

[28] Around 20% of subjects also punished a dictator for a transfer of 50. This punishment can be considered as antisocial, which follows a different motivation.

punishment decisions also vary across the treatments. What stands out from the figure is that when we compare the differences between the treatments, the punishment patterns closely resemble the patterns in social norm perceptions. This suggests a strong correlation which, in the following, we study more formally. First, we examine the patterns of social norm perceptions between and within subjects by pooling all treatments. Subsequently, we conduct a regression analysis. We focus on the second elicitation of norms, but all results can be replicated with the first elicitation of norms. Each individual is a statistically independent observation and hence our unit of analysis.

Figure 3: Social norm perceptions and punishment per treatment

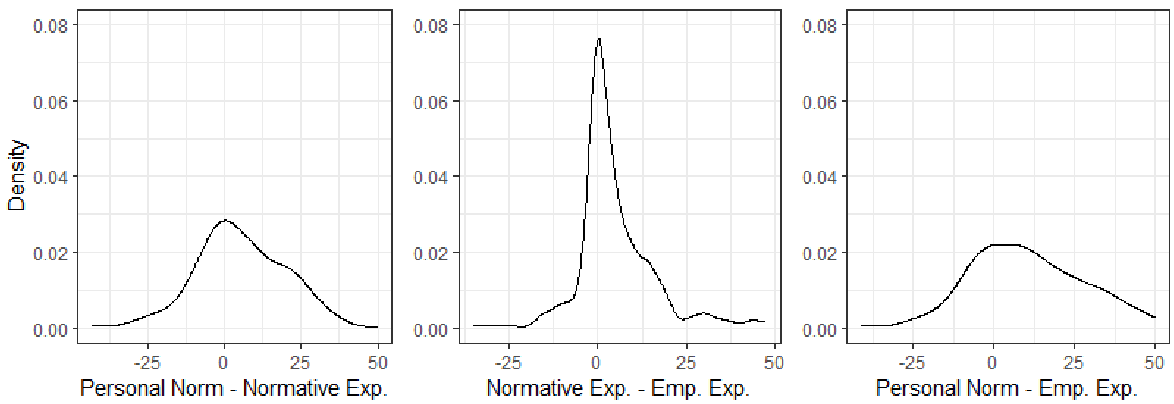


Note: The upper and middle panels show the three social norms perceptions conditional on the treatments of the first or second norm elicitation. The lower panel shows punishment decisions conditional on the transfer of the to-be-punished dictator and the treatments. Error bars show 95% confidence intervals.

We find that average personal norms (mean = 29.80, sd = 19.4) are significantly ($p < 0.001$) higher than normative expectations (mean = 24.44, sd = 12.9), which are significantly ($p < 0.001$) higher than empirical expectations (mean = 19.57, sd = 12.8).[29] That means that, on average, individuals believe that what they themselves perceive as the appropriate transfer is higher than what others, on average, deem appropriate. The expectation of what is usually done is even lower.

The differences in those three norm-related beliefs are substantial also within subjects. Figure 4 depicts the individual-level differences between all pairwise combinations of the three social norm perceptions. While a difference of 0 is the most frequent for all three combinations, most of the density mass is not at 0.[30] Most subjects hold higher personal norms than normative expectations, higher normative expectations than empirical expectations, and also higher personal norms than empirical expectations.[31]

Figure 4: Within-subject differences in norm perceptions



Note: The figures show the Kernel densities of the within-subject differences in all pairwise combinations of the three norm perceptions (second elicitation).

We find substantial differences between the three norm perceptions both aggregated and on a individual level. Thus, we are able to distinguish how each of the norm perceptions correlates with punishment. Now, we will examine whether punishment decisions are driven by what subjects believe should be done, what they believe others believe should be done, or what they think is usually done. We run regressions with all three norm perceptions and all subsets of combinations

[29] All significance tests are non-parametric Wilcoxon signed-rank tests.

[30] The high peek at 0 for the gap between normative and empirical expectations is mostly driven by the Dictator treatment.

[31] The treatments also caused substantially different distributions of those individual differences (see Figure 8 in Appendix A.1.1).

of the three as independent variables.[32] Table 1 shows the results of these Tobit regressions (see Table 3 of Appendix A.1 for the combinations of only two social norm perceptions).[33] In all models, we control for the dummy variables Transfer10 and Transfer40, which indicate the transfer of the to-be-punished dictator in the strategy method.[34] In addition, we control for negative emotions in all models to address potential endogeneity issues due to omitted variables, given that third-party punishment is known to correlate with negative emotions (Jordan et al. 2016, Carpenter & Matthews 2012, Nelissen & Zeelenberg 2009).[35] The results do not substantially change and remain statistically significant if we exclude negative emotions.

In models (1), (2), and (3), we regress punishment decisions on each norm perception individually. In model (4), we include all three norm perceptions simultaneously, as they are highly correlated with each other (personal norms and empirical expectations: $\rho = 0.448$, personal norms and normative expectations: $\rho = 0.642$, and normative expectations and empirical expectations: $\rho = 0.626$, Spearman correlation, $p < 0.001$ all). This is important because the correlation between one of the norm perceptions and punishment without controlling for the other two would pick up the explanatory power of the others. By including all three simultaneously, we can study their relative importance.[36]

In the first three models, we find a positive and significant correlation between punishment and each of the norm perceptions individually. This changes when we include all three norm perceptions simultaneously. In model (4), we observe that the correlation of personal norms and empirical expectations with punishment decisions remains positive and significant. In contrast, the significant positive relationship between normative expectations and punishment decisions reverses to a significant negative relationship.[37] In other words, if subjects hold higher appropriateness views, or believe that higher transfers are more common, they consistently punish more independent

[32] In the robustness checks, we can replicate all results, when we define punishment as a function of how dictators deviate from the respective norms.

[33] About 42% of all punishment decisions were 0, thus we use a Tobit model to account for such corner solutions

[34] Thus, the baseline is for the punishment for a transfer of 0.

[35] We declare a variable 'negative emotions' that consists of the average of anger, irritation, surprise, and envy (Cronbach's α is 0.69 ($CI_{95\%} = [0.66, 0.72]$) confirming the variable is internally consistent). We also include the differences in negative emotions of the second and first elicitation to capture emotional changes on the individual level caused by the Experience Phase.

[36] To deal with potential multicollinearity issues, we run regressions with pairwise subsets of all three norm perceptions in Appendix A.1.2 and run regressions with linear combinations of the variables in Section 4.2.

[37] In model (4), variance inflation factors are between 1.7 and 2.2 for all three norm perceptions. Hence, there is no indication that multicollinearity poses a serious problem for estimation.

Table 1: Tobit regression punishment on social norm perceptions

	Dependent Variable:			
	Punishment			
	(1)	(2)	(3)	(4)
Personal Norm	0.24^{***}			0.23^{**}
	(0.06)			(0.07)
Normative Expect.		0.19^{*}		-0.25^{*}
		(0.08)		(0.12)
Empirical Expect.			0.37^{***}	0.36^{**}
			(0.08)	(0.11)
Neg. Emotions	0.02	0.43	0.50	0.05
	(0.97)	(0.97)	(0.93)	(0.94)
Δ Neg. Emotions	2.11^{+}	2.06	2.46^{+}	2.45^{+}
	(1.26)	(1.35)	(1.36)	(1.26)
Transfer 10	-4.23^{***}	-4.19^{***}	-4.24^{***}	-4.26^{***}
	(0.84)	(0.83)	(0.83)	(0.84)
Transfer 40	-11.52^{***}	-11.45^{***}	-11.50^{***}	-11.56^{***}
	(1.29)	(1.28)	(1.28)	(1.29)
Constant	0.97	2.47	-0.32	0.29
	(3.28)	(3.72)	(3.51)	(3.59)
Observations	888	888	888	888
Log Likelihood	$-2{,}433.93$	$-2{,}450.44$	$-2{,}432.51$	$-2{,}419.70$
Wald Test	96.45^{***}	65.38^{***}	100.35^{***}	123.80^{***}
	(df = 5)	(df = 5)	(df = 5)	(df = 7)

Note: SE clustered at individual level. $+\ p < 0.1$; $*\ p < 0.05$; $**\ p < 0.01$; $***\ p < 0.001$

of the other norm perceptions. This is different for normative expectations: holding personal norms and empirical expectations fixed, individuals who believe others hold higher appropriateness views punish less. This indicates that normative expectations carry a different motivation for punishment than personal norms of appropriateness and empirical expectations. To explore this motivation further, we study how the effect of normative expectations depends on their relative position with the two other norm-related beliefs in Section 4.2.

The observed dynamics remain similar when we regress combinations of only two of the three social norm perceptions (see Table 3 in Appendix A.1.2). The coefficient of negative expectations, however, becomes close to zero and insignificant, when including only either personal norms and empirical expectations. The magnitude of the positive correlations between personal norms and empirical expectations with punishment is very similar across the models. This indicates that both matter for punishment decisions and explain a different part of the variation of punishment.[38]

We replicate our findings in all of the following robustness checks: First, we estimate the relationship between the first elicitation of norm perceptions and punishment decisions. Second, we estimate the relationship between both elicitations of norms and the propensity to punish. Third, we focus on deviations from the respective norms, i.e., the difference between the respective norm and the chosen transfer of the to-be-punished dictator. Fourth, we include the subject's choices of Section B of the experiment, where they themselves made a transfer decision as a dictator.[39] In all robustness checks, we can replicate the positive association of personal norms and empirical expectations with punishment and the negative association of normative expectations with punishment. Details can be found in Appendix A.1.3.

Lastly, we check whether our results are biased by the use of the strategy method. In particular, the relationship of empirical expectations with punishment could be biased because subjects might over-report punishment for transfers that they believe are unlikely to occur.[40] To check whether this is an issue, we run a robustness check (see Appendix A.1.4), where we declare a dummy that is one if the to-be-punished transfer is within a predefined neighborhood of an individual's empirical expectations. We find that empirical expectations are still significantly and positively related to punishment decisions after controlling for any of the neighborhood dummies. Hence, our results remain robust and the strategy method does not seem to affect our results substantially. Summarizing we find that:

[38] Note that we elicit personal norms on a discrete scale (0, 10, 40, 50) and empirical expectations on a continuous scale (0-50). This leads to a measurement error for personal norms. With a positive correlation between personal norms and empirical expectations, this measurement error leads to an underestimation of the correlation between personal norms and punishment and an overestimation of the importance of empirical expectations.

[39] We do not find any significant association between this transfer and their punishment decisions.

[40] With the strategy method, we elicit punishment decisions conditional on all possible transfers, but punishers only have to pay for the punishment of the actual transfer of the matched dictator. As a consequence, subjects with high empirical expectations might over-punish low transfers, as they do not expect to pay for this decision (and vice versa). This would lead to a stronger positive correlation between empirical expectations and third-party punishment.

Result 1 *Personal norms of appropriateness and empirical expectations are positively associated with punishment decisions.*

Result 2 *Normative expectations are negatively associated with punishment decisions.*

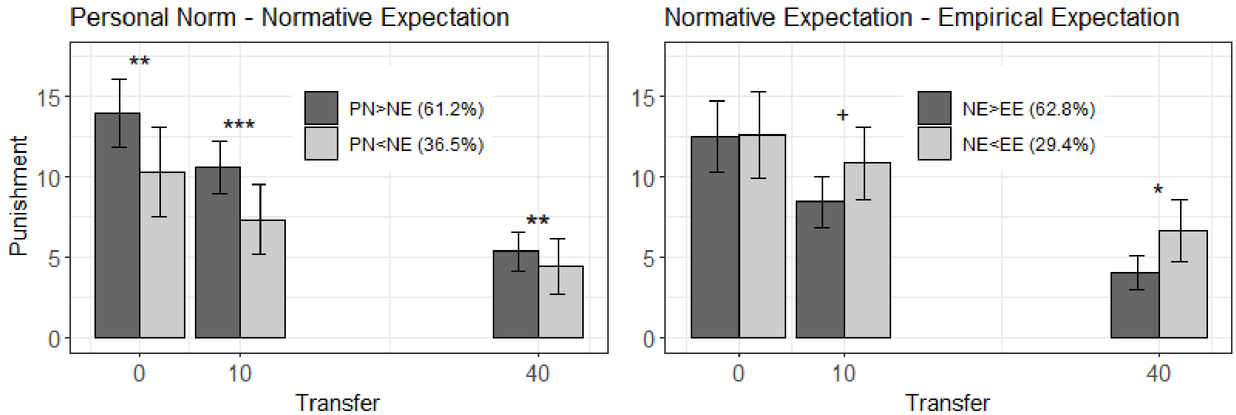
4.2 Normative Expectations Gaps and Punishment

So far, we discovered that, surprisingly, normative expectations (beliefs about what others deem appropriate) are not positively associated with higher punishment decisions. This indicates that subjects do not punish because they want to enforce what they believe society deems appropriate. On the contrary, when they believe society holds higher moral standards, they even punish less. In this section, we explore the rationale behind this behavior and whether individuals consider normative expectations in relation to the two other norm-related perceptions. In particular, we analyze how the differences between normative expectations and the two other norm perceptions relate to third-party punishment decisions.

In a regression analysis (see Appendix A.2), we find a significant positive association between punishment and the difference between personal norms and normative expectations ($Gap_{PN-NE} = PN - NE$) and a significant negative association between punishment and the difference between normative and empirical expectations ($Gap_{NE-EE} = NE - EE$). Figure 5 illustrates those punishment differences conditional on the sign of these gaps, i.e., conditional on whether participants hold higher or lower normative expectations compared to either their personal norms (left graph) or empirical expectations (right graph).

We start with describing the positive relationship between punishment and the gap between personal norms and normative expectations. A positive gap indicates that individuals believe that the appropriate transfer according to themselves is higher than what society views as appropriate. As depicted in Figure 5, subjects with higher personal norms relative to normative expectations punish more than those with higher normative expectations than personal norms. When personal norms are higher than normative expectations, subjects may feel a greater responsibility to punish, as they think societal standards are lower and others are less likely to intervene. Conversely, if normative expectations exceed personal norms, subjects may anticipate others to punish more and consequently free-ride on their punishment decisions. Although subjects in the experiment

Figure 5: Normative expectation gaps and punishment



Note: The left panel shows punishment decisions conditional on the transfer of the to-be-punished dictator for subjects with either higher (PN > NE) or lower (PN < NE) personal norms than normative expectations. The right panel shows these punishment decisions for subjects with either higher (NE > EE) or lower (NE < EE) normative than empirical expectations. The figure omits individuals with equal norm perceptions (PN=NE, 2.4%; NE=EE, 7.8%). All norms are based on the second elicitation. Error bars show 95% confidence intervals. Significance levels based on non-parametric Wilcoxon rank sum tests: + $p < 0.1$; **p* < 0.5; ** $p < 0.01$; *** $p < 0.001$

were not aware of other punishers, experiences in the real world could still induce an instinct of the importance of taking actions, depending on the relation between one's own and others' appropriateness views. In other words, subjects may instinctively feel compelled to take punitive actions if they hold higher normative standards than others, or conversely, may refrain from punitive action if they believe others will uphold higher moral standards.

Second, we explore the negative relationship between punishment and the gap between normative expectations and empirical expectations. If normative expectations differ from empirical expectations, individuals think that others do not behave according to the social norm of appropriateness. Most of the participants (approximately 63%) hold higher normative expectations than empirical expectations.[41] They believe that the transfer, which is considered socially appropriate, is higher than the transfer that is actually sent. They suppose that even though others hold high standards of behavior, they actually do not act like that, and this belief of disconformity leads to lower punishment. For almost 30% of the participants, normative expectations were lower than empirical expectations[42]. These individuals expect others to give more than what is socially ap-

[41]The respective averages of norm perceptions (for those with norm. exp > emp.exp.) are: pers. norm: 31.2, norm. exp.: 27.1, emp. exp.: 17.2

[42]The respective averages of norm perceptions (for those with norm. exp < emp.exp.) are: pers. norm: 28.3,

propriate and punish more than the rest. This could be driven by the fact that those individuals, on average, hold a personal norm that is even higher than their empirical expectations. The exact reasons behind such a combination of beliefs and why it leads to higher punishment, however, can only be speculated upon.

To conclude, normative expectations affect third-party punishment indirectly through its relative position to the other two norm perceptions. We find a strong and stable positive association between the gap of personal norm (what one approves of) and normative expectations (what one believes others approve of) with punishment. The association between punishment and the difference between normative and empirical expectations (what society approves of vs. what one believes is typically done) is less pronounced and seems to matter only for punishment of higher transfers (see Figure 5). Therefore, we conclude that the gap between personal norms and normative expectations matters for third-party punishment.

Result 3 *The gap between personal norms and normative expectations is associated positively with punishment decisions.*

4.3 Heterogeneity in Norm-driven Punishment

In addition, we find that the relative importance of the three norm perceptions for punishment decisions differs between subjects based on the role (treatment) they were assigned to and on gender. For instance, subjects in the role of receivers in the Experience Phase, seem to rely more on their empirical expectations when deciding on punishment compared to the rest of the sample. Subjects in the roles of dictators, observers, and subjects in the baseline hold qualitatively the same relative importance of the norms as shown in the main part (for more details, see Appendix A.4.1). Subjects in the role of dictators hold the lowest personal norms – possibly because of a motivated self-serving belief distortion – and punish significantly less than those in the Baseline treatment. Furthermore, we find substantial differences in the relative importance of the particular social norm perception for punishment between males and females. Males base their punishment decisions on their personal norms, whereas females rely on their empirical expectations (for more details, see Appendix A.4.2). Hence, it seems that males do not care about enforcing a social norm

norm. exp.: 19.6, emp. exp.: 24.2

but rather what they personally deem appropriate, in contrast to females, who seem to care about the enforcement of typical behavior.

4.4 Causality

In this section, we explore whether social norm perceptions *causally* affect punishment, and tackle the challenges of reverse causality and omitted variable bias. In principle, the punishment decision itself could shape social norm perceptions and thus create the challenge of reverse causality. For instance, punishers may want to provide a reason for the way they punish and thus report their social norm perceptions in line with their punishment choices. Additionally, other individual characteristics may be correlated both with the norm-related beliefs and punishment and thus create the challenge of omitted-variable bias.

We will now show that reverse causality and omitted variable bias seem to not play a major role in our analysis. Specifically, we compare social norm perceptions between the first and second elicitation, as well as between punishers and punishees. Additionally, we compare the size of the correlation of social norm perceptions with punishment between the first and second elicitation. Finally, we run an IV analysis and an additional robustness check with individual-level controls.

