EXPERIMENTAL INVESTIGATIONS ON FAIRNESS AND SOCIAL NORMS IN ALLOCATION SETTINGS

DISSERTATION

ZUR ERLANGUNG DES AKADEMISCHEN GRADES
DOCTOR RERUM POLITICARUM

AN DER
FAKULTÄT FÜR WIRTSCHAFTS- UND SOZIALWISSENSCHAFTEN
DER RUPRECHT-KARLS-UNIVERSITÄT HEIDELBERG

VORGELEGT VON
ROBERT JOHANN SEBASTIAN SCHMIDT

HEIDELBERG, OKTOBER 2019

Acknowledgments

During my time at the Alfred-Weber-Institute, I have benefited from interacting with many people, and I would like to thank them.

First and foremost, I want to thank my supervisor Christiane Schwieren. She provided continuous support in all aspects of my work, and I very much enjoyed being part of her team during the last years. In our joint research projects, she offered extraordinary guidance and close cooperation. In addition to that, she provided me with an environment that allowed me to follow my interests, to develop own ideas, and to pursue my own research agenda. Christiane was always available for questions, discussions, proofreading, and for any other kind of problems that junior researchers face. Without her encouragement, and her countless advices, the single-authored projects in this dissertation contained in Chapters 3, 4, and 6, would not have been possible. For all that, I am very grateful to her.

I would also like to thank Stefan Trautmann, who co-authored Chapter 2 of this thesis. Working with Stefan taught me a lot about how to organize research projects and how to push them forward. I gained a lot of insights during our joint project, and these insights were of great help to me in my remaining projects.

I also want to thank Dietmar Fehr and Reimut Zohlnhöfer for agreeing to serve on my thesis committee without hesitation. In addition, I want to thank Dietmar Fehr for several very valuable comments for the projects contained in Chapter 3 and Chapter 6 when I presented these in the Internal Seminar of our institute.

I also want to thank my friends from the institute. I want to thank Martin, with whom I shared an office during my time at the AWI. I very much enjoyed sharing the office with Martin, conducting joint research with him, as well as our morning meetings in the Mildner´s back at the time when I lived in Heidelberg. Also, I want to thank Christian, Christoph, Christoph, Florian, Fuat, Hannes, Illia, Lisa, Marco as well as several further fellow colleagues and friends from the AWI for inspiring discussions, for continuous help when thinking about research ideas or testing experiments, and for the nice time that we had together at the AWI.

Finally, I want to thank my family for unconditional support while I worked on that dissertation.

Contents

Chapter 1

1. Introduction

1.1. Research Context

Perceptions of fairness and social norms are ubiquitous in daily life. They guide how individuals interpret social contexts, and they affect intentions and behavior (e.g., Bolton and Ockenfels, 2000; Elster, 1989; Fehr and Schmidt, 1999; Güth et al., 1982; Kahnemann et al., 1986; Ostrom, 2000; Rabin, 1993). In economic research, this has been acknowledged later than in other fields of social sciences, such as sociology or psychology. Instead, for a long time, economists focused on the effects of monetary incentives, thereby neglecting further potential determinants of decision making. In that sense, economic preferences were usually modeled as "self-regarding", and in the extreme version of that view, decision-makers were assumed to *exclusively* care about their own material outcome.

By now, it is no longer doubted that the spectrum of behavioral determinants shaping economic behavior goes far beyond the motive of material self-interest. That paradigm shift in economic research was accompanied by the establishment of *social* or *other-regarding preferences*, which capture the idea that individuals consider the social context they are embedded when making decisions. Today, there is vast empirical evidence for this class of preferences. For example, individuals have been observed to change the distribution of material outcomes at personal cost, e.g., by rewarding individuals who act cooperatively while punishing those who do not (Fehr and Fischbacher, 2004). This literature is accompanied by a range of further *non-material* aspects, including preferences for fairness, trust, and conformity, the desire to maintain a positive self-image, and the behavioral relevance of (various kinds of) social norms, to name just a few of these concepts.[1]

Acknowledging the fact that these (and many further) factors are necessary to understand, model, and predict economics decisions, led to the emergence of the discipline of *behavioral economics*. Using an interdisciplinary approach that connects economics with other social sciences, behavioral economics attempts to reconcile economic research with topics that have previously been neglected or delegated to other disciplines. My thesis is embedded in that

[1] Note that, economists *still* produce models where subjects exclusively consider their own material outcome. By contrast to previous times, however, this is not due to the fact that other factors are assumed to be irrelevant, but rather constitutes an informed decision resulting from a trade-off between simplicity and accuracy of a model.

research context, but I mainly concentrate on two of the previously mentioned topics: fairness and social norms. In the next section, I elaborate on the research agenda with which I try to contribute to the existing lines of literature concerning these topics.

1.2. Research Agenda and Connections between the Projects

The research agenda that guides this dissertation is characterized by the three aspects specified in the title. First, all of the projects are experimental investigations. That is, we[2] apply the method of incentivized economic experiments (Smith, 1976), and we try to answer research questions through the generation of behavioral data in the "laboratory". This method is by now extensively used in economic research, as it provides an excellent means to study individual behavior in a controlled environment. In fact, laboratory experiments have been a growing field for decades, and the share of lab studies published in general interest journals has risen continuously since the 1980s (Falk and Heckman, 2009). As a result of that, results gained in lab experiments are by now an important source to inform economic theory and public policy (Nikiforakis and Slonim, 2015).[3]

Second, all of the experiments involve the elicitation of fairness perceptions or perceptions of social norms. Importantly, when examining perceptions of fairness and social norms, we are mainly interested in individual statements of participants. In some of the projects, these perceptions are examined along with other traits in a within-subject manner, for example by additionally eliciting *social preferences*, by eliciting *different kinds* of fairness perceptions, or by examining social norm perception in *more than one context*. We then use the data that we gained on the individual level and try to draw inferences on how the different traits relate to one another. In contrast to that approach, fairness and social norms can also be studied on an aggregate level. An example of that approach is examining the interdependence between the economic system in a society and average perceptions about a fair redistributive policy (e.g., Alesina and Angeletos, 2005).[4] Such research would rather lie at the intersection between

---

[2] For reasons of consistency, from now on, until Chapter 7 (the conclusion), the dissertation is written using the first-person plural personal pronoun ("we"), independent from whether the chapters report on single-authored projects or on collaborative projects.

[3] Lab experiments have both advantages and disadvantages, and their magnitudes are regularly discussed in the literature. On the one hand, laboratory experiments provide excellent internal validity. On the other hand, it is argued that lab experiments (at least potentially) suffer from lack of external validity. However, one striking advantage of lab experiments is that they allow to examine a wide range of questions that are otherwise difficult to explore. In Chapter 5, in which we try to contribute to the methodological literature on lab experiments, we elaborate on that method.

[4] Another example is Ariely et al. (2019), who examine long-term effects of living in a specific economic system on the propensity to engage in dishonest behavior.

economics and sociology since the main interest is devoted to analyzing variables that aggregate many individuals. Instead, I see that research to lie at the intersection between *economics and psychology*, as we concentrate on the individual level, attempting to understand how various individual level traits are related to one another.

Third, the experimental paradigms used in the projects represent allocation settings, where some *active* individuals decide about how resources are allocated, while some *passive* subjects depend on these decisions. This is explicitly the case in the projects contained in chapters 2 and 4, in which some subjects decide about how to allocate money between themselves and other participants in the experiment. Also, it is implicitly the case in the projects contained chapters 3, 5, and 6. In these chapters, subjects do not take allocation decisions themselves, but the elicited perceptions about fairness or social norms refer to allocation settings.[5]

Finally, there is a fourth aspect that connects the majority of my projects in this thesis through a methodical aspect. Except for the experiment contained in Chapter 2, all experiments contain coordination games, through which we try to draw inferences about the subjects´ traits on the individual level. The design of these experiments is inspired by the proposition of Krupka and Weber (2013) to use coordination games to measure social norm perception. In that approach, subjects are confronted with descriptions of behavior, and their task is to coordinate on appropriateness ratings. Assuming that social norms reflect shared perceptions about appropriate behaviors (Crawford and Ostrom, 1995), the coordination outcome (i.e., the modal choice of all involved subjects) will reveal social norm perception within the participants´ population.

We adopt their approach in chapters 3, 4, 5, and 6, but we vary the set of available answers, such that it is suited to measure the kind of perceptions that we are interested in. Also, in Chapter 3 and in Chapter 6, we propose variations of their approach, attempting to extend the scope of their idea to use coordination games to measure perceptions. Specifically, in Chapter 3, we argue that their technique is not only suited to identify *social norm perception* on the *aggregate level* but to measure *any kind of subjective information* on the *individual level*. In Chapter 6, we propose an extension of their method that is intended to measure the same kind of perceptions, but in a more precise manner. In that regard, my dissertation also attempts to contribute

[5] More precisely, in Chapter 3, beliefs about behavior in the ultimatum game are elicited (the ultimatum game is a simple bargaining paradigm where behavior is significantly determined by fairness perceptions of the protagonists; Güth et al., 1982). In Chapter 5, social norm perception is elicited from inexperienced and experienced laboratory participants in a series of hypothetical allocation decisions. Finally, in Chapter 6, social norm perception about behaviors in daily life is measured, and these items also implicitly describe how resources (though not always monetary resources) are divided between individuals.

methodological insights to the proposition of Krupka and Weber (2013) to use coordination games to answer research questions.

To provide an overview about the research in that dissertation, I briefly summarize the motivation, research question, main results, and conclusions of each project in the next section.

1.3. The Projects

In Chapter 2, we start with research on fair procedures in allocation settings. Together with Stefan Trautmann, I investigate a problem that is inherent in many real-world situations: Several subjects have justified claims, but goods are indivisible, and therefore, outcome fairness is not feasible (e.g., *one* job but *several* applicants). However, from a prescriptive perspective, it is clearly desirable that the allocation procedure that determines outcomes adheres to process fairness. This concept reflects to what degree a process is deemed just, *independent* from the actual outcome. Our research starts at this point. We experimentally construct a setting where an allocator assigns indivisible resources between participants that have different degrees of similarity with the allocator in terms of political preferences. We then use the observed degree of favoritism (i.e., the extent to which the allocator favors the more similar recipient) as a measure to identify the degree of process fairness.[6] In a subsequent step, we examine the extent of process fairness in different environments that are intended to capture real-world allocation settings. In particular, we are interested in understanding whether lack of process fairness results from subjects being *not willing* or *not able* to implement a fair process.

We draw three main results from our experiment. First, when allocators have full information about recipients, and when they have to determine the allocation themselves, we observe substantial favoritism. Second, when subjects are provided with explicit instruments that help them to avoid favoritism, they extensively use these tools, thereby significantly reducing the extent of favoritism. Third (this follows from the previous result), subjects are not *unwilling* to avoid favoritism and to establish a fair process. Instead, they are mostly *not capable* of implementing a fair process, when this would require to apply (internal) randomization themselves. Our main conclusion is that decision-makers in charge of allocation should be equipped with explicit tools that help implement process fairness, as they are making significant use of them. Specifically, randomization tools seem to be an effective tool in such settings.

[6] Precisely, we use favoritism to indicate *absence* of process fairness. Note that the lack of fairness that we model is different from intentional discrimination, as it is less extreme. While discrimination suggests that a decision-maker actively seeks to discriminate, the kind of unfairness inherent in our setting is more subtle.

In Chapter 3, we make a methodological contribution. We put this project in this chapter because it motivates and helps understand the methodology applied in chapters 4 to 6 of this thesis. In that project, we propose two tractable methods to elicit subjective information, i.e., private data held by a subject. Fairness perceptions or perceptions of social norms are good examples for this kind of data, but it comprises various kind of other data held by individuals, such as beliefs, opinions, preferences, feelings, desires, and the like. Such data is of substantial importance in social science research, but accurately extracting private information is difficult, when the data cannot be verified objectively. This is because conventional experimental methods require to incentivize the subjects´ answers using some (verifiable) ground truth. If such ground truth does either not exist (as the data of interest is of purely subjective nature) or is not verifiable (e.g., for technical reasons or because the question at hand is hypothetical), conventional methods are limited in that regard. In that case, Bayesian revelation mechanisms (Prelec, 2004) are used, which provide a possibility to extract subjective data in the absence of objectively verifiable ground truths.

The motivation for the approaches proposed in Chapter 3 is that Bayesian revelation mechanisms have two problems. First, they require a set of (rather strong) assumptions, and it is questionable whether these hold in practice. Second, the scoring rules used to determine the recipients´ payments in these mechanisms are difficult to understand for participants. As a result of that, the mechanisms are difficult to implement, and experimenters therefore mostly rely on faith-based implementation. That is, they do not exactly explain the payment rules, but simply assure that truthful answering is optimal. This is a problem because the experimenter lacks control about whether subjects actually trust the experimenter (by indicating their true type) or whether they engage in gaming (by misstating their true type) in an attempt to maximize their earnings.

Both methods that we propose capitalize on the false consensus effect (Ross et al., 1977), i.e., individuals use their own traits when predicting the traits of others. In the mechanism *Benchmark*, we propose to use the elicitation of second-order beliefs to predict a subject´s own thoughts. In the mechanism *Coordination*, we propose to have subjects play coordination games, and to use an individual´s coordination choices to predict her actual thoughts. To test the mechanisms, we experimentally investigate the ability of the two mechanisms to detect first-order beliefs about behavior rates in the ultimatum game. On the aggregate level, we find that both mechanisms accurately reveal mean first-order beliefs of the population. On the subject level, we find that the modal difference between probabilities elicited in either mechanism and actual first-order beliefs is zero. We therefore conclude that subjects strongly anchor their statements in *Benchmark* and *Coordination* on their own thoughts.

In Chapter 4, we bring together two prominent topics of behavioral economics: social preferences and social norms. A huge body of literature indicates that social norms play an important role in social and economic decisions. In that context, one important distinction regarding the type of norms refers to *injunctive* social norms and *descriptive* social norms. Injunctive norms indicate perceptions about normatively appropriate behavior, while descriptive norms refer to prevalent or common behavior (Cialdini et al. 1990, 1991). Experimental studies find that both types of norms explain behavior, but also that the two are conceptually different constructs that independently affect intentions and behavior (e.g., Cialdini and Kallgren, 1993; Kallgren et al., 2000; Rivis and Sheeran, 2003). Specifically, the theoretical and empirical evidence on the competing relevance of injunctive and descriptive social norms is inconclusive.

We contribute to this literature by examining whether injunctive or descriptive social norms elicited using coordination games (Krupka and Weber, 2013) are more strongly related to social preferences measured in a series of dictator games. The data from our experiment provides us with three main results. First, we corroborate that the Krupka and Weber (2013) approach is a valid tool to elicit social norm perception on the individual level, as the individuals´ coordination choices (regarding both types of norms) are strongly related to their actual behavior. Second, we find that perceptions about descriptive social norms are significantly more strongly related to social preferences on the individual level than injunctive norms. Third, a comparison of the predictive power on the aggregate level indicates that average descriptive social norms are good predictors for actual behavior on the population level. This suggests that the elicitation of descriptive social norms using coordination games is a powerful approach to predict behavior in settings that are otherwise difficult to explore. We conclude that our data is consistent with the hypothesis that descriptive social norms are behaviorally more relevant than injunctive norms, which is in accordance with the view that changing perceptions about prevalent behavior is a more fruitful behavioral intervention than changing perceptions about appropriate behavior (e.g., Bicchieri and Xiao, 2009).

In Chapter 5, Christiane Schwieren, Alec Sproten, and I use the Krupka and Weber (2013) approach and examine whether inexperienced and experienced subjects in economic laboratory experiments differ. For that sake, we conduct a laboratory experiment and compare the two groups in a series of items that measure social norm perception. Precisely, social norm perception is measured (i) regarding allocation behavior in the lab, (ii) regarding a series of unsocial behaviors in the lab and the field, and (iii) regarding the evaluation of generalizability of behavior from the lab to the field.

We identify substantial differences between the two groups, both regarding injunctive and descriptive social norms in the context of participation in lab experiments. By contrast, social norm perception for the context of daily life does not differ between the two groups. We therefore conclude that learning seems to be more important than selection effects for understanding differences between the two groups. Conducting exploratory analyses, we make the interesting observation that behaving unsocial in an experiment is considered substantially more appropriate than in daily life. This appears inconsistent with the hypothesis that social preferences measured in lab experiments are inflated and indicates a distinction between revealed social preferences as measured commonly and the elicitation of normatively appropriate behavior.

Finally, Chapter 6 is again a methodological contribution. In that chapter, we propose the Point Beauty Contest, a tool that is suited to measure the distribution of focal points in coordination games on the individual level. The contribution is motivated by the fact that conventional coordination, where subjects can only bet on one alternative, only reveals which alternative is considered *most* focal. In contrast to conventional coordination, where subjects coordinate by choosing one alternative, in the Point Beauty Contest subjects are equipped with a budget of points that they can distribute among multiple alternatives. This allows for nuanced coordination strategies, as participants can invest in multiple alternatives at the same time and weigh their choice. We argue that this extension of modeling coordination settings is consistent with many real-world coordination settings.[7]

We experimentally compare the proposed mechanism with conventional coordination and report two main results. First, the data confirms the theoretical predictions regarding coordination behavior and demonstrates that the proposed technique is suited to identify the distribution of focal points on the individual level. Second, we find that the proposed mechanism identifies focal points on the group level more efficiently than conventional coordination. We conclude that the Point Beauty Contest is suited to identify the distribution of

[7] For example, think about a bank run, where depositors might not only think about withdrawing none of their money or all of their money from a bank but want to engage in both strategies at the same time in a weighed manner.

focal points on the individual level and that it represents a framework for modeling coordination settings where fine-grained coordination strategies are feasible.

As shown at the beginning of that section, the projects are connected through their main topics (fairness and social norms), their methodology (experimental methods, coordination games), and the fact that allocation settings are involved in all experiments. In Chapter 7, I draw conclusions from the joint data by (i) connecting the results from the different projects, (ii) by discussing the limitations of the provided research, and (iii) by providing ideas about potentially interesting avenues for future research.

Chapter 2

Implementing (Un)fair Procedures? Favoritism and Process Fairness when Inequality is Inevitable

Robert J. Schmidt & Stefan T. Trautmann

Abstract: We study allocation behavior when outcome inequality is inevitable but a fair process is feasible, as in selecting one person from several candidates for a job or award. We show that allocators may be influenced by inappropriate criteria, impeding the implementation of a fair process. We study four interventions to induce process fairness without restricting the allocator's decisions: Increasing the transparency of the allocation process; providing a private randomization device; allowing the allocator to delegate to a public randomization device; and allowing the allocator to avoid information on inappropriate criteria. All interventions except transparency have positive effects, but differ substantially in their impact.

Highlights:

- We study favoritism and process fairness in a setting where inequality is unavoidable
- Allocators exhibit a substantial degree of favoritism in one-shot decisions
- Providing allocators with explicit tools to implement fair procedures is effective
- We conclude that allocators are not unwilling but unable to implement fair procedures themselves

Acknowledgments: We thank Emin Karagözoğlu as well as seminar audiences in Frankfurt, Heidelberg, Bremen, Vienna, Mannheim, Tilburg, Bocconi, Göttingen, Ulm, Shanghai, Nanjing, Dijon, Lyon, the Max-Planck Institute in Bonn, the ESA World Meeting in San Diego, the ESA European Meeting in Vienna, the HeiKaMax in Karlsruhe, the HeiKaMaxY in Mannheim, the Scandinavian Meeting in Gothenburg, and the Sabe/Iarep Meeting in London for very helpful suggestions.

2.1. Introduction

Consider a committee's choice between two candidates for a job. Both score equally good on various hard measures, thus there is little to guide the decision between the candidates. Other aspects can become influential at this stage, affecting the outcome of the decision. Such aspects may include the candidates' gender, race, or political views. Importantly, the influence of such aspects may be undesirable to the committee itself, or to the organization for which the committee serves to select the candidate. They may still exert influence on the decision. Although the organization and the committee may thus aim to implement a fair selection process, biased outcomes can obtain and render the process unfair ex-post.

We show that when there are no clear economic criteria like performance or effort that could guide an allocation decision, softer criteria (in our case, overlap in political attitudes between allocator and recipients) have an overwhelming influence on how funds are allocated among two people. Despite its prevalence, we argue that the allocation on the basis of the political attitudes is not perceived as an appropriate criterion even by those who apply it. Moreover, irrespective of the decision-makers' perception of the criterion, we take a prescriptive perspective and aim to reduce the influence of the political attitude information on the allocation outcome. To this end, we introduce a set of interventions in our setup and observe whether the adjusted decision environment is able to reduce the prevalence of allocation according to political views. In terms of the above example, even if some committee considers aspects like gender, race, or politics acceptable criteria absent clear differences between the candidates, the organization behind it may not do so, and may wish to reduce the incidence of its use.

Our experimental setup puts decision-makers in the position to allocate two outcomes, €8 and €2, among two recipients. Similar to the case of selecting candidates for a job or to allocating an indivisible good or service to people in need (e.g., a donor organ among to people in need of a donation), there will be ex-post or outcome inequality. Absent clear criteria why one recipient should receive the more favorable payment, outcome inequality is typically assumed to be aversive (e.g., Bolton and Ockenfels, 2000; Fehr and Schmidt, 1999). However, if outcome fairness is not attainable, a fair process may substitute for fair outcomes (Andreozzi et al., 2013; Bolton et al., 2005; Krawczyk, 2011; Trautmann, 2009; Trautmann and van de Kuilen, 2016; Trautmann and Wakker, 2010; Saito, 2013). In particular, randomly allocating the high and the low outcome can be interpreted as a fair process in our setting (Broome, 1990; Konow, 2003). Empirically, random allocation has been shown to be an important mechanism employed by decision-makers (Bastek et al., 2018; Brock et al., 2013; Miao and Zhong, 2018; also Dwenger et al., 2016, in contexts other than fairness). In practical applications, an equal representation of candidates or recipients in terms of sex, race, or political orientation in the absence of clear differences in any hard criteria can be interpreted as evidence of a fair process. Fairness perceptions become especially important from the perspective of the decision-maker if recipients can react by rewarding or punishing fair or unfair processes (suing, filing complaint, recommending). Then, strategic and non-strategic reasons for implementing fair procedures become important to the decision-maker. We implement the possibility for reward and punishment in our setup.

In our allocation decision, there are no easily justifiable criteria (such as effort etc.) for which recipient should receive the higher payment. At the same time, the decision-maker has available a profile of the two recipients' political attitudes (while recipients are uninformed about others' political profiles). We find that in our baseline condition, about 90% of the high payoffs go to the recipient whose political profile is closer to that of the allocator. We then introduce four interventions with the aim to reduce the incidence of such favoritism. Each intervention condition deviates in one aspect from our baseline allocation condition: In *Transparency* we make all three political profiles in a group visible to each group member. Thus, favoritism becomes observable. In *Private Lottery*, the decision-maker allocates probabilities of receiving the high payoff to the two recipients, instead of outcomes directly. Recipients are aware of the procedure and learn about the resulting allocation, but do not know the exact probabilities chosen by the decision-maker. In *Public Delegation*, the decision-maker can publicly delegate the decision to a 50-50 random device; in three subcases, delegation is either for free or costly. Finally, in *Information Avoidance*, the decision-maker can decide not to see the information about the recipients' political attitudes. In two subcases, whether the decision-maker avoided information or not before making her allocation decision is either observed or unobserved by the recipients.

An important feature of all four interventions is that a decision-maker who wants to allocate on the basis of the recipients' political attitudes, has the freedom to do so. That is, the interventions are not trivial in the sense of simply restricting the use of the information or stipulating a decision, which would be unrealistic and undesirable in practice: decision-makers will typically hold specialized information relevant to the selection process, which the organization wants them to be able to use in determining the best allocation. Instead, the interventions provide an institutional environment aimed at reducing favoritism.

In terms of the incidence of favoritism, we find that transparency has no effect. A private lottery works well, with the fair lottery being the modal choice to allocate the payments. Similarly, public delegation is taken up by a majority of decision-makers if for free; costly delegation is still taken up by a substantial share of decision-makers. Information avoidance is taken up by 12.5% of the decision-makers if privately known only, and by 52.3% when it is publicly known. Conditional on acquiring information, decision-makers are equally prone to allocate to the politically closer recipient as in Baseline.

We present the Baseline experimental design, procedures, and empirical measures in Section 2.2. Section 2.3 gives the results of the Baseline treatment. In Section 2.4 we define the interventions and report their effects on the decision-makers' allocation behavior. Data on decision-makers' beliefs are discussed where they help interpreting observed behavior. Data on fairness perceptions, the role of politics, and the behavior of the recipients aid our understanding of how favoritism affects the allocation decisions and how the interventions work. We present these results in the appendix, and refer to them in the main text where needed. Section 2.5 discusses our results in the context of the related literature, and concludes with prescriptions for practical settings.

2.2. Baseline Allocation Game and Experimental Design

2.2.1. Experimental Paradigm

The experimental paradigm is a simple 3-person (=1 group) allocation game in which the decision-maker (DM) allocates the two payments €8 and €2 to two recipients labelled A and B. Subsequently, after learning about the allocation, each recipient can costly reward or punish the DM by paying €0.10 for each €0.50 added or subtracted from the DM's endowment of €5 (which equals the average payoff of the recipients). Each recipient can reward or punish up to a maximum of €2 (at maximum cost of €0.40). Thus, the DM may end up with earnings between €1 where both recipients maximally punish, and €9 where they both maximally reward. Recipients may choose to neither punish or reward, and they make their decisions independently and simultaneously. Reward or punishment costs are deducted from the recipients' earnings. While the DM makes her allocation decision, recipients provide their non-incentivized beliefs about the allocation (high versus low payment); while the recipients make their decisions, the DM provides non-incentivized beliefs about each recipients' reward or punishment in terms of money destroyed/added. After all decisions are made, but before the DM is informed about recipients´ choices, all three group members judge the fairness of the allocation process and the allocation outcome regarding the allocation of the €8 and €2 payments to recipients A and B (exact wording in Appendix A2.1). The DM is not informed about the recipients' actions until after the questionnaire to prevent her fairness judgments reflect the recipients´ views as revealed by their rewarding decisions.

At the very beginning of the experiment, before receiving any instructions for the allocation game and before knowing their role in the game, subjects are asked five questions regarding their political attitudes. We informed subjects that their answers might be shown to other participants during the experiment (while maintaining full anonymity). All questions had binary answer possibilities and we instructed subjects to choose the answer options that best describes their attitudes. The five questions concern (i) their general political orientation ("rather left wing" or "rather conservative"), and whether they support the (ii) unlimited inflow of refugees, (iii) female quotas in organizations, (iv) active euthanasia, and the (v) exit from nuclear energy. The original wording is provided in Appendix A2.1. In the Baseline treatment, before making her allocation decision, the DM is shown the political profile of recipients A and B. In contrast, the recipients do not learn about the profile of any other player. All players in the group are aware of the information available to other players at each stage of the game.

In appendix A2.3, we show that there is substantial variation of political attitudes in our total sample and also at the level of the 3-payer groups: on average, the two recipients within a group disagree on 2.03 of the 5 items of the political attitude questionnaire. We are interested in how similarity in attitudes affects the DM's allocation in the absence of any other relevant information. We therefore define the following *similarity* or closeness measure. We compare the answers of recipient j = {A, B} with the answers of the DM and denote the number of items on which recipient j agrees with the allocator as S_j. Thus, higher S_j indicates higher similarity between the allocator and recipient j. We then define the *relative similarity* $S_{AB} = S_A - S_B$. Thus, if $S_{AB} > 0$ recipient A is more similar to the DM, and if $S_{AB} < 0$ recipient B is more similar to the allocator. For $S_{AB} = 0$ the similarity of recipients with the allocator is identical according to our definition. Note that $S_{AB} = 0$ implies only that recipients agree with the allocator on the same number of items. It does not necessarily imply that they have identical political profiles, and the exact items they agree on may differ between recipients. The average of the absolute value of the relative similarity measure S_{AB} across all groups in our experiment is 1.15 (see Appendix A2.3).

