Social Norm Perceptions in Third-Party Punishment

Katarína Čellárová* Jonathan Stäbler†

August 16, 2023

Abstract

We study whether individuals engage in third-party punishment because they want to enforce social norms. We run an experiment and explicitly measure subjects' beliefs about social norms – social norm perceptions – and associated third-party punishment decisions. Specifically, we focus on personal norms of appropriateness, normative expectations, and empirical expectations for a modified dictator game. We further implement four treatments to exogenously shift norm perceptions and thus identify their causal impact. We find that each of the three norm perceptions has a causal positive impact on punishment on its own. Furthermore, for their joint correlations with punishment, we identify that higher personal norms and empirical expectations are associated with higher punishment decisions, whereas normative expectations are negatively correlated. We observe individual heterogeneity in social norm perceptions and show that their relative importance depends, for instance, on gender or on the role that individuals are assigned to. We conclude that third-party punishment is used for the enforcement of social norms, i.e., beliefs of common behavior, and for the potential creation of new social norms through enforcing the own personal view of appropriateness.

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 it 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). In many cases, this punishment comes at a personal cost to the punisher. Hence, the existence of third-party punishment shows that humans care about how others behave in a specific situation, even when they are not affected by it.

Whatever constitutes wrongdoing, however, is not always clear, is subjective, and 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 – their *personal norms of appropriateness*[1] – and punish those who deviate from it. In this way, they want to enforce their own preferences about how to behave in a specific situation. Second, because humans are a part of society, they potentially base their punishment decisions not only on their personal opinions of appropriateness but may think about what others deem appropriate. *Normative expectations*[2] are individuals' beliefs about what others think is appropriate behavior and can serve to guide punishment decisions to enforce behavior that individuals believe everyone prefers. Third, in addition to thinking about what others think is appropriate, punishers may also rely on their *empirical expectations*[2], their beliefs about what constitutes common behavior. Punishers may take their empirical expectations to guide their 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 explicitly identifies the causal impact of the three types of social norm perceptions on third-party punishment decisions. Furthermore, no other study specifically elicits personal norms, normative expectations, and empirical expectations together and addresses their roles as the underlying motives for third-party punishment.

In this paper, we close this research gap and identify whether and to what extent personal norms of

[1] We follow Bicchieri (2016) to classify the beliefs that matter for norm compliance. For the text to be more concise, we refer to personal norms as a social norm perception, even though it is the belief about the own norm of appropriateness.

[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). The classification that we follow by Bicchieri (2016) can be viewed as the individual's belief of the injunctive norm (normative expectation) and the belief of the descriptive norm (empirical expectation).

appropriateness, empirical expectations, and normative expectations trigger third-party punishment and study the relative importance of the type of norm perception for third-party punishment decisions. Furthermore, we experimentally shift social norm perceptions to identify their causal impact on punishment while controlling for the channel of negative emotions.

Previous studies indicate that social norms and beliefs about social norms matter for third-party punishment. Carpenter & Matthews (2009) test a broad set of different average behavior specifications in public goods game[3] and find that deviations from the average contribution of the session best explain larger third-party punishment decisions. Carpenter & Matthews (2012) confirm this result and observe furthermore 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. Therefore, empirical expectations seem to matter for third-party punishment decisions. However, normative expectations seem to matter as well. House et al. (2020) find that injunctive norm nudges that consist of messages about 'what is wrong and bad behavior' increase third-party punishment decisions for children. Zong et al. (2021) show that punishers react to norm-relevant information, i.e., providing a player's expectations about the behavior in the game. This is especially pronounced if the to-be-punished player is from another country. Dimant & Gesche (2021) show that injunctive and descriptive norm nudges increase third-party punishment decisions. They further show in a subsequent experiment that both nudges increase personal appropriateness ratings of the situation. However, they do not elicit punishment decisions and hence do not show how they are associated with punishment decisions. Finally, personal norms also seem to matter. Bašić & Verrina (2021) show that both normative expectations and personal norms are positively correlated with third-party punishment decisions.

Other studies investigate norms of third-party punishment. 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) shows 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, Reuben & Riedl 2013), as well as the punishment norm itself Li et al. (2021).

In summary, the literature demonstrates that beliefs about 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, i.e., personal norms, empirical expectations, and normative expecta-

[3]Such as the average contribution of the session, of the ingroup, or of the outgroup, the own contribution, or the set of all possible (exogenously set) contributions.

tions, as well as punishment decisions together and, therefore, cannot clearly identify the channels for punishment decisions. In principle, all of the three mentioned norm perceptions inform each other and hence are correlated (Tremewan & Vostroknutov 2021). Therefore, it is important to elicit all of them to disentangle the individual influence of each on third-party punishment decisions. Furthermore, none of the papers provide causal effects of social norm perceptions on punishment. To estimate the causal impact of social norm perceptions, we experimentally shift norm perceptions and exploit the resulting variation.

The contribution of this paper is threefold. First, we observe the correlation between personal norms, normative expectations, and empirical expectations with punishment decisions and their relative importance for punishment. Thus, we disentangle how much prediction value each of them carries when studying the correlation of all of them simultaneously. With our results, we can explain many results of the third-party punishment literature. Second, we identify the causal influence of each norm perception on its own and thus show that social norm perceptions are, in fact, motives for the enforcement of social norms. Third, we show that the importance of the specific social norm perception is heterogeneous and depends, for instance, on the role that individuals are assigned to or their gender.

We run an online interactive experiment that consists of two sections. We measure personal norms, empirical expectations, normative expectations, and emotions before the first section and after the second section. We control for negative emotions, as they are an important driver of third-party punishment (Jordan, McAuliffe & Rand 2016, Carpenter & Matthews 2012, Nelissen & Zeelenberg 2009). In the first section, punishers go through a modified dictator game in different roles, depending on the treatments. In this dictator game, the dictator starts with an endowment of 100 CZK and can decide to transfer either 0, 10, 40, or 50 CZK to the receiver.[4] In the punishment section, subjects choose whether and how much to punish *another* dictator, drawn from a different group of people, conditional on her 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 employ four treatments with an exogenous variation of the first section. Participants are randomly assigned either to the role of dictator (*Dictator treatment*), the role of receiver (*Receiver treatment*), the role of observer (*Observer treatment*), or to the *Baseline*. In the Dictator and Receiver treatment, participants played one round of the modified dictator game. In the Observer treatment, they observed a transfer from a dictator from an earlier session, and in the Baseline, they did not do anything in the first section.[5] In the Dictator treatment, subjects may motivatedly change their

[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 after a transfer.

[5]The Dictator and Receiver treatment took place in the same session, whereas the Observer treatment and Baseline each took place in separate sessions.

norm perceptions because they are in a situation where there may be moral costs for giving a low transfer (see literature on motivated beliefs, e.g., Zimmermann (2020), Epley & Gilovich (2016)). In the Receiver and Observer treatment, subjects get a signal of common behavior through the transfer of the dictator and hence may update their social norm perceptions based on that signal. By having both an Observer and Receiver treatment, we vary the degree to which social norm perceptions get shifted. Moreover, the Observer treatment reduces the triggering of negative emotions that come with receiving a transfer in the Receiver treatment and hence helps to isolate the channel of social norm perceptions on punishment. Nonetheless, we control for the channel of negative emotions in all treatments. The Baseline treatment poses the baseline for norm perceptions and allows us to evaluate the shift of norm perceptions in the other treatments.

Our main findings are the following: We find a significant positive relationship between personal norms of appropriateness and punishment decisions and a significant positive relationship between empirical expectations and punishment. Subjects who believe higher transfers are more appropriate, or those who think that more is usually transferred, punish more. In this sense, third-party punishment is used to enforce social norms when they punish according to their empirical expectations and to potentially create new social norms when punishers punish according to what they personally believe is appropriate which may shift common behavior. Surprisingly, we find a significantly negative relationship between punishment decisions and normative expectations while controlling for personal norms and/or empirical expectations. We find that the difference between personal norms and normative expectations can explain this negative relationship. Punishers use their normative expectations to determine whether they can free-ride on others' punishment decisions or whether their punishment decisions are crucial.

Additionally, we find a positive causal effect of each of the three norm perceptions on third-party punishment on its own. Having been given the role of dictator significantly shifts all three norm perceptions and negative emotions. Observing and receiving a transfer in the first section of the experiment significantly shift normative expectations, empirical expectations, and negative emotions. Therefore, we use the assignment to the role of dictator and the observed and received transfers in the Observer and Receiver treatments as instruments for norm perceptions and negative emotions in a Tobit IV regression.

Furthermore, being assigned the role of dictator significantly decreases punishment, arguably because dictators downward shift all three of their norm perceptions. There is no difference in the overall punishment of the other treatments. However, being in the Receiver treatment makes subjects rely slightly more on their empirical expectations for punishment decisions. All other roles do not differ in

the relative importance of the norm perceptions for punishment.

Moreover, an exploratory heterogeneity analysis reveals that relative importance depends on gender. The positive relationship between personal norms and punishment is stronger for males, whereas the relationship between empirical expectations and punishment is stronger for females.

We conclude that third-party punishment is used to enforce social norms, especially descriptive norms. Additionally, it is used to enforce the own views of appropriateness, which could lead to the creation of new social norms of behavior. We further show that the importance of personal norms or empirical expectations for punishment is heterogeneous and depends, for example, on the role in the dictator game or gender. In this sense, the enforcement and the potential creation of social norms depend on specific characteristics and situations.Finally, the importance of empirical expectations and personal norms, but not normative expectations, provide policy implications that aim to increase third-party punishment. Policies should focus on shifting empirical expectations (and, if possible personal norms) when wanting to increase punishment most efficiently.

The paper is structured as follows: In section 2, we describe the experimental design and the mechanism and identification strategy, section 3 presents the results. Section 4 provides a discussion, and lastly, section 5 concludes.

2 Experiment

We ran an online experiment[6] in March and April of 2021 with subjects from the subject pool 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. 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). The experiment received ethical approval from the GfeW.[7] We analyze the punishment behavior of 296 participants (punishers).[8] Furthermore, we included 120 punishees,[9] who were dictators in the dictator game and were subject to potential sanctions from the punishers. Punishees did not interact in any other way with the punishers. Hence, we ensured impartial third-party punishment decisions and removed any indirect counter-punishment considerations.

[6] 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.

[7] German Association of Experimental Economics

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

[9] We had to exclude one punishee from the analysis because of inattention.

2.1 Experimental Design

The experiment consisted of two parts, A and B, where part A was payoff-relevant with an 80% probability and part B with 20% probability. Part A was the main part of the experiment. In part B, we elicited distributive preferences and demographics. In part A, we implemented four treatments, namely the *Dictator treatment (N=80)*, the *Receiver treatment (N=77)*, the *Observer treatment (N=79)*, and the *Baseline (N=60)*. These treatments differed in what preceded the punishment phase.

Figure 1: Experimental design part A

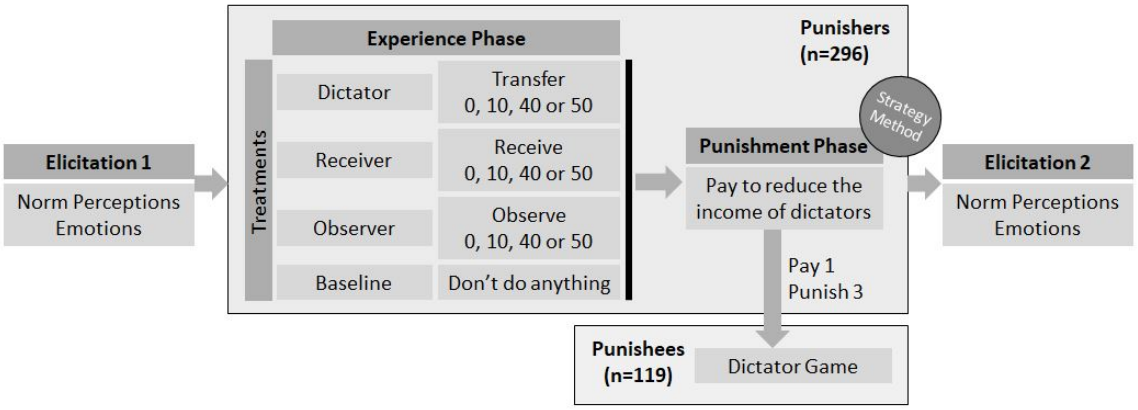


Figure 1 depicts part A of the experimental design. In part A, we elicited norm perceptions at the beginning and at the end.[10] We will explain our elicitation method after we explain the modified dictator game and the subsequent punishment phase.

After the first elicitation of norm perceptions and emotions, punishers went through the experience phase. In the Dictator and Receiver treatment, subjects were randomly assigned to either the role of dictator or the role of receiver. Dictators then made a decision about how much to transfer to a randomly matched receiver. In the Observer treatment, subjects observed the transfer of a randomly matched dictator of the Dictator treatment from a previous experimental session.[11] In the Baseline, the experience phase was simply omitted.

We focus on a dictator game for the experience phase because there is no unique socially optimal allocation as in other games such as public good games. Because of this, social norm perceptions may be more dispersed, as there is not one optimal allocation. In the modified version of the dictator game we introduce, the dictator receives an endowment of 100 CZK, while the receiver starts with zero. The dictator decides to transfer either 0, 10, 40, or 50 CZK to the receiver. By excluding intermediate choices like 20 or 30, we allow for a clearer taxonomy of high (equal split) and low

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

[11] Each dictator was used for one subject from the Observer treatment, who saw the specific decision. Dictators in the Dictator treatment did not know that their choices were used to be shown to other players in a later session.

(unequal split) transfers. In this way, we enforce more extreme transfers that have a larger impact on shifting social norm perceptions.[12] Our modified version has the further advantage that participants are not familiar with what describes common behavior and what ought to be done in this version. This makes participants' norm perceptions more genuine.

To avoid a systematic influence of income effects on punishment across treatments, we paid different show-up fees before the experience phase. Receiver subjects were paid a show-up fee of 50 CZK. In the Observer treatment 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, the distribution of payoffs was replicated.[13] In this way, we established the same wealth level before the punishment phase across all treatments except in the Dictator treatment, where subjects were wealthier by design.[14]

Right after the experience phase, subjects were given the opportunity to 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 that set the transfer in the experience phase and only existed to get punished for their behavior and hence for the punishment decisions of the punishers to have real consequences. The punishment phase was the same for all treatments. All punishers had the opportunity to reduce the earnings of a 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 expressed how many deduction points to assign to the punishee for every possible transfer.[15] 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).[16] Punishers were aware that they themselves would not be punished at any stage of the experiment.

Before the experience phase and after the punishment phase, we elicited subjects' emotions, personal norms, and empirical and normative expectations. To measure personal norms and norm perceptions, we follow 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 that we used throughout the whole experiment: 0, 10, 40, or 50 CZK. Second, we asked them what

[12] In the pilot session, most of the norm perceptions were between 10 an 40 CZK. Thus, we assume that receiving a signal of a higher or lower transfer has the potential to change them.

[13] We could not use the exact same number of payoffs because of a slightly smaller number of subjects in the Baseline.

[14] In principle, a higher wealth should result in higher punishment. However, we find that dictators punish significantly less and that wealth before the punishment decisions did not play an important role.

[15] Jordan, McAuliffe & Rand (2016) find that third-party punishment decisions are not influenced by the strategy method.

[16] 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.

they thought was the *average response* to this question by the other participants of the 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. In this way, we can capture even small changes in individual norm perceptions. We elicited them with a slider, with an initial position at 25 which constitutes half of the highest possible transfer. In the second elicitation, which took part right after the punishment, subjects were shown their choices from the first elicitation. We asked them to consider whether their expectations had changed. The first norm elicitation took place after punishers knew their role in the experience phase. We incentivized the elicitation of normative and empirical expectations. We paid an additional 15 CZK whenever participants were within a range of 6 CZK around the true average. We incentivized this decision-making to obtain a higher accuracy of the responses (Gächter & Renner 2010).

We measured self-reported emotions following the elicitation of Bosman & Van Winden (2002) and Cubitt et al. (2011). We included both positive and negative emotions to prevent experimenter demand effects towards what emotions participants should feel. Specifically, we asked participants about the 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 measured the intensity of emotions both before the experience phase and after the punishment decisions. At the second elicitation, choices from the first elicitation were prefilled. We asked subjects to consider whether the intensity of their emotions had changed.

Next, we describe the procedure for punishees. Punishees started with the same elicitation of emotions and norms as the punishers. Then they played five rounds of the same version of the modified dictator game, only with the addition that dictators could get punished for their choices. In each round, it was determined whether a punishee acted in the role of a punishee dictator or a punishee recipient. To be more economical, we matched 10 punishers to 4 punishees, and punishees played the game multiple rounds. In each round, we matched one punisher with one punishee dictator and implemented the punishment decision. In that way, every punishment decision of a punisher was implemented once if part A was chosen to be payoff relevant for that punisher. The 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.[17]

Lastly, in part B, all punishers played the same dictator game as in part A without the possibility of getting punished. All subjects made decisions as a dictator. Afterward, the computer decided randomly whether their role was a dictator or receiver and who they were matched with. Additionally, subjects received 80 CZK as a show-up fee and 30 CZK for answering the questionnaire with demographic information. This part served as a control for differences in redistributive preferences, which we use as an additional control in a robustness check.

2.2 Mechanism and Identification

Tremewan & Vostroknutov (2021) argue that all three norm perceptions inform each other and therefore are correlated. Therefore, we elicit all norm perceptions to determine which of them is ultimately the driving force of third-party punishment. Previous literature emphasizes 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. 2016, Bicchieri & Xiao 2009) including for punishment decisions (Dimant & Gesche 2021). However, since third-party punishment is considered to serve for the enforcement of social norms (Henrich et al. 2006, Fehr & Fischbacher 2004), the injunctive norm and thus the belief about it - normative expectations - may play a more important role in punishment decisions. Additionally, individuals may want to enforce their own point of view and punish according to their personal norms of appropriateness. Bašić & Verrina (2021) show that personal norms alongside normative expectations matter for economic decision-making, including for third-party punishment decisions. Therefore, we conjuncture that punishers may use their personal norms to make other persons behave in a way that they deem appropriate.