First, we compare the first and second elicitations of norm perceptions in the Baseline treatment, where subjects engage only in punishment between the two elicitations. Hence, any changes in the norm perceptions can be attributed to the Punishment Phase. We do not find significant differences for empirical expectations (from 22.22 to 23.25), but we do find statistically significant ($p < 0.05$) increases in personal norms (from 31.50 to 34.83) and normative expectations (from 25.99 to 28.03) between the two elicitations. When we condition these changes on the levels of punishment, these differences look very similar. Therefore, the Punishment Phase itself seems to slightly influence normative views (personal norms and normative expectations), i.e., it increases how much subjects think should be sent and how much they believe others think should be sent. However, it does not affect subjects' beliefs about typical behavior, i.e., empirical expectations. In the following paragraphs, we will show that reverse causality does not play a major role, as the effect of social norm perceptions on third-party punishment outweighs any effect from punishment on personal norms and normative expectations.

For that, we compare the correlations of norm perceptions with punishment between the norm elicitation before the punishment opportunity (first elicitation) and the norm elicitation after the punishment opportunity (second elicitation). We find that the relationships are in the same direction, similar in size, and at the same significance level for all norm perceptions with punishment between the first and second elicitation (all models reported in Appendix A.1, Table 4).[43]

Furthermore, we compare social norm perceptions of punishers and punishees to study whether the knowledge of the upcoming punishment opportunity changes social norm perceptions before the punishment opportunity. Punishees do not have this punishment opportunity, and thus, their norm perceptions are not influenced by this.[44] We find no significant differences between punishers and punishees norm perceptions, indicating that the knowledge about the upcoming punishment opportunity does not change norm perceptions (see Table 10 in Appendix A.3). This fact, together with the same correlations of the first and second norm elicitations with punishment, demonstrates that reverse causality likely does not play a major role.

In addition, we employ an instrumental variable approach. Here, we exploit the changes in norm perceptions caused by the Experience Phase and the treatment manipulation. We use the treatment assignment to the roles of dictator, receiver, and observer, and the received or observed transfers as instruments for all three norm perceptions. We find that the treatment manipulation (observing/ receiving a specific transfer or being assigned to the role of dictator) significantly shifts norm perceptions compared to the Baseline treatment (see Table 11 in Appendix A.3.2). We find the most prominent changes in empirical expectations, followed by normative expectations and personal norms. For a more formal description and discussion of the IV approach, see Appendix A.3.2. Table 2 shows the results of the second stage of the Tobit IV for all three norm perceptions. In model (1), we use personal norms as an instrumented regressor, in model (2), normative expectations, and in model (3), we instrument for empirical expectations. In model (4), we instrument for all three norm perceptions simultaneously. In all models, we additionally instrument for negative emotions and the change in negative emotions. We replicate all results with models without negative emotions and the change in negative emotions (see Table 12 in Appendix A.3.2).

Models (1), (2), and (3) show a statistically significant positive effect of each norm perception

[43]The same holds true for the associations between the first and second elicitation with a punishment dummy (if the punisher decided to deduct points at all).

[44]Punishees act half of the time as dictators and the other half as receivers. Therefore, their norm perceptions should be the least biased of all subjects.

Table 2: Tobit IV regression punishment on norm perceptions and negative emotions, second stage

	Dependent Variable:			
	Punishment			
	(1)	(2)	(3)	(4)
Personal Norms	0.62^{+} (0.36)			−0.14 (0.70)
Normative exp.		0.75^{*} (0.37)		0.01 (0.87)
Empirical exp.			0.80^{**} (0.30)	0.93 (0.59)
Negative Emotions	−3.93 (5.98)	−2.26 (5.76)	−3.53 (4.45)	−3.40 (5.20)
Δ Neg. Emotions	2.05 (9.20)	3.62 (8.59)	13.03^{+} (7.51)	14.70 (11.01)
Transfer10	-4.26^{***} (0.84)	-4.21^{***} (0.83)	-4.25^{***} (0.83)	-4.27^{***} (0.84)
Transfer40	-11.56^{***} (1.29)	-11.50^{***} (1.28)	-11.51^{***} (1.28)	-11.57^{***} (1.29)
Constant	0.21 (19.46)	−4.04 (18.94)	1.10 (12.56)	1.98 (16.36)
Observations	888	888	888	888
Log Likelihood	−8474.89	−8116.57	−8059.94	−14938.00
Wald χ^2 (df = 5)	88.33^{***}	90.29^{***}	90.04^{***}	90.76^{***}

Note: SE clustered at individual level. $+\ p < 0.1$; $^{*}\ p < 0.05$; $^{**}\ p < 0.01$; $^{***}\ p < 0.001$
Treatments conditional on transfer as instruments for empirical expectations, negative emotions, and Δ negative emotions

individually on punishment (weakly significant in the case of personal norms). As we show in section 4.1, it is important to control for all three norm perceptions when estimating their relative effect. In model (4), however, we do not find a statistically significant impact of any of the three norm perceptions when instrumenting for all of them simultaneously. The reason for the failure in identifying their relative causal impact together is that all norm perceptions are shifted in the same direction in each of the treatments (see Regression Table 11 in Appendix A.3.2). There is not enough induced variability between the three norm perceptions to disentangle the influence of each of them simultaneously in an instrumental variable regression.[45]

Finally, even though the IV analysis indicates that omitted-variable bias does not seem to play a role, we run an additional regression analysis with the following individual-level controls: age, gender, income, field of study, degree of understanding, and degree of concentration (see Appendix A.3.3). The relative correlations between the specific social norm perceptions with punishment remain robust to this additional model specification.

[45]It proves challenging to isolate the specific causal effects of each social norm perception in this IV framework (controlling for all), as information provision leads to an update of all of them in the same direction. Future research should come up with instruments that move only the targeted norm perception while keeping the others constant.

To conclude, given our analyses, we are confident that reverse causality and omitted-variable are rather unlikely to play a major role. Thus, we provide evidence that social norm perceptions affect third-party punishment *causally.*

5 Conclusion

In this paper, we show that perceptions of social norms matter for third-party punishment decisions. We explicitly measure three social norm perceptions about the behavior in a specific situation alongside punishment decisions, in contrast to the existing literature on social norms and punishment (e.g. Bašić & Verrina 2023, Dimant & Gesche 2023, Li et al. 2021, House et al. 2020, Reuben & Riedl 2013, Carpenter & Matthews 2012, 2009, Fehr & Fischbacher 2004). We employ four treatments that manipulate subjects' perceptions of social norms and induce differences among them. This allows us to speak to their relative importance and to identify their causal effects on third-party punishment.

We find a consistent positive effect of personal beliefs about what should be done (personal norms) and beliefs about what is usually done (empirical expectations) on third-party punishment. This means that subjects who hold higher moral standards, or believe that others typically behave more appropriately punish more. On the other hand, beliefs about what others believe should be done (normative expectations) are negatively correlated with punishment, when controlling for either personal norms or empirical expectations. This means that individuals punish less if they believe others to hold higher normative standards.

One explanation for this negative correlation could be that subjects anticipate interventions from others when they hold high normative standards, and thus would not have to intervene themselves. Conversely, when subjects believe others to hold lower normative standards, they anticipate fewer interventions and consequently feel compelled to enforce higher norms themselves. We provide evidence for this rationale, by finding that normative expectations matter in combination with own personal appropriateness views. Specifically, we find that subjects whose normative expectations are higher than their personal beliefs punish less, whereas subjects whose personal norms are higher than their normative expectations punish more. Our argument aligns with the findings of Kamei et al. (2023), who observe that in the presence of other punishers, subjects tend to free-ride on others' punishment decisions.

Overall, our results show that the desire to enforce own beliefs of appropriateness and typical behavior motivates third-party punishment rather than perceived societal appropriateness views. Beliefs about societal appropriateness views seem only to matter in combination with personal norms, and could be used to determine the necessity of one's own punishment decisions. These findings extend and align well with previous literature on third-party punishment and social norms. Unlike existing studies, we provide a more complete picture of how all three norm-related beliefs motivate punishment choices.

For instance, our findings extend Carpenter & Matthews (2012) who find a positive correlation between the belief about average behavior (empirical expectations) with third-party punishment and Carpenter & Matthews (2009) who find that the session's averages best explain punishment compared to own group's averages, in- or out-group averages, own contributions, or the respective medians, minima, or maxima. On the other hand, our finding that third-party punishment is not driven by the belief of the injunctive norms – normative expectations – is in contrast with previous literature, for example with Bašić & Verrina (2023), who show that normative expectations about punishment choices are positively correlated with punishment, or House et al. (2020) and Dimant & Gesche (2023), who show that injunctive norm nudges increase punishment decisions. Instead, we even find a negative correlation when controlling for personal norms or empirical expectations. However, it goes in line with literature that finds that empirical expectations are more important for economic behavior than normative expectations (Bicchieri et al. 2022, Chen et al. 2020, Schmidt 2019, Agerström et al. 2016, Bose et al. 2023, Bicchieri & Xiao 2009).

Our paper also adds to the literature about how normative and empirical information nudges affect third-party punishment. We find that our treatments shift social norm perceptions. For instance, we find that receiver and observers update their beliefs about common behavior depending on the dictator's transfer. Since appropriateness views are correlated with those beliefs, we also find an effect of the experienced transfers on normative expectations and personal norms of appropriateness. This fact can explain the results of House et al. (2020), Dimant & Gesche (2023), and Zong et al. (2021), who find that descriptive or injunctive norm nudges increase punishment decisions. Based on our results, the information about the descriptive or injunctive norm can change not only normative expectations but also empirical expectations and personal norms. Consequently, they impact third-party punishment decisions.

Our results also extend and possibly explain the results of Bašić & Verrina (2023), Kamei (2020), Lois & Wessa (2019), and Fabbri & Carbonara (2017) who show that punishment decisions are correlated with beliefs and information about others' punishment decisions. Others' punishment decisions may also inform the injunctive and descriptive norm of the situation itself. Hence, punishers' beliefs about others' punishment decisions may correlate with their empirical expectations of the situation itself and, through them, correlate with punishment decisions.

As additional results, we find that the reliance on a specific norm-related perception depends on gender. Males punish primarily based on what they personally believe constitutes appropriate behavior. Females, on the other hand, primarily punish according to what they believe constitutes common behavior. The stronger reliance on empirical expectations for females contrasts with Croson et al. (2010) who find that males rely more on their empirical expectations compared to females when donating money. On the other hand, Fišar et al. (2016) do not find any gender differences in the relationship between empirical expectations and third-party punishment decisions in their study of bribing.

Furthermore, we also show that the importance of a specific norm perception and punishment depends on the mere assignment to a role. For example, being assigned to the role of receiver leads to a higher reliance on empirical expectations compared to the rest of the sample. Additionally, being assigned to the role of dictator leads to significantly lower social norm perceptions and, consequently, to lower punishment. This motivated shift of social norm perceptions is in accordance with the literature on motivated beliefs (e.g. Bicchieri, Dimant & Sonderegger 2023, Zimmermann 2020, Epley & Gilovich 2016).

We can draw policy recommendations from our results. Policies that are aimed at changing empirical expectations rather than normative expectations have a higher potential to change third-party punishment decisions. For instance, providing information about common behavior instead of what others deem appropriate may influence empirical expectations more than normative expectations and hence have a higher potential to shift punishment behavior. This would align with Dimant & Gesche (2023), who show that empirical information changes behavior more than normative information (although non-significantly). Additionally, as the reliance on either personal norms or empirical expectations depends on gender and the role that subjects are assigned to, specific information policies should be tailored to the needs of the specific audience in order to increase its effectiveness.

To conclude, we provide consistent evidence that social norm perceptions motivate third-party punishment. Individuals, who hold higher personal appropriateness views and believe that others behave more appropriately, punish more. On the other hand, subjects who believe others to hold higher normative views, punish less. In addition, we find that the initial positive correlation between normative expectations and punishment reverses when controlling for either of the other two norm perceptions. This has important consequences for the overall social norm literature. We show that it is important to consider all norm-related beliefs because otherwise, the effect of one of them might be wrongly attributed to another.

Acknowledgements

We are grateful to Rostislav Staněk, Henrik Orzen, Wladislaw Mill, Marco Faillo, James Tremewan, Ondřej Krčál, Aleksandra Khokhlova, and Dam Thi Anh for their valuable comments. We appreciate comments from participants of the seminars of Ca'Foscari University of Venice, the University of Verona, and the CEEL University of Trento, the ZEW/Uni Mannheim Experimental Seminar, the CDSE Seminar in Mannheim, the International Online Conference RExCon21 on 'Social Preferences and Social Norms', the ESA GOACM 2021, and the Summer School of Behavioural Game Theory in Norwich. The financial support of the grants MUNI/A/0931/2019 and MUNI/IGA/1364/2020 is gratefully acknowledged. This work was supported by the University of Mannheim's Graduate School of Economic and Social Sciences and by the Faculty of Economics and Administration at Masaryk University.

References

Abeler, J., Nosenzo, D. & Raymond, C. (2019), 'Preferences for truth-telling', *Econometrica* **87**(4), 1115–1153.

Agerström, J., Carlsson, R., Nicklasson, L. & Guntell, L. (2016), 'Using descriptive social norms to increase charitable giving: The power of local norms', *Journal of Economic Psychology* **52**, 147–153.

Andre, P., Boneva, T., Chopra, F. & Falk, A. (2024), 'Globally representative evidence on the actual and perceived support for climate action', *Nature Climate Change* pp. 1–7.

Andreoni, J. & Bernheim, B. D. (2009), 'Social image and the 50–50 norm: A theoretical and experimental analysis of audience effects', *Econometrica* **77**(5), 1607–1636.

Arechar, A. A., Gächter, S. & Molleman, L. (2018), 'Conducting interactive experiments online', *Experimental economics* **21**(1), 99–131.

Bašić, Z. & Verrina, E. (2023), 'Personal norms—and not only social norms—shape economic behavior', *MPI Collective Goods Discussion Paper* (2020/25).

Bénabou, R. & Tirole, J. (2006), 'Incentives and prosocial behavior', *American economic review* **96**(5), 1652–1678.

Bicchieri, C. (2016), *Norms in the wild: How to diagnose, measure, and change social norms*, Oxford University Press.

Bicchieri, C., Dimant, E., Gächter, S. & Nosenzo, D. (2022), 'Social proximity and the erosion of norm compliance', *Games and Economic Behavior* **132**, 59–72.

Bicchieri, C., Dimant, E., Gelfand, M. & Sonderegger, S. (2023), 'Social norms and behavior change: The interdisciplinary research frontier', *Journal of Economic Behavior Organization* **205**, A4–A7.

Bicchieri, C., Dimant, E. & Sonderegger, S. (2023), 'It's not a lie if you believe the norm does not apply: Conditional norm-following and belief distortion', *Games and Economic Behavior* **138**, 321–354.

Bicchieri, C. & Xiao, E. (2009), 'Do the right thing: but only if others do so', *Journal of Behavioral Decision Making* **22**(2), 191–208.

Bose, G., Dechter, E. & Ivancic, L. (2023), 'Conformity and adaptation in groups', *Journal of Economic Behavior & Organization* **212**, 1267–1285.

Bosman, R. & Van Winden, F. (2002), 'Emotional hazard in a power-to-take experiment', *The Economic Journal* **112**(476), 147–169.

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

Carpenter, J. P. & Matthews, P. H. (2009), 'What norms trigger punishment?', *Experimental Economics* **12**(3), 272–288.

Carpenter, J. P. & Matthews, P. H. (2012), 'Norm enforcement: anger, indignation, or reciprocity?', *Journal of the European Economic Association* **10**(3), 555–572.

Carpenter, J. P., Matthews, P. H. & Ong'Ong'a, O. (2004), 'Why punish? social reciprocity and the enforcement of prosocial norms', *Journal of evolutionary economics* **14**, 407–429.

Charness, G., Cobo-Reyes, R. & Jiménez, N. (2008), 'An investment game with third-party intervention', *Journal of Economic Behavior & Organization* **68**(1), 18–28.

Chen, H., Zeng, Z. & Ma, J. (2020), 'The source of punishment matters: Third-party punishment restrains observers from selfish behaviors better than does second-party punishment by shaping norm perceptions', *Plos one* **15**(3), e0229510.

Cialdini, R. B., Reno, R. R. & Kallgren, C. A. (1990), 'A focus theory of normative conduct: recycling the concept of norms to reduce littering in public places.', *Journal of personality and social psychology* **58**(6), 1015.

Croson, R. T., Handy, F. & Shang, J. (2010), 'Gendered giving: the influence of social norms on the donation behavior of men and women', *International Journal of Nonprofit and Voluntary Sector Marketing* **15**(2), 199–213.

Cubitt, R. P., Drouvelis, M. & Gächter, S. (2011), 'Framing and free riding: emotional responses and punishment in social dilemma games', *Experimental Economics* **14**(2), 254–272.

d'Adda, G., Drouvelis, M. & Nosenzo, D. (2016), 'Norm elicitation in within-subject designs: Testing for order effects', *Journal of Behavioral and Experimental Economics* **62**, 1–7.

Danilov, A. & Sliwka, D. (2017), 'Can contracts signal social norms? experimental evidence', *Management Science* **63**(2), 459–476.

Dimant, E. & Gesche, T. (2023), 'Nudging enforcers: How norm perceptions and motives for lying shape sanctions', *PNAS nexus* **2**(7), pgad224.

Duch, M. L., Grossmann, M. R. & Lauer, T. (2020), 'z-tree unleashed: A novel client-integrating architecture for conducting z-tree experiments over the internet', *Journal of Behavioral and Experimental Finance* **28**, 100400.

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

Epley, N. & Gilovich, T. (2016), 'The mechanics of motivated reasoning', *Journal of Economic perspectives* **30**(3), 133–140.

Fabbri, M. & Carbonara, E. (2017), 'Social influence on third-party punishment: An experiment', *Journal of Economic Psychology* **62**, 204–230.

Fehr, E. & Fischbacher, U. (2004), 'Third-party punishment and social norms', *Evolution and human behavior* **25**(2), 63–87.

Fišar, M., Kubák, M., Špalek, J. & Tremewan, J. (2016), 'Gender differences in beliefs and actions in a framed corruption experiment', *Journal of Behavioral and Experimental Economics* **63**, 69–82.

Fischbacher, U. (2007), 'z-tree: Zurich toolbox for ready-made economic experiments', *Experimental economics* **10**(2), 171–178.

Gino, F., Ayal, S. & Ariely, D. (2009), 'Contagion and differentiation in unethical behavior: The effect of one bad apple on the barrel', *Psychological science* **20**(3), 393–398.

Henrich, J., McElreath, R., Barr, A., Ensminger, J., Barrett, C., Bolyanatz, A., Cardenas, J. C., Gurven, M., Gwako, E., Henrich, N. et al. (2006), 'Costly punishment across human societies', *Science* **312**(5781), 1767–1770.

Hoeft, L., Mill, W. & Vostroknutov, A. (2023), 'Normative perception of power abuse', *MPI Collective Goods Discussion Paper* (2019/6).