2.2.2. Treatments and Hypotheses

The allocation game as described above forms our *Baseline* condition.[8] We hypothesized that allocation decisions are affected by the political attitude information provided to the DM. In Section 2.3, we show that this was indeed the case. In Section 2.4, we present four interventions for which we hypothesized that they reduce the incidence of favoritism. Importantly, in each of the intervention treatments, the DM in principle has the possibility to allocate according to favoritism just as in the Baseline treatment. Our focus in the current study is prescriptive. That is, we assume that either (i) the DM would like to implement a fair allocation process but does not succeed to do so, or (ii) that some other entity would like the DM to apply a fair process. We aim to design an environment in which the DM then makes choices that imply a fair process. To understand the individual and strategic motivations of the different players, especially the DM, we add a condition *Random*, in which the decision-maker does not make a choice, but outcomes are allocated randomly with equal probability. The condition allows us to observe beliefs and rewards or punishments in the case a fair process is exogenously imposed, but outcomes are still unequal.

2.2.3. Laboratory Procedures

The experiment was programmed using z-Tree (Fischbacher, 2007) and recruitment was done via hroot (Bock et al., 2014) and ORSEE (Greiner, 2015). Across all treatments, 876 subjects took part in the experiment. Subjects were mainly undergraduate students from a wide range of different majors, with 35% having an economics background. A typical session lasted about 35 minutes, and subjects earned on average about €8.10 including a show-up fee of €3. We conducted 48 experimental sessions at Heidelberg University (24 sessions) and Frankfurt University (24 sessions) between November 2016 and November 2017. Table 2.1 shows the number and location distribution of 3-person groups over all treatment conditions, and the number of groups with $S_{AB} \neq 0$ in each condition.

[8] An overview of all experimental conditions and a full set of instructions is available in the webappendix at https://www.dropbox.com/s/yig7lte2pnimcov/ST_2019_procfair_Webappendix_march10.pdf?dl=0.

Table 2.1. Number of 3-person Groups by Treatment and Location

	Heidelberg	Frankfurt	# of groups total	# of groups with $S_{AB} \neq 0$
Baseline	22	18	40	28
Random	24	14	38	26
Transparency	17	24	41	32
Private Lottery	18	23	41	27
Public Delegation	22	18	40	26
Information Avoidance (private)	18	30	48	41
Information Avoidance (public)	18	26	44	32

2.3. Baseline Results: Strong Evidence for Favoritism

In the baseline condition there were 28 out of 40 groups in which the two recipients had an unequal political distance to the DM, that is, $S_{AB} \neq 0$. In 25 of these 28 cases (89.3%), the DM allocated the higher payment to the more similar participant, significantly more often than we would expect under random allocation (p < 0.001, binomial test against 0.5).[9] Using relative similarity indicators based on individual dimensions (that is, defining S_j and S_{AB} for each dimension separately), Table 2.2 shows that all dimensions except female quotas significantly added to this preferential treatment.[10]

Table 2.2. Relative Similarity and Allocation Behavior in Baseline

Dimension of similarity	Assignment of high payoff
Left-wing political attitude	0.387** (2.16)
Support of unlimited refugee inflow	0.269* (1.76)
Support of mandatory female quota	0.221 (1.49)
Support of active euthanasia	0.791** (2.27)
Support of exit from nuclear energy	0.715*** (2.81)
Observations	38

Notes: Marginal effects of probit regression reported, z-values in parenthesis. The number of observations is 38 as in two groups recipients are exactly identical. The explanatory variables are dimension specific relative distance measures. *, **, *** indicates significance at the 10%, 5%, and 1% level.

[9] All test statistics reported in this paper regard two-sided tests.
[10] Analyses of pooled data for all conditions of the experiment replicate this finding with all dimensions significantly adding to favoritism (Appendix A2.3).

Allocation according to similarity is consistent with social identity effects: differences in political views form a basis for interpretations of one recipient as ingroup and the other, less similar one, as outgroup (Chen and Li, 2009). However, to better understand the motivation of the DMs we consider their beliefs and fairness judgments. Table 2.3 shows the DMs' beliefs about the rewards and punishments. We see that DMs expect moderately positive rewards in case of the Random treatment that guarantees a fair procedure. In Baseline, they expect significantly higher reward from the high payoff recipient, and a significantly higher punishment from the low payoff recipient. On average they expect a total reward of €0.25 in Random and €0.00 in Baseline, a marginally significant difference ($p < 0.1$, Mann-Whitney-U test). From a strategic perspective based on their own beliefs, it may thus be desirable for DMs to signal a fair procedure.

Table 2.3. DMs´ Beliefs about Reward and Punishment in Baseline and Random

	High payoff recipient	Low payoff recipient
Baseline	€0.83$^{\#\#\#}$	- €0.83$^{\#\#\#}$
Random	€0.39$^{\#\#}$	0.11
Difference	- €0.44**	€0.94***

Notes: Entries are average beliefs about additions to / reductions from the DM endowment (at cost €0.10 per €0.50 addition / reduction). *, **, *** indicates significant difference between Baseline and Random conditions at the 10%, 5%, and 1% level, Mann-Whitney-U test. $^{\#}$, $^{\#}$, $^{\#\#\#}$ indicates significant difference from zero, two-sided t-test.

Appendix A2.2 discusses fairness perceptions in detail. There we show that DMs perceive the recipients' outcome allocation as rather unfair in both Baseline and Random (2.43 vs. 2.84 on a scale from 1 to 7). At the same time, they perceive the allocation based on their own decision as substantially less procedurally fair than the random allocation (2.45 vs. 4.03, $p < 0.001$, Mann-Whitney-U-test). This suggests, that despite the clear evidence for favoritism, DMs do not seem to consider allocation according to political preferences as appropriate from a fairness perspective. Moreover, in an ex-post questionnaire (Appendix A2.5), DMs indicate that they feel they were influenced by the information, but did not intentionally favor one person. This interpretation is also consistent with findings by Dong and Huang (2018), who show that fairness considerations are stronger than ingroup preferences in an allocation game. In contrast to the current design, however, Dong and Huang's (2018) decision-makers had the opportunity to implement outcomes fairly.

2.4. Interventions to Reduce Favoritism

Given the absence of any strong difference in deservingness or merit, we argue that an allocation that gives equal chances to the two recipients is preferable in the current setting. This also seems to be supported by the DMs' own fairness judgments, and from the perspective of their strategic incentives (beliefs). Still, we found clear evidence that the DMs use the information on political attitudes in their allocation decision. In almost 90% of the groups with differences in political attitudes, we find favoritism. We therefore introduce a set of interventions that aim to reduce the incidence of favoritism, without directly imposing a fair procedure as in Random. The reason is that a direct obligation to use a fair (random) procedure will often not be helpful in practice. In the current setting we know that there are no hard facts guiding the allocation decision, and all deviation from equal chances can be attributed to the effects of political attitudes. However, in practice it will often be less clear whether hard facts were exhausted, and when to impose a fair procedure on the DM. Rather, we would like the DM to freely choose to implement a fair procedure. These considerations guided the design of the following four interventions.

2.4.1. Transparency

2.4.1.1. Design and Hypotheses

In this condition, we establish full transparency with respect to the political profiles between group members. That is, in contrast to Baseline, where only the decision-maker learns the recipients´ profiles, all three political profiles within a group are known by each group member. We hypothesized that the extent of favoritism will be reduced compared to Baseline. Recipients will now be able to assess the potential effect role of politics on the allocation (albeit with only 1 data point). We thus conjectured that some DMs will use a contrarian strategy by choosing the less close recipient in the sense of positive discrimination, and that others may explicitly randomize "in their mind." These two effects would lead to a more balanced allocation in terms of the similarity measure.

2.4.1.2. Results

In Transparency, we have 32 groups in which the recipients have different degrees of similarity with the allocator. The DM chooses the more similar recipient in 30 cases (93.8%), significantly more often than a random choice ($p < 0.001$, binomial test against 0.5). We thus do not find reduced favoritism in an environment where recipients are able to identify it. Recipients anticipate the allocator´s choice behavior, as 89.1% of recipients expect the allocator to choose the more similar recipient (recipients are aware of the political profiles when providing their beliefs).

DMs fairness judgments in Transparency are indistinguishable from the Baseline condition (see Appendix A2.2): they consider both the outcomes and the process as unfair. DMs' beliefs are slightly (insignificantly) less optimistic about the high-payoff recipient and slightly (marginally significantly) less pessimistic about the low-payoff recipient (Appendix A2.4). That is, it seems to be neither the case that DM consider it very appropriate to allocate to the more similar participant (which should have led to substantially more optimistic beliefs), nor do they expect stronger repercussions under transparent allocation towards a more similar recipient. Transparency does not work as an intervention to reduce favoritism.

2.4.2. Private Lottery

2.4.2.1. Design and Hypotheses

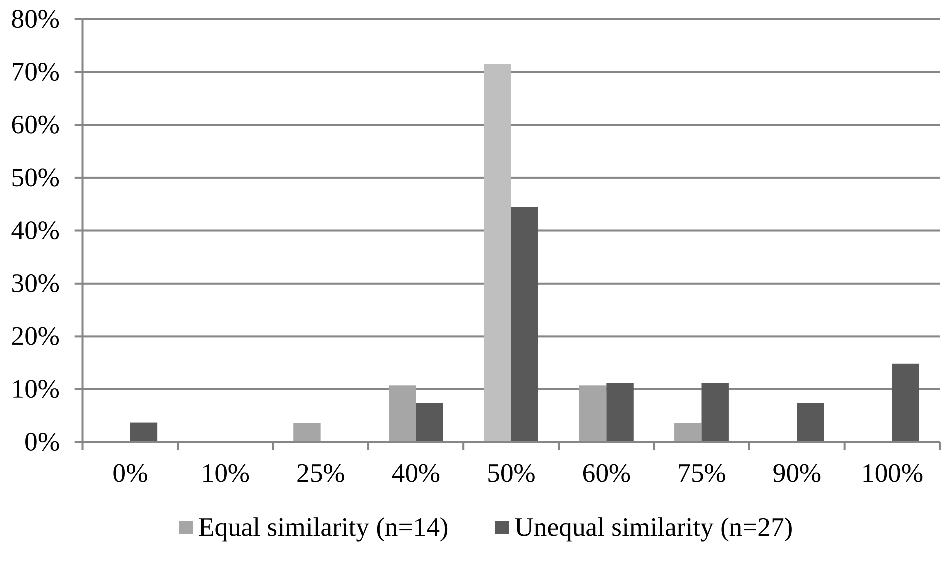
In this condition, we allow allocators to distribute chances to receive the high outcome between recipients, rather than the outcome themselves. The DM is equipped with a randomization device allowing her to allocate a probability *p* of 100%, 90%, 75%, 60%, 50%, 40%, 25%, 10%, 0% for the high payoff to one recipient, with probability *1-p* allocated to the other recipient. Given the selected probabilities, the computer then determines the (still unequal) final outcome allocation. While in Baseline and Transparency only an implicit randomization by "coin flip in their mind" was possible for the DMs, the lottery procedure allows them to explicitly select a fair equal-chance procedure. However, asymmetric allocations of chances are also possible, in particular the allocation of a 100% chance to the politically closer recipient is still available (as in Baseline, only the DM knows the political profiles, and does so before making her allocation decision). Importantly, recipients do not learn which lottery the allocator selected, only the final outcome.

We hypothesized that the extent of favoritism will be reduced compared to Baseline, because a fair procedure (allocation of 50% chance to either participant) is directly available. Implicit randomization in their mind may be more difficult to implement for participants than a direct choice of the fair procedure. Moreover, the availability of the explicit randomization device may make the fair procedure more salient, even in the presence of other unfair allocations (p≠50%). Moreover, for a modest degree of favoritism, a more nuanced way to allocate chances (90%, 75%, or 60%) is now available. That is, although only intrinsic motivations can guide to the DM allocation decisions absent public information about her allocation of chances, we predict that the incidence of "100%" allocation is substantially reduced in comparison to Baseline or Transparency.

2.4.2.2. Results

In Private Lottery, we have 27 groups in which the recipients have different degrees of similarity with the allocator. In these groups, the DM allocates a chance of 100% to the more similar recipient in 4 occasions (14.8%), and to the less similar recipient once (3.7%). Clearly handing out the higher €8-payment to either candidate is not a compelling option for DMs in this treatment. Figure 2.1 shows the full distribution of chance allocations for the 27 groups with unequal similarity measure, and for those 14 groups with equal similarity, i.e., $S_{AB} = 0$.

Figure 2.1. Allocation of Chance to Closer Recipient



Notes: X-axis shows the probability assigned by the DM to the more similar recipient (or one of the equally close recipients in case of equal similarity). Y-axis gives percentage of choices of each of these chance allocations over groups with unequal similarity measures, or groups with equal similarity measures, respectively.

In groups where recipients have different similarities, the modal allocation is the equal-chance allocation, which accounts for almost half of all decisions. Still, the politically closer recipient receives on average a probability weight of 61.7%, which is significantly larger than the equal-chance allocation of 50% (p < 0.05, t-test); moreover, a recipients' similarity score S_j

and her assigned probability of the €8-payoff are positively correlated (ρ = 0.3317, Spearman rank correlation**,** $p < 0.05$). However, favoritism is significantly reduced compared to Baseline and Transparency if we take the realized allocation probabilities of 89.3% and 93.8% in these conditions as a benchmark ($p < 0.001$ against Baseline and $p < 0.001$ against Transparency, two-sided t-tests). That is, there is only a very modest degree of favoritism. This also shows up in the data for the groups with equal similarity measure. Despite the equal similarity measure, in almost 30% of the cases, the DM allocates modestly different chances to the two recipients. These deviations from the equal-chance allocation may either be random, or may derive from differences on individual items of the political attitude questionnaire. The finding suggests that some modest degree of favoritism seems desirable to the DM. On balance, however, providing a private randomization device successfully reduces the incidence of favoritism in our setting by a substantial margin.

We can shed more light on the desirability of favoritism using the fine-grained chance allocations in the current treatment together with the political attitude measures of the DMs. In the 27 groups with $S_{AB} \neq 0$, DMs who are classified as left-wing[11] on average assign a probability weight of 65.8% for the high payoff to the closer participant, while subjects categorized as right-wing on average assign a probability weight of 50% to both recipients. The difference between the two groups is marginally statistically significant ($p < 0.1$, Mann-Whitney-U test).

2.4.3. Public Delegation

2.4.3.1. Design and Hypotheses

In this condition, we give the DM the opportunity to publicly delegate the allocation decision to a random device that assigns the payoffs with equal chances to the two recipients. Delegation may be costly for the allocator with the possible prices being €0.00, €0.50 and €1.00. We apply the strategy method in order to have allocators indicate for each price whether they want to delegate, or whether they want to make the decision themselves (knowing the recipients' political profiles). If they choose to decide themselves for some price, they are also directly asked to indicate their allocation decision for that case. These decisions are all shown on one screen. Subsequently, one of the three prices is randomly selected, and the DM's decision for that price is implemented. Recipients learn only about the payoff relevant decision. That is, for

[11] Based on the answers to our 6 political questions, answering more than 50% of the questions in a typical left-leaning way. For the exact definition of this measure, see Appendix A2.3.

the price that has been selected, they learn the price of delegation, whether the DM delegated or not, and what their payoffs is.

We hypothesized that public delegation allows (i) intrinsically motivated DMs to implement the fair procedure similar to Private Lottery, and (ii) strategically oriented DM to signal fair procedure (see comparison of DMs' beliefs between Random and Baseline). We thus expected a strong reduction in favoritism (indicated by high take-up of delegation). However, we also expect DMs to trade-off the costs of delegation against its benefits (strategic or intrinsic). We thus expect a lower degree of delegation for higher prices. This effect is potentially moderated by the stronger signal to oneself and others of choosing delegation given a higher price.

2.4.3.2. Results

If delegation is costless, 70.0% of the DMs delegate the decision to the random device, a significantly larger share than we would expect if DM randomized ($p < 0.05$, binomial test against 0.5). As the price increases to €0.50, 32.5% of the DM delegate, a significant reduction ($p < 0.001$, sign-test). A further increase to a price of €1.00 leads to 20% of the DM selecting delegation, a significant further reduction ($p < 0.05$, sign-test). However, for 22.5% of the DMs we observe that they delegate if delegation is costless, and in the two costly cases they choose each recipient exactly once for the high payoff. That is, the strong price sensitivity may to some degree be driven by the fact that an alternative (but unobserved by recipients) "implied randomization device" was available through the strategy method.

To better understand the motivation to delegate, for the costless delegation case we asked the DMs' about their expected rewards or punishments in case they delegate and in case they do not. We find that DMs expect significantly lower reward from the high payoff recipient and lower punishment from the low payoff recipient, with a positive but insignificant overall effect on the sum of rewards and punishments. These patterns of beliefs are consistent with those in Baseline versus Random. Regressing political attitudes (a left-wing index) on the delegation decision in the costless case does not suggest any political preference for public delegation (see Appendix A2.3. for details).

Overall, the opportunity for public delegation successfully reduces the incidence of favoritism. Counting delegation as non-favoritism ("50%"), we find overall 65.4% (for price €0), 73.1% (for price €0.50) and 78.8% (for price €1.00) of allocations of high payoffs towards the politically closer recipient (calculations include only groups with $S_{AB} \neq 0$). That is, even costly delegation can have a substantial effect in reducing favoritism. Interestingly, for costless delegation, the effect is comparable to the effect of Private Lottery, which led to a 61.7% likelihood of the closer recipient being allocated the high payoff.

2.4.4. Information Avoidance

2.4.4.1. Design and Hypotheses

In this condition, before the allocation decision takes place, the DM has to decide whether she wants to be informed about the political profiles of the recipients before making her decision, or not. If the DM decides to learn the profiles of the recipients, the decision environment is the same as in Baseline. If she decides to stay uninformed, she does not learn the political profile of recipients and takes the allocation decision without having any knowledge about the recipients (i.e., de-facto allocating randomly)

We conduct two variants of this condition: *private* and *public* avoidance of information. In the first variant, the avoidance decision happens privately, i.e., recipients do not learn whether the DM wanted to remain uninformed about recipients' politics. In the second variant, recipients are informed about the DM's self-selected information status. Recipients always know that information avoidance was possible.

We hypothesized that it may be difficult for DMs to disregard information once they obtained it, but that they may prefer not to receive information that they do not want their decision to be influenced by. Moreover, public delegation or private randomization may sometimes be difficult to implement in practice, possibly because it is at odds with the idea that there exists an optimal allocation and the DM is tasked to identify and implement it. Withholding information that is considered irrelevant or inappropriate to use is potentially easier to justify. We further hypothesized that publicly observable information avoidance is more attractive to DMs as it combines intrinsic and strategic benefits.

2.4.4.2. Results

For private Information Avoidance, we find that 6 of the 48 DMs (12.5%) avoid information about recipients' politics. Conditional on obtaining such information, we find a similar degree of favoritism as in baseline. Of the 42 groups where DMs obtained information, $S_{AB} \neq 0$ in 36 groups. In these groups, 32 DMs (88.9%) allocated the higher payoff to the closer recipient. In public Information Avoidance, we find that 23 of the 44 DMs (52.3%) avoid information about recipients' politics. Conditional on obtaining such information, we find a similar degree of favoritism as in baseline. Of the 21 groups where DMs obtained information, $S_{AB} \neq 0$ in 17 groups. In these groups, 16 DMs (94.1%) allocated the higher payoff to the closer recipient.

The difference in information avoidance between private and public conditions is substantial and statistically significant ($p < 0.001$, Mann-Whitney-U test). Counting information avoidance as non-favoritism ("50%"), we find overall 83.3% allocations of high payoffs towards the politically closer recipient in the private and 68.8% in the public condition (calculations include all DMs who avoid information and DMs who obtain information and observe $S_{AB} \neq 0$). That is, public information avoidance has a somewhat lower but comparable effect to Public Delegation and Private Lottery. The private information avoidance opportunity does not work well. It seems that DMs have a strong preference for information, even if they later end up finding it unfortunate making use of it in the binary allocation decision. It seems more compelling for DMs to later resort to random allocation if given the opportunity, irrespective of its private or public nature, than to avoid information in the first place. We find no link between the political attitude of the DM and the decision to avoid information (Appendix A2.3).

2.5. Discussion and Conclusion

This paper studies a setup common to many allocation problems: (1) a resource is indivisible; (2) objective criteria to guide the allocate decision are unavailable (in practice they may not help distinguish between people after an initial preselection of candidates); and (3) softer subjective criteria may be available to the allocator. We hypothesized that such softer criteria play an important role affecting the allocation. Importantly, we assume that often either the decision-maker or the institution employing or controlling the decision-maker, consider a decision on the basis of such soft criteria inappropriate (Shaw and Olson, 2014; Choshen-Hillel et al., 2015).

Because a fair allocation of outcomes is unattainable, establishing process fairness may be desirable (Moorman, 1991; Organ and Ryan, 1995). Decision-makers or organizations may have an intrinsic preference for process fairness (Krawczyk, 2010; Saito, 2013; Trautmann, 2009). They may also fear repercussions (legal, sabotage) from the affected agents if the allocation process is not perceived as fair (Dickinson et al., 2017; Grosch and Rau, 2017; see also Appendix A2.2), or vice versa, expect higher acceptance of unfavorable outcomes if fairness is signaled (e.g., by delegation, Bartling and Fischbacher, 2012; Coffman and Real, 2018). More generally, process fairness can make undesirable organizational outcomes and procedures more acceptable (Adler et al., 1983; Brockner, 2002; Garonzik et al., 2000; Kessler and Leider, 2016; Skarlicki and Folger, 1997). An organization may therefore want to induce the allocator to employ a fair process. However, any mechanism inducing process fairness should be such that the allocator can freely choose to allocate fairly: in most settings, only the decision-maker will be able to judge which criteria can be used as hard facts, and whether they help to distinguish among candidates.

In a Baseline condition, we replicate empirical evidence that favoritism is strong in situations where an indivisible resource has to be allocated among multiple parties (Hong et al., 2016; Robin et al., 2012). About 90% of the decision-makers allocate according to the soft criterion provided by the political attitude measures. Given that political similarity is the only available information, and that an unfair outcome allocation cannot be prevented, it is not surprising that the attitudes determine the allocation. However, evidence on fairness judgments and expected rewards in comparison to a controlled Random allocation condition suggests that allocation according to political attitudes is not perceived as appropriate. That is, in Baseline, decision-makers do not prefer to allocate according to politics, but there is little to prevent such a result to emerge. Irrespective of whether the allocators find the use of politics to guide their decision appropriate, we emphasize that the entity designing the allocation process may find it undesirable. We therefore design four interventions with the goal to reduce favoritism, that is, aiming for a 50% benchmark for the allocation of the better outcome to the politically closer participant, but without enforcing random choice. Moreover, given the above considerations regarding repercussions if processes are perceived as unfair, the intervention should also improve fairness judgments.

A simple measure to compare the interventions is provided by the implied share of favoritism. The politically closer recipient is allocated the better outcome in the following share of groups: 93.8% when full transparency is provided; 61.7% if a private random lottery can be employed to distribute chances rather than outcomes directly; between 65.4% and 78.8% if public delegation is available (depending on the cost of signaling the fair procedure); and 68.8% respectively 83.3% if information can be avoided before making a decision and this is either communicated to recipients or not.

We make a few observations. All interventions except Transparency have some effect reducing the incidence of favoritism, but none of the interventions fully eliminates favoritism. The possibility to signal the intention to apply a fair procedure is important for its success.

Higher cost of implementing a fair procedure reduce allocators willingness to apply a fair procedure. Making an indivisible resource flexibly “divisible” by using allocation according to probability shares has a stronger effect than any of the process-fairness signaling interventions, despite the private nature of the mechanism. This is consistent with the result that none of the interventions attains a clearly positive judgment of process fairness: unfair outcomes matter for the assessment of the process (Baron and Hershey, 1988; König-Kersting et al., 2018). That is, although there is a process-fairness effect on outcome fairness evaluations (which we may dub an *inverse outcome bias*, see Appendix A2.2), the effect is too modest to make a compelling case for equal-chance random allocation. These findings replicate previous work showing that random allocation of indivisible outcomes may not be as convincing empirically as the theoretical process fairness perspective predicts (Elster, 1989, p. 55; Keren and Teigen, 2010). The success of the flexible Private Lottery condition suggests that equal-chance random allocation (as in Public Delegation or Information Avoidance) may be too strong. Allowing allocators a modest degree of favoritism (by allocating unequal probabilities) works better than the more stringent interventions. As a final observation, we note that the take-up of Information Avoidance, either private or public, seems too modest in comparison to the observed preference for process fairness in Private Lottery and Public Delegation. We interpret the finding as a preference for knowing (e.g., Loewenstein, 1994, but see Hertwig and Engel, 2016), with allocators not anticipating the possibly undesirable degree of influence on their decisions.

Apart from Private Lottery, the interventions discussed here have close counterparts in practical settings. Transparency and delegation to a neutral entity are classic solutions in many domains. Information Avoidance has been proposed in public discourses as a remedy for biased selection in labor or rental markets. Our results suggest that transparency may not work when each decision-maker makes only one decision, and coordination across decision-makers is not easily attainable. How would companies coordinate to attain a close to equal share of male and female CEOs? The possibility to publicly delegate or avoid certain types of information is supported by our results as an effective tool. Interestingly, the most successful intervention, namely a weighted random selection if candidates are very close, has been proposed as a tool to implement justice in distribution processes in an early contribution by Edgeworth (1890). He suggested using *graduated* lotteries in the context of permission to the civil service, i.e., lotteries calibrated with regard to merit, for example, examination scores. Merit would be accounted for, and feelings of injustice by the candidates, stemming from the possibly arbitrary character of the examination, would be mitigated. Boyle (1998) argues that such lotteries help implement fairness in selection processes in organizations, incorporating both efficiency and equity considerations, and avoiding biases and corruption. Despite these theoretical foundations, weighted lotteries are, to the best of our knowledge, not a tool used in practice. We believe that this is due to the possibly questionable normative status of the mechanism. If one candidate is preferable, she should be allocated the better outcome. If there is no clear ordering of candidates, an equal share allocation seems most compelling.[12] A weighted random allocation allowing for inclusion of soft criteria is not convincing from this perspective. However, if the mechanism works best for imperfect, worldly decision-makers, it may be the right intervention to achieve our empirical goal: procedural fairness.

[12] Alternatively, more information should potentially be collected to find the best candidate. While this may not always be feasible in practice, the possibility of random allocation may be problematic if it leads to too little information collected in the first place.

Appendix 2

A2.1. Questionnaire Details

A2.1.1. Original German wording in political-attitudes questionnaire

The original wording of the political attitudes questionnaire taken at the beginning of the experiment was as follows:

Bitte beantworten Sie zunächst den folgenden Fragebogen zu verschiedenen politischen Themen, indem Sie die Antwortmöglichkeit auswählen, die am ehesten auf Sie zutrifft (die Fragen beziehen sich jeweils auf Deutschland).

Ihre Antworten werden möglicherweise an andere Teilnehmer bei diesem Experiment weitergeleitet. Ihre Anonymität bleibt jedoch selbstverständlich während des gesamten Experimentes und auch bei der Auswertung der Daten gewahrt.