To address a potential source of endogeneity, we control for negative emotions, as the literature finds that they affect third-party punishment decisions (Jordan, Hoffman, Bloom & Rand 2016, Carpenter & Matthews 2012, Nelissen & Zeelenberg 2009). For this, we include both the second elicitation of negative emotions as well as the change of negative emotions from the first to the second elicitation.[18] Additionally, we run an instrumental variable approach to identify the causal impact of the norm perceptions on punishment. For this, we rely on treatment manipulations as a tool to shift norm perceptions. Receiving or observing a transfer gives a signal of the underlying social norms, and

[17] Because part A was paid in 80% of cases, the punishment decisions of 8 punishers were implemented in the experiment. In each session, the computer created two groups of each two punishees, who played the dictator game four times. In each of the four rounds, the role of the dictator was assigned randomly. We imposed the condition that each punishee would be twice in the role of dictator and twice in the role of receiver. We imposed this condition to provide more equal payoffs for punishee participants. Neither punishers nor punishees were informed about the matching procedure, and all interactions remained anonymous.

[18] We include the change of negative emotions, as we elicit emotions on a 7 Likert scale, which might be interpreted subjectively. By focusing on the change, we capture the individual changes that are less dependent on the interpretation of the scale.

participants may take that signal to update their norm perceptions downwards or upwards (Bicchieri et al. 2022, Hoeft et al. 2019, Gino et al. 2009, Keizer et al. 2008), or to build their norm perceptions around this reference point.[19] We give a signal of the descriptive norm, thus, we expect empirical expectations to shift the most. At the same time, the signal of the descriptive norm may also inform normative expectations and change personal norms as they are closely related and inform each other.[20] Additionally, being assigned to the role of a dictator can shift both personal norms of appropriateness and the beliefs about the social norms because of motivated beliefs (e.g. Zimmermann 2020, Epley & Gilovich 2016). Specifically, a dictator may reason that it is both personally and socially more appropriate and more common to send lower transfers to not feel guilty to send lower transfers. In this way, a dictator lowers her social norm perceptions to be able to transfer lower amounts.

3 Results

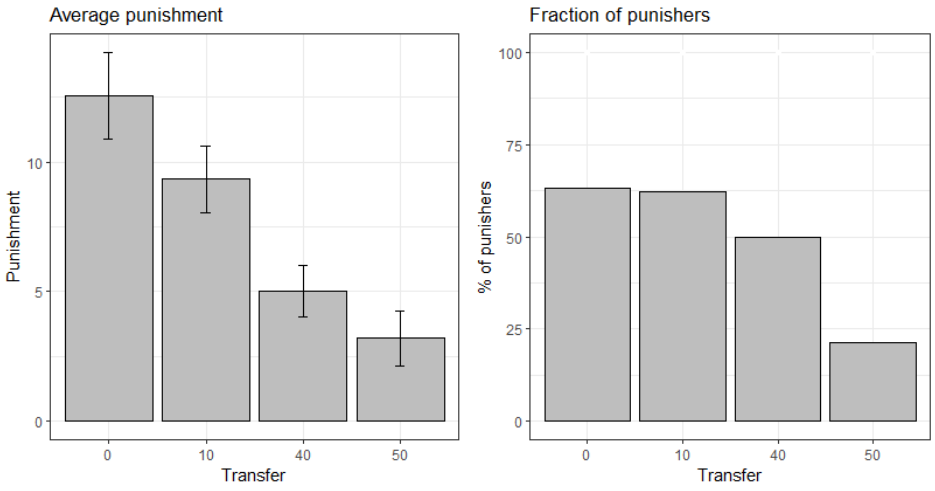
The following sections present the results of our experiment. In particular, we start with the relationship between all three norm perceptions and punishment decisions in section 3.1. We continue with an IV approach to establish the causal influence of norm perceptions on punishment decisions in section 3.2.

Figure 2 depicts the overall punishment and propensity to punish all possible transfers of the dictator game elicited via the strategy method. Notice that subjects mainly punish dictators, who transfer 0, 10, and 40 CZK. Further, the amount of punishment decreases with the more equal splits. That means that subjects want to enforce fairness, also treating the transfer of 40 as something worthy of punishment. In our primary analysis, we focus on prosocial punishment. We treat the transfers of 0, 10, and also of 40 as prosocial punishment. Figure 2 shows that around 60% of all subjects punished a dictator that transfers the low transfers of 0 and 10. Only around 10%-points fewer subjects punished a dictator that transfers 40, and therefore, we treat the punishment of 40 also as prosocial punishment. It can also be argued that only the transfer of 50 provides an equal split, and any deviation from it is, in a sense, selfish and less social, making its punishment prosocial. The punishment of a transfer of 50, however, is clearly antisocial, as it penalizes the most social transfer that was possible. We examine the occurrence of antisocial punishment at the end of section 3.1.

[19] In a pilot session, most of the initial normative and empirical expectations were between 10 and 40. Hence we expect the low transfer to shift norms downwards and the high transfer to shift them upwards.

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

Figure 2: Average punishment decisions and frequency of punishment for different transfers of the dictator elicited via the strategy method.



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 with at least one deduction point.

3.1 Social Norm Perceptions and Punishment

Figure 3 shows the distribution of norm perceptions that subjects hold at the end of part A, i.e., after the punishment phase. Average personal norms (mean = 29.80, sd = 19.40) are significantly ($p < 0.001$) higher than normative expectations (mean = 24.44, sd = 12.89), which are significantly ($p < 0.001$) higher than empirical expectations (mean = 19.57, sd = 12.77).[21] This means that individuals believe that the appropriate transfer according to themselves is higher than what they think others, on average, think is appropriate. The expectation of what is usually done is even lower.[22] Additionally, the figure also reveals that there is no consensus about what subjects believe should be done and what they believe constitutes the injunctive and descriptive norm. In the following, we study the correlations between subjects' punishment decisions and their norm perceptions. In particular, we examine whether the decisions are driven more by what subjects believe should be done, what they believe others believe should be done, or what they think is usually done. For this, we run regressions with all three norm perceptions and all subsets of combinations of the three as independent variables.

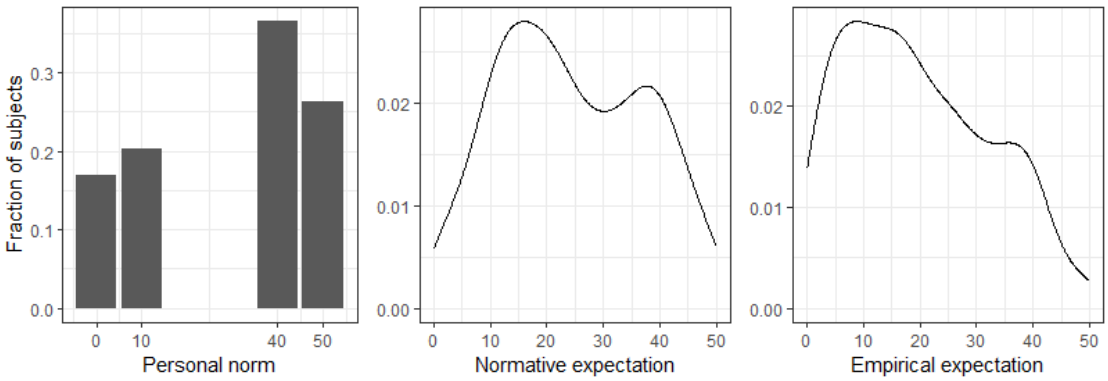
Table 1 shows the results of a Tobit regression.[23] In all models, we control for the dummy variables Transfer10 and Transfer40, which indicate the transfer of the to-be punished dictator in the strategy

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

[22] Note that there is no significant difference ($p = 0.8012$) between how accurate subjects were in predicting what others on average believe is appropriate and in what is usually done. The mean absolute difference between normative expectations and the true mean of everyone's personal norms of appropriateness of the specific session is 12.74 (sd=8.35). The mean absolute difference of empirical expectations and the true mean of the behavior of a session is 13.08 (sd=8.78).

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

Figure 3: Densities of second elicitations of norm perceptions



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

method.[24] All three norm perceptions are z-scored. Because third-party punishment is known to correlate with negative emotions (Jordan, McAuliffe & Rand 2016, Carpenter & Matthews 2012, Nelissen & Zeelenberg 2009), we control for negative emotions in all models. For this, we declare a variable 'negative emotions' that consists of the average of anger, irritation, surprise, and envy[25]. We additionally include the differences in negative emotions of the second and first elicitation to capture individual changes caused by the experience.[26] 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.[27]

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.[28] 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. In appendix A.1, we also include all possible subsets of the combinations of the three norm perceptions.

[24]We find a statistically significant negative influence of both Transfer dummies in all models, with a more pronounced negative impact of 40. This gives evidence that subjects want to enforce fairness because punishment decreases with a higher transfer of the dictator.

[25]For the variables combined in negative emotions, we find a Cronbach's α of 0.69 ($CI_{95\%} = [0.66, 0.72]$). Hence, the variable is internally consistent.

[26]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.

[27]Note that the results do not substantially change (and remain statistically significant) after the exclusion of negative emotions and their differences. We include them nonetheless to already tackle endogeneity problems caused by omitted variables.

[28]Spearman correlation between personal norms and empirical expectations: $\rho = 0.448, p < 0.001$, personal norms and normative expectations: $\rho = 0.642, p < 0.001$, and normative expectations and empirical expectations: $\rho = 0.626, p < 0.001$.

Table 1: Tobit regression of relationship between norm perceptions and punishment

Personal Norms	4.69^{***}			4.42^{**}
	(1.12)			(1.44)
Normative Expect.		2.47^{*}		-3.18^{*}
		(1.06)		(1.55)
Empirical Expect.			4.73^{***}	4.56^{**}
			(1.06)	(1.45)
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)
Transfer10	-4.23^{***}	-4.19^{***}	-4.24^{***}	-4.26^{***}
	(0.84)	(0.83)	(0.83)	(0.84)
Transfer40	-11.52^{***}	-11.45^{***}	-11.50^{***}	-11.56^{***}
	(1.29)	(1.28)	(1.28)	(1.29)
Constant	8.17^{**}	7.16^{*}	6.93^{*}	8.05^{**}
	(2.88)	(2.95)	(2.87)	(2.83)
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 $+\ p < 0.1$; $*\ p < 0.05$; $**\ p < 0.01$; $***\ p < 0.001$

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 norms 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 relationship between normative expectations and punishment decisions becomes negative.[29] The relationship between normative expectations and punishment becomes insignificant when regressing combinations of only two of the three social norm perceptions in Appendix A.1 (table 3), underlining the importance of personal norms of appropriateness and empirical expectations for punishment decisions, whereas normative expectations are not positively associated with punishment when controlling for either of above mentioned.

We replicate our findings in all of the following robustness checks (see Appendix A.1). 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 with 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. Lastly, we include the subject's choices of part B of the experiment, where they themselves made a transfer decision as a dictator. We do not find any significant association between this transfer and their punishment decisions. Summarizing we find that:

Result 1.1 *Personal norms of appropriateness are positively associated with punishment decisions.*

Result 1.2 *Empirical expectations are positively associated with punishment decisions.*

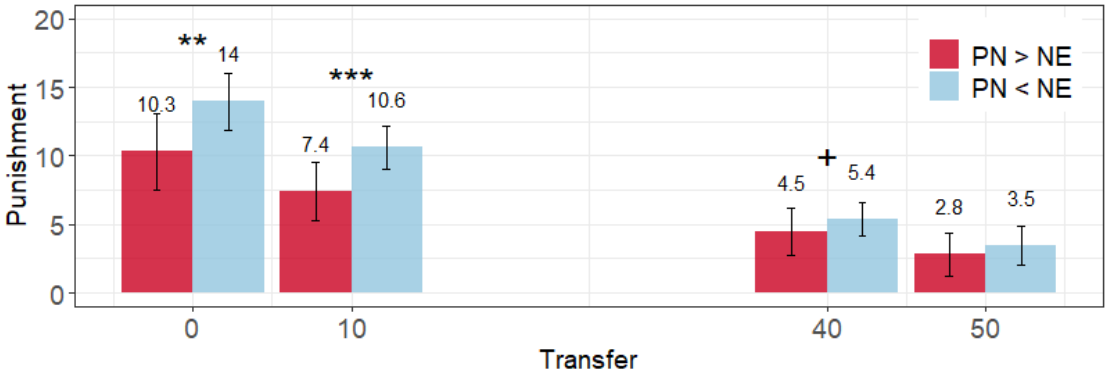
Result 1.3 *After controlling for personal norms and/ or empirical expectations, normative expectations are negatively associated with punishment decisions.*

Normative Expectations and the Gap to Personal Norms

Surprisingly, beliefs about what everyone else deems appropriate are not associated with higher punishment decisions. On the contrary, they are associated with even smaller punishment decisions after controlling for either personal norms or empirical expectations. One explanation for why the estimate of normative expectations becomes negative is the discrepancy between personal norms (what one personally believes is right) and normative expectations (what one believes about what others believe is right). If normative expectations are higher, subjects can free-ride on the punishment of others and punish less. If personal norms are high and normative expectations low, subjects may perceive their

[29] Normative expectations are also significantly negatively correlated with the propensity to punish when studying the extensive margin in appendix A.1 table 4.

Figure 4: Gap between personal norms and empirical expectations and punishment



*Note: The figure shows punishment decisions conditional on the transfer of the to-be-punished dictator (Transfer) for subjects with either a higher (PN > NE) or lower (PN < NE) personal norms than normative expectations. Error bars show 95% confidence intervals and are based on standard errors that are clustered on individual level. + $p < 0.1$; ** $p < 0.01$; *** $p < 0.001$ based on Wilcoxon rank sum tests.*

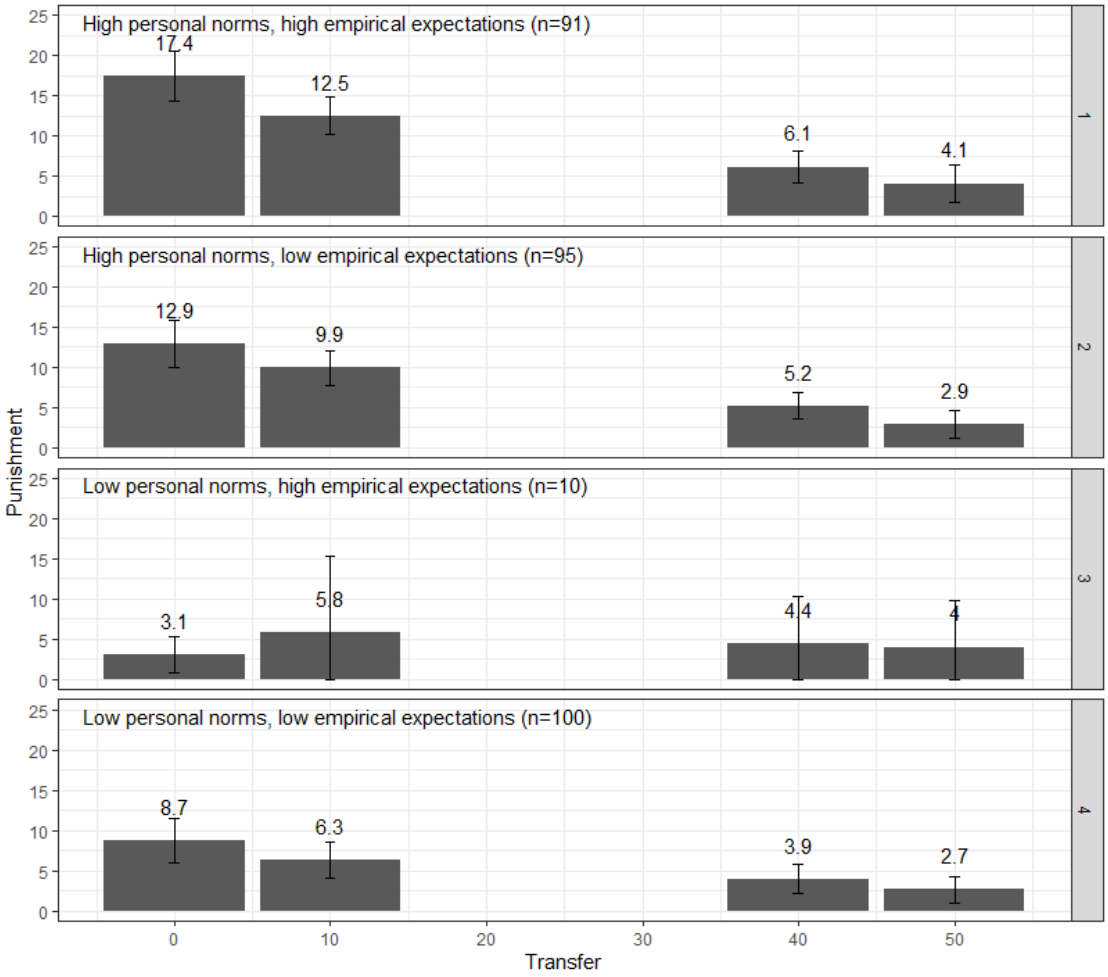
role in punishment as crucial and punish more. Figure 4 illustrates how this gap between personal norms and normative expectations affects punishment. When subjects hold higher personal norms than normative expectations (when what they personally believe is the right transfer is higher than what they believe others believe is the right transfer) they significantly (at least $p < 0.1$) punish more than when they hold lower personal norms than normative expectations (red vs. blue bars). Model (4) of table 3 in the appendix A.1 confirms that punishment is higher when the gap between personal norms and normative expectations increases.

Illustration of the Importance of Personal Norms and Empirical Expectations

In the following, we further explore the positive correlation of personal norms and empirical expectations with punishment decisions, as well as their relative importance for punishment. The magnitude of the positive correlations between personal norms and empirical expectations with punishment is similar (see table 1), suggesting that they both have the same importance for punishment. However, it is not clear which of the two is more important for punishment decisions. This is because we elicit personal norms on a discrete scale (0, 10, 40, 50) while empirical expectations on a continuous one (0-50), which introduces 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. Therefore, it seems that personal norms are slightly more important for punishment than empirical expectations.