House, B. R., Kanngiesser, P., Barrett, H. C., Yilmaz, S., Smith, A. M., Sebastian-Enesco, C., Erut, A. & Silk, J. B. (2020), 'Social norms and cultural diversity in the development of third-party punishment', *Proceedings of the Royal Society B* **287**(1925), 20192794.

Jordan, J., McAuliffe, K. & Rand, D. (2016), 'The effects of endowment size and strategy method on third party punishment', *Experimental Economics* **19**(4), 741–763.

Kamei, K. (2018), 'The role of visibility on third party punishment actions for the enforcement of social norms', *Economics letters* **171**, 193–197.

Kamei, K. (2020), 'Group size effect and over-punishment in the case of third party enforcement of social norms', *Journal of Economic Behavior & Organization* **175**, 395–412.

Kamei, K., Sharma, S. & Walker, M. J. (2023), 'Sanction enforcement among third parties: New experimental evidence from two societies', *Available at SSRN 4429802* .

Keizer, K., Lindenberg, S. & Steg, L. (2008), 'The spreading of disorder', *Science* **322**(5908), 1681–1685.

Kessler, J. B. & Leider, S. (2012), 'Norms and contracting', *Management Science* **58**(1), 62–77.

Kölle, F. & Quercia, S. (2021), 'The influence of empirical and normative expectations on cooperation', *Journal of Economic Behavior & Organization* **190**, 691–703.

Kromer, E. & Bahçekapili, H. G. (2010), 'The influence of cooperative environment and gender on economic decisions in a third party punishment game', *Procedia-Social and Behavioral Sciences* **5**, 250–254.

Krysowski, E. & Tremewan, J. (2021), 'Why does anonymity make us misbehave: Different norms or less compliance?', *Economic Inquiry* **59**(2), 776–789.

Leibbrandt, A. & López-Pérez, R. (2012), ‘An exploration of third and second party punishment in ten simple games’, *Journal of Economic Behavior & Organization* **84**(3), 753–766.

Lergetporer, P., Angerer, S., Glätzle-Rützler, D. & Sutter, M. (2014), ‘Third-party punishment increases cooperation in children through (misaligned) expectations and conditional cooperation’, *Proceedings of the National Academy of Sciences* **111**(19), 6916–6921.

Li, X., Molleman, L. & van Dolder, D. (2021), ‘Do descriptive social norms drive peer punishment? conditional punishment strategies and their impact on cooperation’, *Evolution and Human Behavior* **42**(5), 469–479.

Lois, G. & Wessa, M. (2019), ‘Creating sanctioning norms in the lab: The influence of descriptive norms in third-party punishment’, *Social Influence* **14**(2), 50–63.

Martin, J. W., Martin, S. & McAuliffe, K. (2021), ‘Third-party punishment promotes fairness in children.’, *Developmental Psychology* **57**(6), 927.

Mathew, S. & Boyd, R. (2011), ‘Punishment sustains large-scale cooperation in prestate warfare’, *Proceedings of the National Academy of Sciences* **108**(28), 11375–11380.

McAuliffe, K., Jordan, J. J. & Warneken, F. (2015), ‘Costly third-party punishment in young children’, *Cognition* **134**, 1–10.

Merguei, N., Strobel, M. & Vostroknutov, A. (2022), ‘Moral opportunism as a consequence of decision making under uncertainty’, *Journal of Economic Behavior & Organization* **197**, 624–642.

Nelissen, R. M. & Zeelenberg, M. (2009), ‘Moral emotions as determinants of third-party punishment: Anger, guilt and the functions of altruistic sanctions’, *Judgment and Decision making* **4**(7), 543.

Piardini, P., Drouvelis, M. & Di Cagno, D. (2017), ‘Gender effects and third-party punishment in social dilemma games’.

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

Schmidt, R. J. (2019), Do injunctive or descriptive social norms elicited using coordination games better explain social preferences?, Technical report, Discussion Paper Series.

Tremewan, J. & Vostroknutov, A. (2021), An informational framework for studying social norms, *in* 'A research agenda for experimental economics', Edward Elgar Publishing, pp. 19–42.

Zimmermann, F. (2020), 'The dynamics of motivated beliefs', *American Economic Review* **110**(2), 337–363.

Zong, J., De Jong, E., Qiu, J. & Li, J. (2021), 'Socially appropriate intervention: A cross-country investigation of third-party norm enforcement', *Available at SSRN 3943578* .

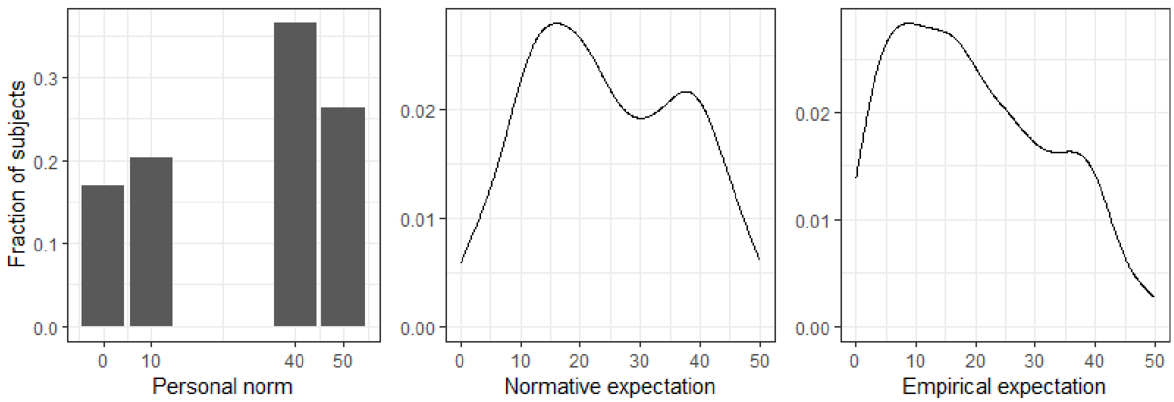
A Appendix

A.1 Social Norm Perceptions and Punishment

A.1.1 Distribution of Norm Perceptions

Figure 6 shows the distribution of each norm perception at the second elicitation. It reveals that there is no clear consensus between subjects in all of the three norm perceptions.

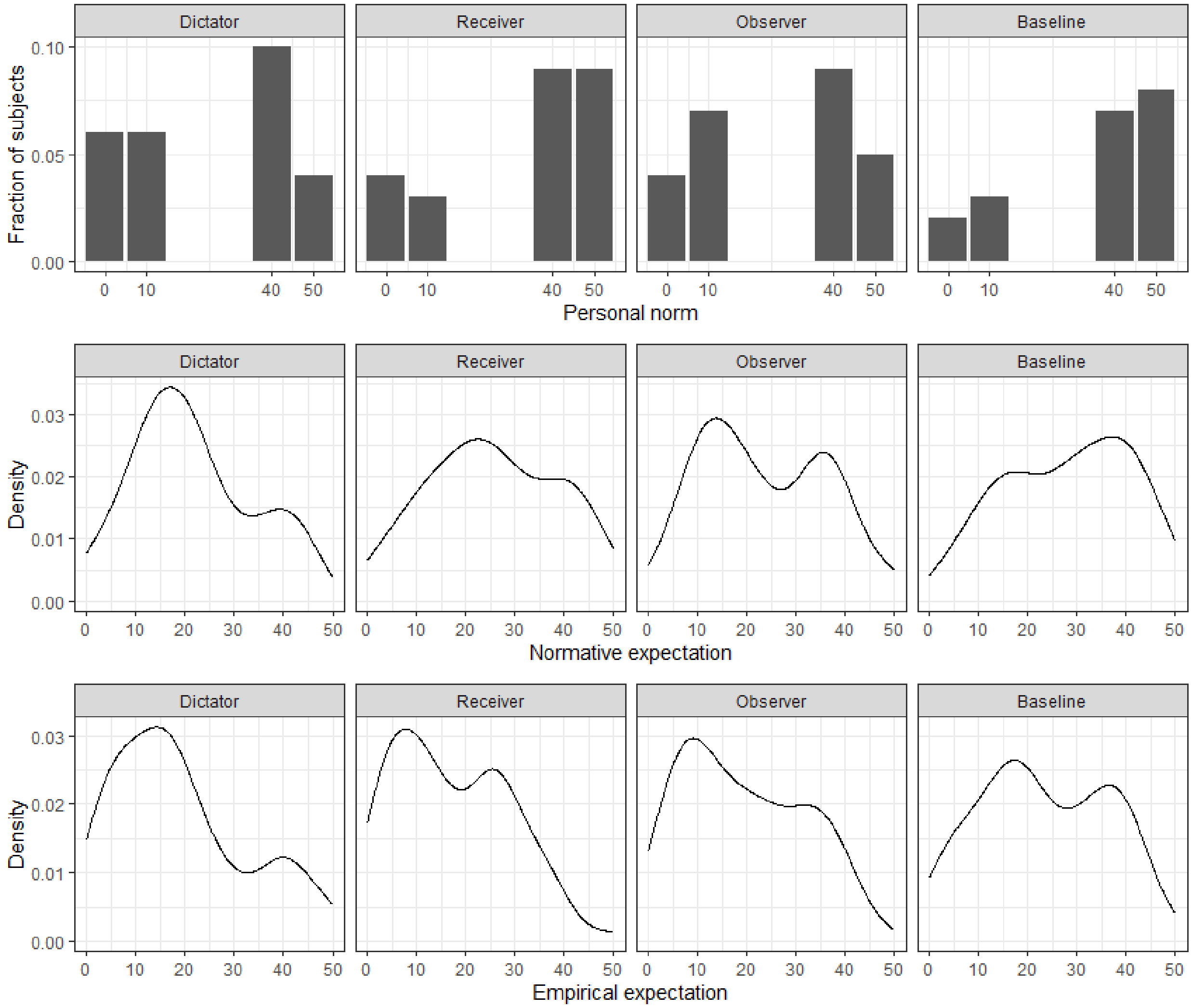
Figure 6: Distributions of social norm perceptions in second elicitation



Note: The left plot shows the fractions of subjects who hold a particular personal norm. The plots in the middle and on the right show the distribution of normative and empirical expectations via Kernel densities.

Figure 7 shows the distribution of each norm perception at the second elicitation conditional on the treatments. It shows that the treatment assignment leads to differences in the distributions of all three norm perceptions.

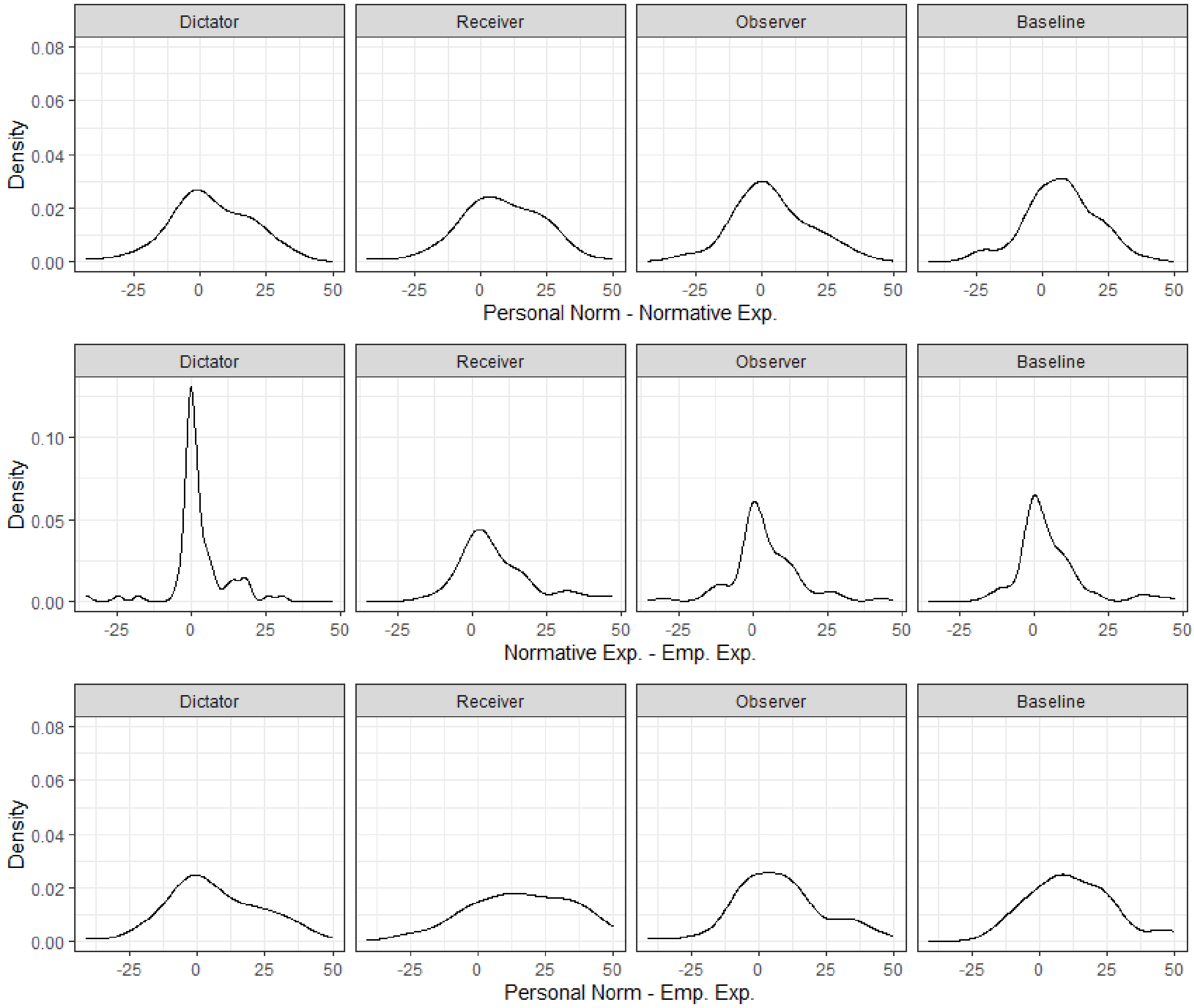
Figure 7: Distributions of social norm perceptions in second elicitation per treatment



Note: The upper panels show the fractions of subjects who hold a particular personal norm conditional on the treatments. The middle and lower panels show the distribution of normative and empirical expectations via Kernel densities conditional on the treatments.

Figure 8 shows the distributions of within-subject differences in the norm perceptions per treatment. It shows that the treatment assignment leads to differences in those distributions.

Figure 8: Within-subject differences in norm perceptions per treatment



Note: The panels show the Kernel densities of the within-subject differences in the three norm perceptions (second elicitation) for all treatments separately.

A.1.2 Combinations of Norm Perceptions and Punishment

In this section, we regress punishment on combinations of social norm perceptions and punishment, where we take only two out of the three social norm perceptions (see Table 3). We replicate the results from the main text. Empirical expectations are significantly positively correlated with punishment, personal norms are significantly positively correlated, and normative expectations are (however non-significantly) negatively correlated with punishment.

Table 3: Tobit regression punishment on social norm perception combinations

	Dependent Variable:		
	Punishment		
	(1)	(2)	(3)
Personal Norm	0.27^{***}	0.16^{*}	
	(0.07)	(0.07)	
Normative Expect.	-0.06		-0.07
	(0.10)		(0.11)
Empirical Expect.		0.26^{**}	0.41^{***}
		(0.10)	(0.11)
Constant	1.89	-2.18	0.56
	(3.63)	(3.50)	(3.65)
Transfer	✓	✓	✓
Neg. Emotions	✓	✓	✓
Δ Neg. Emotions	✓	✓	✓
Observations	888	888	888
Log Likelihood	$-2{,}433.54$	$-2{,}424.92$	$-2{,}432.03$
Wald Test (df = 6)	97.09^{***}	114.17^{***}	101.17^{***}

Note: SE clustered at individual level. $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$

A.1.3 Robustness Norm Perceptions and Punishment

Next, we replicate the overall association between norm perceptions and punishment decisions. In Table 4, in model (1), we regress punishment on the first elicitation of norm perceptions and in model (2) on the second elicitation. In models (3) and (4), we regress a punishment dummy on the first and second elicitation of norm perceptions. All models replicate the results from the main section: personal norms and empirical expectations are significantly positively associated with punishment decisions, while normative expectations are significantly negatively associated.

Table 4: Tobit and Logit regression punishment on norm perceptions

	Dependent Variable:			
	Punishment		*Punishment Dummy*	
	Tobit		*Logistic*	
	(1)	(2)	(3)	(4)
	1st elicitation	2nd elicitation	1st elicitation	2nd elicitation
Personal Norm	0.18^{*}	0.23^{**}	0.02^{*}	0.03^{**}
	(0.08)	(0.07)	(0.08)	(0.07)
Normative Expect.	-0.36^{*}	-0.25^{*}	-0.04^{*}	-0.04^{*}
	(0.14)	(0.12)	(0.14)	(0.12)
Empirical Expect.	0.44^{***}	0.36^{**}	0.05^{***}	0.04^{**}
	(0.13)	(0.11)	(0.13)	(0.11)
Constant	2.07	0.29	-0.22	-0.24
	(3.76)	(3.59)	(3.76)	(3.59)
Transfer	✓	✓	✓	✓
Neg. Emotions	✓	✓	✓	✓
Δ Neg. Emotions	✓	✓	✓	✓
Observations	888	888	888	888
Log Likelihood	$-2{,}425.14$	$-2{,}419.70$	-558.21	-545.02
Akaike Inf. Crit.			1,132.41	1,106.03
Wald Test (df = 7)	113.88^{***}	123.80^{***}		

Note: SE clustered at individual level. $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$

Next, we regress punishment on the deviations of the transfer to the norm perceptions. For this, we take the difference of each of the three norm perceptions to the to-be-punished transfer. Table 5 reveals that the results remain robust to this model specification.

Finally, we include the subjects' own transfer decisions as a dictator in Section B of the experiment. Model (1) in Table 6 reveals that the transfer in Section B is not significantly correlated with punishment decisions. Model (2) adds norm perceptions and shows that even after controlling for the transfers in Section B, the results from the main text prevail: the correlations between norm perceptions and punishment do not substantially change after including the own transfer.