1. Wie ist Ihre politische Ausrichtung? [„Eher Links“ / „Eher rechts“]

2. Sind Sie für oder gegen die unbegrenzte Aufnahme von Flüchtlingen? ["Dafür" / "Dagegen"]

3. Wie stehen Sie der Einführung von Frauenquoten in Unternehmen gegenüber? ["Positiv" / "Negativ"]

4. Sind Sie für oder gegen die Einführung der aktiven Sterbehilfe? ["Dafür" / "Dagegen"]

5. Wie beurteilen Sie den Ausstieg aus der Atomenergie? ["Sinnvoll" / "Nicht sinnvoll"]

A2.1.2. Wording of the fairness-perception questionnaire

The wording of the outcome and process fairness evaluations regarding the allocation of payments to player A and B was follows:

Outcome Fairness [German wording] / [translation]

[„Als wie fair beurteilen Sie die Auszahlungen, die für Spieler A und Spieler B aus dem Verteilungsprozess resultieren?“] / [„How fair do you evaluate the payments that result for Player A and Player B?”]

Process Fairness [German wording] / [translation]

[„Als wie fair beurteilen Sie den Sie den Verteilungsprozess?“] / [„How fair do you evaluate the allocation process?”]

Participants answered the questions on a 7-point-Likert scale, with point 1 labeled as "very unfair" and point 7 labeled as "very fair", and point 4 representing a neutral judgment).

A2.2. Fairness Perceptions

This appendix provides details on the players' perceptions of process fairness and outcome fairness. Table A2.1 shows fairness perceptions in the Baseline condition, and compares them to the perceptions in the forcedly fair Random condition, the Transparency condition, the Private Lottery condition, and the Private Information Avoidance condition. Results are shown by player type and pooled over all player types.

Table A2.1. Fairness Perceptions

Condition	**DM**		**High Payoff Recipient**		**Low Payoff recipient**		**All Players**	
	Process Fairness	Outcome Fairness	Process Fairness	Outcome Fairness	Process Fairness	Outcome Fairness	Process Fairness	Outcome Fairness
Baseline	$2.45^{###}$	$2.43^{###}$	$3.15^{###}$	$2.55^{###}$	$2.10^{###}$	$1.85^{###}$	$2.57^{###}$	$2.28^{###}$
Random	4.03	$2.84^{###}$	3.53	$2.66^{###}$	4.32	$2.50^{###}$	3.96	$2.67^{###}$
Δ *Random*	*1.58****	*0.41*	*0.38*	*0.11*	*2.22****	*0.65***	*1.39****	*0.39***
Transparency	$2.44^{###}$	$2.20^{###}$	$2.85^{###}$	$2.37^{###}$	$2.56^{###}$	$2.37^{###}$	$2.62^{###}$	$2.31^{###}$
Δ *Transparency*	*-0.01*	*-0.23*	*-0.30*	*-0.18*	*0.46*	*0.52***	*0.05*	*0.03*
Private Lottery	$3.05^{###}$	$2.63^{###}$	$3.37^{##}$	$2.93^{###}$	$2.71^{###}$	$2.20^{###}$	$3.04^{###}$	$2.59^{###}$
Δ *Private Lottery*	*0.60***	*0.20*	*0.22*	*0.38*	*0.61***	*0.35*	*0.47****	*0.31***
Private Information Avoidance	$2.88^{###}$	$2.33^{###}$	$2.98^{###}$	$2.48^{###}$	$2.40^{###}$	$2.17^{###}$	$2.75^{###}$	$2.33^{###}$
Δ *Private Information Avoidance*	*0.43*	*-0.10*	*-0.17*	*-0.07*	*0.30*	*0.32*	*0.18*	*0.05*

Notes: The table contains evaluations about process fairness and outcome fairness measured on a 7-point-Likert scale (1="very unfair"; 7="very fair"). Δ *Name* indicates differences between condition *Name* and Baseline. $^{\#}$, $^{\#}$, $^{###}$ indicates significant difference from the value of 4 (which indicates neutrality on the fairness scale), two-sided t-test, and *, **, *** indicates significant difference between Baseline and *Name* condition, Mann-Whitney-U test, at the 10%, 5%, and 1% level.

Table A2.2 considers the role of public delegation and public information avoidance for fairness perceptions. We find that delegation and information avoidance lead to higher perceptions of process fairness. Note that only for the DM the situation is self-selected; the recipients are exposed to the situation as determined by the DM.

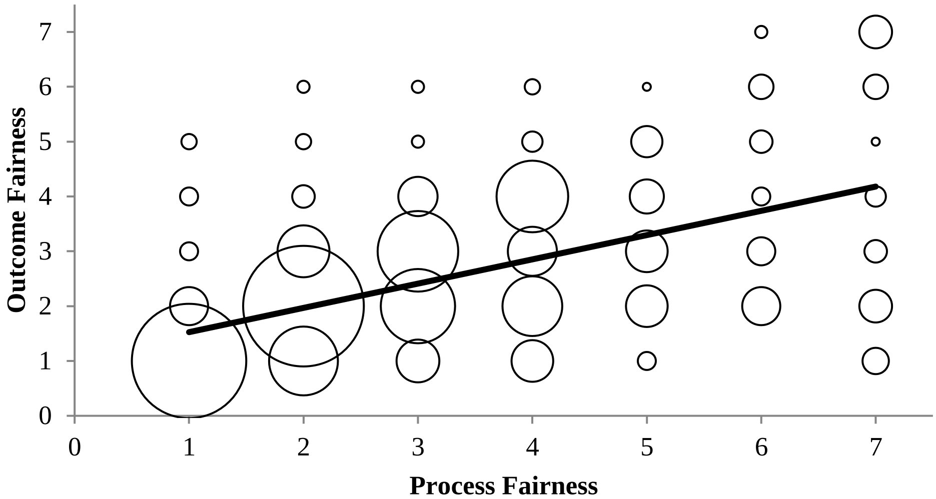
Table A2.2. Fairness perceptions in Delegation and Public Information Avoidance

Condition	**DM**		**High Payoff Recipient**		**Low Payoff recipient**		**All Players**	
	Process Fairness	Outcome Fairness	Process Fairness	Outcome Fairness	Process Fairness	Outcome Fairness	Process Fairness	Outcome Fairness
No Delegation (n=23)	$3.39^{\#\#}$	$2.39^{\#\#\#}$	3.74	$3.04^{\#\#}$	$2.17^{\#\#\#}$	$1.91^{\#\#\#}$	$3.10^{\#\#\#}$	$2.45^{\#\#\#}$
Delegation (n=17)	4.41	$2.47^{\#\#\#}$	4.65	$3.18^{\#}$	4.65	3.24	$4.57^{\#\#}$	$2.96^{\#\#\#}$
Δ *Delegation*	*1.02*	*0.08*	*0.91*	*0.14*	*2.48****	*1.33****	*1.47****	*0.51***
Information publicly not avoided (n=21)	$2.57^{\#\#\#}$	$2.33^{\#\#\#}$	$2.81^{\#\#\#}$	$2.29^{\#\#\#}$	$2.10^{\#\#\#}$	$2.00^{\#\#\#}$	$2.49^{\#\#\#}$	$2.21^{\#\#\#}$
Information publicly avoided (n=23)	3.52	$2.00^{\#\#\#}$	$3.04^{\#\#\#}$	$2.17^{\#\#\#}$	$3.17^{\#}$	$2.83^{\#\#\#}$	$3.25^{\#\#\#}$	$2.33^{\#\#\#}$
Δ *Information publicly avoided*	*0.95*	*-0.33*	*0.23*	*-0.12*	*1.07**	*0.83*	*0.76***	*0.12*

Notes: The table contains evaluations about process fairness and outcome fairness measured on a 7-point-Likert scale (1="very unfair"; 7="very fair"). Δ *Delegation* and Δ *Information publicly avoided* indicates differences between groups in which the DM delegated / avoided information, and those where she did not. $^{\#}$, $^{\#}$, $^{\#\#\#}$ indicates significant difference from the value of 4 (which indicates neutrality on the fairness scale), two-sided t-test, and *, **, *** indicates significant difference between delegation and non-delegation, respectively information avoidance and non-avoidance, Mann-Whitney-U test, at the 10%, 5%, and 1% level.

Table A2.1 and A2.2 show that there is some tendency for self-serving process and outcome fairness evaluations. High payoff recipients report higher process fairness judgments than low payoff recipients in conditions Baseline ($p < 0.01$, Mann-Whitney-U test), Private Lottery ($p < 0.1$, Mann-Whitney-U test), Private Information Avoidance ($p < 0.05$), and in Delegation (but only when the DM does not delegate, $p < 0.01$, Mann-Whitney-U test). Low payoff recipients perceive higher process fairness in Random ($p < 0.1$, Mann-Whitney-U test). Likewise, high payoff recipients report higher outcome fairness judgments than low payoff recipients in conditions Baseline ($p < 0.01$, Mann-Whitney-U test), Private Lottery ($p < 0.01$, Mann-Whitney-U test) and Delegation (but only if the DM does not delegate, $p < 0.05$, Mann-Whitney-U test). Overall the evidence for self-serving fairness perceptions suggests the existence of an *outcome bias* in fairness judgments. Table A2.1 and A2.2 also show that increased perception of process fairness often coincides with higher perceptions of outcome fairness, despite the constant degree of outcome inequality in payoffs across all conditions of the experiment. Pooling data of all treatments and all players, we find a significant positive correlation between outcome and process fairness evaluations at the individual level ($\rho = 0.5716$, Spearman rank correlation, $p < 0.001$). Figure A2.1 shows the relationship graphically for all players pooled (effects are very similar for all three player types). With process fairness judgments affected by treatment conditions and outcome inequality fixed, we observe what we may term a *process bias* in outcome fairness judgments.

Figure A2.1. Relationship Between Process and Outcome Fairness Perceptions (all subjects pooled)



We finally assess whether fairness perceptions correlate with recipients' punishment or retaliation behavior. Table A2.3 shows the results of simple regression analysis of the rewarding behavior across all conditions, as a function of the recipients' process and outcome fairness judgments. We find no significant relationship of process or outcome fairness judgments of high payoff recipient with their behavior. For low payoff recipients, rewarding behavior correlates strongly with their process fairness judgments, but not with outcome fairness judgments.

Table A2.3. Regression analysis of fairness perception on reward/retaliation behavior

	Impact on DM´s payoff from high payoff recipient	Impact on DM´s payoff from low payoff recipient
Process Fairness Judgment	0.002 (0.042)	0.211*** (0.044)
Outcome Fairness Judgment	-0.061 (0.050)	0.001 (0.056)
Constant	1.000*** (0.130)	-0.844*** (0.131)
N	292	292

Notes: Process Fairness and Outcome Fairness are judged on a 7-point-Likert scale, with the value of 1 being labeled as "very unfair" and the value of 7 being labeled as "very fair." The impact on DM´s payoff ranges from –€2 to +€2, with incremental steps of €0.50. OLS regressions; all conditions pooled; * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$, standard errors in parenthesis.

A2.3. The role of politics

This appendix sheds more light on the role of political attitudes in the current allocation setting. Pooling the data from all treatments (876 subjects), Table A2.4 shows the distribution of answers for each of the 5 political attitude questions. The data shows a large degree of variation over recipients, as 92.5% of recipients within a group differ on at least one item. On average, recipients within one group chose different answers in 2.03 items.[13]

Table A2.4. Summary Statistics on Political Attitudes

Item	Distribution among respondents (n=876)				Correlation
General political attitude	Left wing:	61.5%	Conservative:	38.5%	1
Support of unlimited inflow of refugees	Yes:	50.3%	No:	49.7%	0.26***
Support of female quota in organizations	Yes:	63.6%	No:	36.4%	0.15***
Support of active euthanasia	Yes:	77.7%	No:	22.3%	0.14***
Support of exit from nuclear energy	Yes:	86.6%	No:	13.4%	0.09***

Notes: The last column shows the correlation with the general political attitude question. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. The exact wording of the questionnaire is provided in Appendix A2.1.

Table A2.5 shows the distribution of the similarity score S_j between recipients and DM, and the distribution of $|S_{AB}|$ over all groups in the experiment. The distribution of $|S_{AB}|$ provides

[13] We note that in the post-experimental questionnaire, 96.0% of the respondents indicate having answered the questions in the political questionnaire truthfully.

a direct measure for the potential for favoritism. The larger $|S_{AB}|$, the more recipients differ in their similarity to the DM.

Table A2.5. Distribution of S_j and $|S_{AB}|$

Distribution of similarity score S_j						
S_j	0	1	2	3	4	5
#	6	61	116	207	144	50
Distribution of absolute value of relative similarity score S_{AB}						
$\|S_{AB}\|$	0	1	2	3	4	5
#	80	121	64	22	4	1

We next look at the role of political similarity for allocation decisions. Table A2.6 replicates Table 2.2 for the whole sample of decisions for all conditions with an active allocation and known attitudes (i.e., DM does not delegate or avoid information; probabilities in Private Lottery larger 50% coded as assigning higher payoff). We find that all dimensions contribute to the preferential treatment of more similar recipients by the DM.

Table A2.6. Relative Similarity and Allocation Behavior (all Conditions with Active Allocation and Known Attitudes)

Dimension of similarity	Assignment of high payoff
Left-wing political attitude	0.461*** (5.51)
Support of unlimited refugee inflow	0.389*** (5.13)
Support of mandatory female quota	0.272*** (3.62)
Support of active euthanasia	0.584*** (5.15)
Support of exit from nuclear energy	0.543*** (4.08)
Observations	167

Notes: Marginal effects of probit regression reported, z-values in parenthesis. The explanatory variables are dimension specific relative distance measures. *, **, *** indicates significance at the 10%, 5%, and 1% level.

We next study how the DM's political attitudes correlate with her allocation behavior. For this purpose, we define a *left-wing index* as the share of left-wing stereotype answers in the political attitudes questionnaire (left answers in Table A2.4). The index can take the values 0, 0.2, 0.4, 0.6, 0.8, and 1. Its median in the whole sample is 0.8. Subjects with index 0, 0.2, 0.4 or 0.6 are classified as below median left-wing (n=441 in the total sample), and subjects with index 0.8 or 1 are classified as equal/above median left-wing (n=435 in the total sample).

Table A2.7 shows that there is tendency for above-median left wing DMs to show more pronounced favoritism. This effect is insignificant for the Baseline condition on the basis of only 28 groups, but highly significant in the whole sample with 195 groups considered.

Table A2.7. Left Wing Index and Favoritism

	Favoritism	No favoritism	
Panel a: Baseline			
Left-wing index below median	12	3	15
Left-wing index equal/above median	13	0	13
Fisher exact test, $p = 0.226$	25	3	28
Panel b: All conditions			
Left-wing index below median	54	48	102
Left-wing index equal/above median	69	24	93
Fisher exact test, $p < 0.01$	123	72	195

Notes: Contingency table shows the number of DMs by political attitude choosing the closer recipient in Baseline and Transparency, not delegating and then choosing the closer recipient in Delegation, not avoiding information and then choosing the closer recipient in Information Avoidance, for the higher payment; assigning higher probability for the higher payment to the closer participant in Private Lottery; only groups with $S_{AB} \neq 0$.

We finally study how the DM's political attitude affects her decision to delegate (if delegation is costless) and to avoid information (Table A2.8). We find that there is no effect of DM's politics on the decision to delegate or to avoid information (irrespective of whether avoidance is publicly revealed).

Table A2.8. Left Wing Index and Delegation and Information Avoidance

	Delegation (if costless)	Avoid information (if private)	Avoid information (if public)	All three conditions pooled
Left-wing index	-0.134 (-0.33)	-0.221 (-1.49)	0.194 (0.57)	-0.162 (-0.83)
Female	-0.139 (-0.86)	0.086 (1.14)	-0.202 (-1.29)	-0.088 (-0.97)
Age	0.001 (0.04)	0.011 (0.74)	-0.007 (-0.38)	0.011 (0.81)
Economics	-0.131 (-0.73)	0.322* (1.72)	-0.223 (-1.29)	0.051 (0.49)
# Observations	40	48	44	132

Notes: Marginal effects of probit regressions reported, z-values in parenthesis. The left-wing index indicates the share of left-wing stereotype answers in the questionnaire. *, **, *** indicates significance at the 10%, 5%, and 1% level.

A2.4. Recipients' behavior and DM´ beliefs

This appendix summarizes recipients' rewarding and punishment behavior, and the DM' beliefs regarding the recipients' behavior. We first report recipients' behavior in the different conditions, and then the DM's beliefs regarding the recipients' behavior. All analyses report the (belief in the) monetary impact on DM's payoffs caused by each recipient (i.e., not the costs involved by the recipient for punishing or rewarding).

Table A2.9. Recipients´ Impact on DM's Payoff

Condition	High payoff recipient	Low payoff recipient
Baseline	1.41###	- 0.30#
Random	0.43###	0.43##
Δ *Random*	*- 0.98****	*0.73****
Transparency	1.00###	- 0.60###
Δ *Transparency*	*- 0.41**	*- 0.30*
Private Lottery	0.56###	- 0.13
Δ *Private Lottery*	*- 0.85****	*0.17*
Private Information Avoidance	0.88###	- 0.49###
Δ *Private Information Avoidance*	*- 0.53****	*- 0.19*

Notes: Entries are average additions to / reductions from the DM endowment (at cost €0.10 per €0.50 addition / reduction). *, **, *** indicates significant difference between Baseline and the respective intervention at the 10%, 5%, and 1% level, Mann-Whitney-U test. #, #, ### indicates significant difference to the value of zero, two-sided t-test.

Table A2.10. Recipients´ Impact on DM's Payoff by Delegation and by Information Avoidance

Condition	High payoff recipient	Low payoff recipient
No Delegation (n=23)	1.02###	- 0.67###
Delegation (n=17)	0.26	0.26
Δ *Delegation*	*- 0.76****	*0.93****
Information publicly not avoided (n=21)	0.90###	-0.38#
Information publicly avoided (n=23)	0.93###	-0.17
Δ *Information publicly avoided*	*0.03*	*0.21*

Notes: Entries are average additions to / reductions from the DM endowment (at cost €0.10 per €0.50 addition / reduction). *, **, *** indicates significant difference between no delegation and delegation, respectively information non-avoidance and avoidance, at the 10%, 5%, and 1% level, Mann-Whitney-U test. #, #, ### indicates significant difference to the value of zero, two-sided t-test.

Table A2.9 shows rewarding and punishment by recipient and condition, and differences with Baseline. In all conditions, high payoff recipients reward less than in Baseline. For low payoff recipients there are few statistically significant differences; the only effect observed is lower punishment in Random than in Baseline. Table A2.10 shows rewarding and punishing depending on the publicly known delegation and information avoidance. We observe that high payoff recipients reward less, but low payoff recipients punish less (send insignificantly positive rewards) when DM’s delegate. At the end of the experiment, we ask recipients, about their maximum willingness-to-pay for delegation, if they were to be in the role of DM and find that low outcome recipients state a significantly higher WTP ($p < 0.05$, Mann-Whitney-U test). Finally, Information avoidance has no effect on rewarding behavior.

Table A2.11. DM’s Beliefs about Recipients´ Impact on Their Payoff by Delegation and by Information Avoidance

Condition	High payoff recipient	Low payoff recipient
Baseline	0.83$^{\#\#\#}$	- 0.83$^{\#\#\#}$
Random	0.39$^{\#\#}$	0.11
Δ *Random*	*- 0.44****	*0.94****
Transparency	0.73$^{\#\#\#}$	- 0.54$^{\#\#\#}$
Δ *Transparency*	*- 0.10*	*0.29**
Lottery	0.49$^{\#\#\#}$	- 0.50$^{\#\#\#}$
Δ *Lottery*	*-0.34*	*0.33**
Avoidance (private)	0.66$^{\#\#\#}$	- 0.41$^{\#\#\#}$
Δ *Avoidance*	*- 0.17*	*0.42***

Notes: Entries are average beliefs about additions to / reductions from the DM endowment (at cost €0.10 per €0.50 addition / reduction). *, **, *** indicates significant difference between Baseline and the respective intervention at the 10%, 5%, and 1% level, Mann-Whitney-U test. $^{\#}$, $^{\#}$, $^{\#\#\#}$ indicates significant difference to the value of zero, two-sided t-test.

Table A2.12. DM’s Beliefs about Recipients´ Impact on Their Payoff

Condition	High payoff recipient	Low payoff recipient
No Delegation	$0.50^{\#\#\#}$	$- 0.83^{\#\#\#}$
Delegation	0.08	$- 0.22^{\#}$
Δ *Delegation*	*- 0.42****	*0.61****
Information publicly not avoided	$0.57^{\#\#\#}$	$- 0.68^{\#\#\#}$
Information publicly avoided	$0.31^{\#\#\#}$	$- 0.31^{\#\#\#}$
Δ *Information publicly avoided*	*- 0.26**	*0.37***

Notes: Entries are average beliefs about additions to / reductions from the DM endowment (at cost €0.10 per €0.50 addition / reduction). *, **, *** indicates significant difference between no delegation and delegation, respectively information non-avoidance and avoidance, at the 10%, 5%, and 1% level, two-sided t-test. $^{\#}$, $^{\#}$, $^{\#\#\#}$ indicates significant difference to the value of zero, two-sided t-test. All DM were asked about both scenarios, i.e., delegation *and* no delegation as well as information avoidance *and* no information avoidance.

Tables A2.11 and A2.12 show the DM’s beliefs for the behavior shown in tables A2.9 and A2.10. DM’s anticipate the general pattern that high payoff recipients become less rewarding and low payoff recipients become less punishing. However, they are too pessimistic in their judgment, expecting too strong a decline in rewards and too weak a decline in punishment. Table A2.12 shows a similar pattern. DM’s believe that both delegation and information avoidance will have an effect on rewards and punishment. They do not fully anticipate the strength of the effects for the low payoff recipients. Moreover, they seem to be less well calibrated for the case of no delegation and no avoidance. We report the following additional result not shown in the tables. We test whether beliefs are different for those DMs who delegate and those who do not. We do not find substantial qualitative or quantitative differences. That is, expected monetary awards cannot explain differences between these two groups of players.

At the end of the experiment, we also ask DMs whether they would appreciate delegation if they were to be in the role of a recipient. We find that stating that they would appreciate delegation in that case strongly correlates with opting for costly delegation ($p < 0.001$ for both prices). This further supports the conclusion that the motivation to pay for delegation is rather intrinsic than extrinsic.

A2.5. Post-Experimental Questionnaire

This appendix provides the post-experimental questionnaire questions in Table A2.13, and the results in Table A2.14.

Table A2.13. Post-Experimental Questionnaire – Questions

	Question	Answer options
(1)	Do you think that the DM´s decision was influenced by the information about the recipient? *(Recipient version of the question)* Was your allocation decision influenced by the information about the recipient? *(DM version of the question)*	Yes/No
(2)	Do you think that the DM intentionally favored one of the two recipients? *(Recipient version of the question)* Did you intentionally favor one of the two recipients? *(DM version of the question)*	Yes/No
(3)	Do you think that the [recipient with the high payoff] feels intentionally favored or disadvantaged? *(Only DM)*	Favored/ Disadvantaged/ Neither nor
(4)	Do you think that the [recipient with the low payoff] feels intentionally favored or disadvantaged? *(Only DM)*	Favored/ Disadvantaged/ Neither nor
(5)	How important is it to implement a fair process, even when this is costly? *(All)*	Likert-Scale from 1-7 (1=unimportant; 7=important)

Table A2.14. Post-Experimental Questionnaire – Results

	Type	(1): Yes/No	(2): Yes/No	(3): Favored/ Disadvantaged/ Neither	(4): Favored/ Disadvantaged/ Neither	(5)
Baseline	DM	36/4	21/19***	31/3/6	2/33/5	5.7
	High	35/5	19/21***			5.7
	Low	38/2	29/11***			5.8
Transparency	DM	37/4	11/30***	25/1/15	0/34/7	5.2
	High	37/4	14/27***			5.5
	Low	38/3	20/21***			5.8
Private Lottery	DM	22/19	13/28***	14/3/24	5/15/21	5.3
	High	34/7	22/19***			5.5
	Low	31/10	27/14			5.6
Private Info Avoidance	DM	36/6	17/25***	27/0/21	0/32/16	5.3
	High	47/1	21/27***			5.4
	Low	46/2	28/20***			5.6
Delegation	DM	9/8	5/12	1/1/15	1/1/15	5.1
	High	10/7	5/12**			5.5
	Low	10/7	4/13***			6.2
No Delegation	DM	17/6	12/11*	12/0/11	0/11/12	5.4
	High	20/3	14/9**			5.6
	Low	21/2	17/6**			5.6
Public Info Avoidance	DM	-	-	3/1/19	0/2/21	5.3
	High	-	-			5.9
	Low	-	-			5.2
No Public Info Avoidance	DM	19/2	12/9***	17/0/4	0/17/4	6.0
	High	21/0	8/13***			5.5
	Low	21/0	13/8***			5.4

Notes: *, **, *** indicates significant difference between (1) and (2), at the 10%, 5%, and 1% level, two-sided t-test.

We test whether items 1 and 2 differ from one another and find that in almost all cases, both the recipients as well as DMs do more strongly agree to the statement that the DM is influenced by the information about recipients, than to the statement that the DM explicitly and intentionally favors one of the recipients. This corroborates the idea that it is not intentional nepotism that drives the strong degree of favoritism, but the inability to disregard available information even if the DM does not intend to use it.

Chapter 3

Capitalizing on the (False) Consensus Effect: Two Tractable Methods to Elicit Private Information

Robert J. Schmidt

Abstract: We propose and experimentally test two tractable methods to incentivize the elicitation of private information: *Benchmark* and *Coordination*. Both mechanisms capitalize on the false consensus effect, a well-documented phenomenon that follows Bayesian reasoning. That is, individuals use their own type when predicting the type of others. Since it is not feasible to incentivize the elicitation of private information using facts, when these are not verifiable, we incentivize the respondent to reveal her perceptions about others and use that statement to predict the subject´s private information. The stronger the relationship between a subject´s type and her perception about the type of others, the more effective the mechanisms are in revealing the subject´s privately held information. In an experiment, we apply the mechanisms to reveal beliefs about probabilities. On the aggregate level, both mechanisms accurately reveal mean first-order beliefs of the population. On the subject level, the modal difference between probabilities elicited in either mechanism and actual first-order beliefs is zero. The results indicate that subjects strongly anchor their statements in *Benchmark* and *Coordination* on their private information.

Highlights:

- Two tractable methods to elicit private information are proposed
- The methods serve as (simple) alternatives to Bayesian revelation mechanisms
- In an experiment, both mechanisms accurately reveal first-order beliefs

Acknowledgments: I thank Aurelien Baillon, Dietmahr Fehr, Christian König, Christiane Schwieren, Christoph Vanberg, Jens Witkowski as well as seminar audiences in Frankfurt, Heidelberg, the HeiKaMaX in Heidelberg, the HeiKaMaXY in Heidelberg, and the Bayesian Crowd Conference in Rotterdam for very valuable comments and suggestions.

3.1. Introduction

The elicitation of private information, such as preferences, beliefs, feelings or opinions, is key for social sciences (Turner and Martin, 1985), policymakers (Veenhoven, 2002), corporations (Monroe, 1973) and for public opinion research (Price and Neijens, 1997).[14] The value of such data requires that subjects exert effortful thinking when the question at hand is non-trivial and that subjects do not bias their answer towards social desirability (Li, 2007; Manski, 2004; Zizzo, 2010). Therefore, various methods have been proposed that condition a respondent´s answer on some observable fact and monetarily reward the subject for accuracy.[15] If the monetary incentive is sufficient, the mechanism is incentive compatible, as the subject is induced to honestly report her *type* (Smith, 1976).[16]

Incentivizing accurate reporting by conditioning on facts, however, limits a mechanism to the elicitation of private information about *verifiable* questions.[17] By contrast, the elicitation of *unverifiable* questions lies beyond the scope of this approach. This comprises questions that are hypothetical, counter-factual, technically unverifiable, or which refer to subjective tastes.[18] Therefore, so-called truth serums have been developed, which aim at increasing the quality of the elicitation of private data compared to non-incentivized procedures (Prelec, 2004).[19] The core assumption of these methods is that individuals are subject to the false consensus effect (Ross et al., 1977), a well-documented phenomenon that follows Bayesian reasoning (Dawes, 1989).[20] That is, individuals use their own private information when predicting private information of others.[21] Given that a set of behavioral assumptions holds, truthfully answering the question is a Bayesian-Nash equilibrium and maximizes the recipient´s payoff.