Figure 5 illustrates the importance and the interplay of personal norms and empirical expectations

Figure 5: Norm types and punishment



Note: The figure shows punishment decisions conditional on the transfer of the to-be-punished dictator (Transfer) for the four types of punishers w.r.t. to their norm classification (rows 1-4). Subjects are classified as having high personal norms (high empirical expectation) if their personal norms are above 25 (25), and as low otherwise. Error bars show 95% confidence intervals and are based on standard errors that are clustered on an individual level.

for punishment decisions. For this, we classify subjects according to both their personal norms and empirical expectations. Specifically, we consider subjects holding a *low* personal norms if their personal norms are below 25 (i.e., if it is 0 or 10) and holding *low* empirical expectations if they are below 25.[30] Figure 5 shows punishment decisions conditional on the type and the transfer of the to-be-punished dictator (Transfer). It reveals that punishment decisions are highest for subjects with a high personal norms and a high empirical expectations (row 1), followed by subjects with high personal norms but low empirical expectations (row 2). We cannot say much about subjects with low personal norms and high empirical expectations (row 3) because only 10 subjects are classified as this, and these subjects

[30] 25 is the mean of all possible transfers.

seem to be a special case, as there seems to be an increase in the antisocial punishment of these subjects. Finally, subjects with low personal norms and low empirical expectations punish relatively less (row 4). To illustrate the importance of social norm perceptions (as an upper bound), we compare row 1 to row 4 for a Transfer of 0. Punishment is significantly ($p < 0.001$) and substantially (Cohen's d = 0.60) larger for subjects with high personal norms and high empirical expectations compared to subjects with low personal norms and low empirical expectations. All other pair-wise comparisons of the rows for a Transfer of 0 yield significant differences (except row 3 to 4 is not significant, and row 2 to 3 only marginally at $p = 0.052$, otherwise at least $p < 0.05$) and Cohen's d of 0.31 (row 1 and 2), 1.01 (row 1 and 3), 0.71 (row 2 and 3), 0.29 (row 2 and 4), and 0.41 (row 3 and 4). This shows that personal norms and empirical expectations are substantially and significantly correlated with punishment decisions.

The figure further illustrates that both personal norms and empirical expectations are important for punishment decisions. It suggests that high personal norms are more important for punishment than high empirical expectations. However, high empirical expectations almost always come with high personal norms and therefore may be equally important.

3.2 The Causal Effect of Norm Perceptions

So far, we have uncovered a statistically significant relationship between norm perceptions and punishment decisions. This relationship, however, may not be causal because of a potential reverse causality issue of the elicited norm perceptions and punishment decisions. Specifically, punishment itself may have an influence on the second elicitation of norm perceptions because, for instance, punishers may want to provide a reason for why they punished or not. To understand whether this is an issue, we compare the first to the second elicitation of social norm perceptions in the Baseline. In this treatment, only the punishment itself lies in between the two elicitations, hence any changes in the elicitation can be attributed to the punishment section. We do not find significant differences for empirical expectations (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 (see appendix A.2.1). This indicates that the punishment does not significantly influence empirical expectations, and the potential influence of punishment on empirical expectations is negligible. Yet, it may be an issue for normative expectations and personal norms. Specifically, in the correlation analysis, we may have overestimated both the effect of personal norms and normative expectations on punishment. It is comforting, though, that 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 (see appendix A.1, table 4), with a slightly larger negative coefficient for normative expectations and a lower positive coefficient for personal norms, but similar coefficients for empirical expectations. Furthermore, we do not find any significant difference between the first elicitations of all norm perceptions of the punishers and the punishees (see appendix A.2.1). Punishees are the subjects, whose behavior was punished by the punishers and therefore, the knowledge about the upcoming punishment opportunity did not change the social norm perceptions. Hence, reverse causality should be less of a concern.

Nonetheless, to overcome a potential problem of reverse causality and other potential endogeneity issues, we use 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:

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

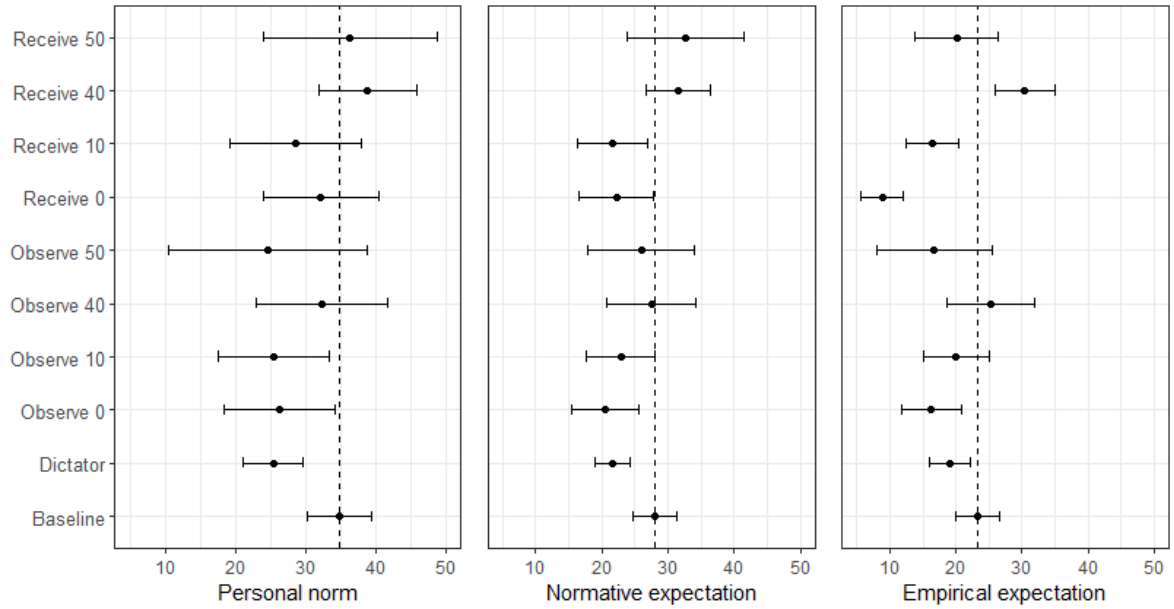
Further note that (ν_i, ϵ_i) are assumed to be distributed multivariate normal. Therefore, the first and second stages are estimated together by Maximum Likelihood.[31] Note that $punishment_i^*$ denotes one single punishment decision for a transfer of either 0, 10, or 40 of one subject. We cluster standard 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.[32]

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

[32] While the first measure shows the emotions of subjects after the punishment, the second variable captures whether emotions were triggered in the experience and punishment part. Hence, the difference accounts for individual changes

Figure 6 shows the second elicitation of norm perceptions conditional on the exogenous treatments: subjects were assigned either to the role of the dictator, or received or observed one of the following transfers from the dictators (% of cases): 0 (33.75%), 10 (30%), 40 (22.5%), or 50 (13.75%) out of 100 CZK.

Figure 6: Instruments and norm perceptions



Note: The figure shows the second norm elicitation for personal norms, normative expectations, and empirical expectations conditional on the specific treatment (instruments). The dotted line shows the average elicitation of the baseline, which serves as the point of comparison for the instruments. Error bars show 95% confidence intervals.

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 (see figure 6 for an illustration of this shift. For a linear regression that also illustrates the shift of negative emotions, see appendix A.2.2). 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

that are produced by the experience (and punishment) phase.

significantly. Being assigned to the role of the dictator significantly shifts personal norms, normative expectations, and empirical expectations. We conclude that the instruments significantly 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, McAuliffe & Rand 2016, Carpenter & Matthews 2012, Nelissen & Zeelenberg 2009). We confirm this by finding a (marginally) significant correlation between the difference of negative emotions and punishment (see table 1). Additionally, we observe that the instruments influence negative emotions and the difference of negative emotions in the case of receiving a transfer or being assigned to the role of a dictator ((see appendix A.2.2, table 9). Hence, we include both negative emotions and the difference of negative emotions.[33]

Table 2: Tobit IV regression punishment on norm perceptions and neg. 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)	3.63 (8.59)	13.03^{+} (7.51)	14,7 (11.01)
Transfer10	-4.26^{***} (0.84)	-4.21^{***} (0.83)	-4.25^{***} (0.83)	-4.27^{***} (0.84)
Transfer40	-11.56^{***} (1.29)	-11.5^{***} (1.29)	-11.51^{***} (1.28)	-11.57^{***} (1.29)
Constant	0.21 (19.46)	−4.04 (18.943)	1.10 (12.56)	1.98 (16.36)
Observations	888	888	888	888
Log Likelihood	−8474.89	−8116.57	−8059.94	−14938
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

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 the model (2) normative expectations, and in model (3) we instrument for empirical expectations. In model (4) we instrument for personal norms, normative expectations, and empirical expectations simultaneously. In all models, we additionally

[33]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.

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 appendix A.2.3).

In model (1), we find a mild statistically significant ($p < 0.1$) positive effect of personal norms on punishment. In models (2) and (3) we find a statistically significant ($p < 0.05$ & $p < 0.01$) positive influence of normative and empirical expectations on punishment. We take this as evidence of a causal positive influence of all three norm perceptions individually on punishment.

In model (4), we do not find any statistically significant impact of any of the three norm perceptions. One reason why we cannot identify the influence of each norm perception, while simultaneously instrumenting all of them, is that all three norm perceptions are shifted in the same direction (see figure 6). Therefore, there is not enough variability to disentangle the influence of each of them. Additionally, the instruments mostly shift empirical expectations, and therefore the influence of empirical expectations seems to remain the most stable. Nonetheless, we argue that it is appropriate to conclude that each of the norm perceptions on its own has a positive effect on punishment.

Result 2.1 *Each of personal norms, normative expectations, and empirical expectations has a causal positive effect on its own on punishment decisions.*

3.3 Heterogeneity in Norm-driven Punishment

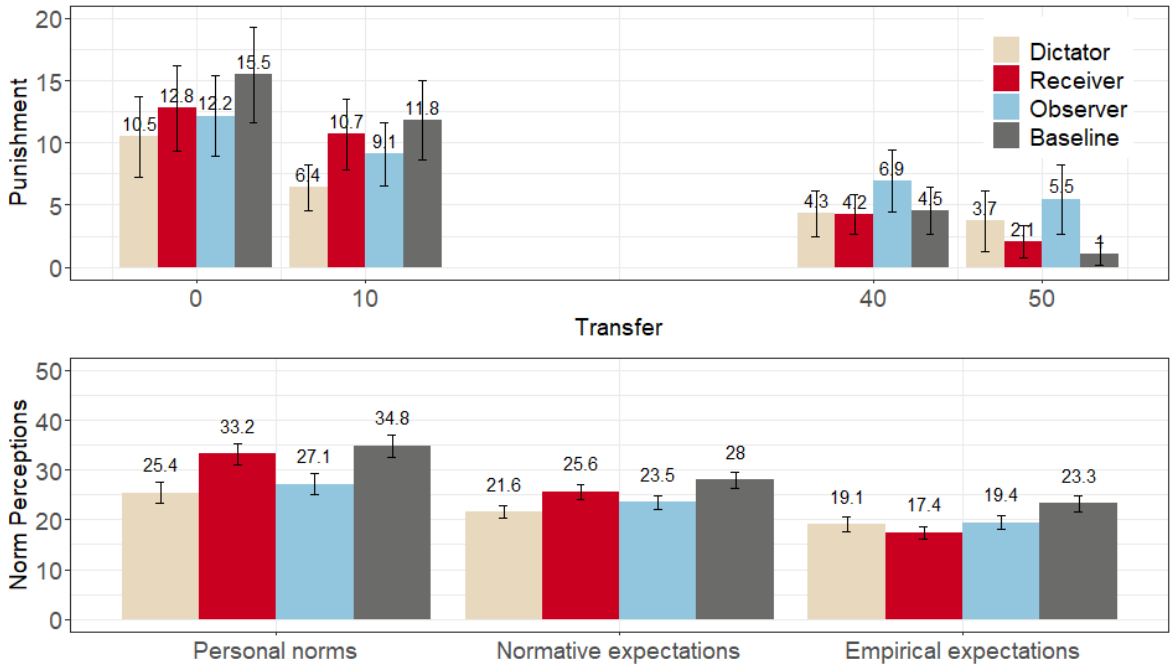
So far, we have identified a causal impact of each norm perception on its own on third-party punishment and uncovered their individual importance for punishment when they are jointly associated with punishment.

In this section, we study how punishment and the relative importance of the three norm perceptions may differ between subjects. For this, 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.

Role Assignment and Norm-driven Punishment

Figure 7 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 as dictators in the experience phase significantly (at least $p < 0.05$) reduces punishment for a transfer of 0 and 10 compared to the Baseline. Additionally, being in the Observer treatment significantly (at least $p < 0.05$) increases the punishment of a transfer of 50 compared to the Baseline and the Receiver treatment. This increase in antisocial punishment is mostly driven by subjects in the Observer treatment, who hold low personal norms and observe a high transfer.

Figure 7: Roles, 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.

More 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.

Next, we explore whether the difference in the norm-driven punishment is entirely due to the treatment assignment or whether the same norm-driven correlations also exist within each treatment. For this, we regress punishment decisions on all norm perceptions for each of the treatments separately (see appendix A.3.1). The regression reveals that norm perceptions are significantly associated in the case of the Dictator and the Receiver treatments. The associations are the same as the overall associations, i.e., positive associations of personal norms and empirical expectations with punishment and a negative association of normative expectations with punishment. In the case of the Observer treatment and Baseline, the associations between the norm perceptions can be replicated qualitatively but not on a statistically significant level. This indicates that the importance of the norm perceptions

is not entirely due to a shift in the norm perceptions because of the treatment assignment but that there exist within-treatment norm heterogeneities that are associated with punishment.

Furthermore, the lower panel of figure 7 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 appendix A.3.1). 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.

Gender and (Norm-driven) Punishment

Next, we explore the importance of the norm perceptions depending on gender (see appendix A.3.2). For this, we first split the sample into only males and only females and run two separate regressions. We find that for males, only personal norms are significantly correlated with punishment decisions, and the estimate for empirical expectations becomes very low. For females, on the other hand, we find the opposite; normative and empirical expectations are significantly correlated with punishment, and the estimate for personal norms becomes low. Running a regression on the full sample with interaction effects between gender and the norm perceptions yields the same results. We conclude that females tend to punish according to their beliefs of what others deem appropriate behavior and common behavior, whereas males tend to do so according to what they personally believe is appropriate, without taking their beliefs about the social norms into account. It seems that males do not care about enforcing a social norm but merely care about what they personally feel would be appropriate in the situation. Females on the other hand care more about enforcing social norms. The importance of gender for norm-driven punishment illustrates that there exist heterogeneities in the importance of each norm perception and punishment.

4 Discussion

In this paper, we show that punishers' perceptions of social norms matter for third-party punishment decisions. Unlike the existing literature on social norms and punishment (Li et al. 2021, House et al. 2020, Kamei 2020, Fabbri & Carbonara 2017, Fišar et al. 2016, Reuben & Riedl 2013, Carpenter &

Matthews 2012, 2009), we explicitly measure three social norm perceptions about the behavior in a specific situation. In addition, we experimentally manipulate them and employ an IV approach to estimate their causal effect on punishment.

We find that all three norm perceptions individually have a positive causal influence on third-party punishment. When studying the correlation of all norm perceptions with punishment simultaneously, we find that both the personal belief about what should be done – personal norms – and the belief about what is usually done – empirical expectations – are positively correlated with third-party punishment decisions. Additionally, we find that the beliefs about what others believe should be done – normative expectations – are negatively correlated with punishment. This negative association also holds after controlling for only either personal norms or empirical expectations.

Thus, we provide evidence that third-party punishment by itself is evidence that humans care for social norms and their enforcement (e.g. Carpenter & Matthews 2012, 2009, Henrich et al. 2006, Fehr & Fischbacher 2004). However, the fact that punishers do not only rely on their empirical expectations but also on their personal norms indicates that third-party punishment is not only used to enforce social norms. It is also used to potentially *create* social norms based on own personal views of what is right. This is because punishers base their punishment decisions also on their personal norms and hence enforce behavior that might be different from what is usually done. Thus, punishment based on personal norms may shift common behavior and create new social norms. Therefore, third-party punishment can be viewed both as evidence for the existence of and as a tool to potentially create particular social norms.

Our finding that empirical expectations causally affect third-party punishment extends Carpenter & Matthews (2012) who find a positive correlation between the belief about average behavior (empirical expectations) with third-party punishment. It can further explain Carpenter & Matthews (2009)'s horse race of punishment-triggering norms. They find that the session's averages best explain punishment compared to the own group's averages, in- or out-group averages, own contributions, or the respective medians, minima, or maxima. Sessions' averages provide the best information to update own empirical expectations and therefore make it the most important information for punishment. We also confirm the result of Bašić & Verrina (2021), who show that personal norms are positively correlated with third-party punishment decisions. We not only find a robust positive correlation with punishment after controlling for all subsets of empirical and normative expectations, but also provide evidence for a positive causal influence of personal norms on punishment.

Our finding that third-party punishment is not driven by the belief of the injunctive norms – normative expectations – is seemingly surprising, given that Bašić & Verrina (2021) show that normative

expectations, together with personal norms, are positively correlated with punishment and that House et al. (2020) and Dimant & Gesche (2021) show that injunctive norm nudges increase punishment decisions. We argue that subjects whose normative expectations are higher than their personal beliefs may anticipate others to punish more since they expect others to hold higher personal norms. They could then free-ride on their peers' punishment decisions and punish less. Conversely, subjects whose personal norms are higher than their normative expectations perceive their own punishment as more crucial. This explanation is supported by the finding that the gap between personal norms and normative expectations positively correlates with punishment. This is in line with Kamei et al. (2023) who show that in the presence of other punishers, subjects free-ride on others' punishment decisions, with a bigger effect in the UK compared to India. Their findings can be explained by generally larger punishment decisions in the UK and, therefore, potentially higher normative expectations than in India, which allows them to free-ride more.