Table 5: Tobit regression punishment on deviations of norm perceptions from transfers

	Dependent Variable:			
	Punishment			
	(1)	(2)	(3)	(4)
Dev Personal Norms	0.26***			0.23**
	(0.04)			(0.07)
Dev Normative Expect.		0.25***		−0.27*
		(0.04)		(0.11)
Dev Empirical Expect.			0.31***	0.34**
			(0.04)	(0.11)
Constant	−0.46	−0.12	0.91	0.97
	(2.81)	(2.87)	(2.76)	(2.70)
Neg. Emotions	✓	✓	✓	✓
Δ Neg. Emotions	✓	✓	✓	✓
Observations	888	888	888	888
Log Likelihood	−2,434.61	−2,451.65	−2,433.95	−2,420.52
Wald Test	94.68***	62.68***	97.52***	122.10***
	(df = 3)	(df = 3)	(df = 3)	(df = 5)

Note: SE clustered at individual level. + $p < 0.1$; * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$

A.1.4 Interaction Empirical Expectations and Strategy Method

In this section, we check whether subjects decrease their punishment for a transfer in the strategy method, which they believe is more likely to occur. We declare the dummy variables Neighborhood5, Neighborhood15, and Neighborhood25, which take the value 1 if the to-be-punished transfer in the strategy method is within the distance of 5 CZK, 15 CZK, or 25 CZK from a subject's empirical expectations, respectively. Table 7 shows Tobit regressions, which include norm perceptions and neighborhood dummies. Models (4) and (5) include an interaction of the neighborhood variable and the Transfer dummies. We allow for such interaction because the effect of the neighborhood variable likely differs depending on the to-be-punished transfer. Specifically, the decrease in punishment may be less pronounced for a to-be-punished transfer of 40 because the initial punishment level is lower already compared to 0 or 10.

All models show a negative association between neighborhood variables and punishment deci-

Table 6: Tobit regression punishment on own transfer in Section B and norm perceptions

	Dependent Variable:	
	Punishment	
	(1)	(2)
Transfer Section B	-0.04	-0.04
	(0.05)	(0.05)
Personal Norm		0.23^{**}
		(0.07)
Normative Expect.		-0.24^{*}
		(0.12)
Empirical Expect.		0.36^{**}
		(0.11)
Constant	7.74^{**}	0.32
	(2.94)	(3.57)
Transfer	✓	✓
Neg. Emotions	✓	✓
Δ Neg. Emotions	✓	✓
Observations	888	888
Log Likelihood	$-2{,}456.35$	$-2{,}418.92$
Wald Test	53.52^{***}	125.08^{***}
	(df = 5)	(df = 8)

Note: SE clustered at individual level. $+\ p < 0.1$; * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$

sions, but only in model (2) the association is significant. This indicates that subjects seem to slightly decrease their punishment in the proximity of their empirical expectations. However, the neighborhood variable may simply capture higher punishment of a transfer of 0, 10, especially when not including the interaction with the Transfer dummies.

Most importantly, the positive significant association between personal norms and empirical expectations and punishment and the negative association with normative expectations (the results from the main text) remain robust and significant in all models. Hence, even though subjects might punish less severely in close proximity to their empirical expectations, empirical expectations still significantly explain punishment behavior.

Table 7: Tobit regression punishment on norm perceptions and neighborhood dummy

	Dependent Variable:				
	Punishment				
	(1)	(2)	(3)	(4)	(5)
Neighborhood5	-1.92			-2.33	
	(1.40)			(3.78)	
Neighborhood15		-2.25^{*}			-0.42
		(1.05)			(3.04)
Neighborhood25			-1.48		
			(1.02)		
Personal Norms	0.23^{**}	0.23^{**}	0.24^{**}	0.23^{**}	0.23^{**}
	(0.07)	(0.07)	(0.07)	(0.07)	(0.07)
Normative Expect.	-0.25^{*}	-0.25^{*}	-0.25^{*}	-0.23^{+}	-0.23^{+}
	(0.12)	(0.12)	(0.12)	(0.12)	(0.12)
Empirical Expect.	0.33^{**}	0.31^{**}	0.32^{**}	0.37^{**}	0.43^{**}
	(0.11)	(0.11)	(0.11)	(0.12)	(0.14)
Constant	1.29	2.25	1.96	0.24	-1.01
	(2.39)	(2.39)	(2.39)	(2.61)	(3.59)
Transfer	✓	✓	✓	✓	✓
Transfer:Neighborhood5	×	×	×	✓	×
Transfer:Neighborhood15	×	×	×	×	✓
Observations	888	888	888	888	888
Log Likelihood	$-2{,}423.00$	$-2{,}422.38$	$-2{,}423.07$	$-2{,}420.82$	$-2{,}420.64$
Wald Test	118.11^{***}	119.65^{***}	118.20^{***}	122.85^{***}	123.05^{***}
	(df = 6)	(df = 6)	(df = 6)	(df = 8)	(df = 8)

Note: SE clustered at individual level. $+\ p < 0.1$; $*\ p < 0.05$; $**\ p < 0.01$; $***\ p < 0.001$

A.2 Normative Expectations Gaps and Punishment

In this section, we run regressions, which estimate the correlations between third-party punishment and the differences between all pair-wise norm perceptions. Table 8 shows the results of these regressions. In model (1), we find a significant positive association between punishment and the difference between personal norms and normative expectations. In model (2), we find a negative association between punishment and the difference between normative and empirical expectations.

In model (3), we do not find any significant association between the difference between personal norms and empirical expectations with punishment. Finally, in model (4), we include both gaps PN-NE and NE-EE to analyze which relationship is more prevalent and stable. We find that the relationship of the gap NE-EE with punishment diminishes, while the significant positive association between punishment and the difference between personal norms and normative expectations prevails.

Table 8: Tobit regression punishment on normative expectation gaps

	Dependent Variable:			
	Punishment			
	(1)	(2)	(3)	(4)
Gap PN-NE	0.26***			0.24**
	(0.07)			(0.07)
Gap NE-EE		−0.24*		−0.19+
		(0.11)		(0.11)
Gap PN-EE			0.10	
			(0.07)	
Constant	7.41*	8.86**	7.02*	8.42**
	(2.90)	(2.91)	(2.96)	(2.85)
Transfer	✓	✓	✓	✓
Neg. Emotions	✓	✓	✓	✓
Δ Neg. Emotions	✓	✓	✓	✓
Observations	888	888	888	888
Log Likelihood	−2,441.13	−2,449.52	−2,454.08	−2,436.64
Wald Test	82.48***	66.68***	57.87***	91.11***

Note: SE clustered at individual level. $+p < 0.1; *p < 0.05; **p < 0.01; ***p < 0.001$

A.3 Causality

In this section, we provide the details of our analyses concerning the causal effect of social norm perceptions on third-party punishment. We first provide details on our analyses for reverse causality, then describe the IV analysis, and lastly provide details on our analysis for omitted-variable bias.

A.3.1 Reverse Causality

Table 9 shows the differences between the first and second elicitation of norm perceptions in the Baseline. It reveals statistically significant higher personal norms and normative expectations in the second elicitation than in the first. Similarly, empirical expectations are higher in the second elicitation however not at a statistically significant level.

Table 9: Comparisons of first and second norm perceptions elicitations in the Baseline treatment

Means (SE)	1st Elicitation	2nd Elicitation	p-value
Personal Norms	31.50 (2.36)	34.83 (2.33)	0.034
Normative Expectations	25.99 (1.59)	28.03 (1.66)	0.034
Empirical Expectations	22.22 (1.60)	23.25 (1.66)	0.915

Note: Non-parametric wilcoxon signed-rank test; N=60

Second, we compare the coefficients of the correlations between first and second norm perceptions elicitation. Table 4 (Section A.1.3) replicates the findings of the main text. We find the same and significant relationships between all norm perceptions with punishment when looking at the first elicitation, i.e., the elicitations before the punishment (and experience) phase with a slightly larger negative coefficient for normative expectations and a lower positive coefficient for personal norms, but similar coefficients for empirical expectations. In addition, the coefficients (and significance levels) in the estimation of whether the punisher decided to punish (*punishment dummy*) are very similar between the first and second elicitation.

Third, we compare the first elicitation of norm perceptions between punishers and punishees. As Table 10 depicts, there are no significant differences between the norm perceptions of punishers and punishees.

Table 10: Comparisons of norm perceptions (first elicitation) between punishers and punishees

Means (SE)	Punishers *(N=296)*	Punishees *(N=119)*	p-value
Personal Norms	27.74 (1.15)	28.74 (1.85)	0.585
Normative Expectations	23.97 (0.70)	24.34 (1.19)	0.724
Empirical Expectations	19.53 (0.70)	20.66 (1.15)	0.492

Note: Non-parametric wilcoxon rank sum test

A.3.2 Instrumental Variable Approach

We employ a Tobit IV regression to estimate the causal influence of norm perceptions on punishment decisions. We use the Receiver and Observer treatments conditional on the transfers (Receive 0, ..., Receive 50, Observe 0, ..., Observe 50), as well as the assignment to the role of dictator (Dictator) as instruments for the second elicitation of each norm perception separately as well as for negative emotions. Specifically, we use the following model specifications:

First stage:

$$y_i = treatments_i \Pi_1 + transfer_i \Pi_2 + \nu_i \tag{1}$$

Second stage:

$$punishment_i^* = y_i \beta + transfer_i \delta + \epsilon_i \tag{2}$$

As punishment is restricted in between 0 and 50, we do not observe *punishment*, but only:

$$punishment_i^* = \begin{cases} 0 & \text{if } punishment_i < 0 \\ punishment_i^* & \text{if } 0 \leq punishment_i \leq 50 \\ 50 & \text{if } punishment_i > 50 \end{cases} \tag{3}$$

Note that $punishment_i^*$ denotes one single punishment decision for a transfer of either 0, 10, or 40 of one subject. We cluster standard errors on the individual level to account for within-individual dependencies. $transfer_i$ is a vector of dummies for the (exogenous) transfers of 10, and 40. $treatments_i$ are the instruments (Receive 0, ..., Receive 50, Observe 0, ..., Observe 50, Dictator). y_i is a vector of the endogenous variables, i.e., personal norms, empirical expectations, normative expectations, negative emotions, and the difference in negative emotions between the first and the second elicitation. Further note that (ν_i, ϵ_i) are assumed to be distributed multivariate normal. Therefore, the first and second stages are estimated together by Maximum Likelihood.[46]

The treatments and transfers are exogenous to the punisher by design and thus fulfill the requirement of instrument exogeneity. To evaluate instrument relevance, we analyze how the instruments shift subjects' norm perceptions. For this, we compare the second elicitation of norm perceptions in the treatments to the Baseline. We show this change by regressing the norm perceptions on

[46] See `https://www.stata.com/manuals/rivtobit.pdf` for more information on the estimation procedure.

the instruments (see the linear regression in Table 11). Note that the first stage in the Tobit IV is estimated simultaneously with the second stage via Maximum Likelihood. We show this linear regression for illustrative purposes only.

Table 11: Linear regression norm perceptions and emotions on instruments

	Dependent Variable:				
	Pers. Norm	Norm. Exp.	Emp. Exp.	Neg. Em.	Δ neg. Em.
	(1)	(2)	(3)	(4)	(5)
Dictator	-9.46^{**}	-6.47^{**}	-4.19^{+}	-0.11	-0.22^{*}
	(3.22)	(2.17)	(2.28)	(0.18)	(0.11)
Receive 0	-2.61	-5.85^{+}	-14.40^{***}	1.05^{***}	0.83^{***}
	(4.61)	(3.21)	(2.28)	(0.29)	(0.20)
Receive 10	-6.20	-6.37^{*}	-6.84^{**}	0.89^{**}	0.21
	(5.08)	(3.03)	(2.55)	(0.28)	(0.17)
Receive 40	3.99	3.58	7.18^{**}	0.82^{**}	0.06
	(4.03)	(2.83)	(2.73)	(0.30)	(0.15)
Receive 50	1.53	4.61	-3.12	-0.08	0.07
	(5.92)	(4.19)	(3.23)	(0.29)	(0.16)
Observe 0	-8.54^{+}	-7.51^{*}	-6.94^{*}	0.27	0.01
	(4.51)	(2.99)	(2.78)	(0.26)	(0.12)
Observe 10	-9.42^{*}	-5.12^{+}	-3.27	-0.01	0.19
	(4.50)	(3.01)	(2.92)	(0.25)	(0.16)
Observe 40	-2.48	-0.52	2.00	-0.49^{+}	-0.10
	(4.95)	(3.57)	(3.53)	(0.27)	(0.12)
Observe 50	-10.29	-2.09	-6.57	0.12	0.02
	(6.62)	(3.92)	(4.16)	(0.39)	(0.13)
Constant	34.83^{***}	28.03^{***}	23.25^{***}	2.47^{***}	0.02
	(2.36)	(1.67)	(1.68)	(0.14)	(0.07)
Observations	296	296	296	296	296
R^2	0.06	0.08	0.15	0.14	0.16
Adjusted R^2	0.03	0.05	0.12	0.12	0.13
Residual Std. Error (df = 286)	19.13	12.54	11.97	1.08	0.67
F Statistic (df = 9; 286)	1.93^{*}	2.84^{**}	5.50^{***}	5.37^{***}	5.85^{***}

Note: SE clustered at individual level. $+p < 0.1$; * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$

In line with the literature on the erosion of social norms (Bicchieri et al. 2022, Keizer et al. 2008, Gino et al. 2009), we find that a norm violation, i.e., a transfer of 0 and 10, has a stronger effect on shifting social norm perceptions compared to high transfers: A low transfer (0, 10) decreases all three types of norm perceptions compared to the Baseline. The effect of a high (40, 50) transfer on norm perceptions is ambiguous and depends on receiving or observing a transfer. Even though the Observer and Receiver treatments give a signal of what is typically done, the instruments do not only significantly shift empirical expectations but also normative expectations - however, to a smaller extent. Personal norms also get shifted, yet mostly not significantly. Being assigned to the role of the dictator significantly shifts personal norms, normative expectations, and empirical expectations. We conclude that the instruments shift all three norm perceptions and hence are relevant.

Finally, a valid instrument has to fulfill the exclusion restriction property. The instruments should exclusively influence punishment through the instrumented variables. To ensure this, we include the channel of negative emotions, as they are known to correlate with punishment (Jordan et al. 2016, Carpenter & Matthews 2012, Nelissen & Zeelenberg 2009). We confirm this by finding a (marginally) significant correlation between the difference between negative emotions and punishment (see Table 1). Additionally, we observe that the instruments influence negative emotions and the difference between negative emotions in the case of receiving a transfer or being assigned to the role of a dictator. Hence, we include both negative emotions and the difference between negative emotions in the IV regression in the main text (Table 2).[47]

Table 12 replicates the results reported in the main text without instrumenting for negative emotions and the change in negative emotions. We find a positive influence of each of the norm perceptions individually.

A.3.3 Omitted-Varibale Bias

In this section, we run an additional robustness check to tackle a potential omitted-variable bias. For doing so, we include the following individual-level controls to the regression: age, gender, income, field of study,[48] degree of understanding, and degree of concentration. As an additional

[47]The second elicitations incorporate the total change of negative emotions compared to the Baseline, however only on an aggregate level. The differences between the first and second elicitations incorporate individual changes. This is particularly important for the self-report of negative emotions because subjects may interpret the 7-Likert scale differently from each other. By focusing on the change of negative emotions, those individual differences in the absolute interpretation of the scale get less pronounced.

[48]The base field of study is ‘Others’.

Table 12: Tobit IV regression punishment on norm perceptions, second stage

	Dependent Variable:			
	Punishment			
	(1)	(2)	(3)	(4)
Personal Norms	0.58^{+} (0.34)			0.15 (1.00)
Normative Exp.		0.76^{*} (0.36)		0.35 (1.45)
Empirical Exp.			0.50^{*} (0.25)	0.26 (0.45)
Transfer10	-4.24^{***} (0.84)	-4.20^{***} (0.83)	-4.23^{***} (0.83)	-4.27^{***} (0.84)
Transfer40	-11.50^{***} (1.30)	-11.45^{***} (1.29)	-11.45^{***} (1.28)	-11.55^{***} (1.29)
Constant	-8.78 (10.18)	-10.12 (8.83)	-1.41 (4.85)	-9.67 (8.35)
Observations	888	888	888	888
Log Likelihood	-6299.94	-5939.23	-5885.93	-12771.59
Wald χ^2 (df = 3)	84.89^{***}	87.38^{***}	86.58^{***}	89.30^{***}

Note: SE clustered at individual level. $+\ p < 0.1$; $*\ p < 0.05$; $**\ p < 0.01$; $***\ p < 0.001$
Treatments conditional on transfer as instruments for each social norm perception separately.

robustness check, we omit negative emotions. Table 13 shows these regressions. Model (1) depicts the model without negative emotions, model (2) includes negative emotions, and finally, model (3) includes the individual-level controls. All models indicate the same correlations between social norm perceptions with punishment. The negative correlation of normative expectations with punishment becomes only marginally significant. Nonetheless, the robustness check replicates the relationships, giving further evidence for a causal influence of social norm perceptions on third-party punishment.

Table 13: Tobit regression punishment on social norm perceptions and controls

	Dependent Variable:		
	Punishment		
	(1)	(2)	(3)
Personal Norm	0.23**	0.23**	0.22**
	(0.07)	(0.07)	(0.07)
Normative Expect.	−0.25*	−0.25*	−0.20+
	(0.12)	(0.12)	(0.12)
Empirical Expect.	0.34**	0.36**	0.35**
	(0.11)	(0.11)	(0.11)
Neg. Emotions		0.05	0.04
		(0.94)	(0.90)
Δ Neg. Emotions		2.45+	2.16+
		(1.26)	(1.20)
Age			0.57
			(0.56)
Female			−1.55
			(2.15)
Income			0.0000
			(0.0001)
Highest Degree			−0.93
			(2.04)
Economics/ Business			2.19
			(4.07)
Engineering/ IT			7.02
			(5.92)
Humanities/ Medicine/ Education			5.34
			(4.35)
Natural Sciences			8.17
			(5.31)
Social Sciences			7.55
			(6.20)
Degree of Understanding			−2.55*
			(1.13)
Degree of Concentration			0.37
			(1.15)
Constant	0.82	0.29	−5.71
	(2.41)	(3.59)	(12.00)
Transfer	✓	✓	✓
Observations	888	888	888
Log Likelihood	−2,423.60	−2,419.70	−2,405.91
Wald Test	116.63*** (df = 5)	123.80*** (df = 7)	150.21*** (df = 18)

Note: SE clustered at individual level. $+p < 0.1; *p < 0.05; **p < 0.01; ***p < 0.001$

A.4 Heterogeneity in Norm-driven Punishment

In this section, we study how punishment and the relative importance of the three norm perceptions may differ between subjects. To do so, we first explore the impact of the exogenous assignment to the different roles (treatments) in the Experience Phase. Additionally, to illustrate subject-specific heterogeneities, we study how norm-driven punishment decisions change with gender.

A.4.1 Role and the Importance of Social Norm Perceptions for Punishment

Figure 9 shows punishment decisions conditional on the role in the Experience Phase and their respective norm perception averages. Non-parametric Wilcoxon rank sum tests reveal that the mere assignment to the role of dictators in the Experience Phase significantly (at least $p < 0.05$) reduces punishment for a transfer of 0 and 10 compared to the Baseline.