[14] In particular, this kind of data is essential for corporations that offer online platforms to gather and provide customer evaluations, e.g., about restaurants, hotels, or other services.

[15] See Schlag et al. (2015) and Schotter and Trevino (2014) for literature reviews.

[16] The term *type* is meant to represent the respondent's trait that is of interest to a researcher. This might be a respondent´s preference, belief, taste, opinion, and the like.

[17] For example, a health scientist might ask a respondent about the risk of smoking or an economist about the current inflation rate. The respondent is then paid based on the *objective precision* of the answer.

[18] In these cases, interviewers are usually limited to the elicitation of non-incentivized statements.

[19] Indeed, there is evidence that incentives for truth-telling induce subjects to report socially desirable behavior less often (Barrage and Lee, 2010), admit wrong-doings more often (John et al., 2012; Loughran et al., 2014), state their future behavior more accurately (Howie et al., 2011), increase accuracy in recognition tasks (Prelec and Weaver, 2006), and increase the coherence between elicited beliefs and observed behavior (Trautmann and van der Kuilen, 2014).

[20] In the title, we put the word "false" in parentheses, in order to remark that it is *not necessarily false* to derive beliefs about others using the signal that stems from one´s own type (Dawes, 1989). Instead, rational belief formation requires that (to some degree) subjects use their own type as a valid source when predicting the type of others (see also Prelec, 2004). Section 3.3.1.2 elaborates on that.

[21] This assumption is also common in various Bayesian settings (e.g., Cremer and McLean, 1988; d'Aspremont and Gerard-Varet, 1979; Johnson et al., 1990; McAfee and Reny, 1992; McLean and Postlewaite, 2002).

However, the application of this group of mechanisms comes with two difficulties. First, they require several behavioral assumptions, such as common knowledge about a shared common prior belief, common knowledge about respondents updating their belief in an impersonally informative manner, subjects being able to identify that truth-telling represents a Bayesian Nash equilibrium and trusting others to play according to that Bayesian Nash equilibrium, too.[22] Second, the scoring rules applied in Bayesian revelation mechanisms are complicated. This makes it hard to implement the mechanisms transparently by informing subjects about the exact scoring system. As a result of that, most empirical applications have relied on faith-based implementation, without explaining the actual scoring rules in detail, but by assuring participants that truth-telling would be optimal (e.g., Barrage and Lee, 2010; Howie et al., 2011; John et al., 2012; Prelec and Weaver, 2006; Shaw et al., 2011).

In this paper, we propose two tractable methods that intend to solve these problems: *Benchmark* and *Coordination*. Like Bayesian revelation mechanisms, both methods rely on the false consensus effect. While it is not possible to incentivize the elicitation of private information when there is no ground truth, it is feasible to incentive compatibly elicit a subject´s perception about others. Therefore, both mechanisms provide direct incentives for respondents to make statements that depend on their beliefs about the type of their peers. Following the idea of the false consensus effect, the elicited statements are then used to predict the respondent´s *own* type. The stronger the relationship between a subject´s type and her beliefs about others, the more effective the mechanisms are in revealing the subject´s private information.

In the first mechanism, *Benchmark*, a two-step approach is applied. First, a subject is asked about her private information in an non-incentivized manner. That statement is then used to serve as a benchmark for the elicitation of second-order beliefs (Perner and Wimmer, 1985) from another respondent, who has to guess the private information of the previously asked subject. The existence of a benchmark allows incentivization using ordinary scoring rules, such that the second respondent is induced to engage in second-order reasoning in an effortful manner. The stronger the relationship between a subject´s own thought and second-order belief, the better of a proxy the elicited statement will be for her type.

In the second mechanism, *Coordination*, respondents are provided with a question and various answer alternatives. The subjects´ task is to coordinate on an answer, and each subject

[22] Impersonal informativeness implies two aspects. First, a subject´s own type is informative, i.e., it provides evidence about population frequencies. Therefore, subjects expect over proportionally large shares of their own type among their peers. Second, this inference is impersonal, i.e., respondents of the same type draw identical inferences about the population, thereby arriving at identical posterior beliefs (Prelec, 2004).

is paid based on the precision with which she anticipates the coordination outcome. The mechanism is inspired by the concept of focal points (Schelling, 1960), an approach to predict behavior in coordination settings with multiple equivalent equilibria. The concept postulates that participants in pure coordination settings exhibit shared perceptions about salient alternatives. Thereby, focal points emerge absent from payoffs and provide an implicit coordination device (Sudgen, 1995). Since the recognition of focal points requires subjects to form beliefs about the perception of saliences in other individuals, it involves higher-order reasoning (Camerer et al., 2004). Consequently, an individual´s coordination choice reflects her belief about the perception of others and, therefore, is informative about her own perception about the question at hand.[23]

The main advantage of the two methods is that they are tractable. The scoring and the payout function are easy to understand, such that participants are provided with a clear task that they have to solve. Therefore, the mechanisms are easy to implement for experimenters. By contrast to Bayesian revelation mechanisms, there is no theoretical necessity that subjects reveal their private information. However, it is reasonable to expect valid signals about the respondent´s thoughts. In section 3.4, we discuss how the mechanisms need to be implemented to maximize the discriminatory power of the two measures.

In an experiment, we mimic the elicitation of beliefs about unverifiable probabilities.[24] Subjects are provided with instructions about an ultimatum game (Güth et al., 1982) conducted by Trautmann and van de Kuilen (2014), and we elicit beliefs about proposer and responder behavior. Applying both a between-subject and a within-subject design, we elicit beliefs using the mechanisms *Benchmark* and *Coordination* as well as actual first-order beliefs by conditioning payments on factual probabilities. In the between-subject comparison, we find that both mechanisms accurately reveal mean first-order beliefs of the population. In the within-subject comparison, we find that the modal difference between probabilities elicited in either

[23] As we will show in section 3.3.2.2., the mechanism represents a generalization of the Krupka and Weber (2013) approach to identify social norms using coordination games. They propose to use coordination games to identify *social norms* on the *aggregate level*, while we argue that coordination games are suited to identify *any kind of private information* on the *individual level*. Indeed, there is evidence that individual coordination choices about social norm perception are related to a subject´s preferences. Schmidt (2019b) finds that injunctive and descriptive social norms elicited using coordination games predict revealed social preferences in a series of dictator games.

[24] Belief about probabilities are an essential form of private information in the social sciences. For example, in psychology beliefs about probabilities are used to understand fear diseases (Slovic et al., 1980), in health sciences to understand risky health behaviors (Khwaja et al., 2006, 2009; Schoenbaum, 1997) and in economics to understand saving and investment behavior (Guiso et al., 1992, 1996).

mechanism and actual beliefs is zero. We therefore conclude that, in the given setting, subjects strongly anchor their statements in *Benchmark* and *Coordination* on their first-order beliefs.

The remainder of the paper is organized as follows. Section 3.2 reviews the literature on methods to elicit private information. Section 3.3 explains *Benchmark* and *Coordination*, and also provides a theoretical background for each mechanism. Section 3.4 illustrates how the mechanisms need to be implemented to maximize the discriminatory power. In section 3.5, we present the experiment to test the mechanisms in the area of probabilistic beliefs, and in section 3.6 we formulate testable hypotheses. Section 3.7 presents the results. Section 3.8 discusses advantages and disadvantages compared to Bayesian revelation mechanisms. Section 3.9 summarizes and concludes.

3.2. Related Literature

The seminal contribution to eliciting unverifiable subjective information is Prelec (2004), who introduces a truth-inducing scoring system that includes two additive parts. First, an information report that refers to information privately owned by a respondent (her type). The information report is scored based on the degree to which it is surprisingly common in the population.[25] Second, subjects make a prediction report. This report refers to the subject´s belief about the distribution of types in the population, and it is scored based on accuracy. Given a set of behavioral assumptions, such as common knowledge about a shared prior belief, impersonal informativeness, Bayesian reasoning and a sufficiently large sample of participants, truthfully reporting the own type represents a Bayesian Nash equilibrium.

Since Prelec´s (2004) innovation, various refinements have been proposed. Prelec and Seung (2006) demonstrate how to use the mechanism even when the majority of respondents are wrong. Witkowski and Parkes (2012a) propose the Robust Bayesian Truth Serum, which corrects Prelec´s (2004) drawback to operate properly only on large samples. In the Robust Bayesian Truth Serum, three subjects are sufficient to establish Bayesian Nash incentive-compatibility, but the mechanism is restricted to the elicitation of binary information. The modifications of Radanovic and Faltings (2013, 2014) allow the elicitation of non-binary

[25] The surprisingly common criterion exploits Bayesian reasoning, as subjects should and usually do make use of their own type, when predicting the prevalence of their own type in the population (Marks and Miller, 1987; Ross et al., 1977). Consequently, subjects anticipate that the actual prevalence of their own type is underestimated by their peers, which renders truthful reporting optimal regarding the surprisingly common principle. Rewarding answers that are more common than predicted avoids to bias a report in the direction of mainstream answers, since it equivalently rewards subjects with minority answers.

signals, while still being incentive-compatible for small populations. Baillon (2017) proposes Bayesian markets, a method that simplifies previous mechanisms. Subjects make only one decision, namely whether or not to trade an asset whose value represents the share of affirmative answers to a question. Bayesian markets are thus more transparent and tractable for participants, but they are suited for binary questions only.

Our paper is also related to the peer prediction method (Miller et al., 2005), a scoring system that is based on the comparison of reports. Subjects state a report and are scored concerning the precision of their implied posterior belief about the report of another subject, such that truth-telling is a Bayesian Nash equilibrium. Unlike the previously mentioned mechanisms, however, the peer prediction method makes the assumption that a common prior belief is not only shared by agents but also known to the mechanism. Witkowski and Parkes (2012b) propose a modification that allows to relax the common prior assumption. Jurca and Faltings (2006) show that paying a subject based on comparison with a sufficiently large number of agents minimizes the budget required for incentive compatibility.

Finally, our paper is related to Carvalho et al. (2017) who discuss mechanisms that are based on the assumption that respondents exhibit social projection, a strong form of the false consensus effect.[26] They theoretically analyze payment structures that reward agreements and demonstrate that risk-neutral agents maximize their expected reward by honestly reporting their private information. In an online experiment involving text content analysis, the subjects´ task is to review short texts under the criteria of grammar, clarity, and relevance. Their results support the hypothesis that agents report more accurate answers than when there are no incentives for honest reporting.

3.3. Benchmark and Coordination

Two methods are proposed to elicit private information in case of unverifiability: *Benchmark* and *Coordination*. In both methods, subjects are incentivized to make statements that depend on perceptions about private information of others. In section 3.3.1, we explain *Benchmark*, and in section 3.3.2, we explain *Coordination*. In both subsections, we provide theoretical backgrounds that illustrate why the methods are suited to predict a respondent´s type.

[26] Social projection implies that subjects believe that their private answer equals the most popular answer of the remaining respondents.

3.3.1. Benchmark

3.3.1.1. The Mechanism

Benchmark consists of two steps and requires at least two subjects, a *benchmarker* and a *respondent*. In step 1, the experimenter asks the benchmarker about some private information in a non-incentivized manner, and her answer is then considered the *benchmark*. In step 2, the respondent (who is of actual interest to the researcher) is asked to guess the answer of the benchmarker, and she receives a payment that depends on the accuracy of her second-order belief.[27] Creating a benchmark in the first place circumvents the problem that scoring against an objective criterion is not feasible when an answer is unverifiable. Using the benchmark to condition the respondent´s guess allows the application of ordinary scoring rules, thereby inducing her to engage in second-order reasoning in an effortful manner. The closer the relationship between the respondent´s first-order and second-order belief, the better the prediction about her private information.

3.3.1.2. Theoretical Background: The False Consensus Effect

Benchmark capitalizes on the false consensus effect, a well-documented phenomenon that describes the tendency to perceive the own traits, such as preferences, habits, behaviors, choices, or opinions to be correlated with the corresponding traits of peers (Ross et al., 1977). As a result, subjects of a particular type expect over proportionally large shares of subjects similar to them in the population. Indeed, there is ample evidence that individuals overestimate the prevalence of their own characteristics (Bellemare et al., 2011; Bennett, 1999; Blanco et al., 2014; Charness and Grosskopf, 2001; Marks and Miller, 1987; Mullen et al., 1985; Toussaert, 2018).[28] Also, there is experimental evidence that the false consensus effect is robust to information provision (e.g., Ambuehl et al., 2019; Engelman and Strobel, 2012).

The term *false* consensus effect accounts for the fact that subjects tend to *overestimate* the similarity between them and others. By now, however, the conclusion that consensus reasoning

[27] We use the common definition that a subject´s first-order belief describes what she thinks about real events, while a second-order belief refers to what a subject believes about another subject´s thought (Perner and Wimmer, 1985).

[28] The false consensus effect is of particular interest for models of psychological game theory. Ellingsen et al. (2010) argue that correlation between behavior and second-order beliefs do not necessarily represent evidence for guilt aversion, but can partially be explained by false consensus. Bellemare et al. (2011) find that controlling for a consensus effect halves the extent of guilt aversion. Blanco et al. (2014) conclude that the false consensus effect explains correlation between first-mover and second-mover cooperation in a sequential prisoner's dilemma. A more general analysis of the implications of false consensus on psychological game theory is done by Vanberg (2019).

would be false per se has been put into perspective (Dawes, 1989, 1990; Engelman and Strobel, 2012; Vanberg, 2019). Since the own type in fact constitutes a valid signal about the population, using that signal reflects a mere facet of Bayesian reasoning and is thus consistent with rational belief formation. Note that it is secondary for the argument made in *Benchmark* whether the false consensus effect is actually an artifact from rational belief formation or whether subjects put irrationally strong emphasis on the informational value steaming from their own type. It is simply necessary that the described relationship between a subject´s own type and her second-order belief does exist. Therefore, the stronger the degree of consensus reasoning inherent in a subject, the better of a proxy the elicited statement in *Benchmark* will be for her type.

3.3.2. Coordination

3.3.2.1. The Mechanism

In *Coordination*, several subjects receive the same question, and they have to coordinate on the answer. Subjects are compensated based on their ability to anticipate the *coordination outcome*, which is determined as a function of all coordination choices. In case of verbal answers, this is usually the modal answer (e.g., Mehta et al., 1994a, 1994b; Krupka and Weber, 2013). In case of coordination with numbers, this could be the average, the median, or the mode.[29] The higher a subject´s accuracy in anticipating the coordination outcome, the higher her payment.

Coordination is different from *Benchmark* in three aspects. First, by contrast to *Benchmark*, it does not require two steps since the elicitation of private information and the creation of the coordination outcome happen simultaneously. Second, while in *Benchmark* two subjects are needed to make the mechanism work, this is not necessarily the case in *Coordination*. Specifically, the mechanism requires that subjects perceive the coordination outcome to be exogenous, i.e., a single participant is not able to influence the coordination outcome.[30] This requires that the number of participants is sufficiently large. Third, *Benchmark* and *Coordination* differ in the potential depth of reasoning required in the settings (Camerer et al., 2004; Nagel, 1995; Stahl and Wilson, 1994, 1995;). *Benchmark* only requires the formation of second-order beliefs (beliefs about the thoughts of others). By contrast, coordination games are complex, and subjects might engage in the formation of even higher-order beliefs, in order to

[29] For example, in Fehr et al. (2019), subjects have to coordinate by stating a number between 0 and 100. The smaller the distance between a respondent's number and the average of all numbers, the higher her payment.

[30] This is relevant, because subjects shall reveal their best guess about the coordination outcome. If they were able to *influence* the outcome, they might engage in strategically affecting it.

anticipate the coordination outcome in a more sophisticated manner. This, however, does not pose a threat to the proposed mechanism, as long as beliefs of higher-order depend on a subject´s first-order belief, i.e., her own thought about the question at hand. The mechanism *Coordination* thus relies on the assumption that a subject´s first-order belief tends to be correlated with beliefs of all orders (Dawes, 1989, 1990; Engelman and Strobel, 2012; Vanberg, 2019).

3.3.2.2. Theoretical Background: Focal Points in Coordination Games

Coordination capitalizes on the theory of focal points, a concept proposed by Schelling (1960) to understand behavior in pure coordination settings. Schelling (1960) argues that in pure coordination games with multiple equivalent equilibria, subjects perceive varying degrees of saliences of the available alternatives, and they assume that their perception is shared by the remaining participants (Sudgen, 1995). As a result, subjects use their own perception about salient choices to make predictions about how saliencies are perceived by other participants.[31] This creates focal points that emerge absent from payoffs, thereby constituting an implicit coordination device.[32]

The proposition to infer private information from coordination choices is a *generalization* of the Krupka and Weber (2013) approach to use coordination games to identify social norms. In their mechanism, subjects are confronted with the description of a particular behavior, and their task is to coordinate on appropriateness ratings. Assuming that social norms reflect shared perceptions about appropriate behaviors (Crawford and Ostrom, 1995), focal points will be determined by social norm perception of subjects. As a result, the coordination outcome indicates the perception of social norms within the players´ population. In their experiment, Krupka and Weber (2013) find that social norms elicited using coordination games predict behavior shifts across different versions of the dictator game. While Krupka and Weber (2013) conclude that coordination games are suited to identify *social norms* on the *aggregate level*, we

---

[31] Importantly, these saliences are assumed to be meaningful (to a researcher), i.e., they are induced by the question at hand. For example, in the original version of the Keynesian beauty contest (Keynes, 1936), respondents have to coordinate on the most attractive pictures of women. According to Schelling´s concept, such choices might reveal prevalent beauty ideals within the guessers´ population.

[32] Since Schelling (1960), both experimental and theoretical work has corroborated the relevance of focal points in a variety of coordination settings, e.g., Binmore and Samuelson, 2006; Casajus, 2000; Crawford et al., 2008; Fehr et al., 2019; Isoni et al., 2013, 2014, 2019; Janssen, 2001, 2006; Metha et al. 1992, 1994a, 1994b; Pope et al., 2015; Sudgen, 1995; Sugden and Zamarrón, 2006. Schmidt (2019a) proposes how to measure the distribution of focal points on the individual level.

argue that their approach is suited to extract *any kind of private information* on the *individual level*.

3.4. Maximizing Discriminatory Power in Benchmark and Coordination

3.4.1. Discriminatory Power

We argue that, based on the phenomenon of false consensus, a subject´s choice in *Benchmark* and *Coordination* yields an informative signal about the respondent´s type. In order to maximize the informativeness stemming from the two mechanisms, the task should be constructed such that subjects which are of different types respond to the task in different ways. In test theory, this feature is referred to as *discriminatory power* and it has been extensively studied in that domain (e.g., Birnbaum et al., 1968; Ferrando, 2012; Hankins, 2007; Loevinger, 1954). Discriminatory power measures the degree to which a test score varies with the level of the measured trait and thus reflects the effectiveness of a test detect differences between participants concerning the respective trait.[33] To illustrate this, imagine a test that is either extremely easy or extremely hard. In both cases, the variance will be low, as the average performance will be either close to the minimum or close to the maximum number of points. To render the distribution of scores informative, the test shall be likely to yield *higher* scores for more capable subjects and *lower* scores for less capable subjects. That is, the test shall yield variable results, *given* that the test-takers are *indeed different*. Therefore, the difficulty of the test needs to be calibrated such that average performances correspond to an expected number of solved tasks lying in the middle of the total number of items. If the difficulty of the task is appropriate, it becomes likely that heterogenous test-takers receive varying scores.

We argue that this design feature is also crucial when applying *Benchmark* and *Coordination*. In particular, we claim that inducing variability in the respondents´ answers is generally feasible, such that a subject´s *choice* is related to her *type* in a meaningful manner. If that holds, then the direction in which the answer of one subject differs from the answer of another subject is informative about the difference between the underlying types of the two recipients.

[33] Therefore, in addition to validity and reliability, discriminatory power is an important feature of the design of tests (Lumsden, 1976).

3.4.2. Example: Eliciting Minority Opinions using Numerical Questions

We illustrate this with the elicitation of minority opinions. Assume that an experimenter wants to elicit beliefs about the probability that, in a sports match, Team A wins against Team B. The experimenter is aware that it is common knowledge among recipients that Team A is significantly weaker than Team B. Let us assume that the participants have diverging opinions about the probability that Team A wins, but the median first-order belief about the probability that Team A wins is 10%. If the experimenter asks whether Team A or Team B would win, then there would not be variation, since no participant in *Benchmark* or *Coordination* would think that Team A is a promising bet in that setting. However, if the experimenter had a reliable prior about the distribution of first-order beliefs, she could calibrate the question accordingly, for example by asking whether or not the probability that Team A wins is smaller or larger than 10%. If the experimenter's prior is accurate, the rephrased question can be expected to induce variability in answers, which would allow to draw discriminatory inferences about the subjects´ types.

That approach, however, requires the experimenter to have a reliable prior about the distribution of first-order beliefs in the population. Another possibility is to provide subjects with numerical answers in a more nuanced way. In the above-described example, the experimenter could have subjects state percentage points for the probability that Team A wins. In the case of *Benchmark*, the experimenter would first elicit the first-order belief of the benchmarker, who states an integer between 0 and 100 that shall represent the probability in percent that the event occurs. The respondent is then asked to guess the integer stated by the benchmarker and is then paid based on the accuracy of her second-order belief. Equivalently, in *Coordination* subjects could coordinate on an integer between 0 and 100. The coordination outcome is determined as a function of the coordination choices, e.g., the mean, the median, or the mode. Each participant is then paid based on the distance between her coordination choice and the coordination outcome.[34] The significant advantage of using numerical scales is that, by design, it is likely to receive variation in the respondents´ answers, since many numbers are a potentially promising bet in the settings of *Benchmark* and *Coordination*.

[34] One potential threat for coordination with numbers results from artefactual focal points, i.e., focal numbers within the set of feasible choices (these could be numbers such as 0, 1, 10, 50, or 100). This, however, is not a problem, when other signals that induce focality, are more prominent. In an experimental setting similar to ours, Fehr et al. (2019) examine whether "sunspots", i.e., external signal about the true state of the world, affect coordination choices, when subjects coordinate on integers between 0 and 100. They find that, when external signals are available, the relevance of artefact focal points diminishes.

3.5. Experimental Design and Procedure

We mimic the elicitation of unverifiable beliefs and examine whether the proposed mechanisms are suited to reveal subjects´ first-order beliefs.[35] The participants´ task is to estimate empirical probabilities of events in an ultimatum game conducted by Trautmann and van de Kuilen (2014), hereafter TK.[36] At the beginning of our experiment, subjects learn that it is their task to estimate probabilities about behavior in an ultimatum game that has already been conducted. For that sake, subjects receive detailed information about TK´s ultimatum game. They are then instructed about their tasks in the respective treatments and the scoring mechanisms. In section 3.5.1, we elaborate on the rules of TK´s ultimatum game. In section 3.5.2, we present the design of our treatments and in section 3.5.3 the procedure of our experiment.

3.5.1. The Ultimatum Game of Trautmann and Van De Kuilen (2014)

In TK´s ultimatum game, the proposer could choose between six alternatives that determined how a fixed pie of 20€ would be divided between herself and a responder. Responders had to indicate via strategy method (Selten, 1967) for each of the possible allocations, whether they would accept that allocation, or not. After every subject took the respective decision, the computer randomly matched proposers and responders. If the responder indicated acceptance for the proposed allocation, the respective allocation was implemented. If the responder indicated rejection, both subjects received nothing.

We pay significant attention to make sure that subjects understand the rules of TK´s ultimatum game and to make clear that it is not their task to play the game themselves but to assess observed behavior rates of others in that game. Subjects are provided with the original wording of TK´s instructions and answer a series of comprehension questions.[37] Also, participants in our experiment receive information about the general setting of TK´s experimental procedure (computerized laboratory experiment, show-up fee of 5€, random

[35] By *mimicking* the elicitation of unverifiable beliefs, we intend to simulate a situation that is equivalent to the measurement of beliefs about unverifiable events. This requires the assumption that subjects are unaware of the factual probabilities that they have to assess.

[36] The experiment of TK consisted of two stages. In stage 1, subjects play the ultimatum game. In stage 2, the authors elicit beliefs from participants using different scoring rules. As subject are paid randomly either for stage 1 (ultimatum game) or stage 2 (belief elicitation), there is no reason to assume that the stages affect each other. Therefore, in our study, we do not mention stage 2 of TK, but only explain the rules of the ultimatum game in stage 1.

[37] Subjects are also explicitly told, that these instructions correspond to the original wording used by TK.

assignment of roles, anonymous interaction, etc.). Table 3.1 shows the available allocations as well as empirical probabilities of proposer choices and responder acceptance rates in TK.

Table 3.1. Ultimatum Game of Trautmann and van de Kuilen (2014)

	Available Allocations in the Ultimatum Game					
	1	2	3	4	5	6
Proposer Payoff	20€	16€	12€	8€	4€	0€
Responder Payoff	0€	4€	8€	12€	16€	20€
	Proposer behavior (n = 103)					
Choice Probability	6%	20%	66%	7%	2%	0%
	Responder behavior (n = 103)					
Acceptance Probability	14%	43%	90%	95%	92%	88%

3.5.2. Treatments and Scoring

Treatments. Four main treatments are conducted: *SURVEY*, *BELIEF*, *BENCHMARK*, and *COORDINATION*. Additionally, we conduct *CONTROL*, a control treatment that is intended to capture the degree of noise inherent in the elicitation of beliefs in the given setting. Subjects are instructed about the task in the respective treatment, i.e., whether their task is to state first-order beliefs, second-order beliefs, or whether their task is to coordinate. Probabilities are separately elicited for the 12 possible events in TK´s ultimatum game. Precisely, subjects state for each of the six possible allocations (i) how probable it was that a proposer chose a particular allocation and (ii) how probable it was that a responder accepted a particular allocation. Our design is intended to compare first-order beliefs, second-order beliefs, and coordination choices both in a between subject-manner and in a within-subject manner (see table 3.2).

• ***SURVEY:*** In treatment *SURVEY*, first-order beliefs are elicited in an non-incentivized manner. Subjects asses the probabilities of the 12 events of TK´s ultimatum game and receive a fixed payment of 12.50€ for their participation in the experiment. Treatment *SURVEY* is intended to yield non-incentivized beliefs that are then used to score second-order beliefs elicited in *BENCHMARK*.[38]

• ***BELIEF:*** In treatment *BELIEF*, first-order beliefs are elicited in an incentivized manner. Subjects are instructed that their payment depends on the precision of their first-order beliefs

[38] Note that, for the purpose of using the results from that treatment for *BENCHMARK*, the number of participants is irrelevant, since the number of participants in a treatment does not affect the expected outcome, as long as subjects are drawn from the same population.

about the factual probabilities in TK. At the end of the experiment, the computer randomly draws one item, and the performance in that item determines a respondent´s payoff.

• ***BENCHMARK:*** *BENCHMARK* consists of two independent parts. In the first part, subjects are instructed, that their task is to assess how *other respondents* previously estimated the results of TK. Also, subjects are informed that their payment depends on the accuracy of their second-order beliefs about the previously elicited estimations of the other respondents. In the second part, subjects have to state their first-order beliefs and are scored as in treatment *BELIEF*, i.e., based on objective accuracy. One randomly drawn item of the two stages determines the payment.

• ***COORDINATION:*** *COORDINATION* consists of two independent parts. In the first part, subjects are instructed that their payment is based on the ability to anticipate the coordination outcome. The coordination outcome is the average number stated by the participants in a session. That is, the closer their stated probability is to the coordination outcome, the higher their payment. In the second part, subjects have to state their first-order beliefs, as is in *BELIEF*. One randomly drawn item of the two stages determines the payment.