By controlling for all three norm perceptions simultaneously, we also confirm the findings of the literature 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. 2016, Bicchieri & Xiao 2009).

Our paper also adds to the literature about how normative and empirical information nudges affect third-party punishment. We show that the provided information about common behavior affects not only empirical expectations but also personal norms and normative expectations. This fact can explain the results of House et al. (2020), Dimant & Gesche (2021), and Zong et al. (2021). House et al. (2020) find that injunctive norm nudges increase punishment decisions. Based on our results, the provided information about the injunctive norm changes possibly also changes empirical expectations, which ultimately increases third-party punishment. In a subsequent study, Dimant & Gesche (2021) show that empirical and normative information increase third-party punishment decisions and personal norms of appropriateness. However, they do not measure the effect of information on empirical and normative expectations, nor measure the correlation of personal norms with punishment. Their information nudges potentially change all three norm perceptions and through these ultimately punishment. In addition, their results of slightly higher third-party punishment decisions after providing empirical compared to normative information can be explained by our finding of a higher relative importance of empirical compared to normative expectations. In Zong et al. (2021), punishers react to the sender's expectation in a trust game. The sender's expectation may give a signal not only about injunctive norms but also about descriptive norms, and hence influence also the punishers' empirical expectations and through them third-party punishment decisions.

We can also explain punishment differences across societies (e.g. Kamei et al. 2023, Henrich et al. 2006) by showing that punishment varies with different norm perceptions. As different societies have different social norms, the beliefs about their social norms likely also differ, and hence punishment decisions do as well.

Finally, our results may also explain the results of Kamei (2020), Lois & Wessa (2019), Fabbri & Carbonara (2017) who show that punishment decisions are correlated with beliefs and information about others' punishment decisions. Others' punishment decisions may not only create own norms of punishment, but they 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.

Furthermore, we explore heterogeneities in the relative importance of norm perceptions and find that it only slightly differs between the specific roles subjects were assigned to. In particular, being a receiver in the dictator game increases the relative importance of empirical expectations for punishment. For all the other roles (treatments), there is no significant difference in the importance of the norm perceptions for punishment. This aligns with Vollan et al. (2020) who do not find any difference in third-party punishment decisions between traditional leaders exogenously chosen by birth and democratically elected leaders in Namibia.

Moreover, we find that males' punishment decisions are significantly more related to their personal norms (what they believe should be done) and significantly less to their empirical expectations (what they believe is common behavior) compared to females. Even more, males' punishment decisions are only positively correlated with their personal norms and not significantly with normative or empirical expectations. Males do not consider their beliefs of the underlying social norms but punish merely based on what they personally believe constitutes appropriate behavior. Females, on the other hand, primarily punish according to what they believe constitutes common behavior. This finding illustrates that the importance of each norm perception may differ based on individual characteristics. The stronger reliance on empirical expectations for females contrasts to Croson et al. (2010) who find that males' donation decisions are more influenced by their empirical expectations of others' donation behavior than females'. On the other hand, Fišar et al. (2016) do not find any gender differences in the positive relationship between the empirical expectations of offering, accepting, or sanctioning a bribe and third-party punishment decisions.

Finally, we also show that the mere assignment to the role of dictator leads to lower punishment decisions. In accordance with the literature on motivated beliefs (e.g. Zimmermann 2020, Epley & Gilovich 2016), we find that being assigned to the role as dictator significantly decreases all of their

social norm perceptions, including their own personal norms of appropriateness. By doing so, they reduce their need to give a higher transfer and, consequently, their need to punish other unrelated dictators. We conclude that using the specific social norm perception is heterogeneous and varies based on gender, the assigned role, and possibly other factors.

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 providing information about 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 (2021), who show that empirical information changes behavior more (but non-significantly) than normative information. Similarly, policymakers should be aware of how signals of common behavior may erode social norm perceptions and hence decrease third-party punishment and should try to balance providing information about both norm violations and norm alignment behavior, for instance in the media.

Our study comes with certain drawbacks: Most importantly, empirical expectations and their effect on punishment might be biased by using the strategy method. Because of the strategy method, only transfers that occur have to be paid. Subjects with higher empirical expectations anticipate higher transfers of a dictator. Therefore, they may punish low transfers because they are less likely to occur, and they would not have to pay for them. In this way, they can signal their willingness to punish and enforce social norms without actually having to pay for it. When empirical expectations are low, subjects expect dictators to transfer less, and hence any punishment of low transfers has to be paid with a higher likelihood. Those subjects may then punish low transfers less. This avoidance strategy may therefore drive the positive association of empirical expectations with punishment decisions.

To check whether this is an issue, we run a robustness check (see appendix A.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 do not find any significant punishment differences if the to-be-punished transfer is within a predefined neighborhood of 5 and 25. We do, however, find that subjects punish significantly less if the to-be punished transfer is within a predefined neighborhood of 15, yet only when we do not condition the dummy on the to-be punished transfer (0, 10, or 40). More importantly, 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 consistently show that empirical expectations matter for punishment decisions.

In addition, our results could be affected by the fact that we elicited norm perceptions before

punishment. The elicitation itself could push subjects to align their punishment decisions to their elicited norm perceptions. However, d'Adda et al. (2016) show that the order of the norm elicitation does not significantly alter punishment behavior.

5 Conclusion

We run an incentivized experiment to study whether and how appropriateness beliefs (*personal norms*), the beliefs about others' appropriateness beliefs (*normative expectations*), and beliefs about common behavior (*empirical expectations*) shape the punishment decisions of third parties. The existing literature suggests that third-party punishment is due to the willingness to costly enforce social norms (Carpenter & Matthews 2012, 2009, Henrich et al. 2006, Fehr & Fischbacher 2004). Yet, to the best of our knowledge, no existing study explicitly measures three different beliefs that matter in the context of social norms – *social norm perceptions* – and identifies their causal effect on third-party punishment. In our paper, we show that individuals use third-party punishment to enforce their beliefs of common behavior but also to enforce their personal beliefs of appropriate behavior. Therefore, we show that third-party punishment is used for the enforcement of social norms, and by the reliance on their personal appropriateness beliefs also for the enforcement of other ways of behavior that might result in new social norms.

We measure personal norms, normative expectations, and empirical expectations and employ four treatments that manipulate the experience subjects make prior to making punishment decisions. In the Dictator and Receiver treatment, subjects play the same variant of the dictator game for which they subsequently make decisions in the role of third-party punishers. In the Observer treatment, subjects observe a single decision from the previously conducted Dictator treatment. The Baseline does not involve any additional tasks before the punishment decisions. The treatments serve to exogenously shift all three norm perceptions, allowing us to identify their causal effects. We employ an Instrumental Variable approach where we use the treatments to instrument the norm perceptions. We also use the treatments to instrument for the influence of negative emotions that are triggered by the experience and have been found to correlate with third-party punishment decisions (Jordan, McAuliffe & Rand 2016, Carpenter & Matthews 2012, Nelissen & Zeelenberg 2009). In this way, we can exclude the channel of negative emotions that may be correlated with the channel of norm perceptions and hence we give more validity to the exclusion restriction.

We find that each of all three social norm perceptions has a positive causal impact on third-party punishment decisions on its own. When studying how they jointly correlate with punishment, we find that both personal norms and empirical expectations are positively associated with punishment

decisions and that normative expectations are negatively correlated. We conclude that individuals punish because they want to enforce what they believe constitutes average behavior, and because they want to enforce what they personally think is appropriate. Surprisingly, individuals do not punish according to what they think constitutes the injunctive norm, i.e., what all believe on average should be done. Instead, their normative expectations are negatively correlated with punishment. One of the reasons is that higher normative expectations imply that individuals think that others hold higher personal norms. Therefore, they may anticipate that others punish more and that they can free-ride on their peers' punishment decisions. We confirm this by finding a positive correlation between the gap between personal norms and normative expectations with punishment decisions.

We conclude that individuals rely both on their personal norms and empirical expectations when deciding on punishment decisions, and seem to use normative expectations to free-ride on others' punishment decisions. The fact that punishers do not only rely on their empirical expectations but also on their personal norms indicates that third-party punishment is not only used to enforce social norms but also to enforce punishers' own personal views which potentially creates new social norms. Therefore, third-party punishment can be viewed both as evidence and as one of the tools to create particular social norms.

We cannot say whether personal norms or empirical expectations are more important for punishment decisions. However, we find that their relative importance depends on gender and illustrates that there are heterogeneities in the use of social norm perceptions. We find that females punish more according to what they believe constitutes common behavior (empirical expectations), whereas males punish more according to what they personally believe constitutes appropriate behavior (personal norms). Furthermore, the situation and the experience that subjects are exogenously put in shape the importance of the relative importance of personal norms and empirical expectations. Being a receiver in the dictator game significantly increases the importance of empirical expectations compared to all other treatments. Finally, we find that being assigned to the role of a dictator significantly decreases punishment decisions, arguably because dictators (motivatedly) downwards shift their norm perceptions to allow themselves to punish less.

Our results have policy implications about how to best trigger third-party punishment. We find that empirical expectations and personal norms positively influence punishment, whereas normative expectations are negatively correlated. Therefore, policies should focus on shifting empirical expectations, and if possible personal norms, to most efficiently increase third-party punishment.

Acknowledgements

We are grateful to Rostislav Staněk, Henrik Orzen, Wladislaw Mill, Marco Faillo, James Tremewan, Ondřej Krčál, Aleksandra Khokhlova, 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

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.

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

Bašić, Z. & Verrina, E. (2021), 'Personal norms—and not only social norms—shape economic behavior', *MPI Collective Goods Discussion Paper* (2020/25).

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. & 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. (2016), 'Conformity and adaptation in groups'.

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

Carpenter, J. & 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.

Dimant, E. & Gesche, T. (2021), 'Nudging enforcers: How norm perceptions and motives for lying shape sanctions'.

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.

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.

Gächter, S. & Renner, E. (2010), 'The effects of (incentivized) belief elicitation in public goods experiments', *Experimental Economics* **13**(3), 364–377.

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. (2019), '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. J., Hoffman, M., Bloom, P. & Rand, D. G. (2016), 'Third-party punishment as a costly signal of trustworthiness', *Nature* **530**(7591), 473–476.

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.
URL: *https://science.sciencemag.org/content/322/5908/1681*

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.

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.

Vollan, B., Blanco, E., Steimanis, I., Petutschnig, F. & Prediger, S. (2020), 'Procedural fairness and nepotism among local traditional and democratic leaders in rural namibia', *Science Advances* **6**(15), eaay7651.

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* .

Appendix

A Additional Models and Robustness Checks

A.1 Social Norm Perceptions and Punishment

A.1.1 Combinations of Norm Perceptions and Punishment

First, we regress punishment on combinations of social norm perceptions and punishment, where we take only two out of the three social norm perceptions. We replicate the results from the main text, namely, that empirical expectations are significantly positively correlated with punishment, that personal norms are significantly positively correlated, and that normative expectations are (however non-significantly) correlated when controlling for any combination of personal norms and empirical expectations.

Table 3: Tobit regression punishment on social norm perceptions combinations

	Dependent variable:			
	punishment			
	(1)	(2)	(3)	(4)
Personal Norms	5.19***	3.05*		
	(1.41)	(1.33)		
Normative Expect.	−0.79		−0.85	
	(1.32)		(1.44)	
Empirical Expect.		3.26**	5.26***	4.41***
		(1.26)	(1.43)	(1.05)
Gap PN-NE				3.41**
				(1.08)
Constant	8.37**	7.49**	7.01*	8.03**
	(2.90)	(2.88)	(2.86)	(2.85)
Transfer	✓	✓	✓	✓
Neg. Emotions	✓	✓	✓	✓
Δ Neg. Emotions	✓	✓	✓	✓
Observations	888	888	888	888
Log Likelihood	−2,433.54	−2,424.92	−2,432.03	−2,419.73
Wald Test (df = 6)	97.09***	114.17***	101.17***	123.72***

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

In model (4), we include the individual differences between normative expectations and personal norms to explain why normative expectations are negatively correlated when controlling for other norm

perceptions. We find a significant positive association between the gap and punishment decisions. The higher the difference between punishers' belief of what is right and what she thinks others think is right, the higher are punishment decisions.

A.1.2 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. In models (2) and (3), we regress a punishment dummy on the first and second elicitation of norm perceptions to study the extensive margin of punishment. Both models replicate the results from the main section: Personal norms and empirical expectations are significantly positively associated with punishment decisions. We further find that normative expectations are negatively associated with punishment decisions at the 10% and 0.1% significance levels.

Table 4: Tobit and Logit regression punishment on standardized norm perceptions

	Dependent variable:		
	punishment	punishment_dummy	
	Tobit	*logistic*	
	(1)	(2)	(3)
	1st elicitation	1st elicitation	2nd elicitation
Personal Norms	3.63^{*}	0.36^{*}	0.63^{***}
	(1.49)	(1.49)	(0.10)
Normative Expect.	-4.40^{*}	-0.47^{*}	-0.55^{***}
	(1.73)	(1.73)	(0.11)
Empirical Expect.	5.26^{***}	0.61^{***}	0.54^{***}
	(1.55)	(1.55)	(0.10)
Constant	6.99^{*}	0.34^{*}	0.50^{*}
	(2.90)	(2.90)	(0.22)
Transfer	✓	✓	✓
Neg. Emotions	✓	✓	✓
Δ Neg. Emotions	✓	✓	✓
Observations	888	888	888
Log Likelihood	−2,425.14	−558.21	−545.02
Akaike Inf. Crit.		1,132.41	1,106.03
Wald Test	113.88^{***} (df = 7)		

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

Additionally, 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 transfer separately. Table 5 reveals that the results remain robust to this model specification.

Table 5: Tobit regression punishment on deviations of norm perceptions with transfer

	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 $+ p < 0.1$; $^{*} p < 0.05$; $^{**} p < 0.01$; $^{***} p < 0.001$

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

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

	Dependent variable:	
	punishment	
	(1)	(2)
Transfer Part B	−0.04	−0.04
	(0.05)	(0.05)
Personal Norms		4.39**
		(1.45)
Normative Expect.		−3.11*
		(1.55)
Empirical Expect.		4.58**
		(1.45)
Constant	7.74**	8.19**
	(2.94)	(2.83)
Transfer	✓	✓
Neg. Emotions	✓	✓
Δ Neg. Emotions	✓	✓
Observations	888	888
Log Likelihood	−2,456.35	−2,418.92
Wald Test	53.52*** (df = 5)	125.08*** (df = 8)
Note: SE clustered at individual level	+ $p < 0.1$; * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$	

A.2 The Causal Effect of Norm Perceptions

A.2.1 Reverse Causality

In this section, we first study whether reverse causality may be an issue for our identification. Table 7 shows 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. We conclude that the punishment phase increases personal norms and normative expectations but does not significantly change empirical expectations.

Table 7: Comparisons of first and second norm perceptions elicitations in the Baseline

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: Wilcoxon signed rank test with continuity correction; N=60

Moreover, we compare the first elicitation of norm perceptions between punishers and punishees. Table 8 shows no significant differences. Hence, the knowledge about the upcoming punishment opportunity does not change norm perceptions.

Table 8: Comparisons of first norm perceptions 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: Wilcoxon rank sum test with continuity correction

A.2.2 Effects of Treatments on Norm Perceptions

Table 9 shows a linear regression, which estimates the effects of the treatments on social norm perceptions and emotions. Being assigned to the role as dictator significantly decreases all three norm perceptions. In the case of receiving or observing a low transfer, norm perceptions are updated downwards, with the most pronounced changes in empirical expectations after receiving a transfer. After observing a transfer of 0, all three norm perceptions significantly decrease. Receiving a low transfer increases negative emotions, and the difference between negative emotions at the end and the beginning of the experiment decreases for dictators and increases for subjects, who receive a transfer of 0.

Table 9: Linear regression experience on norm perceptions and emotions

	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^{*}
	(3.19)	(2.15)	(2.25)	(0.18)	(0.11)
Receive 0	-2.61	-5.85^{+}	-14.40^{***}	1.05^{***}	0.83^{***}
	(4.57)	(3.18)	(2.26)	(0.29)	(0.20)
Receive 10	-6.20	-6.37^{*}	-6.84^{**}	0.89^{**}	0.21
	(5.03)	(3.00)	(2.52)	(0.28)	(0.17)
Receive 40	3.99	3.58	7.18^{**}	0.82^{**}	0.06
	(3.99)	(2.80)	(2.70)	(0.30)	(0.15)
Receive 50	1.53	4.61	-3.12	-0.08	0.07
	(5.87)	(4.15)	(3.20)	(0.29)	(0.15)
Observe 0	-8.54^{+}	-7.51^{*}	-6.94^{*}	0.27	0.01
	(4.47)	(2.97)	(2.76)	(0.26)	(0.12)
Observe 10	-9.42^{*}	-5.12^{+}	-3.27	-0.01	0.19
	(4.46)	(2.98)	(2.90)	(0.24)	(0.16)
Observe 40	-2.48	-0.52	2.00	-0.49^{+}	-0.10
	(4.91)	(3.54)	(3.50)	(0.27)	(0.12)
Observe 50	-10.29	-2.09	-6.57	0.12	0.02
	(6.56)	(3.88)	(4.12)	(0.39)	(0.13)
Constant	34.83^{***}	28.03^{***}	23.25^{***}	2.47^{***}	0.03
	(2.33)	(1.66)	(1.66)	(0.14)	(0.07)
Observations	888	888	888	888	888
R^2	0.057	0.082	0.148	0.144	0.155
Adjusted R^2	0.045	0.070	0.137	0.134	0.145
Res. Std. Error (df = 876)	18.931	12.409	11.847	1.073	0.663
F Statistic (df = 11; 876)	2.37^{*}	3.31^{***}	9.74^{***}	4.92^{***}	3.49^{***}

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

A.2.3 Robustness Check of Causal Effect of Norm Perceptions

Table 10 replicates the main results without instrumenting for negative emotions and the change in negative emotions. We find a positive influence of each of the norm perceptions individually.