Importantly, the figure indicates that the changes in punishment decisions (for a transfer of 0 and 10) closely follow the changes in the norm perceptions induced by the treatments. For instance, in the Baseline, both punishment decisions and norm perceptions are the highest of all treatments. Similarly, in the Dictator treatment, both punishment decisions and personal norms and normative expectations are the lowest. This illustrates the importance of norm perceptions for punishment decisions, also if the differences are induced by the exogenous assignment to a specific treatment.

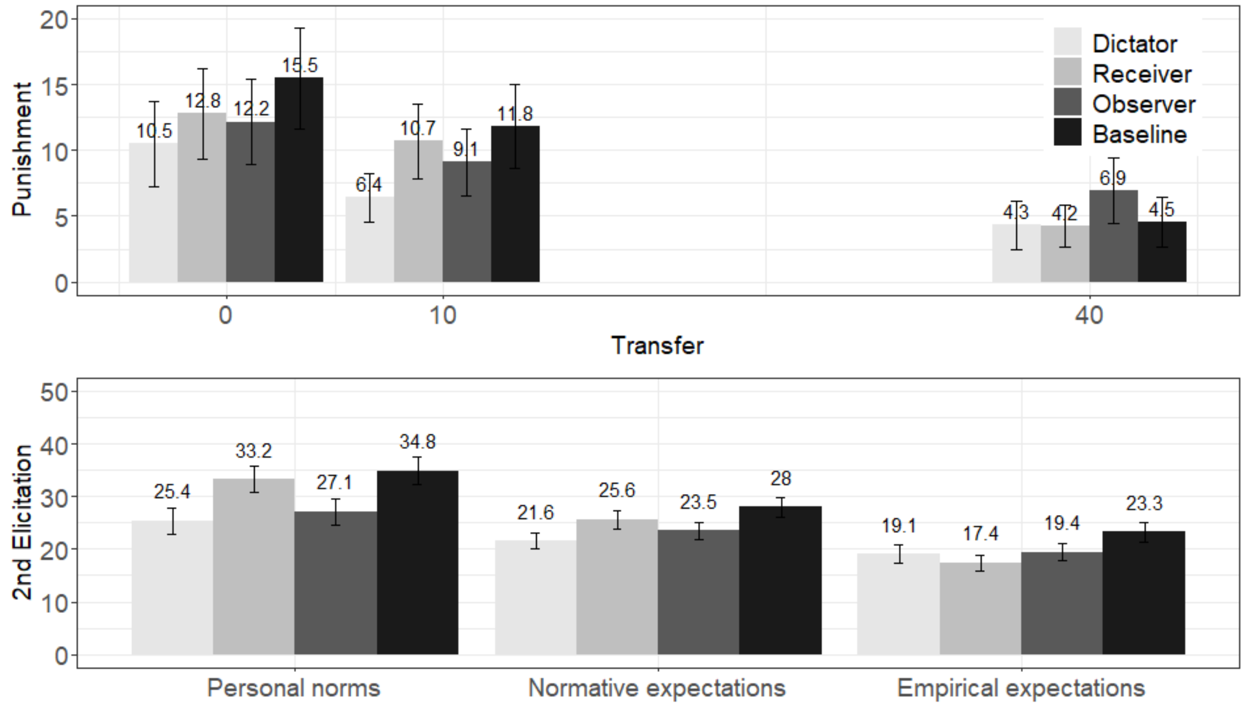
Furthermore, the lower panel of Figure 9 reveals that the between-treatment differences in empirical expectations do not follow the same pattern as the differences in personal norms and more importantly, punishment decisions. This indicates that the importance of the three norm perceptions may differ depending on the role. To study whether this is the case, we regress punishment on the norm perceptions and interact them with the treatments (see Table 14). We find that being in the Receiver treatment (marginally) significantly ($p < 0.1$) increases the importance of empirical expectations for punishment compared to all other treatments. Apart from this, there is no other significant interaction between the treatment assignment and one of the norm perceptions. We conclude that the exogenous assignment into a specific role may change the importance of empirical expectations, however, the relative importance seems to be rather stable for the different roles.

Table 14: Tobit regression punishment on the role and interaction with norm perceptions

	Dependent Variable:			
	Punishment			
	(1)	(2)	(3)	(4)
Personal Norm	0.20*	0.22*	0.24**	0.25**
	(0.09)	(0.09)	(0.08)	(0.08)
Normative Expect.	−0.19	−0.18	−0.28*	−0.31*
	(0.14)	(0.15)	(0.14)	(0.13)
Empirical Expect.	0.39**	0.22	0.38**	0.41***
	(0.13)	(0.14)	(0.14)	(0.12)
Dictator	3.06			
	(5.20)			
Receiver		−8.22		
		(5.50)		
Observer			0.95	
			(5.36)	
Baseline				5.62
				(7.25)
Personal Norm:Treatment	0.08	0.06	−0.01	−0.16
	(0.15)	(0.14)	(0.18)	(0.21)
Normative Expect.:Treatment	−0.37	−0.14	0.15	0.34
	(0.29)	(0.25)	(0.27)	(0.31)
Empirical Expect.:Treatment	−0.002	0.46+	−0.12	−0.30
	(0.25)	(0.25)	(0.25)	(0.32)
Constant	0.14	1.86	−0.23	−0.64
	(4.06)	(3.68)	(3.82)	(3.94)
Interaction:Treatment	Dictator	Receiver	Observer	Baseline
Transfer	✓	✓	✓	✓
Neg. Emotions	✓	✓	✓	✓
Δ Neg. Emotions	✓	✓	✓	✓
Observations	888	888	888	888
Log Likelihood	−2,414.39	−2,413.75	−2,418.54	−2,415.53
Wald Test (df = 11)	134.16***	135.12***	125.97***	131.48***

Note: SE clustered at individual level. $+\ p < 0.1$; * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$

Figure 9: Treatments, social norm perceptions, and punishment



Note: The upper panel shows punishment decisions conditional on the transfer of the to-be-punished dictator (Transfer) and on the specific role in the Experience Phase. The lower panel shows norm perception averages conditional on the specific role. Error bars show 95% confidence intervals and are based on standard errors that are clustered on individual level.

A.4.2 Gender

In this section, we look closer at gender differences in the importance of norm perceptions for punishment decisions. In the existing literature, the results about the effect of gender on punishment remain mixed. Kromer & Bahçekapili (2010) find that males punish selfish behavior more often than females and McAuliffe et al. (2015) confirm this results among children. In contrast, Carpenter & Matthews (2012) find that females punish more than males and Leibbrandt & López-Pérez (2012) show that females engage in more antisocial punishment. Piardini et al. (2017) study different gender compositions of punisher and punishee and find that males punish females significantly more than females punish males and that same-sex groups do not differ in punishment. Moreover, there is only little evidence on how norm perceptions are related to economic behavior with respect to gender. Croson et al. (2010) find that males rely more on their empirical expectations when deciding about their own donations. Krysowski & Tremewan (2021) find that females find giving unfair amounts in a dictator game less acceptable compared to males when the dictator is unidentified. Fišar et al. (2016) study gender differences in bribing behavior and do not find a gender difference in the positive association between accepting bribes and beliefs about how often others accept bribes.

In our experiment, we do not find significant differences in third-party punishment decisions between males and females. However, we find differences in the relative importance of norm perceptions for punishment decisions across genders. For studying this, we first split the sample into males and females. Table 15 shows the results of a Tobit model, where we regress punishment on norm perceptions. Model (1) includes only females, model (2) only males, and model (3) the full sample with a dummy indicating females and the interaction of females with all norm perceptions. In model (3), the interaction effect shows the importance of norm perceptions for punishment for females, whereas the baseline effect shows this relationship for males. In model (1), we find a positive relationship between empirical expectations and punishment for females and a negative relationship with normative expectations. Personal norms seem not to matter for females. In model (2), we find the opposite for males. We find a statistically significant positive association between personal norms and punishment, whereas empirical and normative expectations seem not to matter for their punishment decisions. Model (3) confirms this pattern, as there is a significant negative interaction effect between females and personal norms and a significant positive interaction effect between females and empirical expectations.

We find that this difference in the importance of the norms for punishment decisions is not driven by different initial (first elicitation) levels of norm perceptions between females and males. Table 16 shows average norm perceptions for females and males. The only substantive difference in norm perceptions is between empirical expectations (20.79 for males vs. 18.57 for females), yet the difference is not statistically significant.

We conclude that females want to enforce typical behavior, whereas males punish according to what they personally believe is appropriate. This result illustrates that there exist heterogeneities in the importance of each of the norm perceptions for third-party punishment.

Table 15: Tobit regression punishment on norm perceptions and gender

	Dependent Variable:		
	Punishment		
	(1)	(2)	(3)
Personal Norm	0.08	0.48^{***}	0.45^{***}
	(0.09)	(0.13)	(0.12)
Normative Expect.	-0.29^{*}	-0.22	-0.21
	(0.15)	(0.19)	(0.17)
Empirical Expect.	0.55^{***}	0.05	0.06
	(0.14)	(0.18)	(0.18)
Female			2.95
			(4.76)
Female:Personal Norm			-0.36^{*}
			(0.15)
Female:Normative Expect.			-0.09
			(0.23)
Female:Empirical Expect.			0.49^{*}
			(0.22)
Neg. Emotions	-0.32	0.24	0.04
	(1.14)	(1.70)	(0.94)
Δ Neg. Emotions	4.39^{*}	0.73	2.33^{+}
	(2.00)	(1.44)	(1.20)
Constant	1.67	-0.75	-1.03
	(4.46)	(5.73)	(4.32)
Gender	Female	Male	Both
Transfer	✓	✓	✓
Observations	489	399	888
Log Likelihood	$-1,337.91$	$-1,063.89$	$-2,406.57$
Wald Test	72.39^{***} (df = 7)	77.82^{***} (df = 7)	147.31^{***} (df = 11)

Note: SE clustered at individual level. $+\ p < 0.1$; $*\ p < 0.05$; $**\ p < 0.01$; $***\ p < 0.001$

Table 16: Differences first elicitation norm perceptions between gender

Means (SE)	Males *(N=163)*	Females *(N=133)*	p-value
Personal Norms	30.00 (1.70)	29.63 (1.51)	0.922
Normative Expectations	24.85 (1.17)	24.11 (0.97)	0.668
Empirical Expectations	20.79 (1.14)	18.57 (0.97)	0.145

Note: Non-parametric Wilcoxon rank sum test

B Experimental Instructions

The experiment took place online with the subject pool of the Masaryk University Experimental Economics Laboratory (MUEEL). We enclose the experimental protocol and instructions with experimental screens for punishers and punishees. The subjects went through pages independently, and the experimenter was present at the Zoom meeting, communicating with participants through the Zoom chat. If some participants did not show any activity for more than 2 minutes (apart from the planned waiting time within instructions), the experimenter contacted them through the chat or called them on the phone in case they did not respond.

Protocol and Verbal Instructions for Online Sessions for EXPERIMENTER

Main Experimenter - checks in participants, gives all verbal instructions + number, runs ZTU from VM, sends them the individual links to participants, solves ZTU issues if they come up, handles private chat if necessary.

20 - 30 MINUTES BEFORE SESSION BEGINS:

<Host/ Experimenter 1 checks in participants one by one, once their audio connects, says>

*EXPERIMENTER 1: "Can you hear me? If you can, please unmute yourself and let me know. < if full name is not clear from zoom name, ask for it – Can you tell me your full name? – thank you>. **Now I will assign you the number, which we will use at the beginning of the experiment. Your number is X.** I will now direct you to a breakout room where you should wait for the experiment to begin. You don't have to be at computer now, just be fully prepared at least 5 minutes before the beginning of the experiment. Once you see join, please click on it, and you can mute yourself and turn the camera off now.*

<Direct participant to breakroom where on second computer shares screen (see Experiment_Lobby_Screen), >

Welcome to the experiment lobby! The experiment will begin shortly…

- The random number you got assigned is your participant label. We will use it in case we have more participants than we need in this experiment. If that is the case, we will randomly draw a number – if your number will be drawn, you won't participate in today's experiment, but we will pay you 50 CZK for your show up.
- Please turn off your video and put yourself on mute. It is not allowed to communicate with other participants for the whole duration of the experiment.
- Please keep **your unique ID (or code)** ready as we will ask you to enter it during this session. (You can find this in the email you received with the zoom link for this session).
- Make sure you are in a calm, quiet area without any distractions or people around. While the experiment is running, please stay focused. Please make sure to close all applications on your computer and any websites that are open.
- If you have any questions at any point, please type them into the private chat feature on Zoom and one of the experimenters will respond.
- You have to be fully prepared at least 5 minutes before the beginning of the experiment.

Thank you for your cooperation.

At the beginning - ASKING RESERVES TO LEAVE:

<Experimenter enters breakout room and inform participants, that it will be closed and they will get instructions in main meeting >

MAIN EXPERIMENTER: "Hello again and thanks for coming. For this experiment, we require 14 people. All of you have been assigned a number between 1 and X. I will now share my screen with random number generator. If your number is randomly generated, I will ask you to leave this zoom meeting. If you have entered your bank details in hroot then you will receive CZK 50 for this experiment, and you can register for another session of this experiment. Thanks for coming"

<repeat for all numbers, afterward makes sure that the reserves have left the zoom meeting room->

STARTING THE SESSION:

MAIN EXPERIMENTER: "I will now be sending you your individual links to the experiment. I need to do this one by one so please keep an eye on the chat and you should receive your link in a few minutes. After receiving the link, please paste it in your browser. The experiment will begin shortly after that. If you have any questions during the experiment, please use the chat feature on zoom to ask the question and we will respond to it there.>

<send individual links to all via private chat>

MAIN EXPERIMENTER: "You have now all received the link to the online experiment, so we can start soon. You should see grey screen with small green leaf on side, if you do not, please write to me through chat. If you have not done so already, please now minimize this zoom meeting (without closing it) and move it away from your screen. Since this is an interactive experiment, you might have to wait while other participants make decisions but it is important that you do not engage in any other activity during this time. Please do not open up any tabs on your browser. We will begin shortly."

<Experimenters go on mute, we make sure background is set to number of links that were sent out, press F5 in the VM and start >

ENDING THE SESSION

MAIN EXPERIMENTER: "Now you see payment screen. This is the last screen of this experiment. If you have any question or feedback, please, write to us. Thank you for participating in this experiment. Your full payment will be transferred to your accounts until the end of two working days. After you click on proceed, you can close the tab and leave the zoom meeting room. Thanks again and goodbye."

--DONE--

Instructions – punishers

[screen 1]

Experimental instructions

You are now taking part in an economic experiment. You can earn a considerable amount of money depending on the decisions that you and other participants make. Therefore, it is very important that you read the instructions carefully. It is important to us that you stay concentrated and in front of your computer. Communication with any of the participants is strictly forbidden and can lead to withholding of the payment.

This experiment consists of a part A and a part B. Either part A or part B is going to be paid out to you. Part A will be paid with a probability of 80% and part B with a probability of 20%. You will receive your payoff on your bank account within two working days from the end of the experiment.

If you have questions or technical problems, please write to us through the chat in zoom.

[screen 2 – dictators and receivers]

Experimental instructions - Part A - Stage 1

We will now explain part A. After reading the instructions for the entire part A, you will start to make decisions. Therefore, carefully read the instructions and if you have any questions, please write to us through the chat in zoom. Instructions for part B will be shown after you have finished part A. Part A consists of two stages.

In the first stage, you will be paired randomly and anonymously with another participant. One of you will be randomly assigned to be Player A and the other to be Player B.

Player A will receive 100 CZK and Player B will not receive anything. Player A can then decide to transfer either 0, 10, 40, or 50 CZK to Player B.

Your role will be Player A. [Your role will be Player B.]

[screen 3 – all treatments]

Experimental instructions – Part A [- Stage 2 – dictators and receivers]

[In stage 2, – dictators and receiver] There are players C and D, who are like you real human participants of this experimental session. You will make decisions that will affect the payoff of Player C, therefore your choices have real consequences for your own payoff and for the payoff of Player C.

Player C receives 100 CZK and Player D does not receive anything. Player C can then decide to transfer either 0, 10, 40, or 50 CZK to Player D.

You can assign deduction points to participant C. Each deduction point you transfer to participant C diminishes your income by 1 CZK and participant C's income by 3 CZK. You can assign a number of deduction points between 0 and 50. You will decide how many deduction points to assign to Player C for any possible choice of him/her. Specifically, you will decide how many deduction points to assign if Player C transfers either 0, 10, 40, or 50 CZK.

You won't know how Players C have decided until the end of the experiment. Your choice will be implemented and the number of deduction points you chose will be assigned to Player C and his income will be reduced accordingly, depending on Player C's chosen transfer. Therefore, all your choices potentially have a real impact on the payoff of Players C. Note that nobody has the opportunity to assign deduction points to you at any point in the experiment.

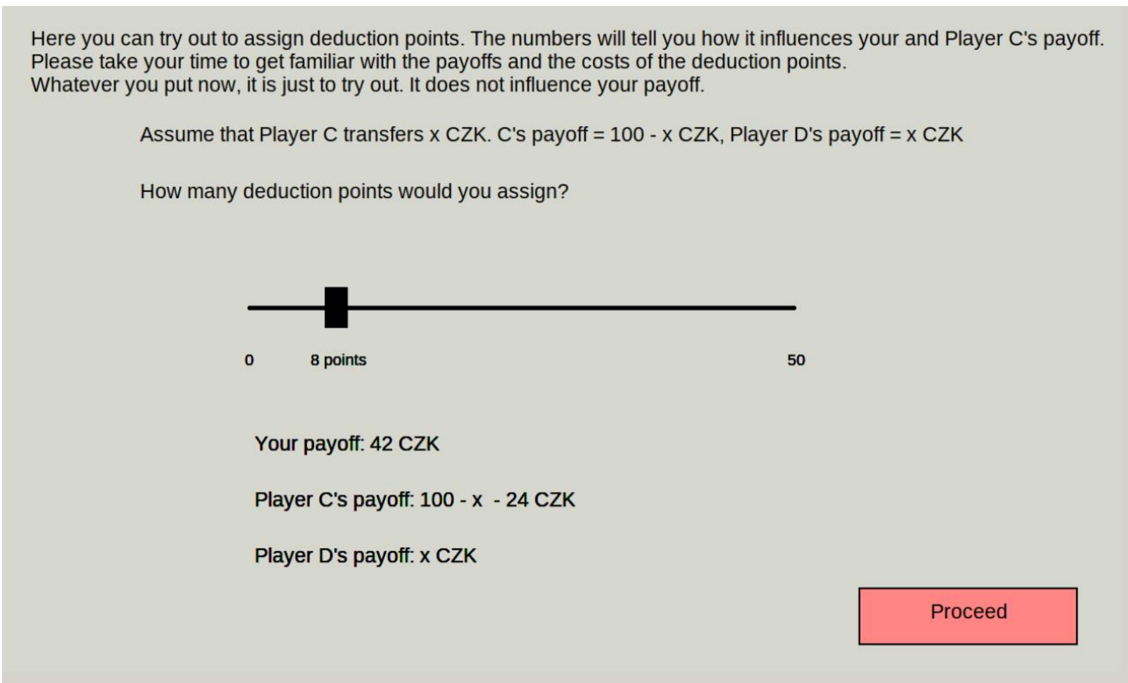
[screen 4 – tryout stage]

Here you can try out to assign deduction points. The numbers will tell you how it influences your and Player C's payoff. Please take your time to get familiar with the payoffs and the costs of the deduction points.