• ***CONTROL:*** Treatment *CONTROL* is identical to *BELIEF* except that the treatment consists of two stages, both of which elicit first-order beliefs. That treatment serves as a control condition for the treatments *BENCHMARK* and *COORDINATION*, in order to identify the degree of noise that is inherent in the elicitation of beliefs in the given setting.

Scoring. In each treatment (except *SURVEY*) subjects are paid based on accuracy regarding the respective task, and performance is evaluated relative to the other subjects within a session. Subjects within a session are ranked from highest to lowest accuracy regarding the respective task. In *BELIEF*, subjects are ranked according to the accuracy of their first-order belief in one randomly drawn item. In *BENCHMARK*, subjects are ranked according to the accuracy of their second-order belief. In *COORDINATION*, subjects are ranked according to their ability to anticipate the coordination outcome. The subject with the highest accuracy earns 15€. Payment then linearly diminishes by 0.75€ by each rank. That is, the subject with the second-highest performance earns 14.25€, the subject with the third-highest performance earns 13.50€, and so forth.[39] Since all sessions were conducted with 20 participants, the incentives created through

[39] We opted for this payment scheme for the sake of simplicity for participants. In the experiment, subjects are handed a sheet of paper showing which relative rank yields which payoff. Another advantage of the relative scoring regime we apply is that the incentives for accuracy are high. By contrast, in static scoring rules incentives for being accurate are limited. In the quadratic scoring rule, for example, moderate inaccuracies have only relatively small effects on the respondent´s payoff, while the subject´s payoff diminishes exponentially when the degree of inaccuracy becomes large.

the relative payment scheme are identical. In addition to that payment, subjects receive a show-up fee of 5€. By contrast to these treatments, in treatment *SURVEY*, subjects receive a show-up fee of 12.50€.[40]

After instructing subjects about their specific task and the scoring mechanisms, they answer a series of control questions. This way, we make sure that they understand their task, and how their compensation would be determined in the respective treatments. Table 3.2 summarizes the structure of treatments and illustrates the between-subject and the within-subject comparisons that the experiment allows.

Table 3.2. Treatment Overview and Content

Treatment	n	Stage 1	Stage 2
Survey	20	First-order belief (non-incentivized)	-
Belief	60	First-order belief	-
Benchmark	60	Second-order belief	First-order belief
Coordination	60	Coordination	First-order belief
Control	40	First-order belief	First-order belief

3.5.3. Procedure

The computerized experiment (z-Tree; Fischbacher, 2007) was conducted at the experimental laboratory of Heidelberg University (Germany). 240 participants were recruited from the general student population via hroot (Bock et al., 2012) and participated in 12 experimental sessions between January and June 2019. Mean age was 23.4 years, 53.8% were female, and 30.4% had an economics background in their studies. Pairwise Mann-Whitney-U tests indicate that the composition of participants´ gender, age, and field of study does not differ between treatments. A typical session lasted about 45 minutes, and subjects earned on average about €12.80 including a show-up fee of €5.[41]

3.6. Hypotheses

A simple model of second-order beliefs in *Benchmark* and coordination choices in *Coordination* is set up to formulate testable hypotheses. Denote subjects as $i = 1, \dots, n$ and

[40] In expectancy, the payment between the five treatments is (almost) identical. The expected payoff in the treatments with relative payments is 12.88€.
[41] A replication package, including instructions in German and English language, raw data, and data analysis files is available at the repository for research data of Heidelberg University: https://heidata.uni-heidelberg.de.

events as $j = 1, \dots, m$. Subject i´s first-order belief about the probability that event *j* materializes is FB_{ij}. Second-order beliefs elicited in *Benchmark* are denoted as SB_{ij} and coordination choices elicited in *Coordination* as C_{ij}. All statements FB_{ij}, SB_{ij} and C_{ij} are expressed as integers between 0 and 100, representing the probability in percent that an event materializes. Accordingly, average first-order beliefs of the population about the probability that event *j* materializes are $\overline{FB_j} = (\sum_{i=1}^{n} FB_{ij})/n$, average second-order beliefs are $\overline{SB_j} = (\sum_{i=1}^{n} SB_{ij})/n$ and average coordination choices are $\overline{C_j} = (\sum_{i=1}^{n} C_{ij})/n$.

We model second-order beliefs and coordination choices as a function of first-order beliefs: $SB_{ij} = FB_{ij} + \varepsilon_{ij}^{SB}$ and $C_{ij} = FB_{ij} + \varepsilon_{ij}^{C}$. The error terms ε_{ij}^{SB} and ε_{ij}^{C} capture the difference between a respondent´s statement in the respective mechanism and her actual first-order belief. One way to interpret these error terms is that they result from an anchoring and adjustment procedure (Tversky and Kahnemann, 1974).[42] That is, subjects *anchor* their statements in *Benchmark* and *Coordination* on their first-order belief, and they then *adjust* it deepening on their perception about the coherence between their own perception and their best guess about the perception of others (Epley et al., 2004).[43] Accordingly, averages of the population can be formulated as $\overline{SB_j} = \overline{FB_j} + \overline{\varepsilon_j^{SB}}$ and $\overline{C_j} = \overline{FB_j} + \overline{\varepsilon_j^{C}}$.[44] The model illustrates when the mechanisms Benchmark and Coordination work best, namely when ε_{ij}^{SB} and ε_{ij}^{C} are small. The following hypotheses formulate how ε_{ij}^{SB} and ε_{ij}^{C} as well as $\overline{\varepsilon_j^{SB}}$ and $\overline{\varepsilon_j^{C}}$ look like.

Hypothesis 1A refers to *average* second-order beliefs $\overline{SB_j}$ made in *Benchmark* and Hypothesis 1B refers to *average* coordination choices $\overline{C_j}$ made in *Coordination*. We hypothesize that the average statements made in the two mechanisms about a particular event *j* do not differ from average first-order beliefs $\overline{FB_j}$ elicited in *Belief*. This implies that $\overline{\varepsilon_j^{SB}}$ and $\overline{\varepsilon_j^{C}}$ do not differ from zero.

Hypothesis 1A. Average second-order beliefs about the probability of event *j* do not differ from average first-order beliefs: $\overline{SB_j} = \overline{FB_j}$.

[42] The "anchoring and adjustment heuristic" describes the strategy to make judgments under uncertainty by anchoring on information that comes to mind and adjust it until a plausible estimate is reached.

[43] Epley et al. (2004) propose to model perspective taking as an anchoring and adjustment procedure. People derive beliefs about others by initially anchoring their beliefs in an egocentric manner, and subsequently accounting for potential differences between themselves and others. In a series of experiments, the authors find evidence for this hypothesis.

[44] Note that the model is intended to be simple and yield tractable hypotheses, therefore it is not the aim to model what kind of processes shape error terms.

Hypothesis 1B. Average coordination choices about the probability of event *j* do not differ from average first-order beliefs: $\overline{C_j} = \overline{FB_j}$.

Hypothesis 2A refers to *individual* second-order beliefs SB_{ij} elicited in *Benchmark* and Hypothesis 2B refers to *individual* coordination choices C_{ij} elicited in *Coordination*. We hypothesize that the individual statements made in the two mechanisms about a particular event *j* do not differ from individual first-order beliefs FB_{ij} elicited in *Belief*. This implies that ε_{ij}^{SB} and ε_{ij}^{C} do not differ from zero.

Hypothesis 2A. Individual second-order beliefs about the probability of event *j* do not differ from individual first-order beliefs: $SB_{ij} = FB_{ij}$.

Hypothesis 2B. Individual coordination choices about the probability of event *j* do not differ from individual first-order beliefs: $C_{ij} = FB_{ij}$.

3.7. Results

3.7.1. Between-Subject Analysis of Averages

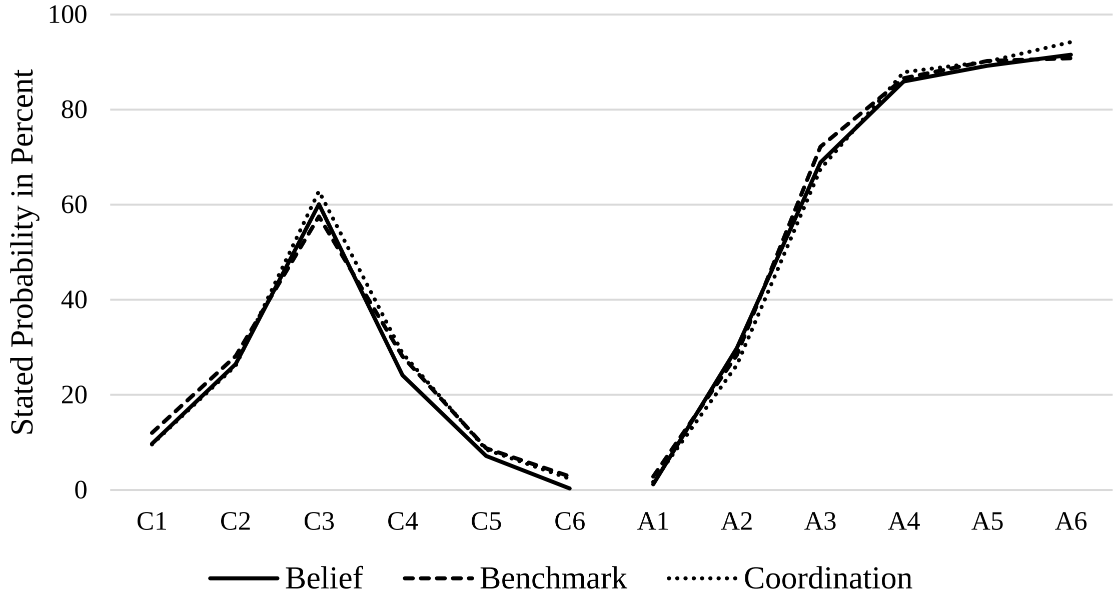
We start with aggregate level analysis by comparing average first-order beliefs with (i) average second-order beliefs and (ii) average coordination choices. Figure 3.1 shows average first-order beliefs elicited in *BELIEF*, average second-order beliefs elicited in stage 1 of *BENCHMARK*, and average coordination choices elicited in stage 1 of *COORDINATION*. Items C1 to C6 refer to probabilities of proposers-choices and A1 to A6 refer to acceptance-rates of responders. Mann-Whitney-U tests are conducted to test for equality of averages. Before correcting for multiple testing, item C6 differs between *BENCHMARK* and *BELIEF* ($p < 0.05$) and the same item differs between *COORDINATION* and *BELIEF* ($p < 0.01$).[45] Both significances vanish when correcting for multiple testing using the Bonferroni procedure.[46] We therefore cannot reject hypotheses 1A and 1B, which state that the average probabilities elicited in *BENCHMARK* and *COORDINATION* are identical to average first-order beliefs elicited in *BELIEF*.

[45] In Appendix A3.2, the reader finds a graph with the results from treatment *SURVEY*. Graphical analysis suggests that non-incentivized beliefs elicited in *SURVEY* tend to differ from incentivized beliefs elicited in *BELIEF*. This is not implausible, given the lack of incentivization to carefully read the instructions and exert effortful thinking in that treatment, since subjects were informed about their fixed compensation at the beginning of the experiment. Note, however, that our study is not intended to test whether non-incentivized elicitation differs from incentivized elicitation of beliefs.

[46] We account for the fact that multiple items are used to detect treatment differences. In order to take care of the inflation of the overall type-I-error rate, we therefore multiply the *p*-values by the number of items (i.e., by twelve).

Result 1. In a between-subject analysis, average second-order beliefs and average coordination choices do not differ from average first-order beliefs.

Figure 3.1. Between-Subject Comparison of Elicited Probabilities



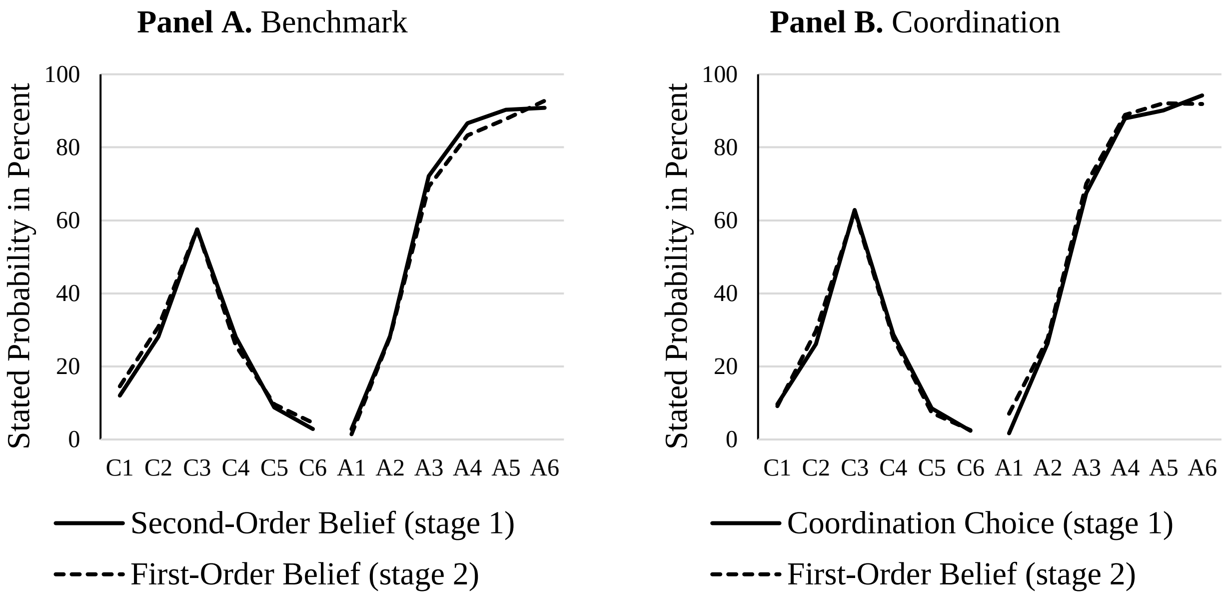
Notes: Numbers are percentage points. C1-C6 refer to probabilities for choice behavior of proposers and A1-A6 refer to probabilities for acceptance behavior of responders. The numbers in *BENCHMARK* and *COORDINATION* are elicited in the first stage of the treatments, i.e., using the respective mechanisms.

3.7.2. Within-Subject Analysis of Averages

To examine differences in *BENCHMARK* and *COORDINATION* to first-order beliefs on the individual level, we start by comparing average outcomes between stage 1 and stage 2 in these treatments. Remember that in stage 1, the respective mechanisms are applied, i.e., subjects state their second-beliefs in stage 1 of *BENCHMARK*, and they coordinate in stage 1 of *COORDINATION*. In stage 2, first-order beliefs are elicited in the same manner as in *BELIEF*. By contrast to the above section, we now compare the outcomes of the mechanisms with first-order beliefs in a within-subject manner. Panel A of Figure 3.2 compares averages of stage 1 and stage 2 in *BENCHMARK*, and Panel B compares averages of stage 1 and stage 2 in *COORDINATION*. Wilcoxon Signed-Rank tests are conducted to detect differences between averages. Before correcting for multiple testing, item C1 ($p < 0.1$) and item C4 ($p < 0.05$) differs between stage 1 and stage 2 in *BENCHMARK*. In *COORDINATION*, item C2 ($p < 0.05$), item A1 ($p < 0.05$) and item A4 ($p < 0.1$) differ between stage 1 and stage 2. Again, these significances vanish after the correction procedure. The results are thus consistent with those reported in the previous section, i.e., average second-order beliefs elicited in *BENCHMARK* and average coordination choices elicited in *COORDINATION* do not differ from average first-order beliefs of subjects.

Result 2. In a within-subject analysis, average second-order beliefs and average coordination choices do not differ from average first-order beliefs.

Figure 3.2. Within-Subject Comparison of Elicited Probabilities



Notes: Numbers are percentage points. C1-C6 refer to probabilities for choice behavior of proposers and A1-A6 refer to probabilities for acceptance behavior of responders. The straight line in Panel A indicates average second-order beliefs elicited in stage 1 of *BENCHMARK* and the straight line in Panel B indicates average coordination choices elicited in stage 1 of *COORDINATION*. The dashed lines in both panels indicate average first-order beliefs elicited in stage 2 from the same participants.

3.7.3. Correlations

We now analyze to what degree probability statements elicited in *BENCHMARK* and *COORDINATION* are related to first-order beliefs of individuals by conducting correlation analyses. Looking at the combined data of all items, we find that the statements in stage 1 and stage 2 are strongly and statistically significantly correlated in both treatments (r = 0.87 in *BENCHMARK*; r = 0.90 in *COORDINATION*; p < 0.00001 in both treatments; Pearson correlation).[47] That result is consistent with the idea promoted in section 3.4, i.e., that the statements extracted in *BENCHMARK* and *COORDINATION* vary with the underlying first-order belief of individuals. Likewise, the correlation between stage 1 and stage 2 is strongly

[47] The correlation coefficients are based on 720 observations in *BENCHMARK*, 720 observations *COORDINATION* and 480 observations in *CONTROL*. Conducting correlation analyses separated by items also yields strongly positive and significant correlations.

positive and statistically significant in treatment *CONTROL* ($r = 0.89$; $p < 0.00001$; Pearson correlation), but the correlation is not higher than in *BENCHMARK* and *COORDINATION*.

Result 3. Second-order beliefs, as well as coordination choices, are significantly positively correlated with first-order beliefs.

3.7.4. Analysis of Error Terms

We proceed by analyzing the congruence between numbers stated in stage 1 and stage 2 of *BENCHMARK* and *COORDINATION*. For that sake, we examine the distribution of error terms ε_{ij}^{SB} and ε_{ij}^{C} defined in section 3.6, which emerge when a subject states different numbers in stage 1 and stage 2 for the same item.[48] It is reasonable to expect that subjects will exhibit noise when stating their beliefs for 12 items two times in a row. To have a baseline to compare the distribution of error terms with, we use the error terms observed in treatment *CONTROL*, which provide a measure for the degree of noise that occurs when subjects state first-order beliefs.

In order to get an impression about that measure, Figure 3.3 shows the distribution of individual error terms based on the combined data of all items.[49] The distribution is centered around zero, and the modal error term, as well as the median error term in each treatment, equal zero (see Table 3.3). Two-sided t-tests are conducted to test if mean error terms on the item level differ from zero.[50] We do not find that error terms in any item differ from zero, neither in *BENCHMARK* nor *COORDINATION*.[51] The fact that error terms do not differ from zero is consistent with Hypotheses 1A and 1B.

To investigate Hypotheses 2A and 2B, we analyze means of *absolute* error terms: $|\varepsilon_{ij}^{SB}|$ and $|\varepsilon_{ij}^{C}|$.[52] Two-sided t-tests are conducted to test if mean absolute error terms on the item level differ from zero.[53] We find that in all three treatments, in most items mean absolute error terms are significantly different from zero on the 5%-level.[54] After the correction procedure, still,

[48] In *BENCHMARK*, error terms are defined as the difference between a subject´s second-order belief and first-order belief: $\varepsilon_{ij}^{SB} = SB_{ij} - FB_{ij}$. In *COORDINATION*, error terms are defined as the difference between a subject´s coordination choice and first-order belief: $\varepsilon_{ij}^{C} = C_{ij} - FB_{ij}$.

[49] The number of data points per treatment equals the number of participants multiplied by the number of items.

[50] In Appendix A3.1, Table 3.5, Panel A, we report mean error terms on the item level.

[51] Likewise, error terms in *CONTROL* do not differ from zero.

[52] Absolute error terms $|\varepsilon_j|$ are the absolute values of error terms ε_j. The *average* absolute error term $\overline{|\varepsilon_j|}$of item j is calculated as $\overline{|\varepsilon_j|} = (\sum_{i=1}^{n} |\varepsilon_{ij}|)/n$.

[53] In Appendix A3.1, Table 3.5, Panel B, we report mean absolute error terms on the item level.

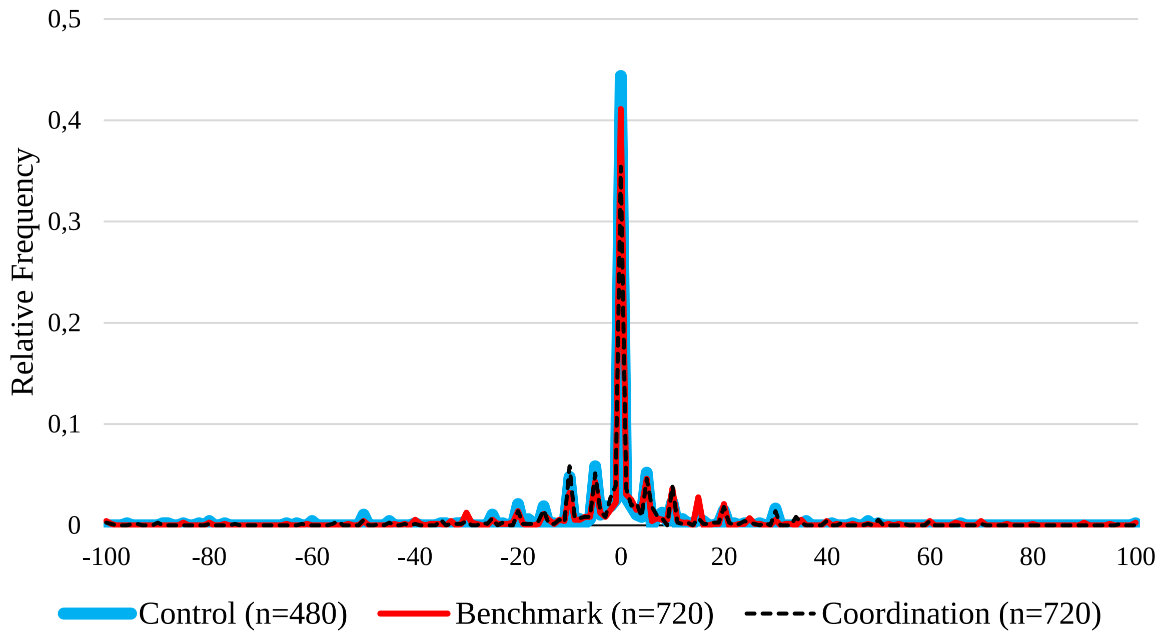
[54] Precisely, in *BENCHMARK* and *CONTROL*, in 11 of the 12 items mean absolute error terms differ from zero with $p < 0.05$; in *COORDINATION*, in 10 items mean absolute error terms differ from zero with $p < 0.05$ (see Appendix A3.1).

most items remain significantly different from zero on the 5%-level. This result is not consistent with Hypotheses 2A and 2B. In order to identify what part of these differences is due to noise, we compare mean absolute error terms in *BENCHMARK* and *COORDINATION* with mean absolute error terms observed in treatment *CONTROL*. Conducting Mann-Whitney-U tests to identify differences between treatments, we do not find that *BENCHMARK* or *COORDINATION* differs from *CONTROL* in terms of absolute error terms.

Result 4. Mean error terms do not differ from zero in either treatment.

Result 5. Mean absolute error terms significantly differ from zero in each treatment. However, mean absolute error terms observed in *BENCHMARK* and *COORDINATION* do not differ from mean absolute error terms in *CONTROL*.

Figure 3.3. Distribution of Individual Error Terms (all Items)



Notes: Error terms are percentage points stated in stage 1, minus the percentage points stated in stage 2 for identical items. The graph indicates the relative frequency of each possible value of error terms. The data of the graph comprises all 12 items of treatments.

Table 3.3. Error Terms and Absolute Error Terms (all Items)

	Error Terms ε_{ij}			**Absolute Error Terms $\lvert\varepsilon_{ij}\rvert$**		
	Mode	Median	Mean	Mode	Median	Mean
Benchmark (*n* = 720)	0	0	0.25	0	2	9.57
Coordination (*n* = 720)	0	0	-0.86	0	3	9.03
Control (*n* = 480)	0	0	-2.21	0	2	8.45

Notes: Numbers are percentage points. The data of the table comprises all 12 items of treatments. In Appendix A3.1, we report mean error terms and mean absolute error terms on the item level.

3.7.5. External Validity

Figure 3.4 (Appendix A3.2) indicates a high external validity of all mechanisms, as subjects are accurately assessing the patterns of proposer choices and responder acceptance rates in each treatment. To evaluate external validity, mean Brier scores (Brier, 1950), i.e., average squared deviations between factual data and elicited beliefs, are calculated and reported in Table 3.4. Before the correction procedure, Brier scores of item C6 differ between *BENCHMARK* and *BELIEF* (p < 0.05), and the same item differs between *COORDINATION* and *BELIEF* (p < 0.01). None of these differences survive the correction procedure. That is, external validity in *BENCHMARK* and *COORDINATION* does not differ from the degree of external validity that is obtained when first-order beliefs are elicited in an ordinary manner.

Table 3.4. Mean Brier Scores

	C1	C2	C3	C4	C5	C6	A1	A2	A3	A4	A5	A6
Belief	0,05	0,06	0,04	0,06	0,01	0,00	0,02	0,10	0,12	0,05	0,05	0,06
Benchmark	0,06	0,06	0,06	0,10	0,02	0,02*	0,03	0,08	0,08	0,05	0,04	0,06
Coordination	0,04	0,05	0,04	0,09	0,02	0,01**	0,02	0,09	0,10	0,04	0,05	0,04

Notes: The table contains mean Brier scores. Items C1-C6 refer to choice-probabilities of proposers and items A1-A6 refer to acceptance-probabilities of responders. Lower scores represent higher levels of accuracy. *, ** indicates significance at the 5%, and 1% level compared to the respective score in treatment *BELIEF*.

3.7.6. Discussion of Results and Evaluation of Hypotheses

We cannot reject the hypotheses 1A and 1B that the average outcomes in *Benchmark* and *Coordination* correspond to average first-order beliefs. This result holds both in a between-subject analysis and in a within-subject analysis. In accordance, mean error terms do not differ from zero in either treatment. The gathered evidence therefore supports the idea that both methods are effective in revealing mean beliefs on the population level. The correspondence of averages between mean second-order beliefs elicited in *BENCHMARK* and coordination choices elicited in *COORDINATION* implies that the mechanisms yield an unbiased measure about a subject´s first-order belief. Still, when comparing individual choices made in *BENCHMARK* and *COORDINATION* with first-order beliefs on the individual level (i.e., mean absolute error terms), we find them to be significantly larger than zero.

However, two considerations put the results on absolute error terms into perspective. First, the mean of absolute error terms is significantly larger than the median of absolute error terms in both treatments (see Table 3.3). Almost half of the estimations extracted in *BENCHMARK* and *COORDINATION* are identical with first-order beliefs, and also the median differences indicate negligible deviations between statements extracted in the two mechanisms and actual first-order beliefs. In the given setting, the median difference is more informative, since the mean is strongly affected by few subjects that enter strongly diverging numbers in the two stages (thereby strongly increasing the mean of absolute error terms). Second, as seen in treatment *CONTROL*, the degree of noise inherent in the setting equals the extent of error terms in *BENCHMARK* and *COORDINATION*. The deviations on the individual level thus seem to be driven by subjects being ambiguous about their actual first-order belief, thus creating noise.

Altogether, the observed differences between stage 1 and stage 2 in *BENCHMARK* and *COORDINATION* are not distinguishable from treatment *CONTROL*. Also, the correlation in *CONTROL* between the two stages is not higher than in *BENCHMARK* and *COORDINATION*. We therefore cannot reject hypotheses 2A and 2B that the statements extracted in the two mechanisms correspond to first-order beliefs on the individual level.

3.8. Advantages Compared to Bayesian Revelation Mechanisms

Compared to Bayesian Revelation Mechanism, we see three main advantages of *Benchmark* and *Coordination*. First, they require fewer behavioral assumptions. Precisely, it is sufficient to assume that a subject´s perceptions about others are correlated with her own type. Second, the scoring systems of both mechanisms are less complicated. This makes it easier for participants to understand the scoring system and, therefore, it simplifies a tractable and transparent implementation for experimenters. Third, it is easier for subjects to understand their "challenge" in the game, i.e., to comprehend the task necessary to maximize earnings. Subjects learn that their specific challenge is to foresee a particular outcome (either a statement by another person or a coordination outcome). Therefore, respondents know that their payment is conditioned on that particular value and that their payments monotonically increase with the precision of their guess about that specific value. This makes the task tangible for respondents.