Table 10: 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^{*} (0.24)	0.26 (0.45)
Transfer10	-4.24^{***} (0.84)	-4.20^{***} (0.83)	-4.23^{***} (0.83)	-4.27^{***} (0.84)
Transfer40	-11.50^{***} (1.3)	-11.45^{***} (1.29)	-11.45^{***} (1.28)	-11.55^{***} (1.28)
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^{***}	89.3^{***}

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.

A.3 Heterogeneity in Norm-driven Punishment

A.3.1 Role and the Importance of Social Norm Perceptions for Punishment

This section shows the regression analysis of the role in the experience phase and its interaction with the social norm perceptions. Table 11 shows the results of such a Tobit regression. First of all, model (1) reveals that being assigned to the role of a dictator significantly decreases punishment decisions compared to the Baseline. All other roles also lead to smaller, yet non-significant punishment decisions compared to being in the Baseline. Models (2) to (5) split the data set into the respective treatments. They reveal that the direction of the correlation remains the same for all treatments. Only the effect sizes and the standard errors differ between the treatments. It is noteworthy that in the Receiver treatment, the correlation between empirical expectations and punishment decisions is the strongest.

Table 12 shows the interaction of the role (treatments) with the social norm perceptions. It reveals that indeed being in the role of the receiver increases the importance of empirical expectations for punishment compared to the baseline. In all other treatments, there is no significant difference in the correlation of the norm perceptions with punishment compared to the baseline.

Table 11: Tobit regression punishment on the role and norm perceptions

	Dependent variable:				
	punishment				
	(1)	(2)	(3)	(4)	(5)
Transfer10	-4.16^{***}	-5.62^{**}	-2.94^{+}	-4.44^{**}	-4.17^{***}
	(0.82)	(2.02)	(1.66)	(1.72)	(0.97)
Transfer40	-11.45^{***}	-10.38^{***}	-13.69^{***}	-8.82^{***}	-14.33^{***}
	(1.29)	(2.90)	(2.40)	(2.56)	(2.29)
Dictator	-6.73^{*}				
	(2.93)				
Receiver	-3.28				
	(3.03)				
Observer	-2.43				
	(2.95)				
Personal Norms		5.23^{*}	5.63^{*}	4.52	1.95
		(2.25)	(2.29)	(3.28)	(3.51)
Normative Expect.		-7.45^{*}	-4.29^{+}	-1.84	-0.02
		(3.35)	(2.54)	(3.09)	(3.44)
Empirical Expect.		5.70^{+}	9.17^{***}	3.36	1.24
		(2.97)	(2.62)	(2.60)	(3.86)
Neg. Emotions		3.39^{+}	-1.59	-0.42	-0.89
		(1.94)	(1.71)	(1.60)	(2.52)
Δ Neg. Emotions		-0.61	6.62^{**}	1.96	-0.25
		(1.60)	(2.14)	(3.65)	(3.10)
Constant	11.79^{***}	-3.25	11.83^{*}	9.71^{+}	14.36^{*}
	(2.34)	(5.58)	(5.84)	(4.98)	(6.46)
Treatment	All	Dictator	Receiver	Observer	Baseline
Transfer	✓	✓	✓	✓	✓
Neg. Emotions	✓	✓	✓	✓	✓
Δ Neg. Emotions	✓	✓	✓	✓	✓
Observations	888	240	231	237	180
Log Likelihood	$-2{,}453.59$	-589.12	-594.70	-669.42	-542.13
Wald Test	59.10^{***}	34.34^{***}	75.74^{***}	25.14^{***}	23.92^{**}
	(df = 5)	(df = 7)	(df = 7)	(df = 7)	(df = 7)

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

Table 12: Tobit regression punishment on the role and interaction with norm perceptions

	Dependent variable:			
	punishment			
	(1)	(2)	(3)	(4)
Personal Norms	3.91*	4.29*	4.56**	4.87**
	(1.75)	(1.74)	(1.56)	(1.58)
Normative Expect.	−2.50	−2.34	−3.65*	−4.05*
	(1.74)	(1.98)	(1.80)	(1.67)
Empirical Expect.	4.99**	2.87	4.90**	5.29***
	(1.68)	(1.82)	(1.76)	(1.49)
Dictator	−3.80			
	(2.42)			
Receiver		−1.04		
		(2.45)		
Observer			1.88	
			(2.33)	
Baseline				3.29
				(2.91)
Personal Norm:Treatment	1.47	1.09	−0.19	−3.03
	(2.93)	(2.77)	(3.52)	(4.15)
Normative Expect.:Treatment	−4.78	−1.86	1.90	4.35
	(3.69)	(3.17)	(3.50)	(3.99)
Empirical Expect.:Treatment	−0.03	5.90+	−1.56	−3.87
	(3.18)	(3.25)	(3.16)	(4.14)
Constant	9.06**	8.42**	7.35*	7.29*
	(2.96)	(2.86)	(2.88)	(3.03)
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 + $p < 0.1$; * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$

A.3.2 Gender

In this section, we look closer at gender differences in the importance of norm perceptions for punishment decisions. The existing literature remains ambiguous as to the effect of gender on punishment. 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 gender. For this, we first split the sample into males and females. Table 13 shows the results of a Tobit model, where we regress punishment on standardized norm perceptions for females in the model (1), males in the model (2), and on the full sample in the model (3), with a dummy indicating female, and the interaction of female with all norm perceptions. The interaction effect then 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, whereas personal norms seem not to matter. In model (2), we find the opposite for males; we find a statistically significant positive association between personal norms and punishment for females, whereas empirical and normative expectations seem not to matter for their punishment decisions. Model (3) confirms this pattern.

The difference in the importance of the norms for punishment decisions is not driven by different initial (first elicitation) levels of the norm perceptions between females and males. Table 14 shows average norm perceptions for females and males. The only substantive difference in norm perceptions is between empirical expectations (20.79 for females vs. 18.57 for males), yet the difference is not statistically significant.

Table 13: Tobit regression punishment on interaction norm perceptions conditional on gender

	Dependent variable:		
	punishment		
	(1)	(2)	(3)
Personal Norms	1.65	9.31^{***}	8.76^{***}
	(1.76)	(2.58)	(2.29)
Normative Expect.	-3.75^{*}	-2.83	-2.66
	(1.89)	(2.40)	(2.21)
Empirical Expect.	6.99^{***}	0.67	0.78
	(1.74)	(2.32)	(2.24)
Female			-0.54
			(2.02)
Female:Personal Norms			-7.07^{*}
			(2.97)
Female:Normative Expect.			-1.20
			(2.97)
Female:Empirical Expect.			6.29^{*}
			(2.86)
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	7.80^{*}	9.21^{+}	8.57^{**}
	(3.39)	(4.91)	(3.04)
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 14: Differences first elicitation norm perceptions between gender

Means (SE)	Females *(N=163)*	Males *(N=133)*	p-value
Personal Norms	30 (1.7)	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

We conclude that females punish according to their beliefs of the descriptive norm, whereas males do so according to what they personally believe is appropriate, without taking their beliefs about the injunctive and descriptive norm into account. This result illustrates that there exist heterogeneities in the importance of each of the norms.

A.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 that 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 15 shows a Tobit regression, which includes 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 decisions. However, only in model (2) is the relationship statistically significant, giving some indication that subjects decrease their punishment in the proximity of their empirical expectations. However, the neighborhood variable may just capture higher punishment of a transfer of 0, 10, especially when not including the interaction with the Transfer dummies. Additionally, using of the strategy method is also evaluated by Jordan, McAuliffe & Rand (2016), who find that it does not bias third-party punishment decisions.

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) remains robust and significant for 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 15: 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. + $p < 0.1$; * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$

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 and 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>. N**ow 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 – only experience]

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]

Experimental instructions - Part A [- Stage 2 - experience]

[In stage 2, – experience] 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.

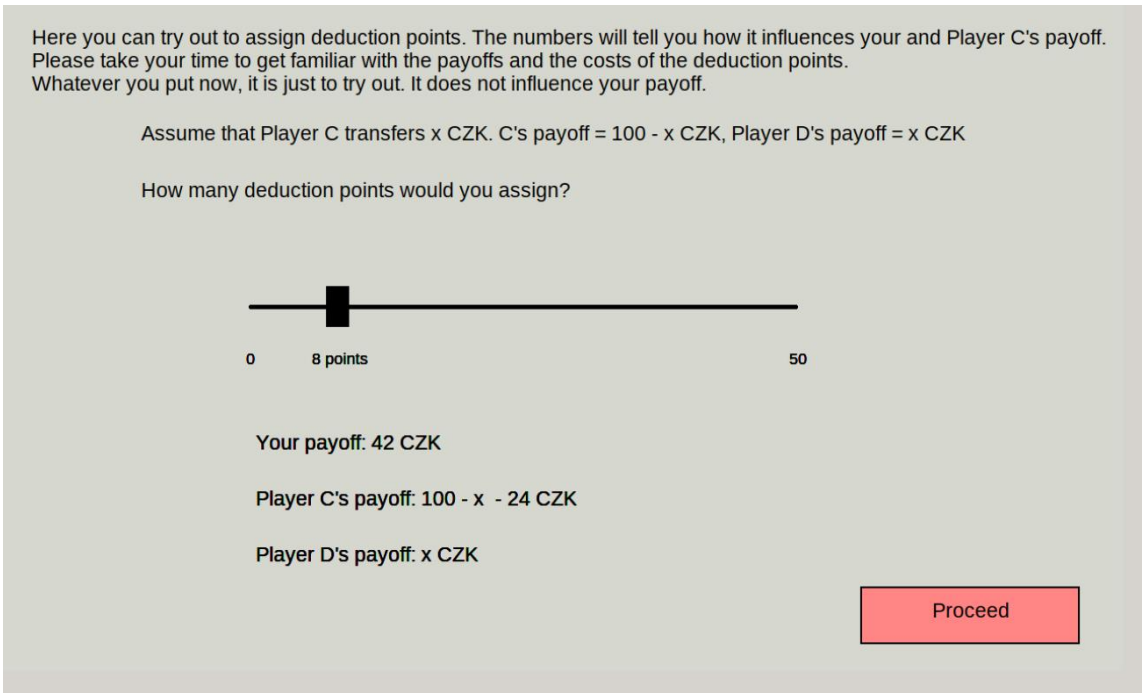
[screen 4 - tryout stage]

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.

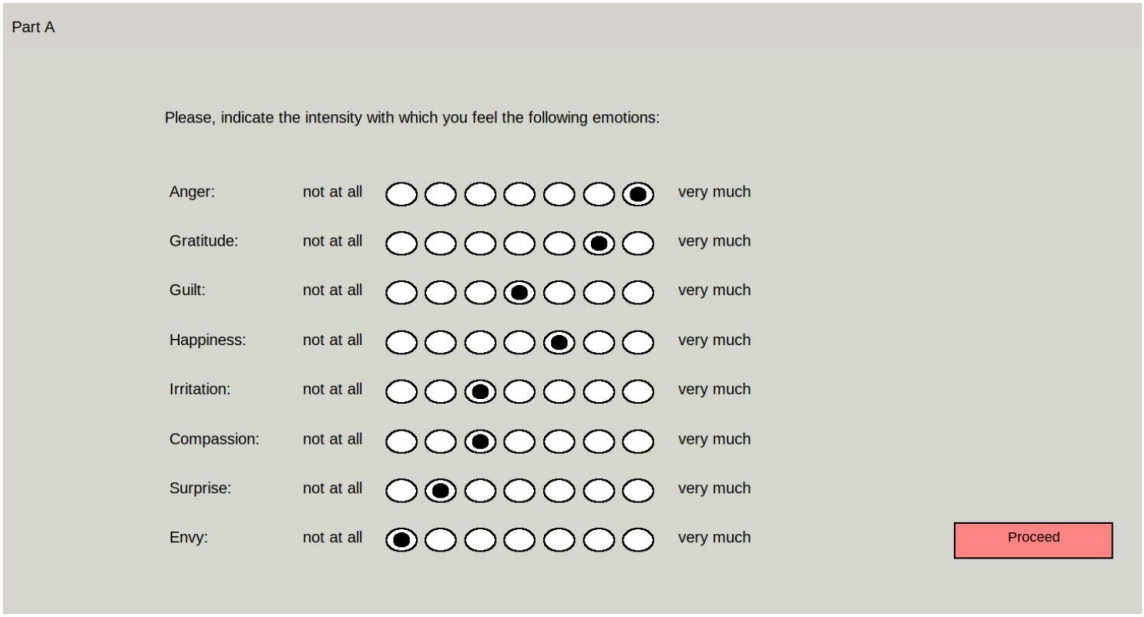
[screen 6 all]

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.

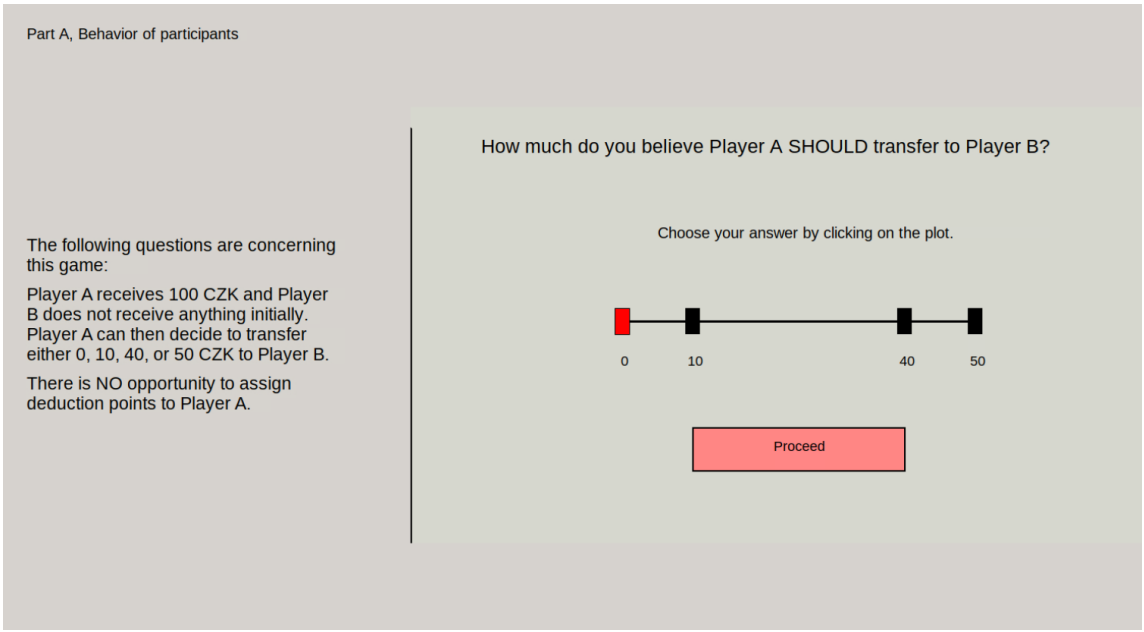
[Your role will be Player A [B] – experience]

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.

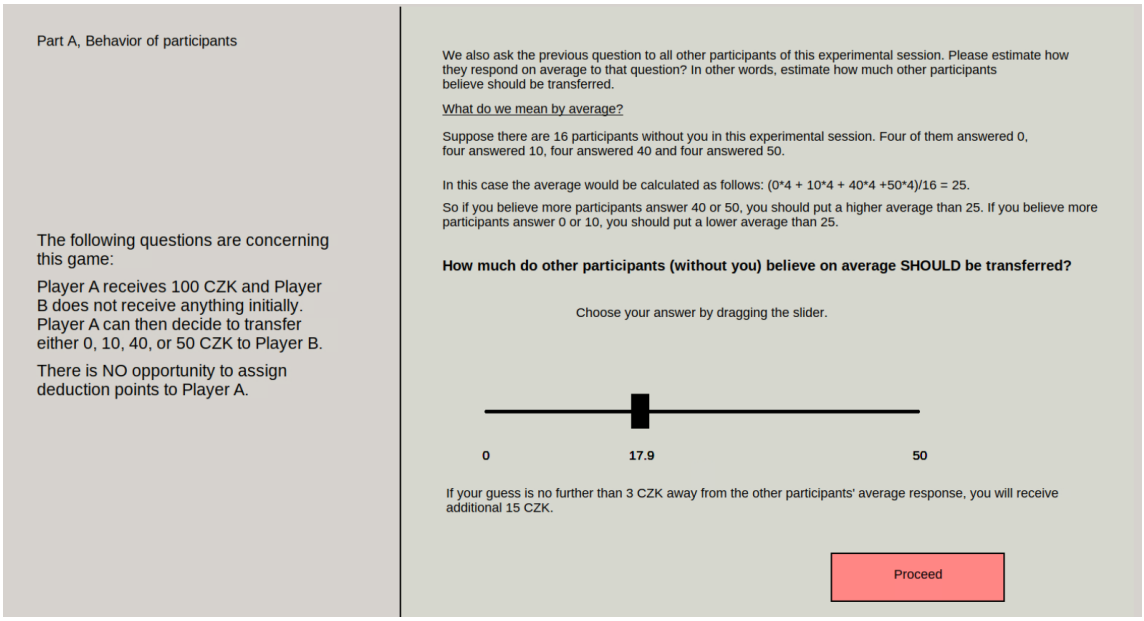
[screen 7 – 1st Emotions elicitation]