Whatever you put now, it is just to try out. It does not influence your payoff.

Assume that Player C transfers x CZK. C's payoff = 100 - x CZK, Player D's payoff = x CZK

How many deduction points would you assign?



[screen 5 – observers]

Before your decisions, we will give you an impression on how participants decided about the transfers. We will show you the decision and consequences of a randomly chosen participant, Player A, that participated in an earlier session of this experiment. That player is randomly chosen from all the players that participated in the earlier session and that were making a decision as Player A.

Player A received 100 CZK and Player B did not receive anything. Player A could then decide to transfer either 0, 10, 40, or 50 CZK to Player B. There was no opportunity to assign deduction points to Player A.

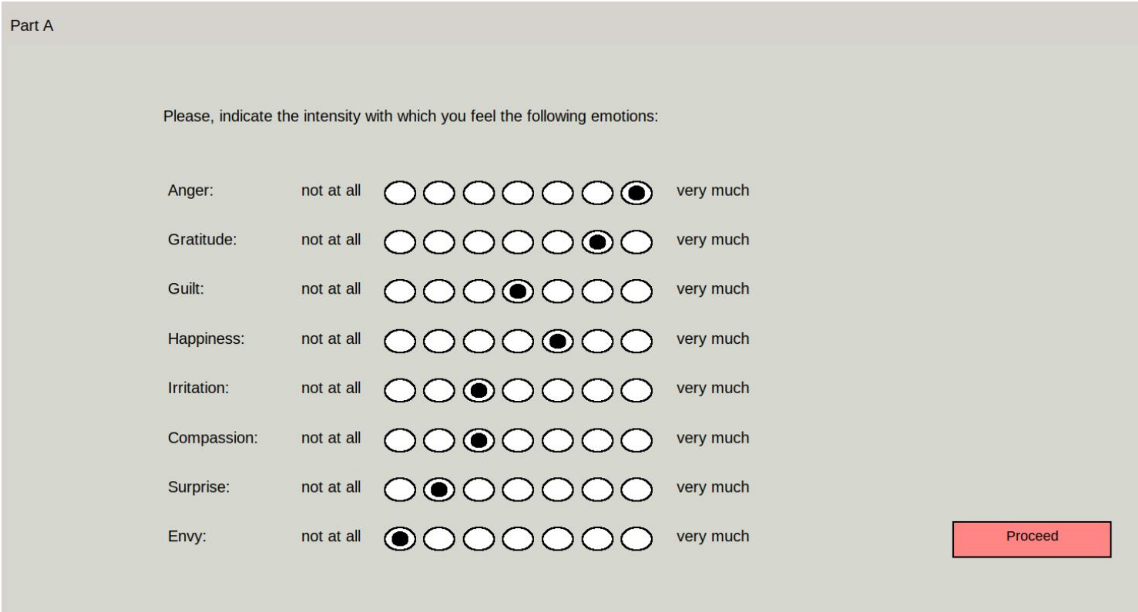
[screen 6 – all treatments]

Now we will start with part A. Remember that there is an 80% chance that this part is going to be payoff relevant for you. In this part A, you receive a base payment of 50 CZK [if observer or inactive 50 + X CZK; X randomly chosen payoff of subject from experience treatment in earlier session] that is independent of your future decisions and will be paid out for sure if part A is going to be picked for your payoff.

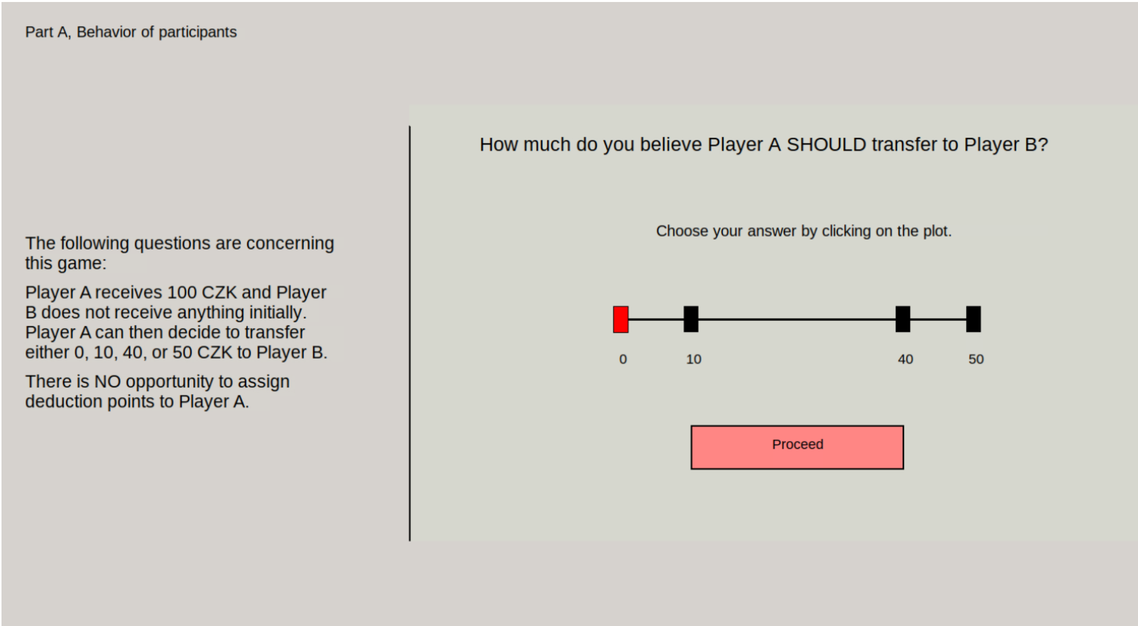
[Your role will be Player A [B] – experience]

Before we start with the first stage, we will ask you about your emotions and opinions on the behavior of participants in this [a previous – observe & inactive] experimental session.

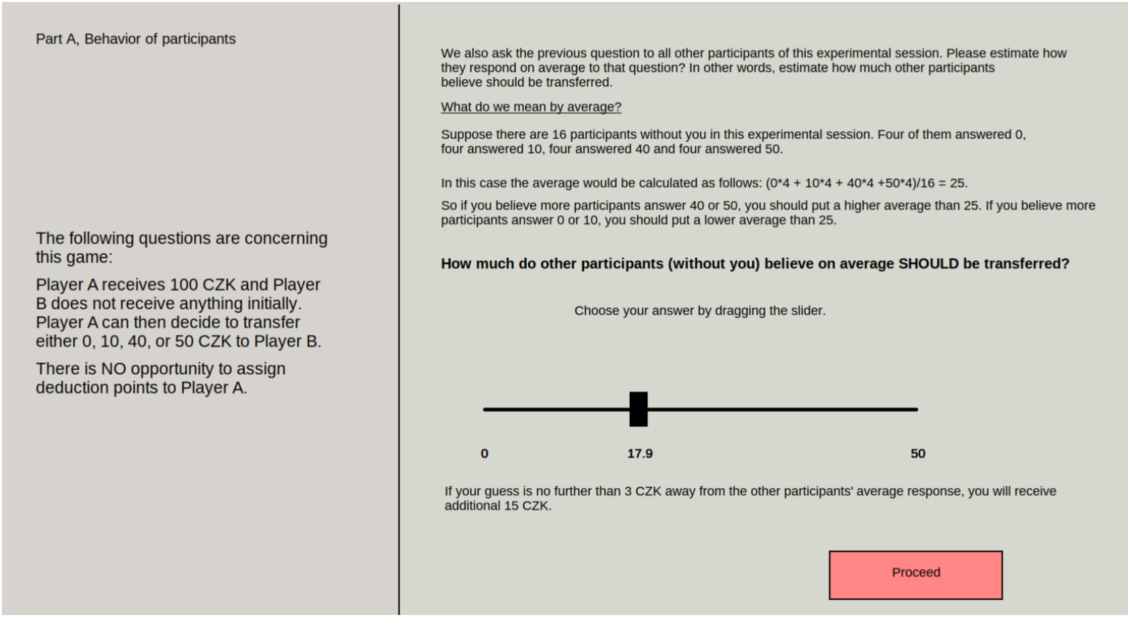
[screen 7 – emotions]



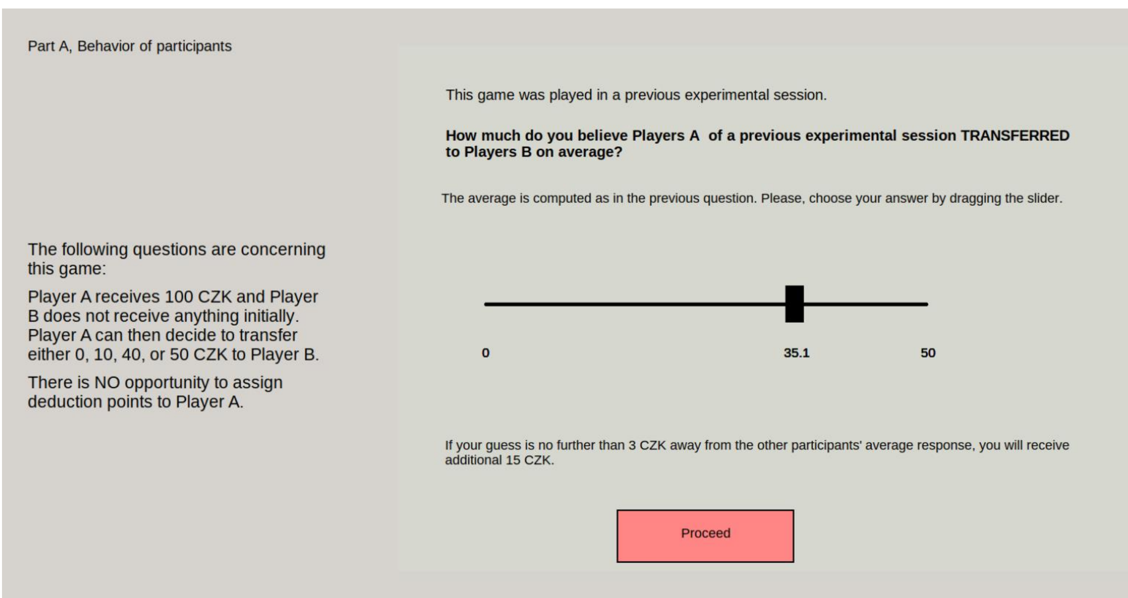
[screen 8 – norm elicitation – personal norm]



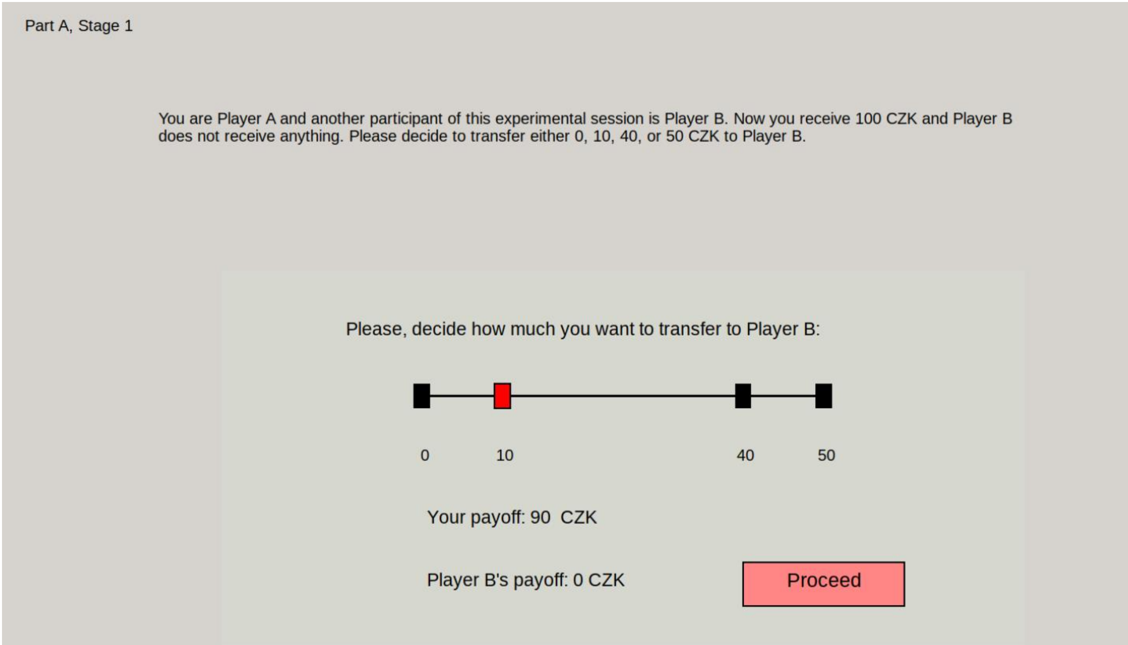
[screen 9 – norm elicitation – normative expectation]

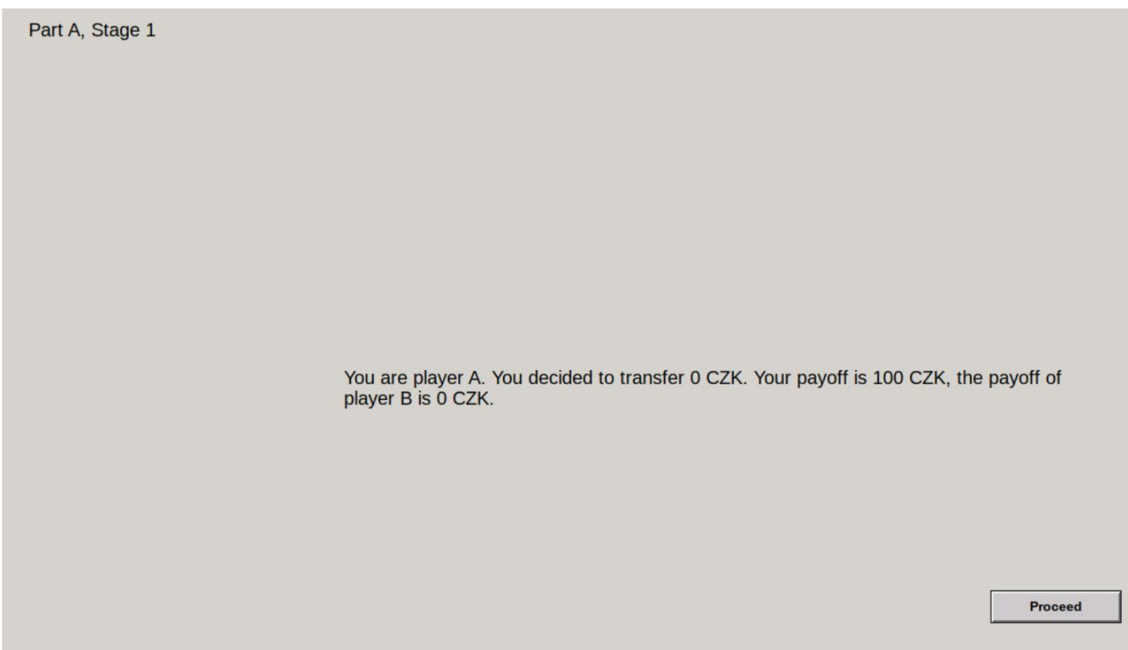


[screen 10 – norm elicitation – empirical expectation]

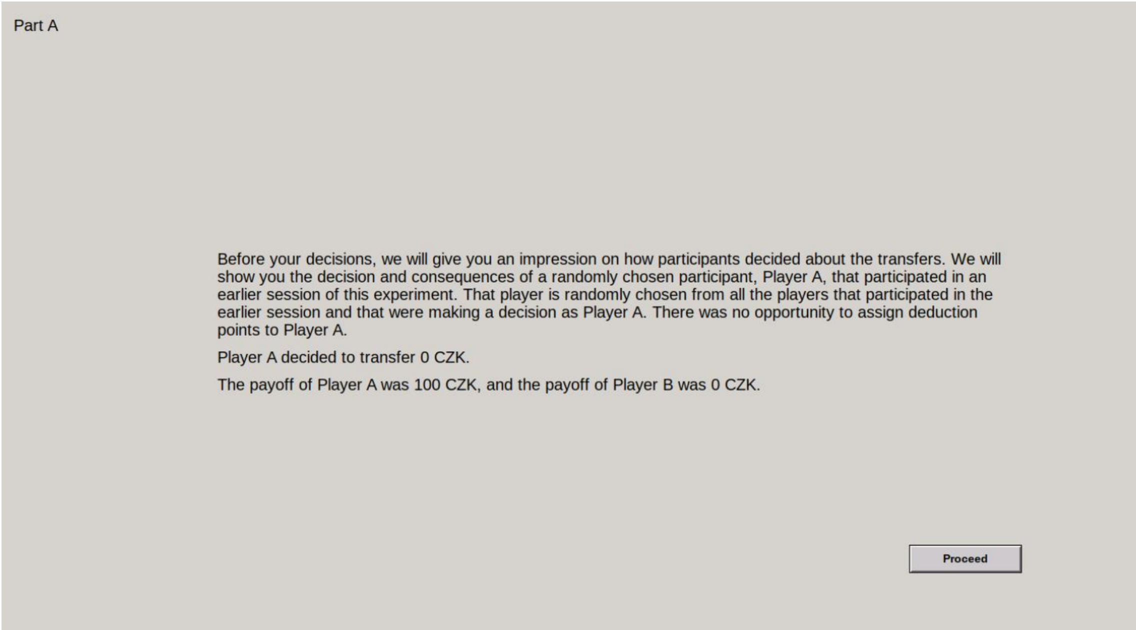


[screen 11 - Experience Phase – dictators]