By contrast, in empirical applications of Bayesian revelation mechanisms, subjects often do not learn how the calculation of their score, and thus their payoff, exactly look like. If subjects lack comprehension of the underlying mechanisms, it is plausible that subjects deviate from their true thought if they believe that they might "know better" how the profit-maximizing statement looks like. This might lead subjects to engage in an attempt to game the mechanism, which is problematic because it is not observable by the experimenter and thus cannot be controlled for. Likewise, even if participants fully understand the mechanisms and are aware of the Bayesian Nash equilibrium inherent in the setting, it is unclear whether they trust in other subjects to play Bayesian Nash, too. Obviously, it is rational to play according to the Bayesian Nash equilibrium only if one is confident that the remaining players also play according to that concept. Therefore, as in weakest-link games, lack of trust regarding Bayesian Nash play of other subjects might refrain a subject from playing Bayesian Nash herself (Knez and Camerer, 1994).

One further advantage is that the mechanisms, especially *Benchmark*, might be suited to elicit questions about shameful traits. In Bayesian Revelation Mechanisms, subjects are usually directly asked about their *own* type. Therefore, submitting shameful answers comes at a cost when *admitting* one´s own (shameful) type, either to oneself or to the experimenter. By contrast, this is avoided, when subjects are asked about potentially shameful traits of *others* (as is done in *Benchmark*).

The fact that the proposed methods are more tractable and transparent comes at the cost that truth-telling is not a theoretical necessity. By contrast, this is the case in Bayesian revelation mechanisms, given that all assumptions hold. Therefore, Bayesian revelation mechanisms potentially yield more accurate information if the subjects´ behavior adheres to all necessary assumptions.

3.9. Summary and Conclusion

We propose two tractable methods to incentivize the elicitation of unverifiable private information: *Benchmark* and *Coordination*. In both mechanisms, participants are incentivized to reveal their perception about others, and these statements are then used to predict the subjects´ own thoughts. The stronger the relationship between a subject´s type and her perception about others, the more effective the mechanisms are in revealing the subject´s private information.

The main advantage of the two methods is that scoring and payout functions are simple to understand, such that participants are provided with a clear task that they have to solve. This makes the mechanisms easy to implement for experimenters. The methods thus provide simple alternatives to Bayesian revelation mechanisms, when an experimenter is interested in eliciting non-verifiable, private information from subjects.

In an experiment, we mimic the elicitation of beliefs about unverifiable probabilities. In a between-subject comparison, we find that both mechanisms accurately reveal mean first-order beliefs of the population. In a within-subject comparison, we find that the modal difference between probabilities elicited in either mechanism and actual beliefs is zero. We therefore conclude that subjects strongly anchor their statements in *Benchmark* and *Coordination* on their first-order beliefs.

The paper also contributes to the literature on the elicitation of social norms using coordination games, initiated by Krupka and Weber (2013). Our results suggest that the two methods *Benchmark* and *Coordination* yield identical results, which indicates that incentivized elicitation of social norms using coordination games is also feasible through the elicitation of second-order beliefs. As a result, it allows eliciting such data without the necessity to establish an infrastructure for coordination. This simplifies data collection in contexts other than laboratory experiments, for example in (online) polls with laypeople, while still maintaining the feature of incentivization.

Appendix 3

A3.1. Error Terms and Absolute Error Terms on the Item Level

A3.1.1. Mean Error Terms

Panel A of Table 3.5 shows means of error terms ε_{ij} on the item level. The average error term $\bar{\varepsilon_j}$ of item j is calculated as $\bar{\varepsilon_j} = (\sum_{i=1}^{n} \varepsilon_{ij})/n$. Two-sided t-tests are conducted to test if mean error terms on the item level differ from zero. We do not find that error terms in any item differs from zero neither in *BENCHMARK* or *COORDINATION* nor in *CONTROL*.

A3.1.2. Mean Absolute Error Terms

Panel B of Table 3.5 shows means of absolute error terms $|\varepsilon_{ij}|$ on the item level. The average absolute error term $\overline{|\varepsilon_j|}$of item j is calculated as $\overline{|\varepsilon_j|} = (\sum_{i=1}^{n} |\varepsilon_{ij}|\ )/n$. Two-sided t-tests are conducted to test if mean absolute error terms on the item level differ from zero. We find that in all three treatments, in most items mean absolute error terms are significantly different from zero on the 5%-level. Precisely, in *BENCHMARK* and *CONTROL*, in 11 of the 12 items mean absolute error terms differ from zero with $p < 0.05$; in *COORDINATION*, in 10 items mean absolute error terms differ from zero with $p < 0.05$. Mann-Whitney-U tests are conducted to test for differences between treatments. Before the correction procedure, item C6 differs between *BENCHMARK* and *CONTROL* ($p < 0.05$) and items C3 ($p < 0.1$), C5 ($p < 0.01$) and A1 ($p < 0.05$) differ between *COORDINATION* and *CONTROL*. None of these differences survives the correction procedure.

Table 3.5. Analysis of Error Terms on the Item Level

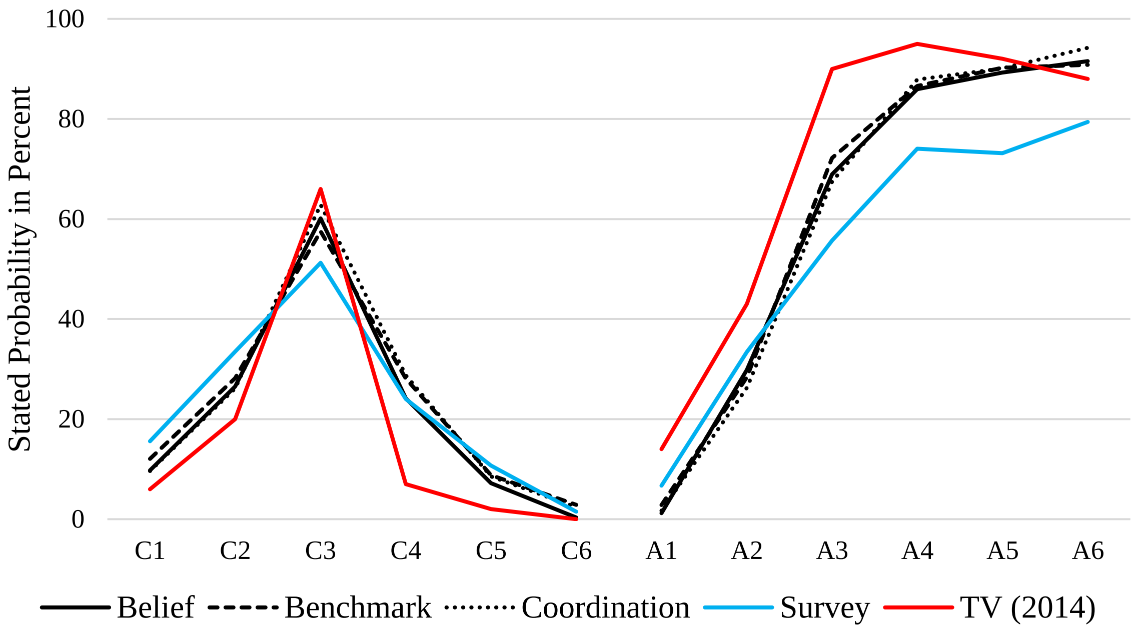
Panel A. Mean Error Terms $\overline{\varepsilon_j}$												
	C1	C2	C3	C4	C5	C6	A1	A2	A3	A4	A5	A6
Benchmark	-2,5	-2,7	0,0	2,4	-0,9	-1,7	1,4	0,3	3,0	3,3	2,5	-1,9
Coordination	0,5	-3,6	0,5	1,0	1,2	-0,2	-5,4	-1,2	-2,6	-1,0	-2,0	2,3
Control	-1,6	-0,6	-1,5	1,4	0,1	-0,2	-0,2	-4,5	-3,1	-3,3	-5,2	-7,9
Panel B. Mean Absolute Error Terms $\lvert\varepsilon j\rvert$												
	C1	C2	C3	C4	C5	C6	A1	A2	A3	A4	A5	A6
Benchmark	10,0	12,8	10,8	9,0	6,6	3,7	1,8	13,3	14,6	12,6	9,7	10,0
Coordination	7,9	15,0	12,7	15,5	8,1	3,7	6,1	12,0	11,0	8,1	5,3	3,0
Control	5,4	13,8	9,7	8,2	3,6	0,5	0,5	10,6	12,6	11,4	11,1	13,9

Notes: Numbers are percentage points. Error terms are defined as the difference between statements in stage 1 and stage 2 in treatment *BENCHMARK*, *COORDINATION*, and *CONTROL*. Absolute error terms are the absolute values of error terms. Panel A and Panel B report the means of these two measures on the item level.

A3.2. Results of Treatment Survey

Graphical analysis (figure 3.4) as well as mean Brier scores (Table 3.6) indicate a lower external validity of *SURVEY*, compared to *BELIEF*, *BENCHMARK*, and *COORDINATION*. The number of participants, however, is not sufficient to draw statistical inferences on that question.

Figure 3.4. Extracted Beliefs and Factual Data of TK (2014)



Notes: Numbers are percentage points. C1-C6 refer to probabilities for choice behavior of proposers and A1-A6 refer to probabilities for acceptance behavior of responders. The numbers in *BENCHMARK* and *COORDINATION* are elicited in the first stage of the treatments, i.e., using the respective mechanisms.

Table 3.6. Mean Brier Scores

	C1	C2	C3	C4	C5	C6	A1	A2	A3	A4	A5	A6
Belief	0.05	0.06	0.04	0.06	0.01	0.00	0.02	0.10	0.12	0.05	0.05	0.06
Benchmark	0.06	0.06	0.06	0.10	0.02	0.02	0.03	0.08	0.08	0.05	0.04	0.06
Coordination	0.04	0.05	0.04	0.09	0.02	0.01	0.02	0.09	0.10	0.04	0.05	0.04
Survey	0.05	0.06	0.06	0.09	0.04	0.00	0.05	0.10	0.20	0.13	0.17	0.13

Notes: The table contains mean Brier scores on the item level. Items C1-C6 refer to probabilities for choice behavior of proposers and A1-A6 refer to probabilities for acceptance behavior of responders. Lower scores represent higher levels of accuracy.

Chapter 4

Do Injunctive or Descriptive Social Norms Elicited Using Coordination Games Better Explain Social Preferences?

Robert J. Schmidt

Abstract: We experimentally study the relationship between social norms and social preferences on the individual level. Subjects coordinate on injunctive and descriptive norms, and we test which type of norm is more strongly related to behavior in a series of dictator games. Our experiment yields three insights. First, both injunctive and descriptive norms explain dictator behavior and recipients´ guesses, but perceptions about descriptive social norms are behaviorally more relevant. Second, our findings corroborate that coordination games are a valid tool to elicit social norm perception on the subject level, as the individuals´ coordination choices are good predictors for their actual behavior. Third, average descriptive norms on the population level accurately predict behavior on the population level. This suggests that the elicitation of descriptive social norms using coordination games is a potentially powerful tool to predict behavior in settings that are otherwise difficult to explore.

Highlights:

- The relationship between social norms and social preferences is examined
- Both injunctive and descriptive norms explain revealed social preferences
- Descriptive social norms are more strongly related to social preferences

Acknowledgments: I thank Fuat Ecer, Christian König, Franziska Lembcke, Gerhard Minnameier, Hannes Rau, Christiane Schwieren, Bertil Tungodden, Martin Vollmann as well as seminar audiences in Frankfurt and the Friedrich-Ebert-Foundation in Bonn for valuable comments and suggestions.

4.1. Introduction

Perceptions about social norms influence how individuals interpret social contexts, and they affect both intentions and behavior.[55] Traditionally, the study of social norms has received less attention in economics than in other fields of social sciences, such as sociology (Coleman, 1990; Merton, 1957) or psychology (Cialdini et al., 1990; Sherif, 1936). During the last decades, however, social norms became a vital topic of research in economics (e.g., Elster, 1989; Ostrom, 2000). By now, it is no longer disputed that social norm perception influences economic decisions, for example saving rates (Cole et al., 1992), consumer behavior (Bagozzi and Warshaw, 1990), financial reporting (Dyreng et al., 2012), job search (Stutzer and Lalive, 2004), or energy consumption (Allcott, 2011), to name just a few.[56]

Cialdini et al. (1990, 1991) argue that it is essential to differentiate between *injunctive* and *descriptive* social norms. Injunctive norms indicate perceptions about normatively appropriate behavior in a specific context. They reflect what kind of behavior is approved or disapproved by the community and thereby motivate actions through the anticipation of social rewards or punishment. By contrast, descriptive social norms refer to prevalent or common behavior, and they reflect perceptions about the likelihood that others engage in the normative behavior themselves.[57] Experimental studies find that both types of norms explain behavior, but also that the two norms are conceptually different constructs that independently affect intentions and behavior (e.g., Kallgren et al., 2000; Reno et al., 1993; Rivis and Sheeran, 2003).[58]

There is also research on which type of norm has more explanatory power for actual behavior. Some studies argue that injunctive social norms are more influential because they refer to broader underlying principles. Therefore, they motivate behavior across a spectrum of situations, while descriptive social norms are to a stronger degree context-dependent (e.g., Cialdini et al., 2006; Manning, 2009; Reno et al., 1993). It is also argued that descriptive norms are associated with a boomerang effect (Cialdini, 2003), i.e., that salient descriptive social

[55] We refer to social norms as shared perceptions about behavior. As Crawford and Ostrom (1995) formulate, this might be shared understandings about actions that are obligatory, permitted, or forbidden.

[56] As a result of that, the relevance of social norms is also often explicitly considered in economic models of human behavior (e.g., Bénabou and Tirole, 2006; Bolton and Ockenfels, 2000; Fehr and Schmidt, 1999; Rabin, 1993).

[57] Cialdini et al. (1990) summarizes injunctive norms as “norms of ought" and descriptive norms as “norms of is”.

[58] A large part of studies on the behavioral relevance of social norms is dedicated to pro-environmental behavior (e.g., Cialdini et al., 2006; Göckeritz et al., 2010; Goldstein et al., 2008; Nolan et al., 2008; Schultz, 1999; Schultz et al., 2008; Smith et al., 2012) and health behavior (e.g., Borsari and Carey, 2003; Elek et al., 2006; Larimer et al., 2004; Lee et al., 2007; Neighbors et al., 2008; Walker et al., 2011). Typically, behavior rates are highest when injunctive and descriptive norms are aligned.

norms increase, rather than decrease, problematic behaviors. Therefore, it has been hypothesized that the manipulation of injunctive social norms is a more powerful intervention to affect behavior (e.g., Blanton et al., 2008).

However, there is also ample evidence that the manipulation of descriptive social norms, through the provision of information about peers, affects behavior. Changing descriptive norms can be powerful because of preferences for conformity (Asch, 1956). Also, the provision of information about descriptive norms is potentially effective when subjects tend to overestimate the prevalence of problematic behaviors (e.g., Baer and Carney, 1993; Baer et al., 1991; Carey et al., 2006). Indeed, both lab and field experiments show that the provision of information about peers significantly affects behavior in the desired direction (e.g., Frey and Meier, 2004; Gerber and Rogers, 2009; Goeschl et al., 2018; Mair and Bergin-Seers, 2010; Reese et al., 2014).

The theoretical and empirical evidence on the competing relevance of injunctive and descriptive social norms is inconclusive. One problem with the mentioned evidence is that most studies examine aggregate effects of the provision of information or the manipulation of social norms. This approach helps to understand the behavioral effect of interventions, which in turn sheds light on the competing relevance of different types of norms. However, the approach to examine aggregate or treatment effects does only indirectly explain the association between the perception of a *specific norm* and a *specific action* on the *individual level*.

One important study in that context is Bicchieri and Xiao (2009). They consider Bicchieri (2006), who differentiates two types of *expectations*, that are conceptually related to injunctive and descriptive norms. *Normative* expectations refer to what an individual believes others think she ought to do and *empirical* expectations refer to what an individual expects others to do.[59] Bicchieri and Xiao (2009) conduct a series of treatments and exogenously manipulate dictators' expectations in the direction of either selfishness or fairness. They find that when normative and empirical expectations conflict, empirical expectations significantly predict a dictator's own choice, while normative expectations do not have a significant impact on dictator behavior after controlling for empirical expectations.

We contribute to this literature by examining under controlled conditions, whether injunctive or descriptive social norms elicited using coordination games are more strongly

[59] Note that, although expectations and social norms are closely related, they are not identical. Instead, according to Bicchieri (2006), normative and empirical expectations are a building block for social norms to emerge, including norms for fairness, reciprocity, or cooperation.

related to social preferences measured in a series of dictator games. In a laboratory experiment, we elicit injunctive and descriptive social norms from dictators and recipients as well as beliefs about social norms held by others.[60] That design differs from Bicchieri and Xiao (2009) in three aspects. First, instead of eliciting expectations, subjects coordinate on social norms according to the approach proposed by Krupka and Weber (2013).[61] Second, the subjects´ perceptions are not exogenously manipulated through the provision of information beforehand. Third, injunctive and descriptive social norms are elicited in a between-subject design, which allows for separately assessing and comparing their explanatory power for individual decision-making.

Another paper that we relate to is Kimbrough and Vostroknutov (2016). They also study social preferences and social norms on the individual level, by examining whether revealed preferences are driven by heterogeneous *sensitivity* to social norms. That hypothesis is motivated by the observation that differences in payoffs hardly explain behavioral shifts across seemingly similar allocation settings (List, 2007). In an experiment, they elicit individual norm-sensitivity and relate that measure to actual choices in a series of standard experimental paradigms.[62] Their results demonstrate that observed behavior is consistent with norm-dependent preferences, i.e., a preference per se to obey a social norm, independent from social preferences. They conclude that the substantial degree of behavioral variation across contexts does not represent inconsistent preferences, but a consequence of the fact that people care about norms and that norms fundamentally differ across contexts.[63] We contribute to that analysis by examining whether perceptions regarding the above-described differentiation (*injunctive* versus *descriptive* norms) better explain variations in revealed social preferences.

Finally, our paper is strongly related to the experiment conducted by Krupka and Weber (2013). They elicit injunctive social norms regarding behavior in different versions of the dictator game, and their results demonstrate that average coordination choices about injunctive norms predict behavioral changes between the different versions of the dictator game.[64] The

[60] By beliefs about social norms held by opponents, we mean that dictators (recipients) state their beliefs about social norms held by recipients (dictators).

[61] In the approach of Krupka and Weber (2013), subjects are confronted with the description of a particular behavior and they have to coordinate on appropriateness ratings. Their approach assumes that social norms are constituted through shared perceptions (Crawford and Ostrom, 1995), which thereby determine focal points in the coordination setting (Schelling, 1960; Sudgen, 1995). Consequently, subjects´ coordination choices reveal perceptions about prevailing social norms.

[62] Specifically, they examine the public goods game, trust game, dictator game, and the ultimatum game.

[63] As a result of that, social norms are considered to be a potentially powerful tool for nudging (Thaler and Sunstein, 2008). For an experimental analysis of using social norms as an instrument to affect behavior via nudging, see Bicchieri and Dimant (2019).

[64] Krupka and Weber examine four variants of the dictator game: Dana et al. (2007), Lazear et al. (2012), List (2007), and Bardsley (2008).

analysis that we conduct is therefore similar to their analysis, as we attempt to explain changes in revealed social preferences by social norm perception elicited using coordination choices, but we differ from their experiment in three aspects. First, we do not apply a between-subject design to predict *average* changes across environments. Instead, preferences and norms are measured in a within-subject design, and they are related to one another on the *individual level*.[65] Second, we do not use variations of the standard dictator game, but a series of varying mini-dictator games.[66] Third, Krupka and Weber (2013) focus on the predictive power of injunctive social norms. Our experimental setup extends that analysis to the measurement of injunctive *and* descriptive social norms.

Our results show that both injunctive and descriptive social norms are significantly related to dictator behavior and recipients´ guesses on the subject level. Likewise, beliefs about social norms of others significantly predict social preferences. Comparing the relative importance of injunctive and descriptive norms shows that descriptive norms are significantly more strongly related to social preferences on the individual level in almost all specifications. We also conduct aggregate level analysis by comparing whether average injunctive or average descriptive norms better predict average behavior on the population level. While the relationship between average injunctive social norms and average allocation behavior is loose, we observe that average descriptive social norms accurately predict average allocation behavior.

Three main insights can be drawn from these results. First, perceptions about descriptive social norms are significantly more strongly related to social preferences on the individual level, than injunctive norms. This supports the idea that changing perceptions about prevalent behavior is a more fruitful behavioral intervention than changing perceptions about appropriate behavior (e.g., Bicchieri and Xiao, 2009). Second, the paper corroborates that the Krupka and Weber (2013) approach is a valid tool to elicit social norm perception on the individual level, as the individuals´ coordination choices in both types of norms are strongly related to their actual behavior. This indicates that an individual´s coordination choice in that approach represents a good estimator for their actual perception of social norms.[67] Third, comparing the

[65] The approach has already been used to relate coordination choices to decision making on the individual (e.g., Barr et al., 2018; Burks and Krupka, 2012; Gächter et al., 2013; Krupka et al., 2016).

[66] Using a series of mini-dictator games allows us to vary distributive motives of allocation behavior (such as the degree of efficiency), while this is possible only to a smaller degree in the standard dictator game with fixed pie.

[67] Several studies explain why coordination games are suited to reveal a participant´s own perception about the question at hand (e.g., Dawes, 1989; Epley et al., 2004; Schmidt, 2019c; Vanberg, 2019). This literature shows that, in order to successfully coordinate with others, subjects use their own type, when making predictions about the type of others (Prelec, 2004). In doing so, they overestimate the degree to which others perceive the question in a similar way as they do (Ross et al., 1977). Consequently, an individual´s coordination choice is indicative for

predictive power on the aggregate level indicates that average descriptive social norms are good predictors for behavior, while injunctive norms are almost unrelated to average behavior rates. This suggests that the elicitation of descriptive social norms using coordination games potentially is a powerful approach to predict behavior in settings that are otherwise difficult to explore.

The remainder of the paper is organized as follows. In section 4.2, we present the experiment, and in section 4.3, we report the results. Section 4.4 summarizes and concludes.

4.2. Experimental Design and Procedure

4.2.1. Experimental Design

All treatments consist of three stages: an allocation stage, a norm elicitation stage, and a belief elicitation stage. The allocation stage is identical in all treatments and consists of a series of ten mini-dictator games. In the norm elicitation stage, injunctive and descriptive social norms are elicited using coordination games (Krupka and Weber, 2013). The norm elicitation stage is varied regarding the type of norm and the reference group for coordination, resulting in a 2×2 factorial design. In the belief elicitation stage, beliefs about social norms held by others are elicited. Subjects earn money in each stage and receive the earnings from one randomly drawn stage at the end of the experiment.

Allocation stage: At the beginning of the allocation stage, subjects are randomly assigned to the roles of dictator or recipient, and subsequently matched in pairs.[68] The dictator´s task is to decide in a series of ten mini-dictator games (MDG) how money is divided between herself and the recipient (see Table 4.1). The MDG are designed such that different distributive motives are varied between the two options.[69] The subjects´ earnings in that stage are determined by the dictator´s decision in one randomly drawn MDG. While the dictators make the allocation decisions, recipients state their guesses about the dictators´ allocation behavior in each of the ten MDG.[70]

her own perception about the question at hand. In an experiment on the elicitation of beliefs, Schmidt (2019c) finds that coordination choices are suited to reveal first-order beliefs about probabilities in an ultimatum game.

[68] In the instructions, the dictator is labeled as "Player A" and the recipient as "Player B". Subjects are informed that they remain in their role throughout the whole experiment.

[69] Note that the MDG 1-5 correspond to MDG 6-10 in terms of distributive motives.

[70] Recipients are asked to state their guess about the behavior of the dictator that they are matched with. In order to keep the instructions simple, the elicitation of these beliefs is unincentivized. If the recipients´ beliefs were incentivized, their payment in that stage would need to be randomly determined either by the dictators´ decisions or by the accuracy of the recipients´ beliefs.

Table 4.1. Mini-Dictator Games used in the Allocation Stage

Decision	Option 1	Option 2	Efficiency	Egalitarianism	Profit
1	7, 4	5, 5	Option 1	Option 2	Option 1
2	5, 4	4, 6	Option 2	Option 1	Option 1
3	6, 4	5, 5	-	Option 2	Option 1
4	6, 3	5, 5	Option 2	Option 2	Option 1
5	5, 5	5, 6	Option 2	Option 1	-
6	11, 0	5, 5	Option 1	Option 2	Option 1
7	5, 0	0, 10	Option 2	Option 1	Option 1
8	10, 0	5, 5	-	Option 2	Option 1
9	7, 1	5, 5	Option 2	Option 2	Option 1
10	5, 5	5, 10	Option 2	Option 1	-

Notes: The numbers represent payoffs in Euro. The first payoff refers to the dictator and the second payoff to the recipient.

Norm elicitation stage: After completing the allocation stage, subjects coordinate on social norms regarding dictator behavior in the MDG. Two aspects are varied in a 2×2 between-subject design. The first aspect that is varied is the type of norm. In treatments *INJUNCTIVE*, subjects coordinate on injunctive norms. In treatments *DESCRIPTIVE*, subjects coordinate on descriptive norms. Subjects always evaluate option 1 of an allocation decision. For injunctive social norms, subjects are asked for each MDG: "How appropriate is it to choose option 1 in the role of dictator?", and they are provided with four answer options: "very appropriate", "somewhat appropriate", "somewhat inappropriate" or "very inappropriate". For descriptive social norms, subjects are asked for each MDG: "How many dictators choose option 1 in the role of dictator?", and they are provided with four answer options: "a large majority", "a majority", "a minority", "a small minority". The subjects´ task is to choose the answer option of which they think that it will be chosen by the majority of subjects that participate in the coordination game. Subjects that manage to pick the modal answer in one randomly drawn MDG earn 10€ in that stage (and zero otherwise).

Second, the reference group for coordination is varied. In the current setting, where subjects with different roles coordinate on social norms, two variants of coordination are possible. Dictators and recipients could either *separately* coordinate, or they could *jointly* coordinate on social norms. Both variants are applied in the experiment. In the *SUBJECTIVE* treatments, dictators and recipients coordinate only with participants that have the same role as themselves in a session. In the *OBJECTIVE* treatments, dictators and recipients altogether coordinate on social norms. Table 4.2 summarizes the 2×2 factorial design of the norm elicitation stage.

Table 4.2. 2×2 Factorial Design of the Norm Elicitation Stage

<table>
<tr><th rowspan="2" colspan="2"></th><th colspan="2">Reference Group for Coordination</th></tr>
<tr><th>Subjective</th><th>Objective</th></tr>
<tr><td rowspan="2">Type of Social Norm</td><td>Injunctive</td><td>• Treatment: INJUNCTIVE_SUBJECTIVE
• Subjects are asked about the appropriate behavior of dictators
• Dictators and recipients separately coordinate on the answers</td><td>• Treatment: INJUNCTIVE_OBJECTIVE
• Subjects are asked about the appropriate behavior of dictators
• Dictators and recipients jointly coordinate on the answers</td></tr>
<tr><td>Descriptive</td><td>• Treatment: DESCRIPTIVE_SUBJECTIVE
• Subjects are asked about the most common behavior of dictators
• Dictators and recipients separately coordinate on the answers</td><td>• Treatment: DESCRIPTIVE_OBJECTIVE
• Subjects are asked about the most common behavior of dictators
• Dictators and recipients jointly coordinate on the answers</td></tr>
</table>

Belief elicitation stage: After completing the norm elicitation stage, subjects state their beliefs about the coordination outcomes of their opponents.[71] In the *SUBJECTIVE* conditions, dictators (recipients) state their belief about the coordination outcome of recipients (dictators). In the *OBJECTIVE* conditions, both dictators and recipients state their belief about the modal choice made by dictators and by recipients. That is, each subject states her belief about the modal choice entered by subjects in the role of dictator and her belief about the modal choice entered by subjects in the role of recipient. In the belief elicitation stage, subjects earn 10€ in case of a correct belief in one randomly drawn MDG (and zero otherwise).