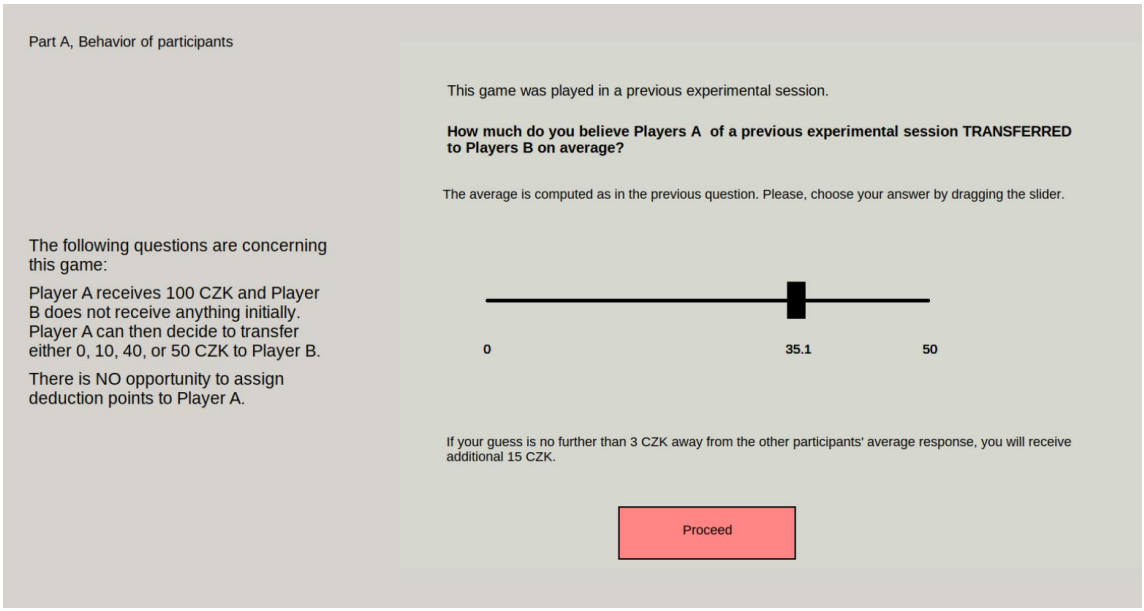
[screen 8 – 1st Norms elicitation – Personal norm]



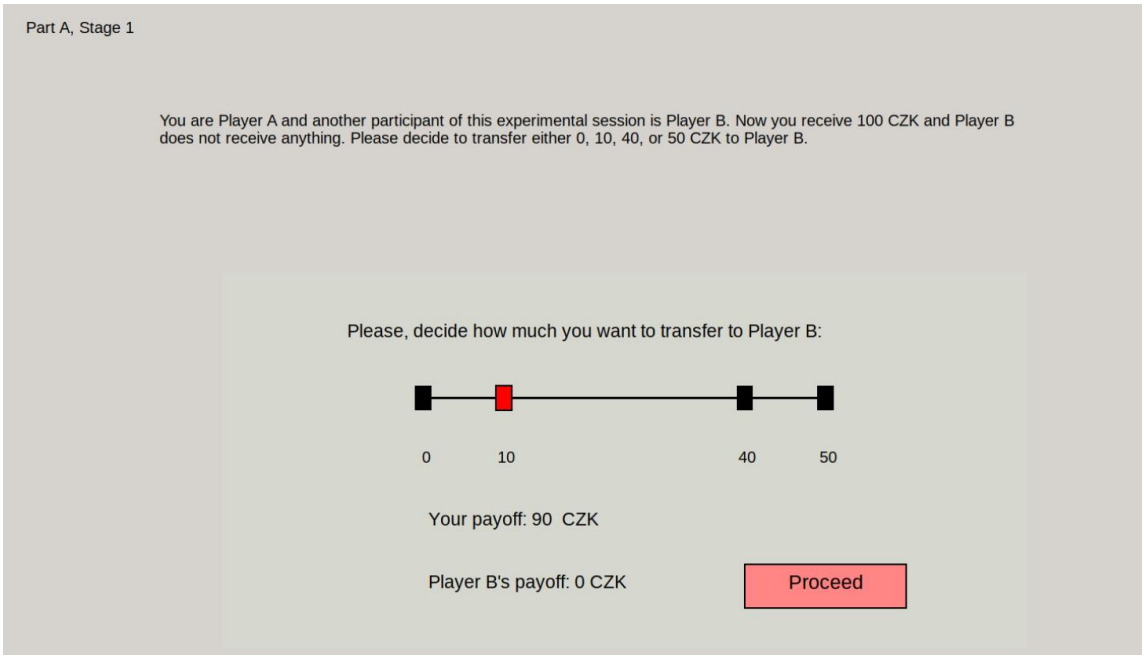
[screen 9 – 1st Norms elicitation – Normative expectations]

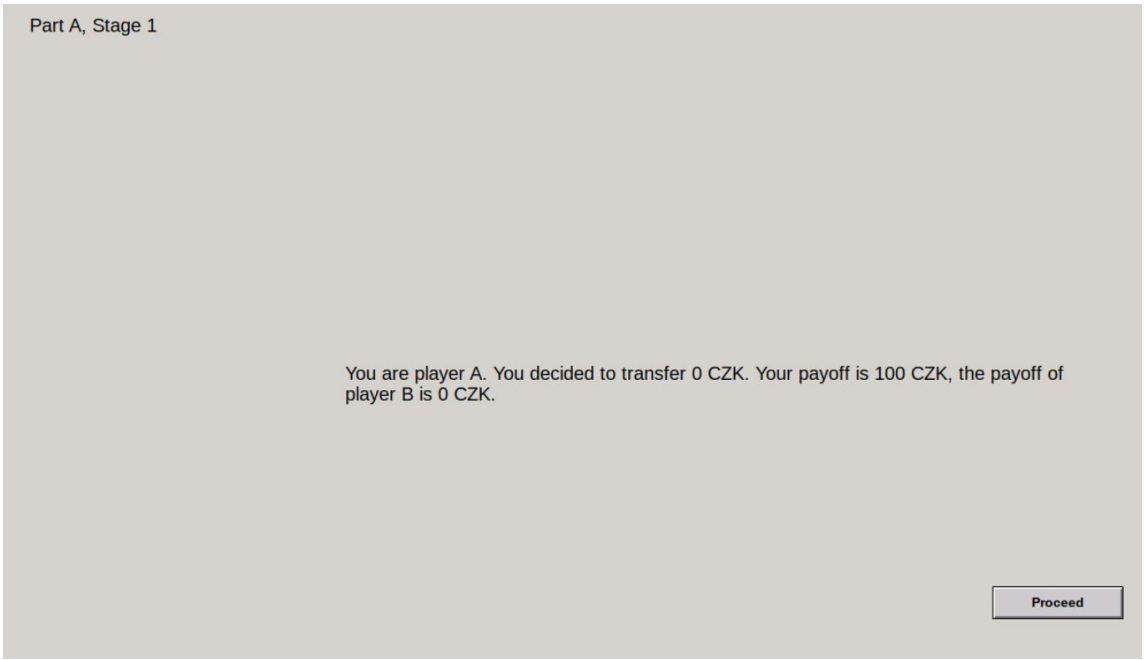


[screen 10 – 1st Norms elicitation - Empirical expectations]

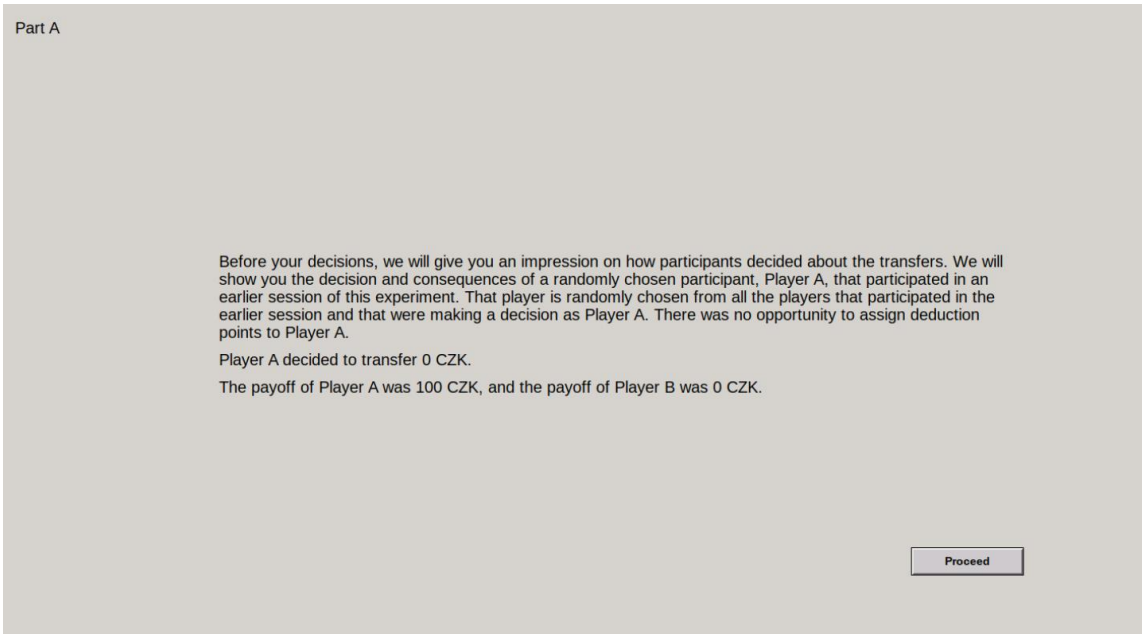


[screen 11 - Treatment Receive – DG game and results]

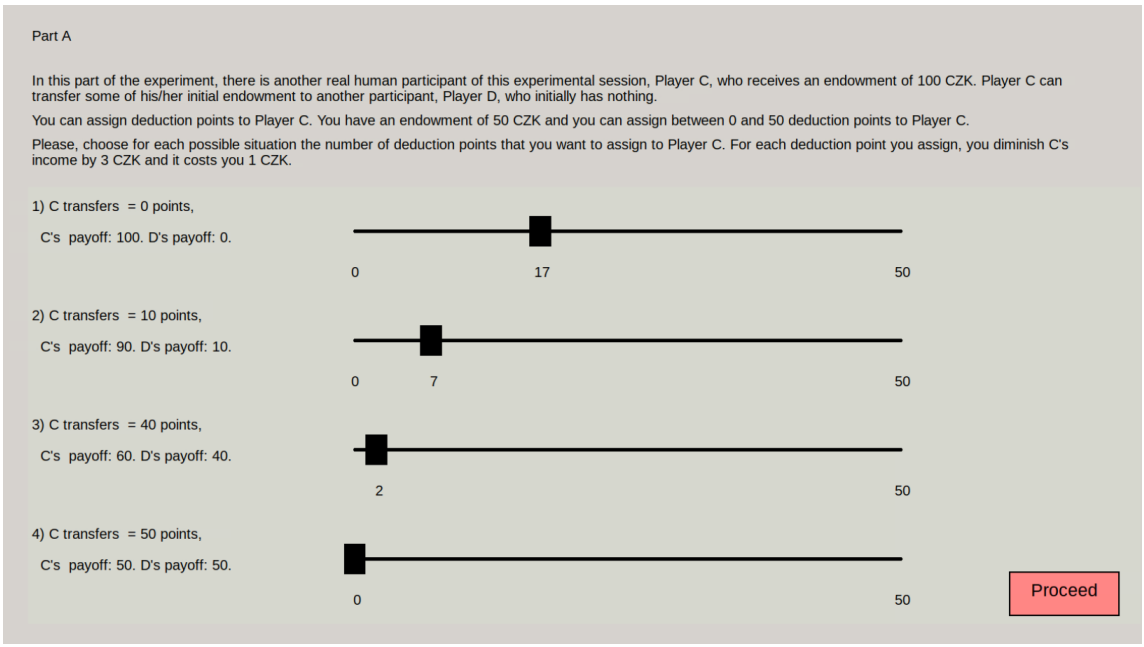




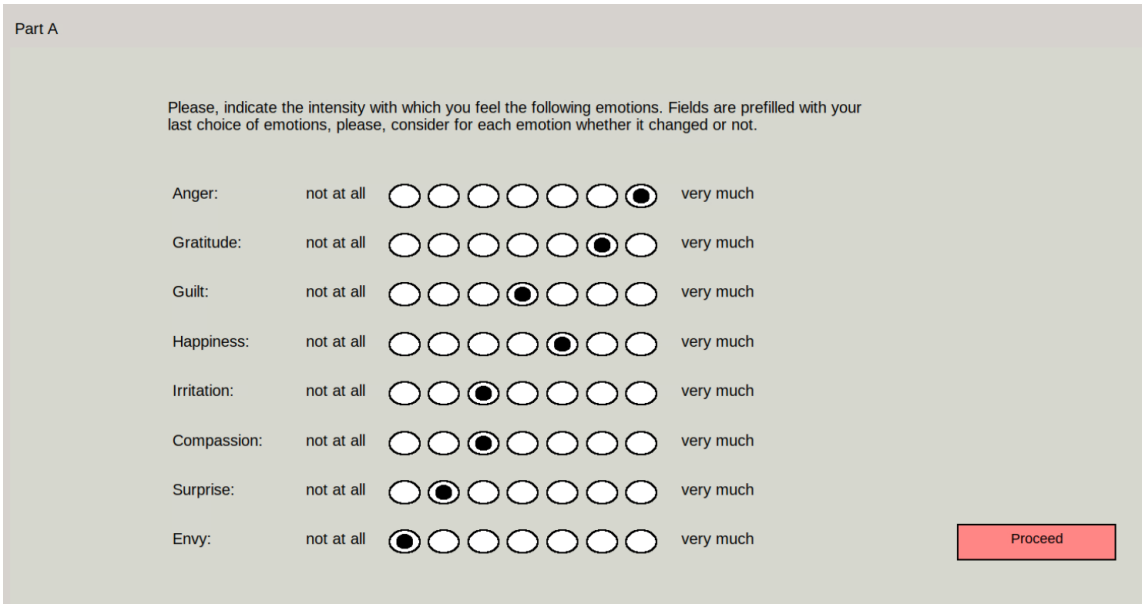
[screen 11 – Treatment Observe]



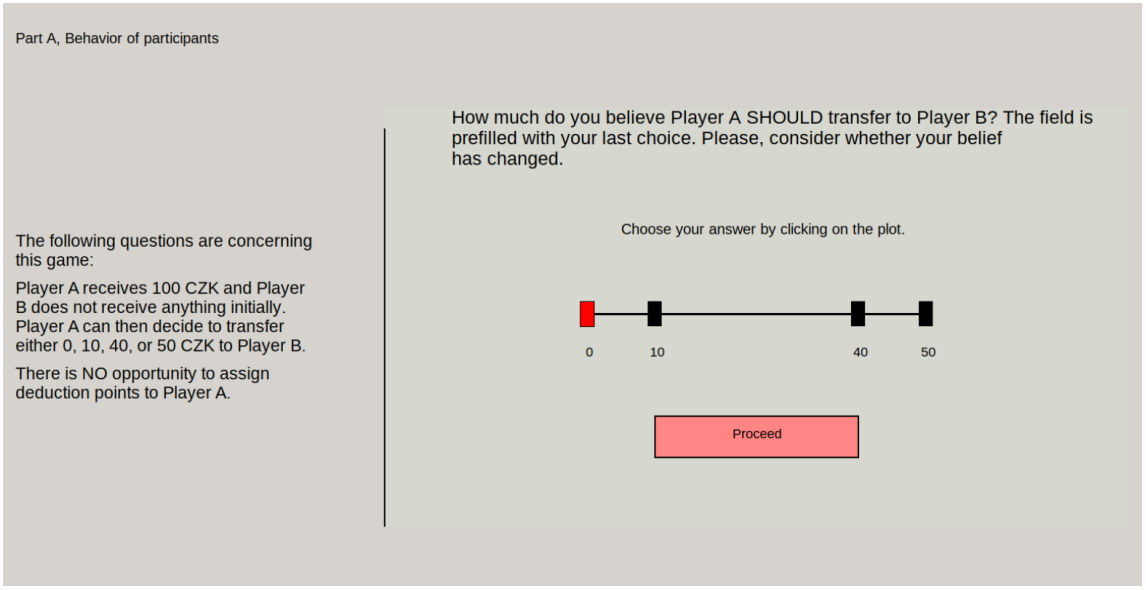
[screen 12 – Punishment phase]



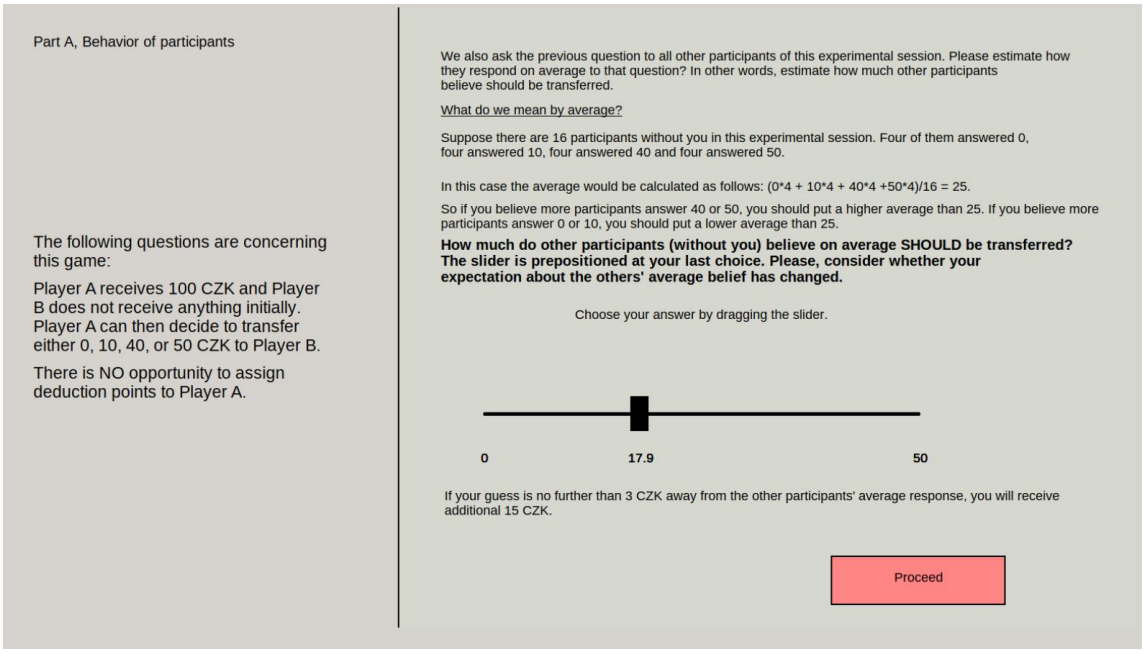
[screen 13 – 2nd Emotions elicitation]