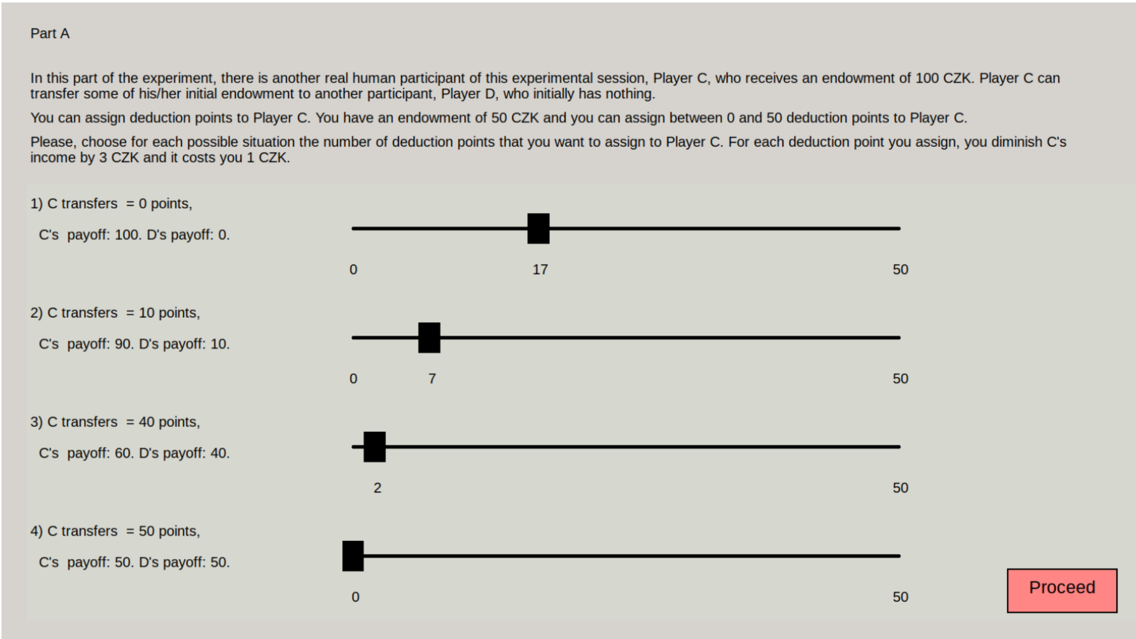




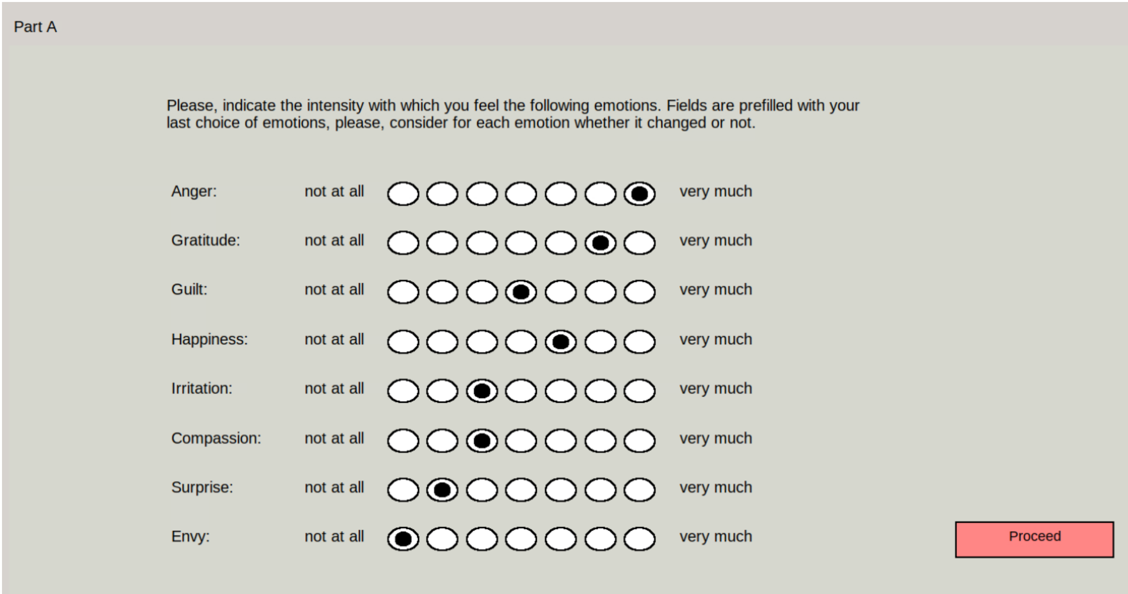
[screen 11 – Experience Phase - observers]



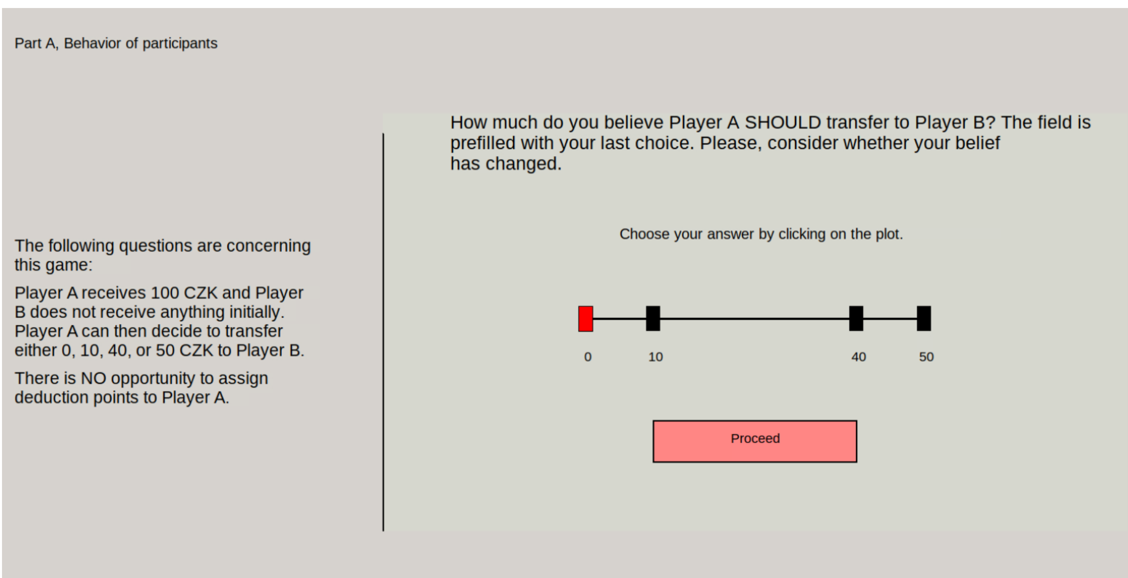
[screen 12 – Punishment Phase]



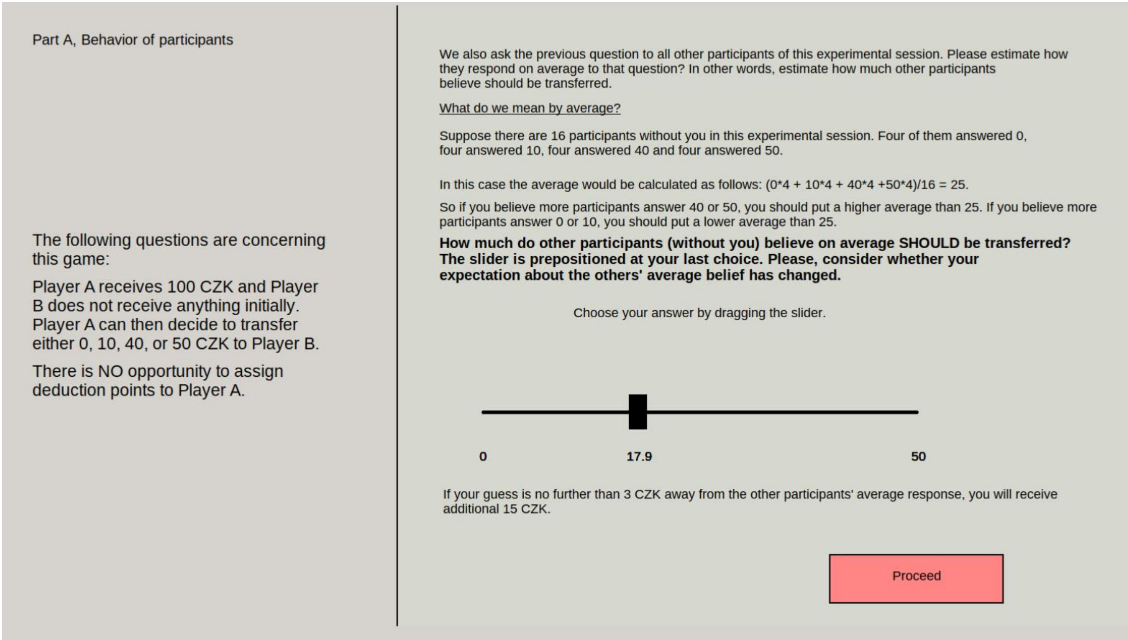
[screen 13 – emotions 2]



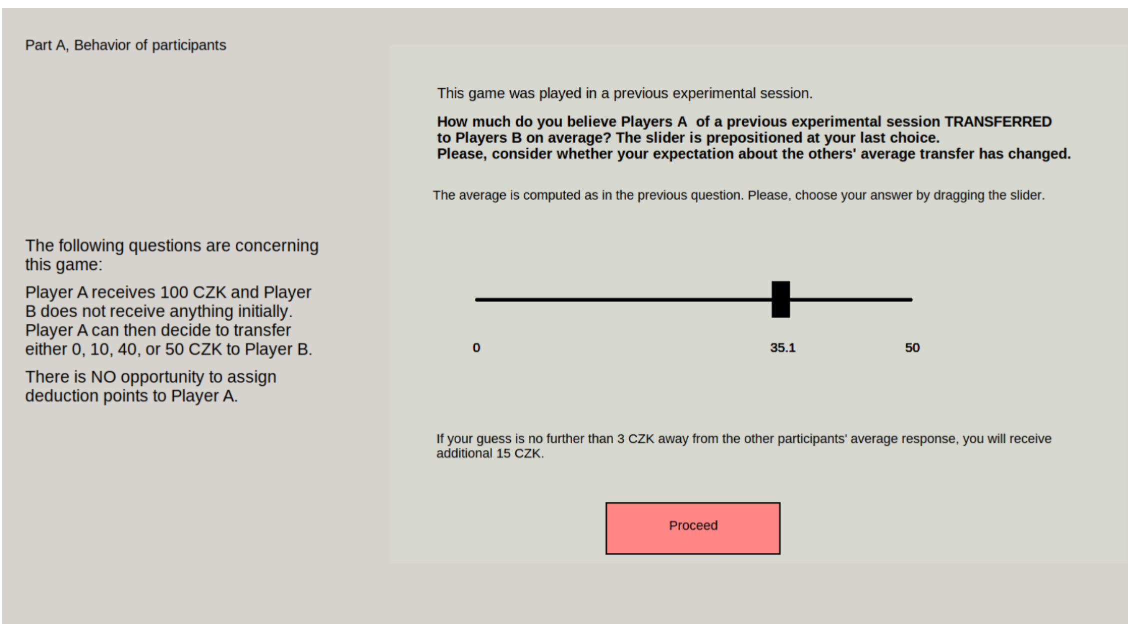
[screen 14 - norm elicitation 2 – personal norm]



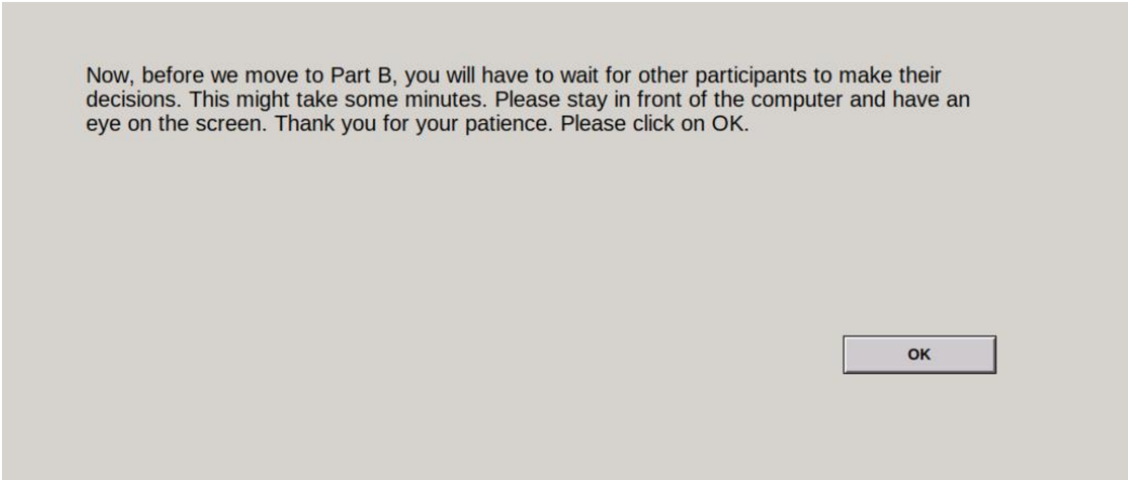
[screen 15 – norm elicitation 2 – normative expectation]



[screen 16 – norm elicitation 2 – empirical expectation]



[screen 17 - waiting stage]



[screen 18 – instructions part B]

Experimental instructions - Part B

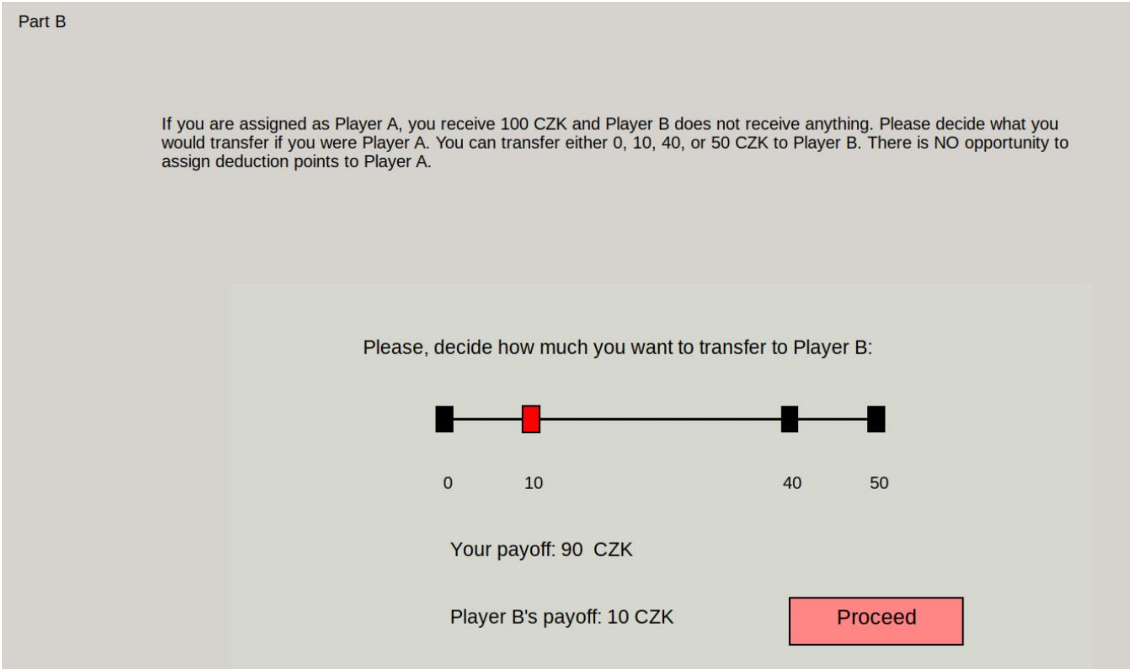
We have completed part A of this experiment. Now we will start with part B. In this part, you receive a base payment of 50 CZK that is independent of your future decisions and will be paid out for sure if part B is going to be picked for your payoff.

You will be paired with another participant of this experimental session. One of you will be randomly assigned to be Player A and the other to be Player B. Player A will receive 100 CZK and Player B will not receive anything. Player A can then decide to transfer either 0, 10, 40, or 50 CZK to Player B.

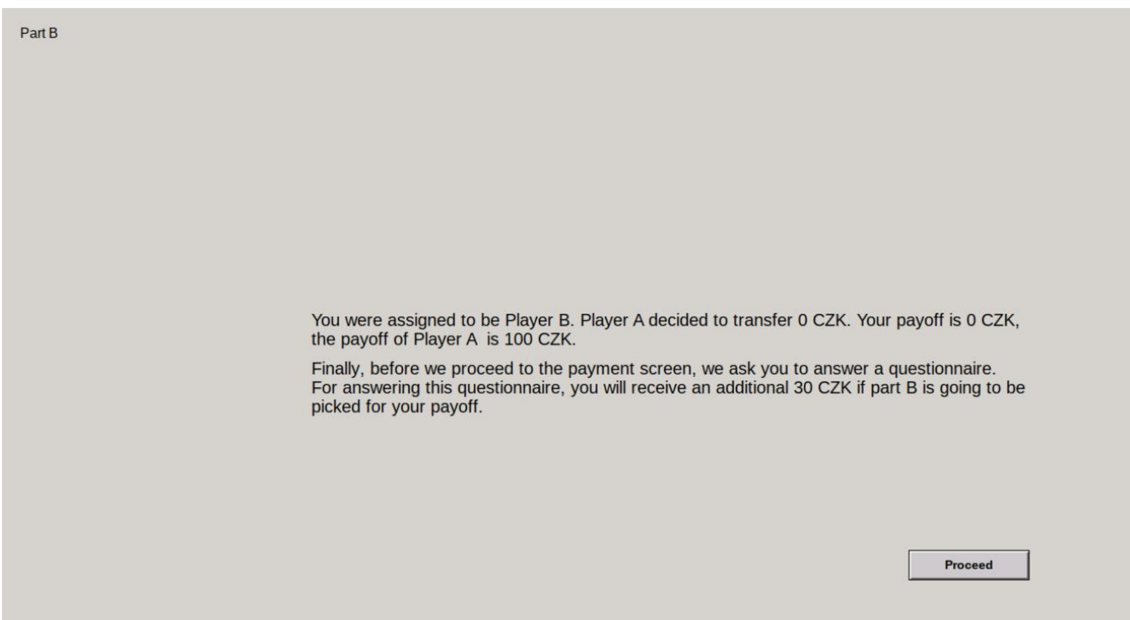
You will make a decision as Player A before you know if you are going to be assigned to be Player A or Player B.

Remember that there is a 20% chance that part B is going to be payoff relevant for you. If it is payoff relevant for you, your transfer as Player A will also be payoff relevant for Player B.

[screen 19 – part B]



[screen 20 – part B results]



[screen 21 – questionnaire]

Questionnaire
Please,fill in this short questionnaire:
Your answers do not influence your payoff.