4.2.2. Procedure

The experiment was programmed in z-Tree (Fischbacher, 2007), and recruitment was done via hroot (Bock et al., 2014) and ORSEE (Greiner, 2015). Experimental sessions were conducted at the experimental laboratories of the University of Heidelberg and the University of Frankfurt (both Germany) between June and December 2016.[72] In total, 328 subjects participated. Sessions lasted about 35 minutes and subjects earned on average 9.01€, including a show-up

[71] The modal choice of participants is considered the coordination outcome.
[72] In each treatment, one session was conducted in Frankfurt. The shares of observations collected in Heidelberg and Frankfurt is thus similar across treatments (between 21% and 29% per treatment).

fee of 4€. Mean age was 22.5 years, 56.1% were female, and 32.0% had an economics background in their studies.[73] Table 4.3 gives an overview of the treatments and the sample.[74]

Table 4.3. Number of Subjects by Treatment and Location

Treatment	Subjects in Heidelberg	Subjects in Frankfurt	Total *N* (Subjects)	Total *N* (Pairs)
INJUNCTIVE_SUBJECTIVE	66	18	84	42
INJUNCTIVE_OBJECTIVE	60	24	84	42
DESCRIPTIVE_SUBJECTIVE	58	22	80	40
DESCRIPTIVE_OBJECTIVE	58	22	80	40
	Σ = 242	Σ = 86	Σ = 328	Σ = 164

4.3. Results

4.3.1. Descriptive Results on the Aggregate Level

To get an impression about social preferences and social norms on the population level, we report average behavior in the allocation stage and the norm elicitation stage. We start by analyzing allocation behavior of dictators and corresponding guesses of recipients (n = 164 pairs of dictator and recipient). Figure 4.1 shows the share of dictators that choose option 1 in the respective allocation decision, and the share of recipients that believe that the dictator matched with them would choose option 1. Conducting Mann-Whitney-U tests, we find that items 1, 2, and 5 marginally differ between dictators and recipients ($p < 0.1$). These differences vanish after applying the correcting procedure à la Bonferroni.[75] The results indicate that recipients are well able to predict allocation behavior of dictators.[76] This suggests that the two groups have a similar prior regarding actual behavior in the given allocation setting, which implies that subjects have a common ground for the evaluation of social norms.

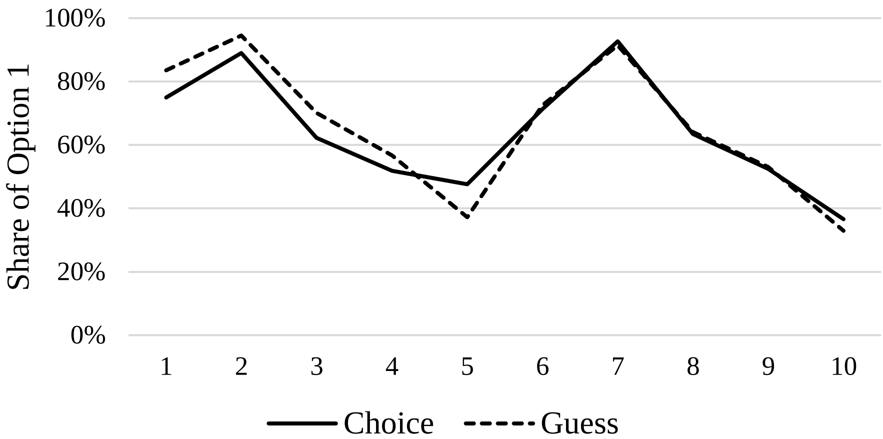
[73] Mann-Whitney-U tests indicate that the two samples (Heidelberg and Frankfurt) do not differ in terms of socio-demographics.

[74] A replication package, including instructions in German and English language, raw data, and data analysis files is available at the repository for research data of Heidelberg University: https://heidata.uni-heidelberg.de.

[75] We account for the fact that multiple items are used to detect differences between dictators and recipients in that test. In order to take care of the inflation of the overall type-I-error rate, we therefore multiply the p-values by the number of items (i.e., by ten).

[76] This indicates that the lack of incentivization of recipients in the allocation stage was not a problem for properly extracting recipients´ beliefs.

Figure 4.1. Allocation Behavior and Guesses



Notes: The figure shows the percentage of dictators choosing option 1 in the mini-dictator games, as well as corresponding guesses from recipients. Recipients are asked to guess the behavior of the dictator that they are matched with.

To shed light on the predictive power of elicited norms on the aggregate level, and to compare injunctive and descriptive norms in that regard, we conduct simple descriptive analyses.[77] Figure 4.2 shows the average results from the allocation stage and the norm elicitation stage of the four treatments. To graphically depict norms, these are quantified such that the resulting scores are normalized between -1 and 1.[78] The more positive (negative) the score for injunctive norms, the more appropriate (inappropriate) it is considered to choose option 1 in the respective decision. The more positive (negative) the score for descriptive norms, the more (less) common choosing option 1 is considered in the respective decision. Dictator choices and recipient guesses depicted in Figure 4.1 are adapted to that scale.[79]

As can be observed in Figure 4.2, in all panels the blue lines (average injunctive norm) are rather loosely related to the black line (average choice/guess), while the red lines (average descriptive norms) are remarkably similar to the black lines. In that simple graphical analysis, we thus observe that averages of descriptive norms much better capture the pattern of allocation behavior. This applies independently from the reference group for coordination (*SUBJECTIVE* vs. *OBJECTIVE*), and it applies both for dictators and recipients.

[77] Note that the comparison of social preferences and social norms on the aggregate level is possible only in a descriptive manner, since the scales used to measure social norms are verbal. This makes it difficult to compare them to behavior rates. Still, the *direction* in which averages of elicited norms vary when actual behavior varies is a sensible comparison in terms of predictive power on the aggregate level.

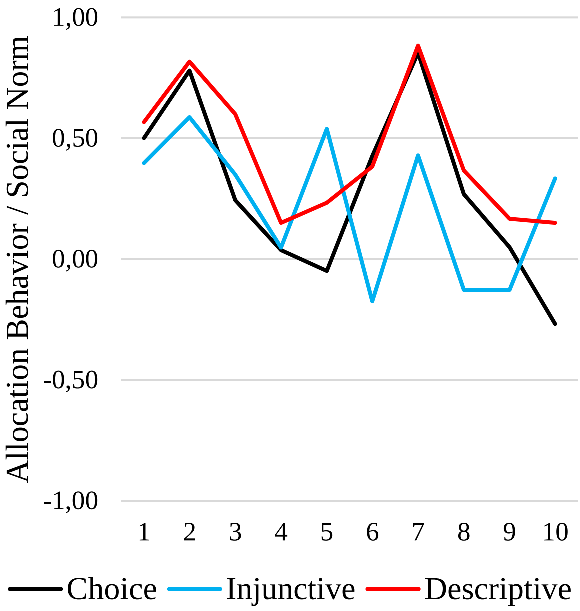
[78] Coordination choices are quantified as follows. For injunctive norms: 1 = ”very appropriate”, 1/3 = ”somewhat appropriate”, -1/3 = “somewhat inappropriate”, -1 = “very inappropriate”. For descriptive norms: 1 = ”a large majority”, 1/3 = ”a majority”, -1/3 = “a minority”, -1 = “a small minority”. Note that subjects always evaluate the choice of option 1.

[79] For that sake, option 2 is coded with the value -1 instead of 0 (as in Figure 4.1).

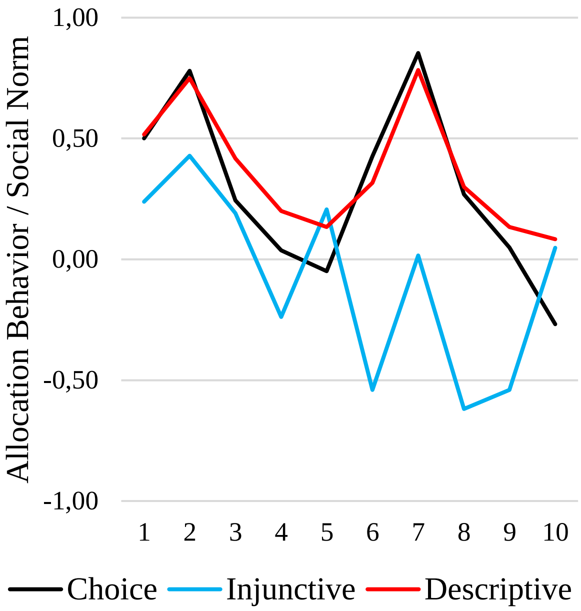
Result 1. Graphical analysis indicates that descriptive social norms better predict average behavior on the population level than injunctive norms.

Figure 4.2. Averages of Allocation Behavior and Social Norms

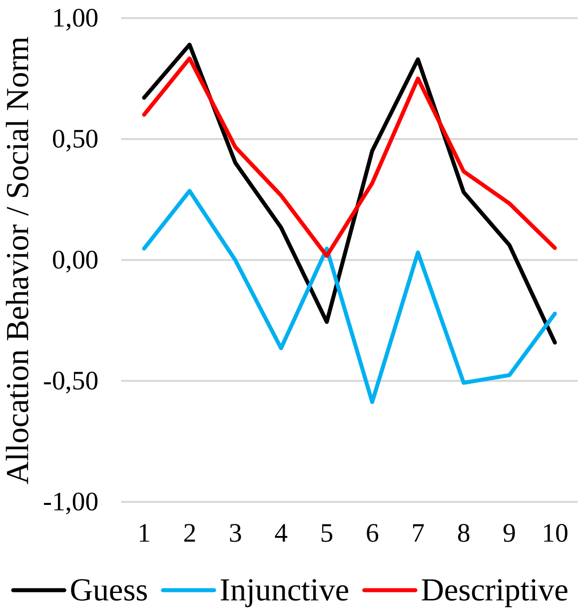
Panel A. Subjective Dictator Norms



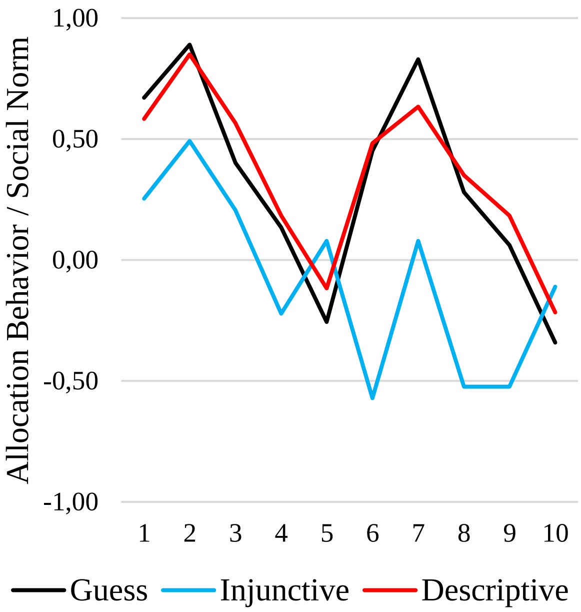
Panel B. Objective Dictator Norms



Panel C. Subjective Recipient Norms



Panel D. Objective Recipient Norms



Notes: "Choice" indicates allocation behavior of dictators, and "Guess" indicates the recipients´ guesses about dictator behavior. In "Choice" and "Guess", option 1 is coded as "1", and option 2 is coded as "-1".

4.3.2. Individual Level Analysis

We proceed by analyzing the relationship between social norm perception and social preferences on the individual level.[80] For that sake, we regress the choices from the allocation stage on the choices made in the norm elicitation stage.[81] In Table 4.4, we analyze dictator choices, and in Table 4.5, we analyze recipient guesses about dictator choices. Panels A of these tables refer to elicited injunctive norms, and panels B refer to elicited descriptive norms. Regression analyses are conducted with (i) a Probit-model and (ii) an OLS-model.[82]

We find that, in each specification, the regressor that refers to elicited norms ("Injunctive Norm" in panels A and "Descriptive Norm" in panels B) is statistically significant, independent from the regression model. This holds in either treatment condition *SUBJECTIVE* and *OBJECTIVE*, and it holds for both types of social norms. We interpret this as evidence that injunctive and descriptive norms elicited in the norm elicitation stage are related to social preferences measured in the allocation stage.

Result 2. Both injunctive and descriptive social norms are statistically significantly related to dictator behavior and recipients´ guesses about dictator behavior.

We proceed by comparing whether injunctive or descriptive norms are more strongly related to choices in the allocation stage. For that sake, one needs to column-wise compare the regressions contained in Table 4.4 and Table 4.5. We find that the size of the *p*-values of the relationship between descriptive norms and allocation behavior (contained in panels B) is smaller in all specifications than the corresponding *p*-values for injunctive norms (contained in panels A). This holds for all specifications that refer to dictators (Table 4.4) and to all specifications that refer to recipients (Table 4.5). This indicates that social norms elicited in the *DESCRIPTIVE* treatments are more strongly related to social preferences than social norms elicited in the *INJUNCTIVE* treatments. In order to test whether these differences are statistically meaningful, we conduct regression analyses with interaction terms. We first pool the observations from the conditions *INJUNCTIVE* and *DESCRIPTIVE.* Then, we perform the same analysis, i.e., we regress allocation behavior on norm perception, but we add an interaction

[80] For simplicity, we refer to "social preferences" as choices made in the allocation stage, i.e., actual dictator choices as well as recipients´ guesses about dictator choices.

[81] We code the decisions made in the allocation stage by a dummy variable which takes a value of "1" if a dictator chooses option 1 in a MDG (and "0" for option 2). Respectively, the dummy indicates that a recipient´s guess in a MDG is that the dictator chooses option 1. In the norm elicitation stage, the evaluation of injunctive and descriptive norms is coded as in the analyses on the aggregate level in section 4.3.1.

[82] We employ a Probit-model in order to account for the binary nature of the dependent variable. The OLS-regressions serve as robustness checks.

term between (i) the variable that indicates norm perception and (ii) a dummy that indicates whether that norm was elicited in the *INJUNCTIVE* or the *DESCRIPTIVE* condition of the respective treatment.[83] The interaction term yields a significance test about whether the relationship between the norm choice and the choice made in the allocation stage is statistically significantly different between injunctive and descriptive norms.

As can be seen in panels C of Table 4.4 and Table 4.5, the interaction term is positive and significant in all specifications, i.e., both for dictator behavior and recipient guesses about dictator behavior (again independent from the regression model). We interpret this as evidence for descriptive social norms being more strongly related to behavior in the allocation stage, than injunctive norms.

Result 3. Descriptive norms are statistically significantly more strongly related to dictator behavior and recipients´ guesses about dictator behavior than injunctive norms.

[83] The dummy takes a value of 0, if the norm is injunctive, and a value of 1, if the norm is descriptive.

Table 4.4. Social Norms and Dictator Choices

Panel A. Injunctive Social Norms

	Treatment *SUBJECTIVE*		Treatment *OBJECTIVE*	
	Dictator Choice	Dictator Choice	Dictator Choice	Dictator Choice
Injunctive Norm	0.488*** (0.162)	0.169*** (0.056)	0.418*** (0.135)	0.136*** (0.044)
Constant	1.336* (0.811)	0.944*** (0.272)	-0.330 (0.686)	0.414* (0.232)
Model	Probit	OLS	Probit	OLS
# Obs.	420	420	420	420

Panel B. Descriptive Social Norms

	Treatment *SUBJECTIVE*		Treatment *OBJECTIVE*	
	Dictator Choice	Dictator Choice	Dictator Choice	Dictator Choice
Descriptive Norm	1.042*** (0.180)	0.367*** (0.051)	1.240*** (0.157)	0.401*** (0.041)
Constant	1.092 (0.913)	0.854*** (0.292)	-1.168 (0.800)	0.188 (0.225)
Model	Probit	OLS	Probit	OLS
# Obs.	400	400	400	400

Panel C. Comparison of Injunctive and Descriptive Social Norms

	Treatment *SUBJECTIVE*		Treatment *OBJECTIVE*	
	Dictator Choice	Dictator Choice	Dictator Choice	Dictator Choice
Norm	0.495*** (0.163)	0.171*** (0.057)	0.412*** (0.133)	0.136*** (0.044)
Norm × Descriptive	0.521** (0.243)	0.187** (0.076)	0.820*** (0.215)	0.262*** (0.061)
Constant	1.611*** (0.594)	1.043*** (0.196)	-0.302 (0.507)	0.453*** (0.157)
Model	Probit	OLS	Probit	OLS
# Obs.	820	820	820	820

Notes: *, **, *** indicates significance at the 10%, 5%, and 1% level. Each participant yields ten observations. Standard errors are clustered on the subject level and reported in parentheses. The independent variable "Dictator Choice" is a dummy variable that indicates whether dictators choose option 1 in a mini-dictator game. In all regressions, we control for gender, age, economics study, and the location of the experimental laboratory (Heidelberg or Frankfurt). In the regressions that analyze interaction effects, we also control for the treatment condition (injunctive or descriptive). As a further robustness check, we conduct logit regressions and find the same results.

Table 4.5. Social Norms and Recipient Guesses about Dictator Choices

Panel A. Injunctive Social Norms

	Treatment *SUBJECTIVE*		Treatment *OBJECTIVE*	
	Recipient Guess	Recipient Guess	Recipient Guess	Recipient Guess
Injunctive Norm	0.641*** (0.131)	0.211*** (0.043)	0.625*** (0.140)	0.206*** (0.042)
Constant	0.722 (1.059)	0.770** (0.371)	-1.352 (0.949)	0.081 (0.277)
Model	Probit	OLS	Probit	OLS
# Obs.	420	420	420	420

Panel B. Descriptive Social Norms

	Treatment *SUBJECTIVE*		Treatment *OBJECTIVE*	
	Recipient Guess	Recipient Guess	Recipient Guess	Recipient Guess
Descriptive Norm	1.366*** (0.211)	0.402*** (0.049)	1.292*** (0.167)	0.417*** (0.034)
Constant	0.341 (0.630)	0.571*** (0.174)	0.260 (0.442)	0.595*** (0.118)
Model	Probit	OLS	Probit	OLS
# Obs.	400	400	400	400

Panel C. Comparison of Injunctive and Descriptive Social Norms

	Treatment *SUBJECTIVE*		Treatment *OBJECTIVE*	
	Recipient Guess	Recipient Guess	Recipient Guess	Recipient Guess
Norm	0.573*** (0.137)	0.184*** (0.044)	0.618*** (0.133)	0.206*** (0.040)
Norm × Descriptive	0.656*** (0.234)	0.212*** (0.064)	0.691*** (0.214)	0.215*** (0.053)
Constant	0.419 (0.604)	0.663*** (0.190)	0.211 (0.473)	0.595*** (0.139)
Model	Probit	OLS	Probit	OLS
# Obs.	820	820	820	820

Notes: *, **, *** indicates significance at the 10%, 5%, and 1% level. Each participant yields ten observations. Standard errors are clustered on the subject level and reported in parentheses. The independent variable "Recipient Guess" is a dummy variable that indicates whether recipients believe that dictators choose option 1 in a mini-dictator game. In all regressions, we control for gender, age, economics study, and the location of the experimental laboratory (Heidelberg or Frankfurt). In the regressions that analyze interaction effects, we also control for the treatment condition (injunctive or descriptive). As a further robustness check, we conduct logit regressions and find the same results.

4.3.3. The Relationship between Beliefs about Social Norms and Social Preferences

We continue by analyzing the relationship between beliefs about social norms and social preferences. We conduct the same analysis as in the previous sections, but instead of using data from the norm elicitation stage, we use the data from the belief elicitation stage. Remember that the belief elicitation stages do slightly differ between the *SUBJECTIVE* and the *OBJECTIVE* conditions. In the *SUBJECTIVE* conditions, dictators (recipients) state their beliefs about the modal choices of recipients (dictators) in the norm elicitation stage. In the *OBJECTIVE* conditions, all subjects (i.e., independent from their roles) state their beliefs about the modal choices of both dictators and recipients. In Table 4.6, we analyze beliefs from dictators, and in Table 4.7, we analyze beliefs from recipients.[84] As in the previous section, panels A of these tables refer to injunctive norms, panels B refer to descriptive norms, and panels C contain the combined data with interaction terms.

The separate analyses in panels A and panels B show the same general pattern as observed in the previous section. In most of the specifications, beliefs about injunctive and descriptive norms are significantly related to allocation behavior of dictators (Table 4.6) and to recipient guesses about allocation behavior (Table 4.7). However, the relationship between behavior in the allocation stage and beliefs about norms is less strong than the relationship between behavior in the allocation stage and actual norms.

Result 4. In most of the specifications, beliefs about injunctive and descriptive social norms are statistically significantly related to dictator behavior and recipients´ guesses about dictator behavior.

Again, we compare whether the relationship between social preferences is stronger with beliefs elicited in the *INJUNCTIVE* or the *DESCRIPTIVE* conditions in panels C of Table 4.6 and Table 4.7. Though the results are less clear than in the previous section, the general pattern is identical. Specifically, most of the interaction terms are positive, and the majority of them are statistically significant. This indicates that beliefs about descriptive social norms are more strongly related to allocation behavior than beliefs about injunctive social norms.

Result 5. In most of the specifications, beliefs about descriptive norms are statistically significantly more strongly related to dictator behavior and recipients´ guesses about dictator behavior than beliefs about injunctive norms.

[84] Table 4.6 and Table 4.7 are contained in the appendix. The analyses are fully equivalent to the analyses conducted in tables 4.4 and 4.5, except that the choices from the allocation stage are not regressed on the data from the norm elicitation stage, but on the data from the belief elicitation stage.

4.4. Summary and Conclusion

We study the relationship between social norms and social preferences in a series of dictator games. Subjects first undergo an allocation stage where dictators decide about the division of money, and recipients state their beliefs about the behavior of dictators. Subsequently, subjects evaluate allocation behavior, by coordinating on injunctive and descriptive social norms as proposed by Krupka and Weber (2013). Finally, both types of players state their beliefs about the coordination outcomes of their opponents. We find that both injunctive and descriptive norms are significantly related to dictator behavior and recipients´ beliefs about dictator behavior. Likewise, beliefs about social norms held by others significantly predict social preferences. Comparing the relative importance of injunctive and descriptive norms shows that descriptive norms are significantly more strongly related to social preferences in almost all specifications.

The paper yields three contributions. The first contribution refers to the literature on the relative importance of different types of social norms as determinants of behavior. While there is mixed evidence on whether injunctive or descriptive social norms are more related to individual decision making, our paper supports the hypothesis that the explanatory power of perceptions about descriptive social norms is behaviorally more relevant than perceptions about injunctive social norms. Apparently, the analysis of this paper does not identify causal effects of injunctive or descriptive norm perception on actual behavior. However, in line with Bicchieri and Xiao (2009), the results support the view that changing perceptions about prevalent behavior is a more fruitful behavioral intervention than changing perceptions about appropriate behavior.

The second contribution is methodological, as the paper provides a direct test on the informativeness of coordination choices à la Krupka and Weber (2013) as a measure for social norm perception on the individual level. Our results suggest that individual coordination choices are a valid tool to elicit social norm perception on the subject level, as the participants´ coordination choices are significantly related to their actual behavior. In line with previous studies, this supports the idea that predictions about others are informative about a subject´s own perception about the question at hand (Dawes, 1989; Epley et al., 2004; Ross et al., 1977; Schmidt, 2019c; Vanberg, 2019), i.e., in this case about the own perception about prevailing social norms. This enlarges the potential scope of the Krupka and Weber (2013) method, as it indicates that not only the aggregate outcome of elicited norms is suited to predict behavioral changes across contexts on the group level. Instead, a subject's coordination choice also explains behavioral changes across different contexts on the individual level.

The third contribution is again methodological. Although the experiment is designed to investigate the relationship between social preferences and social norms on the individual level, we conducted descriptive analyses on the aggregate level. For that sake, we compared average outcomes from the social preference tasks with average behavior from the tasks where subjects coordinate on injunctive and descriptive social norms. While the relationship between average injunctive social norms and average allocation behavior is rather loose, average descriptive social norms accurately predict average allocation behavior. That observation is particularly remarkable as the scale used to measure social norms is verbal, because it was not the focus of the elicitation of descriptive norms to extract *accurate* estimations about behavior rates, which could then serve as a prediction device. That result supports the idea from Krupka and Weber (2013) to use social norms elicited using coordination games as a device to predict how behavior changes across environments. In fact, our data suggest that coordination games are not only suited to make prediction about shifts in behavior but to make *point predictions* about precise behavior rates. This is particularly appealing to predict behavior in contexts that are otherwise difficult to explore. We hope that further experiments are conducted to follow up on that observation and to examine coordination games as a tool to predict behavior, both on the individual and the aggregate level.

Appendix 4

Table 4.6. Beliefs about Social Norms and Dictator Choices

Panel A. Beliefs about Injunctive Social Norms

	Treatment *SUBJECTIVE:* Beliefs about Recipient Norms		Treatment *OBJECTIVE*: Beliefs about Dictator Norms		Treatment *OBJECTIVE*: Beliefs about Recipient Norms	
	Dictator Choice	Dictator Choice	Dictator Choice	Dictator Choice	Dictator Choice	Dictator Choice
Belief Injunctive Norm	0.173 (0.165)	0.060 (0.055)	0.582*** (0.145)	0.193*** (0.046)	0.298** (0.138)	0.098** (0.044)
Constant	1.325 (0.857)	0.964*** (0.294)	-0.657 (0.643)	0.310 (0.215)	-0.310 (0.685)	0.411* (0.235)
Model	Probit	OLS	Probit	OLS	Probit	OLS
# Obs.	420	420	420	420	420	420

Panel B. Beliefs about Descriptive Social Norms

	Treatment *SUBJECTIVE:* Beliefs about Recipient Norms		Treatment *OBJECTIVE*: Beliefs about Dictator Norms		Treatment *OBJECTIVE*: Beliefs about Recipient Norms	
	Dictator Choice	Dictator Choice	Dictator Choice	Dictator Choice	Dictator Choice	Dictator Choice
Belief Descriptive Norm	0.048 (0.148)	0.019 (0.058)	1.684*** (0.161)	0.489*** (0.034)	0.103 (0.153)	0.040 (0.055)
Constant	0.777 (0.857)	0.805** (0.336)	-1.882** (0.820)	0.061 (0.180)	-1.039 (0.756)	0.111 (0.271)
Model	Probit	OLS	Probit	OLS	Probit	OLS
# Obs.	400	400	400	400	400	400

Panel C. Comparison of Beliefs about Injunctive and Descriptive Social Norms

	Treatment *SUBJECTIVE:* Beliefs about Recipient Norms		Treatment *OBJECTIVE*: Beliefs about Dictator Norms		Treatment *OBJECTIVE*: Beliefs about Recipient Norms	
	Dictator Choice	Dictator Choice	Dictator Choice	Dictator Choice	Dictator Choice	Dictator Choice
Belief Norm	0.171 (0.161)	0.059 (0.054)	0.558*** (0.137)	0.186*** (0.044)	0.298** (0.136)	0.100** (0.043)
Belief Norm × Descriptive	-0.121 (0.217)	-0.040 (0.078)	1.120*** (0.207)	0.306*** (0.055)	-0.201 (0.204)	-0.062 (0.069)
Constant	1.215** (0.601)	0.945*** (0.218)	-0.750 (0.491)	0.351** (0.136)	-0.438 (0.506)	0.342* (0.175)
Model	Probit	OLS	Probit	OLS	Probit	OLS
# Obs.	820	820	820	820	820	820

Notes: *, **, *** indicates significance at the 10%, 5%, and 1% level. Each participant yields ten observations. Standard errors are clustered on the subject level and reported in parentheses. The independent variable "Dictator Choice" is a dummy variable that indicates whether dictators choose option 1 in a mini-dictator game. In all regressions, we control for gender, age, economics study, and the location of the experimental laboratory (Heidelberg or Frankfurt). In the regressions that analyze interaction effects, we also control for the treatment condition (injunctive or descriptive). As a further robustness check, we conduct logit regressions and find the same results.