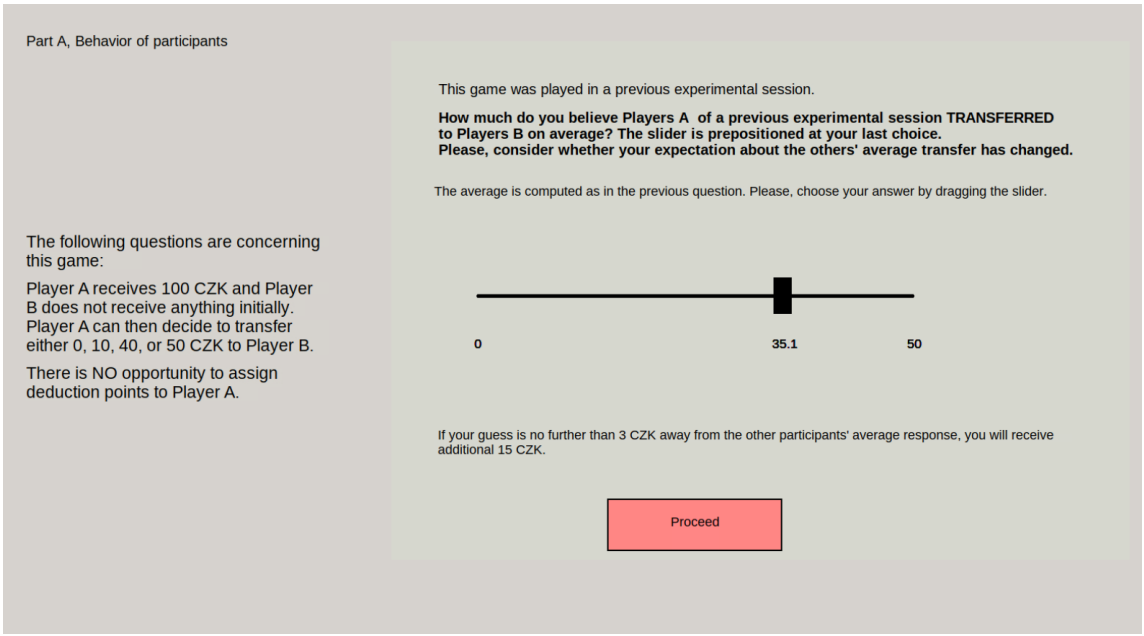
[screen 14- 2nd Norms elicitation – Personal norm]



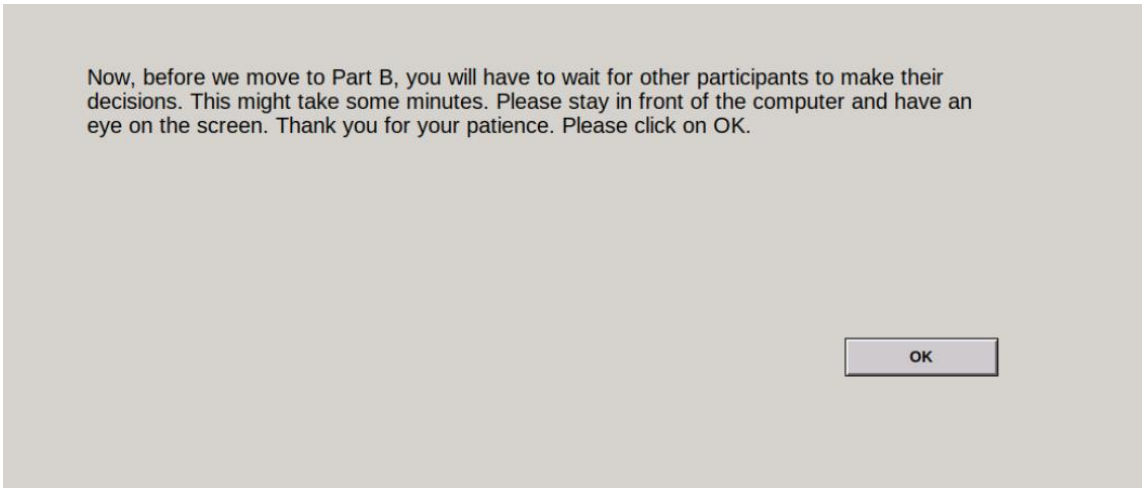
[screen 15- 2nd Norms elicitation – Normative expectations]



[screen 16 - 2nd Norms elicitation – Empirical expectations]



[screen 17 - waiting stage]



[screen 18 – Instructions part B]

Experimental instructions - Part B

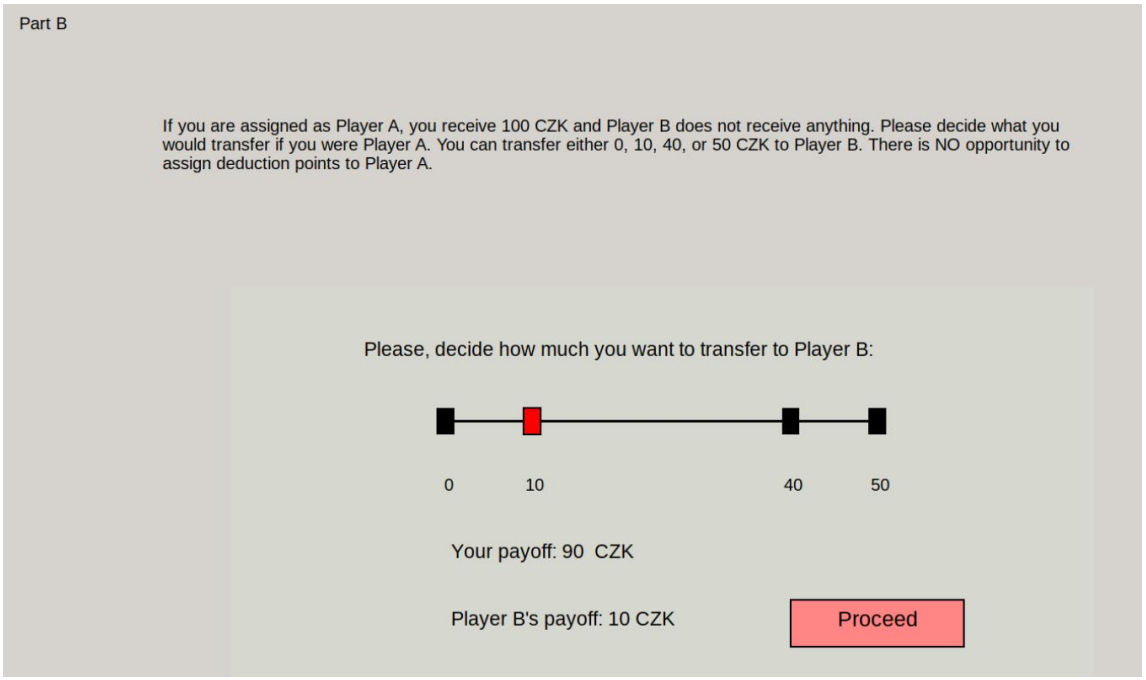
We have completed part A of this experiment. Now we will start with part B. In this part, you receive a base payment of 50 CZK that is independent of your future decisions and will be paid out for sure if part B is going to be picked for your payoff.

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.

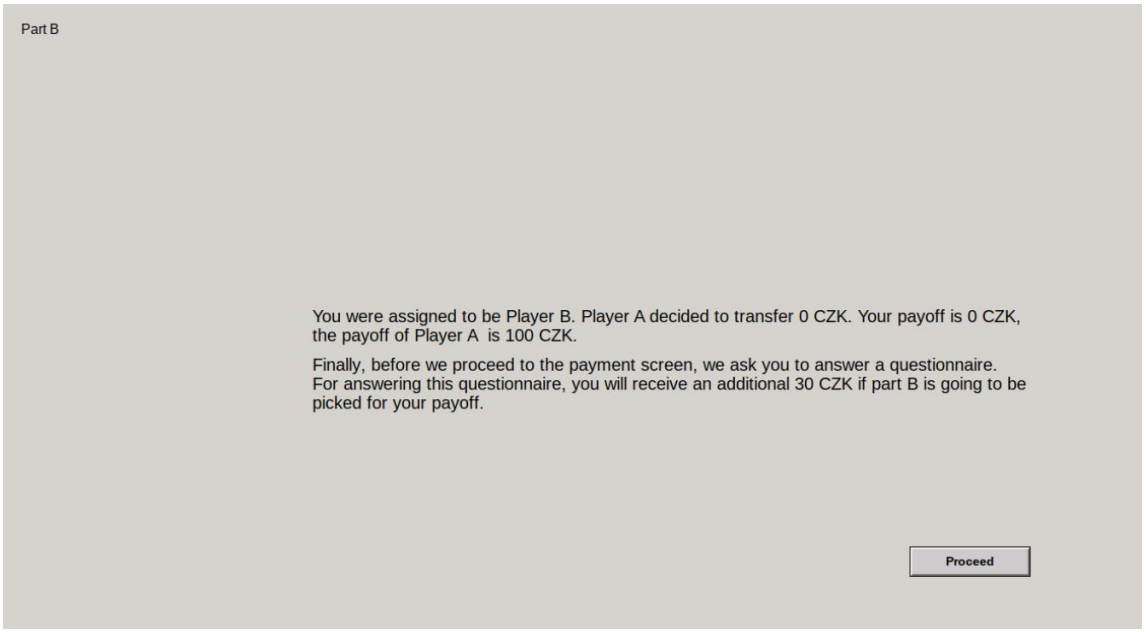
You will make a decision as Player A before you know if you are going to be assigned to be Player A or Player B.

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.

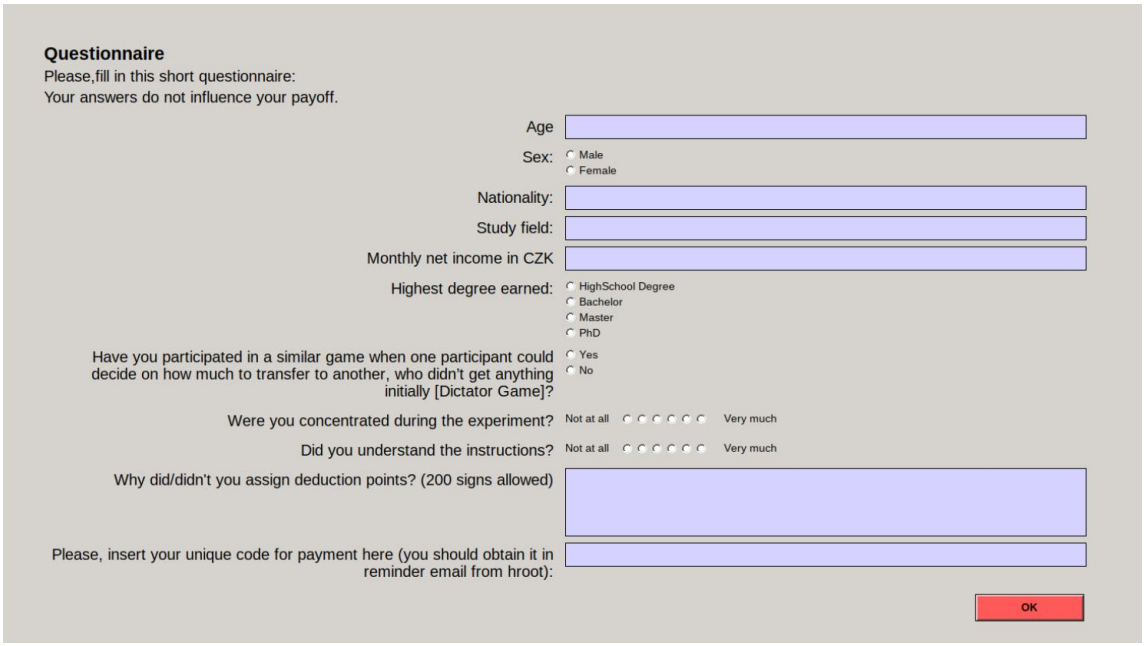
[screen 19 – part B]



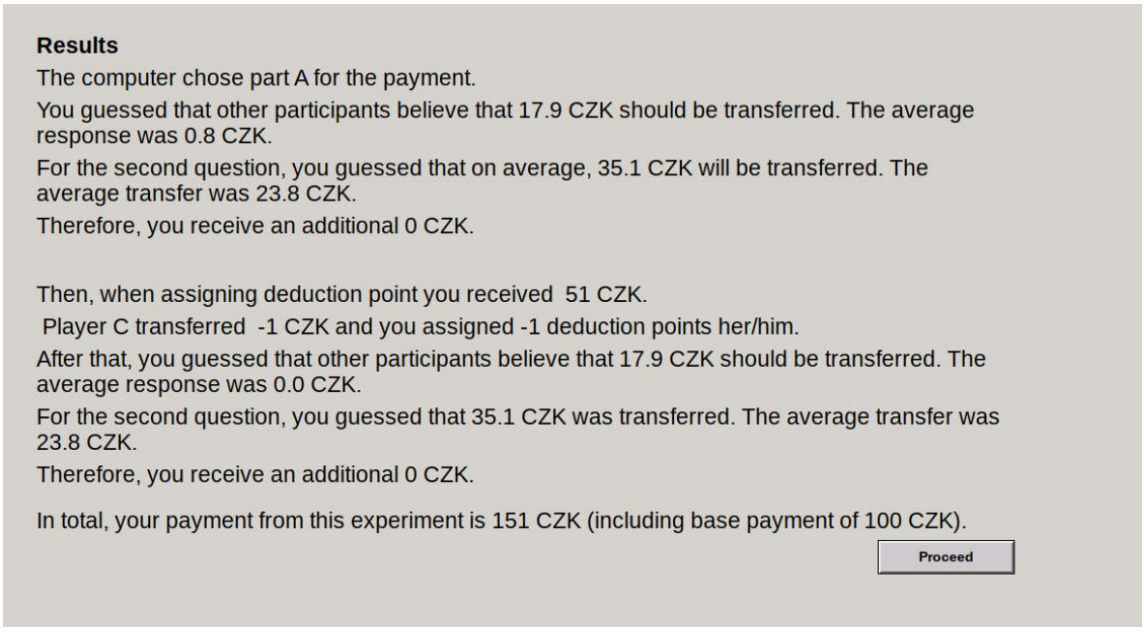
[screen 20 – part B results]



[screen 21 – questionnaire]



[screen 22 – payment screen]



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.

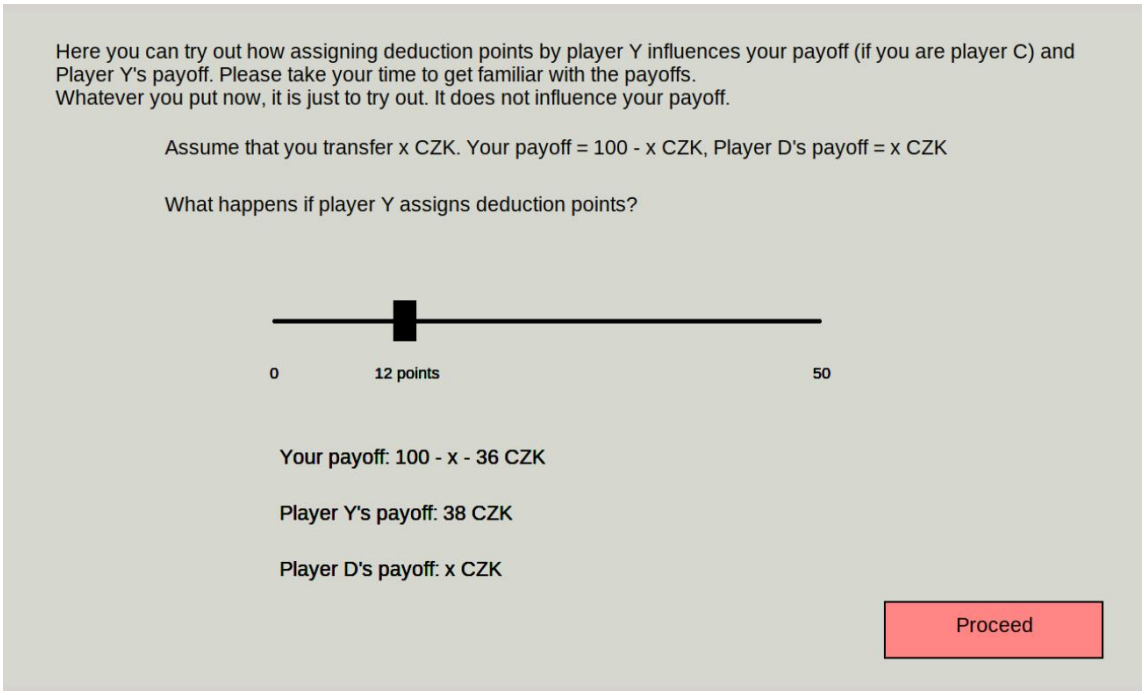
[screen 3 – tryout stage]

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

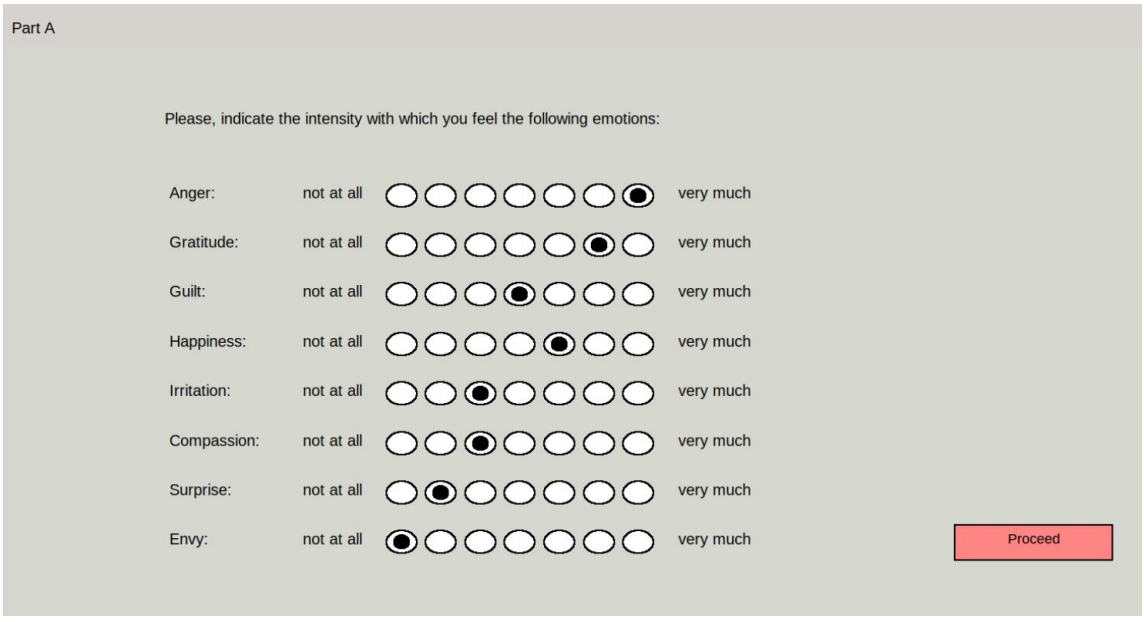
What happens if player Y assigns deduction points?