Age

Sex: Male / Female

Nationality:

Study field:

Monthly net income in CZK

Highest degree earned: HighSchool Degree / Bachelor / Master / PhD

Have you participated in a similar game when one participant could decide on how much to transfer to another, who didn't get anything initially [Dictator Game]? Yes / No

Were you concentrated during the experiment? Not at all ○ ○ ○ ○ ○ ○ Very much

Did you understand the instructions? Not at all ○ ○ ○ ○ ○ ○ Very much

Why did/didn't you assign deduction points? (200 signs allowed)

Please, insert your unique code for payment here (you should obtain it in reminder email from hroot):

OK

[screen 22 – payment screen]

Results

The computer chose part A for the payment.

You guessed that other participants believe that 17.9 CZK should be transferred. The average response was 0.8 CZK.

For the second question, you guessed that on average, 35.1 CZK will be transferred. The average transfer was 23.8 CZK.

Therefore, you receive an additional 0 CZK.

Then, when assigning deduction point you received 50 CZK.

Player C transferred 0 CZK and you assigned 5 deduction points her/him.

After that, you guessed that other participants believe that 17.9 CZK should be transferred. The average response was 0.0 CZK.

For the second question, you guessed that 35.1 CZK was transferred. The average transfer was 23.8 CZK.

Therefore, you receive an additional 0 CZK.

In total, your payment from this experiment is 135 CZK (including base payment of 100 CZK).

Proceed

Instructions – punishees

[screen 1]

Experimental instructions

You are now taking part in an economic experiment. This experiment consists of one game that will be repeated for 4 rounds. You can earn a considerable amount of money depending on the decisions that you and the other participants make. Therefore, it is very important that you read the instructions carefully. It is important to us that you stay concentrated and in front of your computer. Communication with any of the participants is strictly forbidden and can lead to withholding of the payment.

You will receive your payoff on your bank account within two working days from the end of the experiment.

If you have questions or technical problems, please write to us through the chat in zoom.

[screen 2]

Experimental instructions

You will be paired randomly and anonymously with another participant. In each round, one of you will be randomly chosen to be Player C and the other Player D. Before making a decision, you will learn your role, which will be randomly assigned in each round anew.

Player C will receive 100 CZK and Player D will not receive anything. Player C can then decide to transfer either 0, 10, 40, or 50 CZK to Player D.

Other participants from this experiment (Players Y) have the opportunity to assign deduction points to Player C depending on Player C's transfer decisions. Each deduction point assigned to Player C will diminish Player C's income by 3 CZK. Players Y have to pay 1 CZK for each deduction point that they assign. They decide for each possible choice of transfer how many deduction points they want to assign to you. Before we start with the game, we will ask you about your emotions and opinions on the behavior of participants in this session.

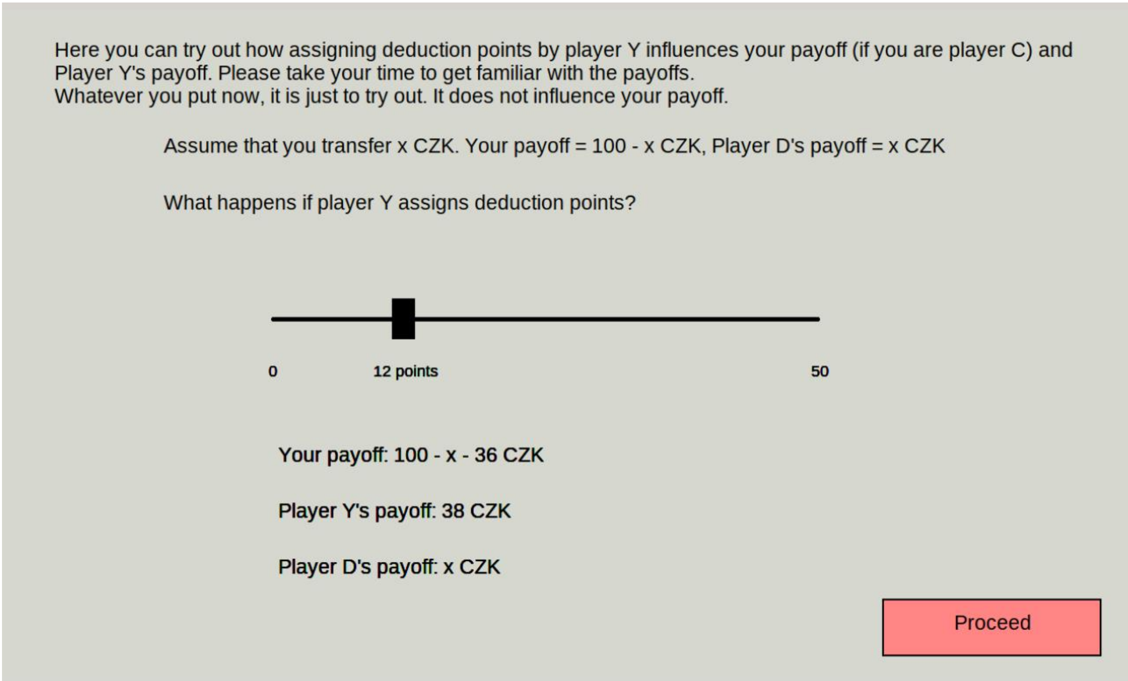
[screen 3 – tryout stage]

Here you can try out how assigning deduction points by player Y influences your payoff (if you are player C) and Player Y's payoff. Please take your time to get familiar with the payoffs.

Whatever you put now, it is just to try out. It does not influence your payoff.

Assume that you transfer x CZK. Your payoff = 100 - x CZK, Player D's payoff = x CZK

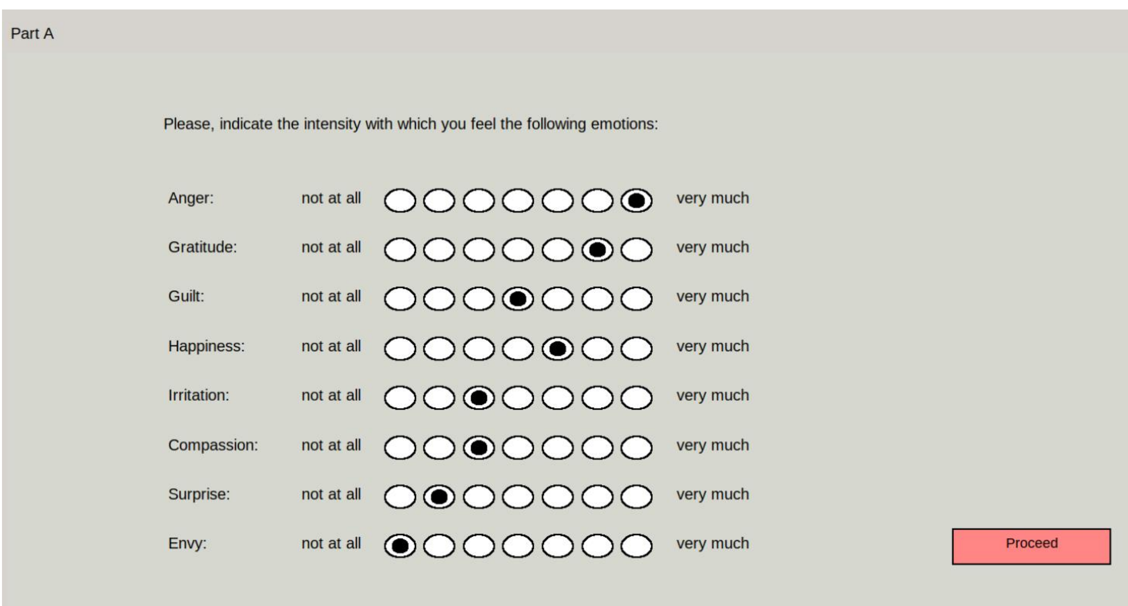
What happens if player Y assigns deduction points?



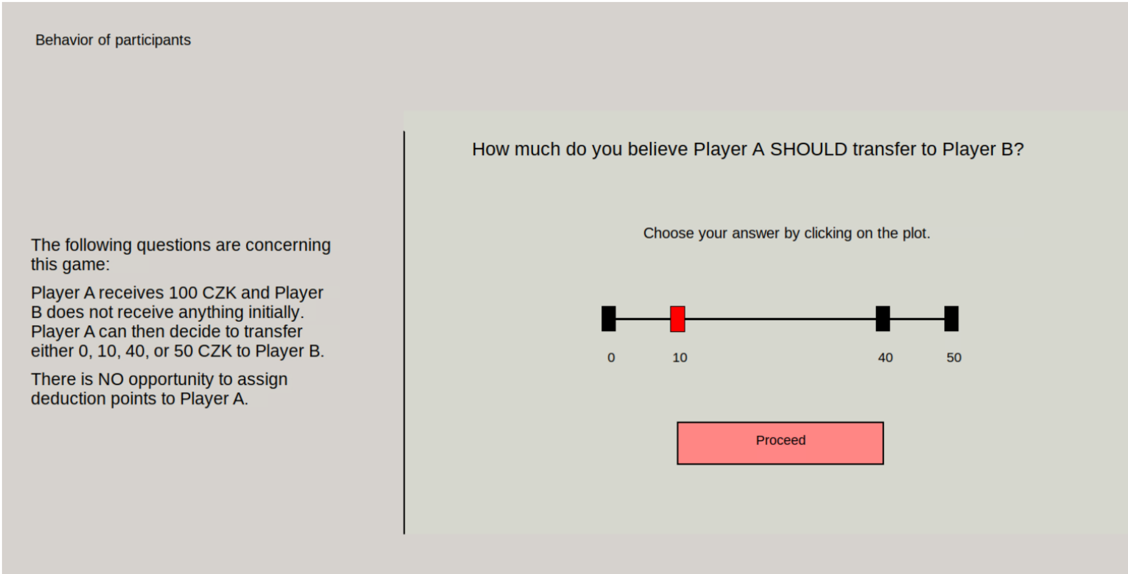
[screen 4]

Before we start with the experiment, we will ask you about your emotions and opinions on the behavior of participants in experimental session.

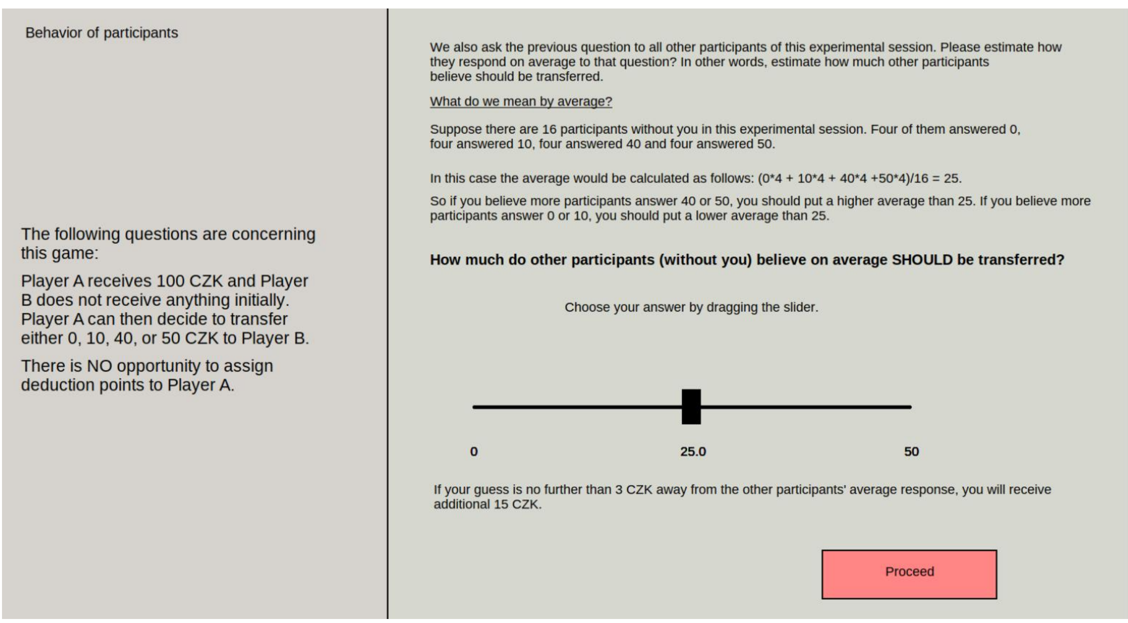
[screen 5 – emotions]



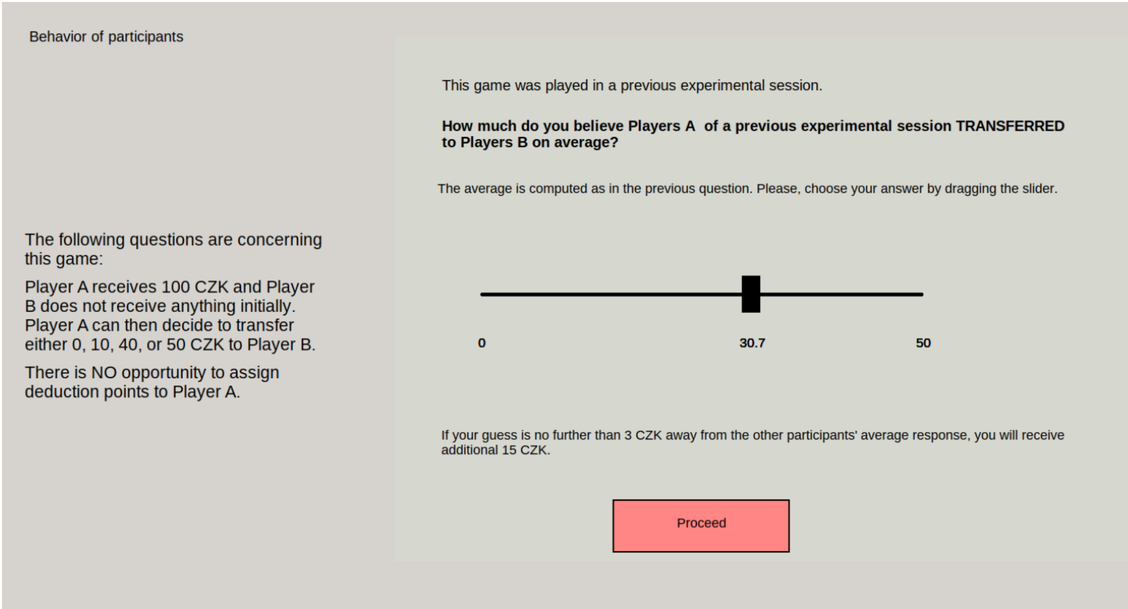
[screen 6 – norm elicitation – personal norm]



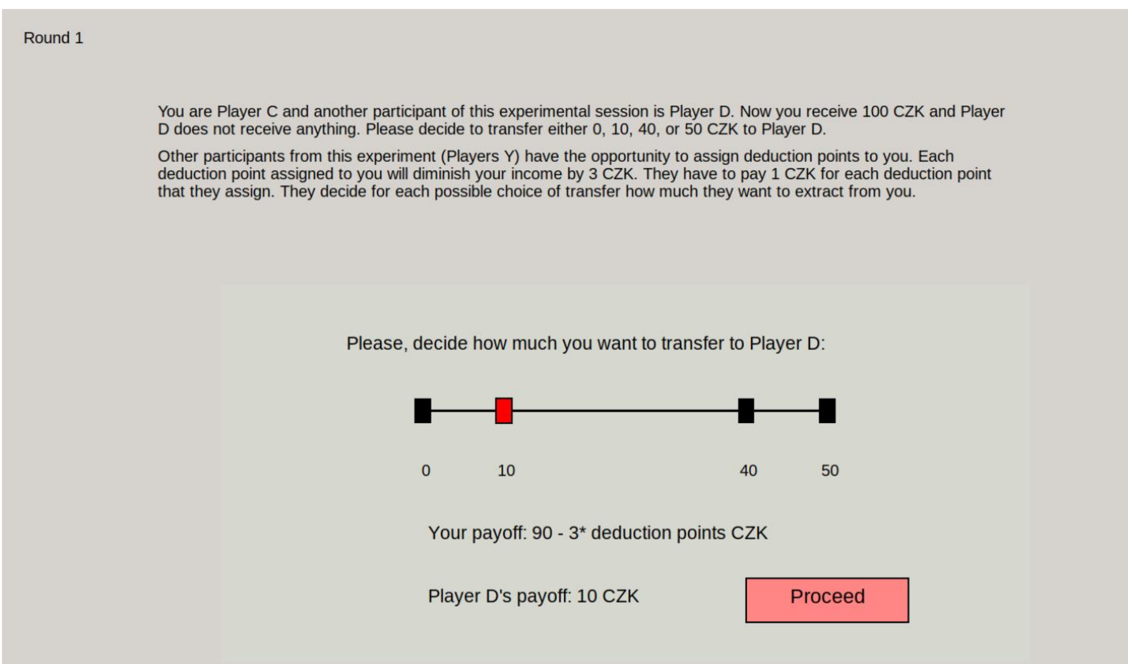
[screen 7 – norm elicitation – normative expectation]



[screen 8 – norm elicitation – empirical expectation]



[screen 9 – dictator game round 1]



[screen 10 – dictator game results]

Round 1

You are Player C. You decided to transfer 10 CZK. Players Y will decide how many deduction points to assign to you for this transfer. At the end of the experiment, the deduction points will be subtracted from your payoff.

Proceed

[screen 11 – waiting stage]

Now, you will have to wait for the decisions of Players Y of this experimental session. This might take around 5 - 10 min. You are allowed to do something else, but it is important that you stay in front of the computer and have an eye on the screen. Once that the experiment continues for you, you will have a time limit of 2 min to do your next decision. If you fail to be back and make a decision within that time, you will not receive any payment from this round. Thank you for your patience. Please click on OK.

OK

[screen 9 and 10 repeated 4 times]

[screen 11 – questionnaire]

Questionnaire

Please,fill in this short questionnaire:
Your answers do not influence your payoff.

Age

Sex: Male / Female

Nationality:

Study field:

Monthly net income in CZK

Highest degree earned: HighSchool Degree / Bachelor / Master / PhD

Have you participated in a similar game when one participant could decide on how much to transfer to another, who didn't get anything initially [Dictator Game]? Yes / No

Were you concentrated during the experiment? Not at all ○ ○ ○ ○ ○ ○ Very much

Did you understand the instructions? Not at all ○ ○ ○ ○ ○ ○ Very much

How much did you adjust your transfers because there were possibly deduction points assigned to you? (200 signs allowed)

Please, insert your unique code for payment here (you should obtain it in reminder email from hroot):

OK

[screen 12 – payment screen]

Results

Your total earnings from four rounds of the experiment are 192 CZK.

You guessed that other participants believe that 36.9 CZK should be transferred. The average response was 0.0 CZK.

For the second question, you guessed that 30.7 CZK was transferred. The average transfer was 23.8 CZK.

Therefore, you receive an additional 0 CZK.

In total, your payment from this experiment is 192 CZK.

Proceed