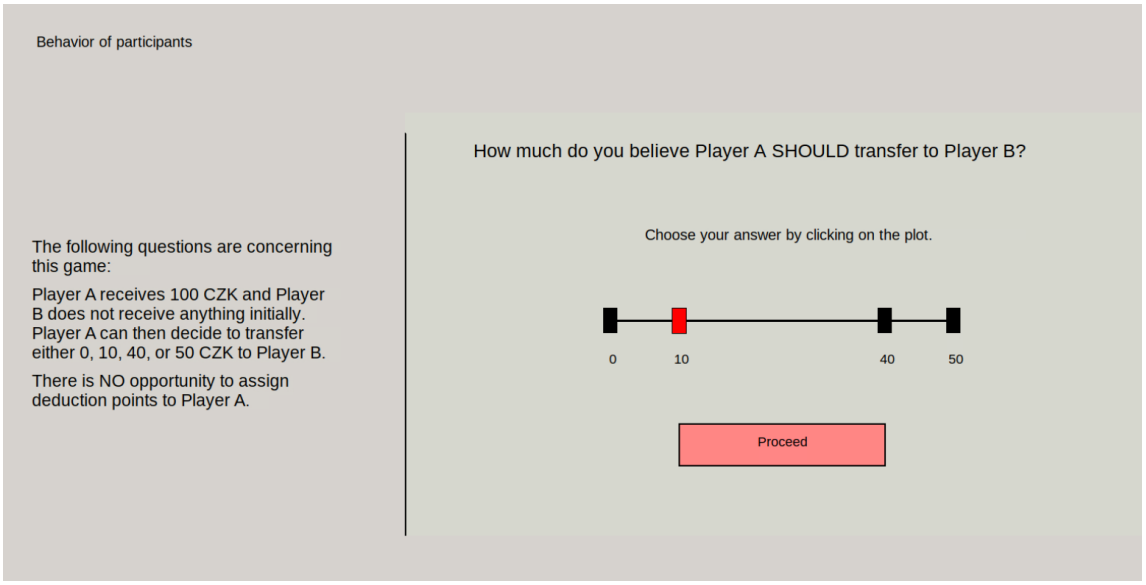
[screen 4]

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

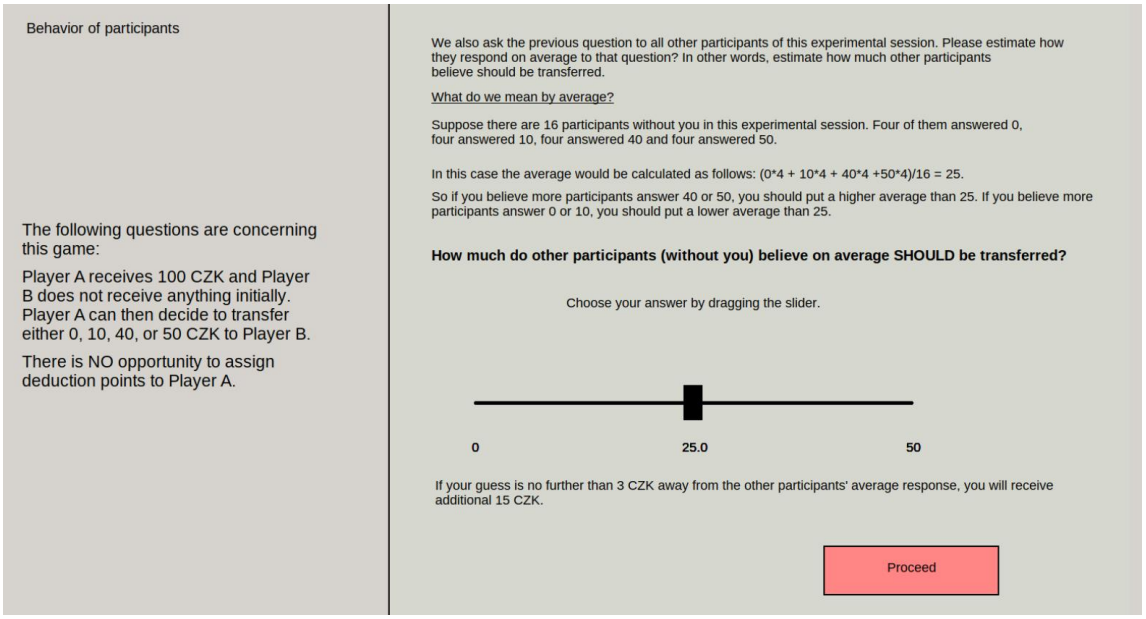
[screen 5 – Emotions elicitation]



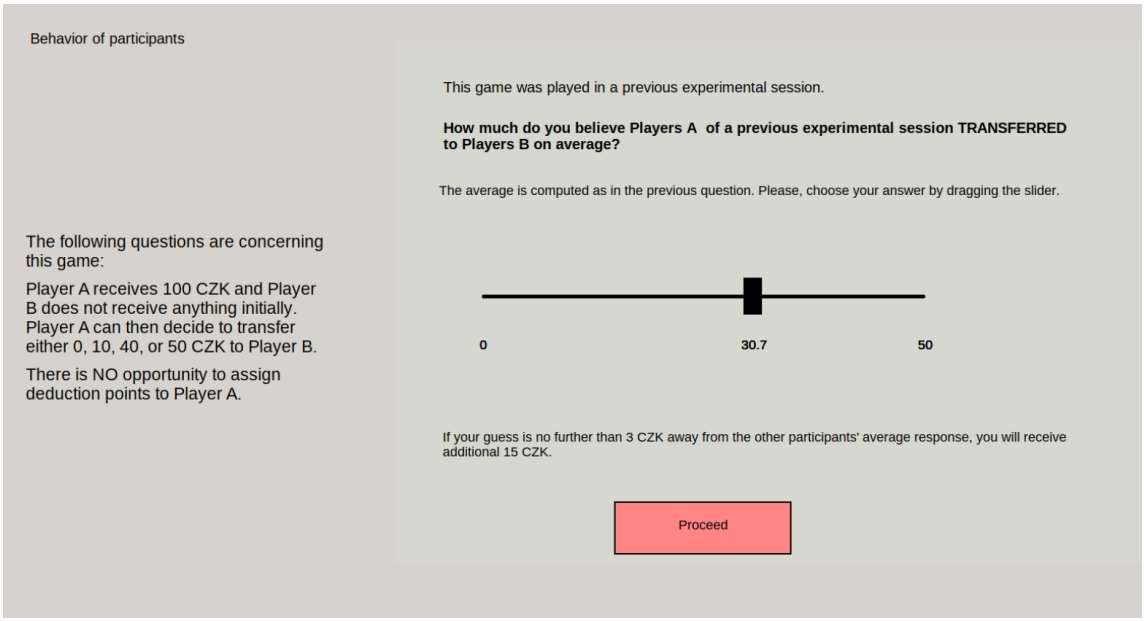
[screen 6 – Norms elicitation – Personal norm]



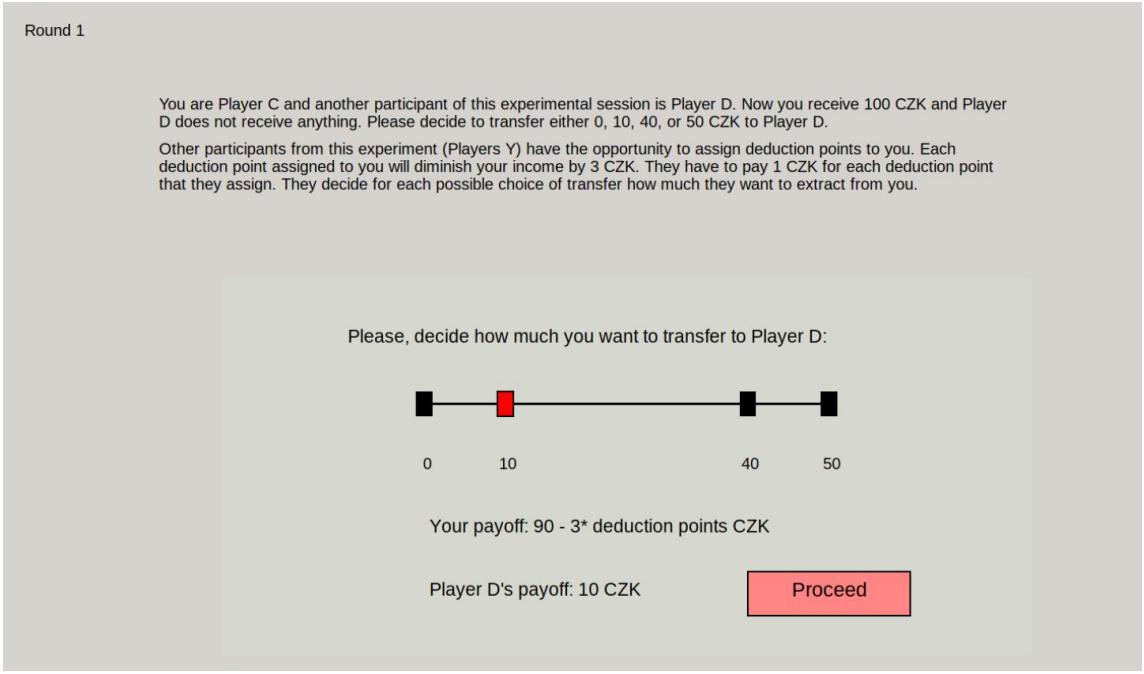
[screen 7 – Norms elicitation – Normative expectations]



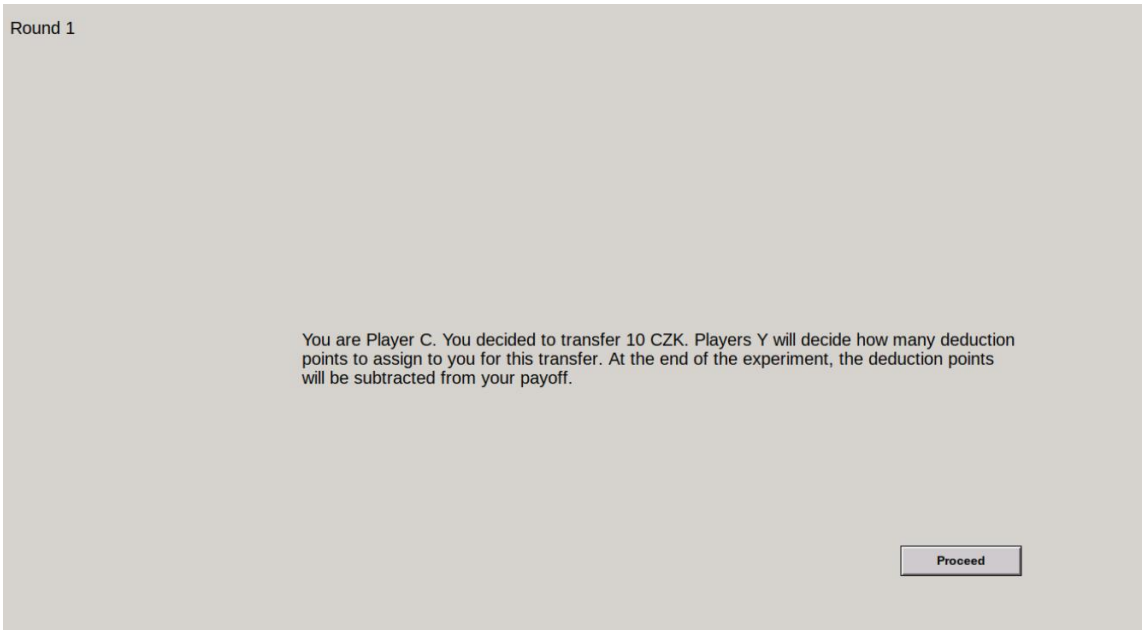
[screen 8 – Norms elicitation – Empirical expectations]



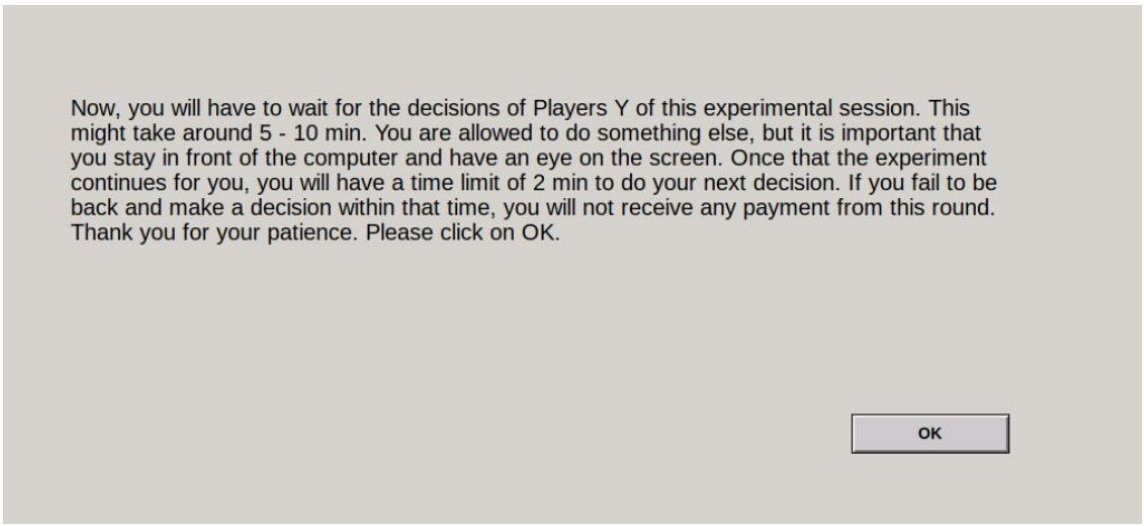
[screen 9 – DG round 1]



[screen 10 – DG results]

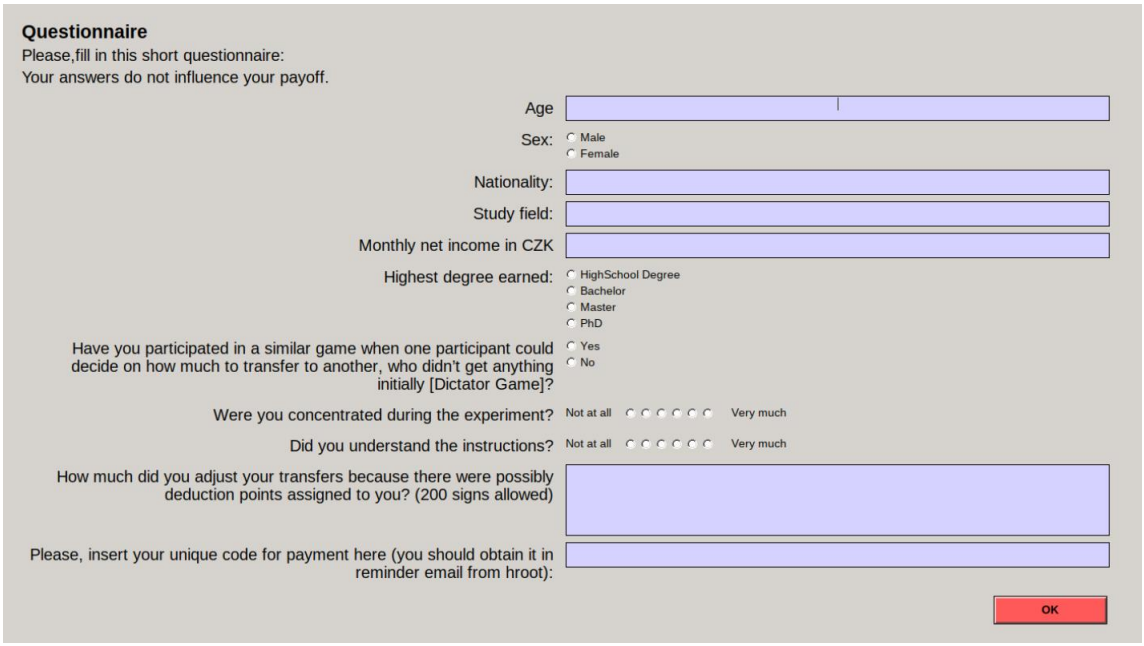


[waiting stage]



[screen 9 and 10 repeated 4 times]

[screen 11 – questionnaire]



[screen 12 – payment screen]