Table 4.7. Beliefs about Social Norms and Recipient Guesses

Panel A. Beliefs about Injunctive Social Norms

	Treatment *SUBJECTIVE:* Beliefs about Dictator Norms		Treatment *OBJECTIVE*: Beliefs about Dictator Norms		Treatment *OBJECTIVE*: Beliefs about Recipient Norms	
	Recipient Guess	Recipient Guess	Recipient Guess	Recipient Guess	Recipient Guess	Recipient Guess
Belief Injunctive Norm	0.356** (0.161)	0.130** (0.058)	0.680*** (0.167)	0.232*** (0.054)	0.564*** (0.168)	0.188*** (0.050)
Constant	0.156 (1.170)	0.581 (0.426)	-1.773* (0.997)	-0.052 (0.296)	-0.996 (0.936)	0.179 (0.286)
Model	Probit	OLS	Probit	OLS	Probit	OLS
# Obs.	420	420	420	420	420	420

Panel B. Beliefs about Descriptive Social Norms

	Treatment *SUBJECTIVE:* Beliefs about Dictator Norms		Treatment *OBJECTIVE*: Beliefs about Dictator Norms		Treatment *OBJECTIVE*: Beliefs about Recipient Norms	
	Recipient Guess	Recipient Guess	Recipient Guess	Recipient Guess	Recipient Guess	Recipient Guess
Belief Descriptive Norm	1.395*** (0.195)	0.416*** (0.041)	1.289*** (0.160)	0.416*** (0.034)	0.824*** (0.144)	0.292*** (0.043)
Constant	0.592 (0.679)	0.640*** (0.187)	0.532 (0.433)	0.686*** (0.133)	0.175 (0.438)	0.566*** (0.146)
Model	Probit	OLS	Probit	OLS	Probit	OLS
# Obs.	400	400	400	400	400	400

Panel C. Comparison of Beliefs about Injunctive and Descriptive Social Norms

	Treatment *SUBJECTIVE:* Beliefs about Dictator Norms		Treatment *OBJECTIVE*: Beliefs about Dictator Norms		Treatment *OBJECTIVE*: Beliefs about Recipient Norms	
	Recipient Guess	Recipient Guess	Recipient Guess	Recipient Guess	Recipient Guess	Recipient Guess
Belief Norm	0.281* (0.154)	0.106* (0.055)	0.654*** (0.167)	0.227*** (0.054)	0.581*** (0.160)	0.193*** (0.048)
Belief Norm × Descriptive	0.999*** (0.226)	0.308*** (0.064)	0.619*** (0.231)	0.185*** (0.064)	0.263 (0.213)	0.105 (0.063)
Constant	0.321 (0.654)	0.634*** (0.212)	0.330 (0.478)	0.614*** (0.158)	0.312 (0.405)	0.610*** (0.135)
Model	Probit	OLS	Probit	OLS	Probit	OLS
# Obs.	820	820	820	820	820	820

Notes: *, **, *** indicates significance at the 10%, 5%, and 1% level. Each participant yields ten observations. Standard errors are clustered on the subject level and reported in parentheses. The independent variable "Recipient Guess" is a dummy variable that indicates whether recipients believe that dictators choose option 1 in a mini-dictator game. In all regressions, we control for gender, age, economics study, and the location of the experimental laboratory (Heidelberg or Frankfurt). In the regressions that analyze interaction effects, we also control for the treatment condition (injunctive or descriptive). As a further robustness check, we conduct logit regressions and find the same results.

Chapter 5

Norms in the Lab: Inexperienced versus Experienced Participants

Robert J. Schmidt, Christiane Schwieren & Alec N. Sproten

Abstract: Using coordination games, we study whether social norm perception differs between inexperienced and experienced participants in economic laboratory experiments. We find substantial differences between the two groups, both regarding injunctive and descriptive social norms in the context of participation in lab experiments. By contrast, social norm perception for the context of daily life does not differ between the two groups. We therefore conclude that learning through experience is more important than selection effects for understanding differences between the two groups. We also conduct exploratory analyses on the relation between lab and field norms and find that behaving unsocial in an experiment is considered substantially more appropriate than in daily life. This appears inconsistent with the hypothesis that social preferences measured in lab experiments are inflated and indicates a distinction between revealed social preferences as measured commonly and the elicitation of normatively appropriate behavior.

Highlights:

- Social norm perception of inexperienced and experienced lab participants is compared
- Substantial differences are observed in lab norms, but not in field norms
- The evidence suggests that learning is more important than selection-effects

Acknowledgments: We thank seminar participants at the UPF Barcelona, the GATE Lyon Saint Etienne, the Barcelona GSE Summer Forum, the 9th International Conference of the ASFEE in Nizza, the ESA-World Meeting in Berlin, and the ESA Asia-Pacific Meeting in Abu Dhabi for very valuable comments.

5.1. Introduction

Economic research makes extensive use of laboratory experiments for studying individual behavior in a controlled environment. Since the 1980s, the share of experimental studies published in general interest journals has risen continuously (Falk and Heckman, 2009). By now, lab experiments are an important source to inform economic theory and public policy (Nikiforakis and Slonim, 2015). However, methodological limitations of lab experiments, in particular the generalizability from the lab to the field, are regularly discussed (e.g., Dana et al., 2007; Galizzi and Navarro-Martínez, 2018; Levitt and List, 2007, 2008; Zizzo, 2010).

Recently, more specific aspects of the recruitment process have been examined, such as the representativeness of registered students for the underlying student population (Abeler and Nosenzo, 2015; Cleave et al., 2013; Eckel and Grossman, 2000; Falk et al., 2013; Krawczyk, 2011; Slonim et al., 2013) or whether participants behave differently depending on the number of previous participations (Benndorf et al., 2017, Matthey and Regner, 2013).

In this study, we contribute to these literatures by examining two questions. First, we test whether social norm perception differs between inexperienced participants and experienced participants.[85] Second, by comparing differences between the two groups both for the lab and the field context, we attempt to investigate whether potential differences between the two groups are rather caused by learning through experience (when participating repeatedly) or by selection effects (through systematic differences in the probability to drop out from the pool depending on the participants´ characteristics).

To investigate these questions, we conduct a laboratory experiment and compare the two groups in a series of items that measure social norm perception. Precisely, we elicit social norm perception (i) regarding allocation behavior in the lab, (ii) regarding a series of unsocial behaviors in the lab and the field and (iii) regarding the evaluation of generalizability of behavior from the lab to the field. All questions are examined using the approach proposed by Krupka and Weber (2013) to elicit social norm perception via coordination games. In that approach, subjects are confronted with descriptions of behavior, and their task is to coordinate on appropriateness ratings. Assuming that social norms reflect shared perceptions about appropriate behaviors (Crawford and Ostrom, 1995), the coordination outcome (i.e., the modal choice) reveals social norm perception within the participants´ population. We adopt that

[85] In our study we classify the degree of experience as follows: Inexperienced subjects did not yet participate in any economic (or psychological) experiment and experienced subjects participated at least ten times in an economic laboratory experiment.

approach and sometimes adapt the set of available answers, such that the questions are suited to measure the kind of perceptions that we are interested in.[86]

A crucial assumption of the methodology applied in this experiment is that a coordination choice in the Krupka and Weber (2013) method is informative about a subject´s *own* perception about the question at hand. This view is justified if a respondent uses her own perception about a question when predicting the perception of others, in order to successfully anticipate the coordination outcome. Such behavior is implied by the false-consensus effect (Ross et al., 1977; Marks and Miller, 1987; Mullen et al., 1985), a well-documented phenomenon that follows Bayesian reasoning (Dawes, 1989, 1990; Engelman and Strobel, 2012; Schmidt, 2019c; Vanberg, 2019). Building on that assumption, we consider coordination choices as *informative signals* about the subjects´ actual *types*.[87] This allows us to draw conclusions on how the subjects themselves perceive the questions that they "answer" in the coordination games.

Our experiment yields three insights. First, social norm perception regarding behavior in the context of participation in a laboratory experiment differs significantly between inexperienced and experienced subjects. In a series of hypothetical dictator games, inexperienced subjects pronounce egalitarianism, while experienced subjects pronounce efficiency and the maximization of their earnings. Consistent with Matthey and Regner (2013), this indicates that behavior in experiments that involve allocation settings systematically depends on a subject's number of previous participations in lab experiments. Moreover, in the lab context, experienced subjects consider exploitation and deception as more appropriate than inexperienced participants, indicating that the two groups differ in experiments that involve these behaviors.

Second, by contrast to lab norms, neither field norms nor perceptions about the relation between the lab and the field differ between the two groups. This indicates that learning through experience is more important than selection effects for understanding the observed differences regarding lab norms.

Third, conducting exploratory analyses on whether norm perception in the lab corresponds to norm perception in the field, we find that norm perception between the lab and the field is correlated. However, using the same items to evaluate unsocial behaviors, once framed to the

[86] For example, in some items, we do not do ask whether a particular behavior is appropriate or not. Instead, we state that a particular behavior would be appropriate and the subjects´ task then is to coordinate on the degree of consent with that statement.

[87] For example, Schmidt (2019b) shows that injunctive and descriptive social norms elicited using coordination games are strongly related to revealed social preferences in a series of mini-dictator games on the subject level.

lab and once framed to the field context, shows that these contexts differ substantially. Specifically, behaving unsocially is considered significantly more appropriate in the lab than in the field. This appears inconsistent with the hypothesis that social preferences measured in economic experiments are inflated (e.g., Bardsley, 2008; Levitt and List, 2007; List, 2007) and indicates a distinction between revealed social preferences as measured commonly and the elicitation of normatively appropriate behavior using coordination games.

Our main conclusion is that, when conducting economic laboratory experiments, the degree of experience of participants needs to be taken care of. Since it is difficult to control explicitly for the exact number of previous participations by eliciting that information from subjects (as is done with gender or age), that characteristic needs to be properly randomized between treatments, when sessions are organized.[88] To ensure this, the recruitment bias identified by Benndorf et al. (2017), i.e., that the share of inexperienced subjects tends to be lower in early recruitment waves, needs to be considered.

The remainder of the paper is organized as follows. Section 5.2 provides a literature overview and derives hypotheses. Section 5.3 presents the experiment and section 5.4 the experimental results. Section 5.5 contains exploratory analyses on generalizability from the lab to the field and on socio-demographics. Section 5.6 summarizes and concludes.

5.2. Related Literature and Hypotheses

5.2.1. Related Literature

Our paper relates to studies which examine selection effects associated with recruitment to lab experiments. Two types of selection effects need to be distinguished in that regard: selection *into* the subject pool and selection *out of* the subject pool. Selection into the subject pool results if subjects with specific characteristics have a higher probability of entering the subject pool through registration.[89] Selection out of the subject pool results if registered subjects with

[88] Note that it is difficult to simply use the number of participations of a subject as recorded in the data base, since this would require to give up the anonymity of participants towards the experimenter. This is because it would be necessary to connect a participant's profile from the database with the data she produced in the experiment, which would require to identify which individual is sitting at which slot in the laboratory. Making that connection, however, is not in accordance with the usual policy to have subjects participate in economics experiments in a fully anonymous manner.

[89] The evidence on selection into the subject pool is mixed, although the majority of studies examining that kind of selection report null results. While Eckel and Grossman (2000) do identify differences in social preferences between registered and non-registered subjects, Abeler and Nosenzo (2015), Cleave et al. (2013), Falk et al. (2013), and Slonim et al. (2013) do not identify meaningful differences. Krawczyk (2011) examines optimal advertisement of participation in experiments and finds that recruitment is more effective when emphasizing pecuniary benefits of participation. Subjects that were recruited with advertisement of pecuniary benefits were less altruistic.

specific characteristics vary in the probability to drop-out after having participated once. In our study, we contribute to the topic of selection out of the subject pool by comparing first-time participants with those who participated many times, i.e., with the group of subjects that tends to remain in the pool.

For example, Casari et al. (2007) find that subjects are more likely to participate in a follow-up study, the more successful they were in monetary terms in a previous experiment. Similarly, Guillén and Veszteg (2012) find that earnings in previous experiments positively correlate with the probability of participating in future experiments. Thus, it has been hypothesized that more selfish subjects, which consequently earn more money in experiments, are more likely to regularly participate. As a result, it might be that common subject pools contain over-proportionally large shares of selfish individuals.

Another literature related to our study examines differences between inexperienced and experienced participants. Matthey and Regner (2013) use data about participants´ behavior in previously conducted dictator games, ultimatum games, and trust games and find that the number of participations is negatively correlated with sharing behavior in all three games. Based on post-experimental questionnaires, they conclude that repeated participation in experiments involving allocation decisions leads to learning effects through negative experiences. Benndorf et al. (2017) directly test for behavioral differences between participants with extensive lab experience and first-time participants across four one-shot two-player games (trust game, beauty contest, ultimatum game, and traveler's dilemma) and two individual decisions (lying task and risk preferences). In the trust game, experienced subjects trust less often, and they also behave significantly more selfish as second movers. In the risk elicitation tasks, experienced participants submit fewer non-monotonic strategies. The authors also document a recruitment bias as the share of inexperienced subjects was lower in early recruitment waves (i.e., in initial sessions of an experiment).

5.2.2. Hypotheses

We elicit social norm perception both for the lab and the field context. Based on the results from Benndorf et al. (2017) and Matthey and Regner (2013), which identify that behavior in the lab is related to experience, we hypothesize that this relationship is reflected in social norm perception.

Hypothesis 1. The perception of social norms for the context of lab experiments differs between inexperienced and experienced participants.

Regarding real-world norms, we again test the hypothesis that inexperienced and experienced participants differ. This hypothesis follows the idea that selection effects lead to an over-proportionally large share of selfish participants in the subject pool, as suggested by the literature on selection effects that result from drop-out (Casari et al., 2007; Guillén and Veszteg, 2012). Differences in field norms would thus indicate that selection also potentially explains differences in lab norms between inexperienced and experienced participants. By contrast, little or no differences in field norms would support the hypothesis that such differences mainly result from learning through repeated participation.

Hypothesis 2. The perception of social norms for the context of daily life differs between inexperienced and experienced participants.

5.3. Experimental Design and Procedures

5.3.1. Experimental Design

Conceptually, the experiment is divided into three parts and structured in five modules. In part 1, injunctive social norms and descriptive social norms are elicited in a series of hypothetical mini-dictator games. In part 2, injunctive social norms regarding unsocial behaviors in the lab and the field are measured. In part 3, subjects evaluate the generalizability from the lab to the field. Each module contains five items, and we use coordination games to measure social norm perception throughout the whole experiment (Krupka and Weber, 2013). In each item, subjects are asked a question, and they coordinate on one of four answer possibilities. At the end of the experiment, one of the 25 items is selected at random. If a subject´s answer in that item matches the modal choice in the current session, the subject earns 10€ (and 0€ otherwise).[90]

Modules 1 and 2: Allocation decisions in the lab. We elicit *injunctive* social norms (module 1) and *descriptive* social norms (module 2) in a series of hypothetical allocation decisions. Injunctive norms indicate perceptions about normatively appropriate behavior in a specific context. They reflect what kind of behavior is approved or disapproved by the community and thereby motivate actions through the anticipation of social rewards or

[90] We take care to make sure that subjects understand the coordination mechanism by reminding them before each item that their task is *not to state their own opinion*, but to *coordinate* with the remaining participants in the room. Only one item is paid in order to avoid hedging and ensure incentive compatibility (Azrieli et al., 2018).

punishment. In contrast, descriptive social norms refer to what kind of behavior is assumed to be common or prevalent (Cialdini et al., 1990).[91]

Five mini-dictator games are used that allow differentiating between competing distributional motives: efficiency, egalitarianism, and profit maximization. At the beginning of the modules, subjects learn the rules of the classical dictator game paradigm[92] used in economic lab experiments and they are instructed to imagine that these allocation decisions would be used in an actual lab experiment.[93] Table 5.1 shows the five hypothetical mini-dictator games and how the choices correspond to the distributional motives.

Table 5.1. Hypothetical Mini-Dictator Games used in Module 1 and 2

	Option 1		Option 2		Distributive Motives		
Game	Dictator	Recipient	Dictator	Recipient	Efficiency	Egalitarianism	Profit max.
1	15€	5€	11€	11€	Option 2	Option 2	Option 1
2	10€	10€	10€	15€	Option 2	Option 1	-
3	15€	5€	9€	9€	Option 1	Option 2	Option 1
4	10€	9€	9€	11€	Option 2	Option 1	Option 1
5	12€	8€	10€	10€	-	Option 2	Option 1

For simplicity, we always ask to evaluate option 1.[94] For injunctive norms in module 1, subjects indicate "how appropriate it would be, to choose option 1" in the role of the dictator by coordinating on these answer options: "very appropriate", "somewhat appropriate", "somewhat inappropriate" or "very inappropriate". For descriptive social norms in module 2 subjects indicate "which of the two options would be chosen more often" by coordinating on these answer options: "option 1 much more often", "option 1 somewhat more often", "option 2 somewhat more often" or "option 2 much more often".

Modules 3 and 4: Evaluation of unsocial behaviors. We study social norm perception in the lab and the field by eliciting injunctive social norms regarding a series of unsocial behaviors: selfishness, exploitation, spitefulness, deception, and willful ignorance. Subjects are confronted

[91] Cialdini et al. (1990) summarizes injunctive norms as "norms of ought" and descriptive norms as "norms of is". Experimental studies find that both types of norms explain behavior, but also that the two norms are conceptually different constructs that independently affect intentions and behavior (e.g., Cialdini and Kallgren, 1993; Kallgren et al., 2000; Rivis and Sheeran, 2003). This research also shows that behavior rates are typically highest when injunctive and descriptive norms are aligned.

[92] Each step of the classical dictator game paradigm (anonymity, randomization of roles, matching, decision rights, and payout function) is explained in detail. We use the term „Player A" for the dictator and „Player B" for the recipient.

[93] Precisely, they should imagine that the allocations would take place in an experiment like the one they are located in at that moment.

[94] Note that option 1 is always dominant for the dictator in terms of profit maximization.

with the statement that the respective behavior would be appropriate and they then indicate the degree of consent with the respective statement by coordinating on: “fully agree”, “somewhat agree”, “somewhat disagree” or “fully disagree”. We use the identical set of items and frame them once to the lab (module 3) and once to the field (module 4). Table 5.2 shows the statements we use in the two modules.

Table 5.2. Items used in Modules 3 and 4

As a participant in a laboratory experiment, it is appropriate to / In daily life, it is appropriate to
1) … mainly consider the own well-being.
2) … take advantage of other subjects, when this leads to a material advantage for oneself.
3) … harm other subjects, even when this does not lead to a material advantage for oneself.
4) … deceive other subjects, in order to materially gain from it.
5) … remain ignorant about the consequences that the own decisions have on other people.

Notes: The wording “As a participant in an experiment, it is appropriate to…” refers to module 3 and “In daily life, it is appropriate to…” refers to module 4.

Module 5: Generalizability of lab behavior. In module 5, we elicit perceptions about the generalizability of lab behavior. Again, we confront subjects with a set of statements and have them coordinate on the degree of consent. Table 5.3 contains the items used in module 5.

Table 5.3. Items used in Module 5

1) As a participant in an experiment, I have **the same** moral standards regarding my own behavior as in daily life.
2) As a participant in an experiment, I have **the same** moral standards regarding the behavior of others as in daily life.
3) Selfishness in the lab is **not the same** as selfishness in daily life.
4) Social norms in the laboratory are **not the same** as social norms in daily life.
5) My behavior as a participant in an experiment is **representative** of my behavior in daily life.

Notes: In the experiment, none of the words were printed boldly. The bold print here is to illustrate which of the statements are affirmations (suggesting similarity between the two contexts) and which of the statements are negations (suggesting dissimilarity between the two contexts).

Order of modules. To mitigate order effects, we vary the order of modules as well as the order of items within modules. Moreover, we avoid that those modules which are subject to comparison (module 1 and 2; module 3 and 4) appear consecutively, in order to reduce spillover effects between modules. Also, we elicit norms of daily life always at the end to avoid priming field context before eliciting perceptions about the lab context. We test for order effects but do not find an interaction between the different order variants and the subjects´ choices.

5.3.2. Procedures

We conducted sessions either only with inexperienced subjects (no prior participation) or only with experienced subjects (at least 10 participations). Subjects were not informed about the fact that they were recruited as a specific subpart of the participant pool. In total, we recruited 82 inexperienced subjects and 68 experienced subjects. From the 82 inexperienced participants, 9 were excluded from the analysis because they stated in a post-experimental questionnaire that they already participated in at least one economic or psychological lab experiment before. Thus, 73 inexperienced and 68 experienced subjects remained in the analysis, leaving us with a total *N* of 141 observations and fairly balanced sample sizes for the two groups. The experiment was programmed in z-Tree (Fischbacher 2007), recruitment was done via hroot (Bock et al., 2014), and the sessions were conducted at the experimental laboratory of the University of Heidelberg, Germany, between November 2016 and May 2017. A typical session lasted about 35 minutes and subjects earned on average 10.30€ including a show-up fee of 3€.[95]

5.4. Results

5.4.1. Part 1: Allocation Decisions in the Laboratory

To analyze modules 1 and 2, we quantify the answers such that the resulting scores are normalized between -1 and 1. For injunctive norms in module 1, we quantify the answers as: 1 = ”very appropriate”, 1/3 = ”somewhat appropriate”, -1/3 = “somewhat inappropriate”, -1 = “very inappropriate”. Thus, the more positive (negative) the score in module 1, the more appropriate (inappropriate) it is considered to choose option 1 in the respective decision. For descriptive norms in module 2, we quantify the answers as: 1 = “option 1 much more often”, 1/3 = “option 1 somewhat more often”, -1/3 = “option 2 somewhat more often”, -1 = “option 2 much more often”. The more positive (negative) the score in module 2, the more common choosing option 1 (option 2) is considered in the respective decision. Figure 5.1 provides descriptive analysis to give an impression about coordination outcomes.

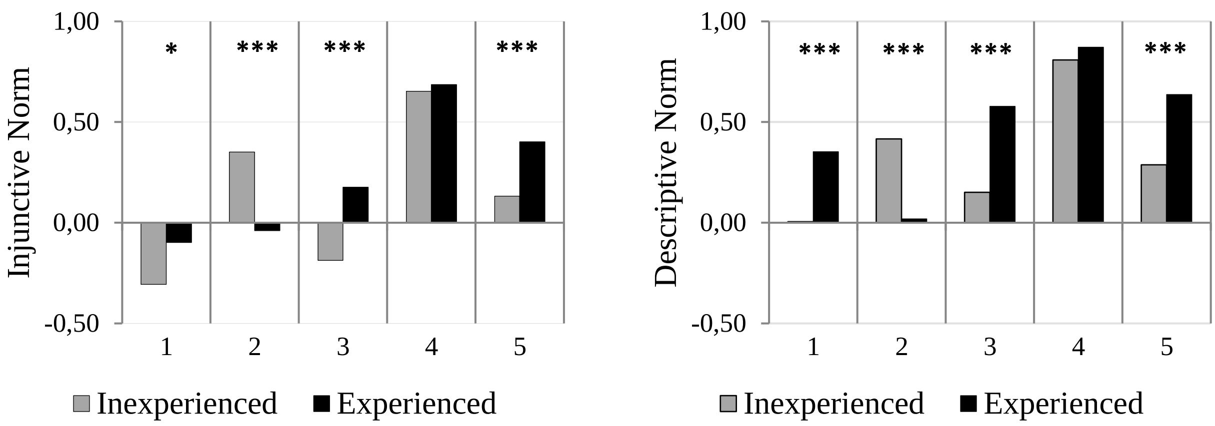
In order to draw statistical inferences on differences between inexperienced and experienced participants and to control for potential confounds, we conduct regression analyses (Table 5.4). Regression results suggest that inexperienced and experienced subjects differ both concerning injunctive norms and descriptive norms in most of the allocation decisions, with the

[95] A replication package, including instructions in German and English language, raw data, and data analysis files is available at the repository for research data of Heidelberg University: https://heidata.uni-heidelberg.de.

latter diverging more strongly. In module 1, the two groups differ in three items before the correction procedure (items 2, 3, and 5), but two of these differences vanish after applying the Bonferroni correction.[96] The results regarding injunctive norms indicate that the two groups differ in particular regarding their evaluation of the competing motives captured in item 3. In that item, a dictator chooses between advantageous efficiency and egalitarianism. While opting for advantageous efficiency is considered rather inappropriate by inexperienced subjects (score for injunctive norm of -0.19), it is evaluated as rather appropriate by experienced subjects (score for injunctive norm of 0.18). In module 2, the differences are considerably stronger than in module 1. Four items (1, 2, 3, and 5) differ between the two groups, and all of these significances survive the correction procedure.

Figure 5.1. Allocation Decisions in the Laboratory

Panel A. Injunctive Norms (Module 1) **Panel B.** Descriptive Norms (Module 2)



Notes: *, **, *** indicates significance at the 10%, 5%, and 1% without correcting for multiple testing; Mann-Whitney-U tests. Positive (negative) values on Panel A represent that it is considered to be appropriate (inappropriate) to choose option 1 in the respective mini-dictator game. Positive (negative) values on Panel B represent that option 1 (option 2) is considered to be chosen more often in the respective mini-dictator game.

[96] We account for the fact that multiple items are used within modules to detect differences between inexperienced and experienced participants. In order to take care of the inflation of the overall type-I-error rate, we therefore multiply the *p*-values by the number of items within a module (i.e., by five).

Table 5.4. Regression Analysis on Injunctive and Descriptive Social Norms

	Panel A. Injunctive Social Norms (Module 1)				
	1	2	3	4	5
Experienced	0.302 (0.194)	-0.404* (0.211)	0.608***/### (0.195)	0.123 (0.213)	0.447** (0.197)
# Obs.	141	141	141	141	141
	Panel B. Descriptive Social Norms (Module 2)				
	1	2	3	4	5
Experienced	0.555***/## (0.196)	-0.545***/## (0.203)	0.668***/### (0.202)	0.145 (0.254)	0.524***/## (0.205)
# Obs.	141	141	141	141	141

Notes: Ordered probit regressions. *, **, *** indicates levels of significance at the 10%, 5%, and 1% level before correcting for multiple testing. #,##, ### indicates levels of significance at the 10%, 5%, and 1% level after correcting for multiple testing according to the Bonferroni method. Standard errors are clustered on the individual level and reported in parentheses. Controls are gender, age, and field of study (economics or not).

We further analyze the distributive motives reflected in social norm perception. To this end, we calculate scores that reflect the relative importance of efficiency, egalitarianism, and profit maximization on the individual level and run regression analyses (Table 5.5).[97] The results show that experience is systematically related to these motives. Consistent with Benndorf et al. (2017) and Matthey and Regner (2013), inexperienced subjects pronounce egalitarianism, while experienced subjects pronounce efficiency and profit orientation. The results from modules 1 and 2 support hypothesis 1.

Result 1. Inexperienced and experienced participants differ both concerning injunctive and descriptive social norms in allocation decisions. Regarding both types of norms, egalitarianism is more pronounced in inexperienced subjects, while efficiency and profit orientation are more pronounced in experienced subjects.

[97] Precisely, a score on the individual level for a particular distributive motive (efficiency/egalitarianism/profit) is calculated as follows. We take the choice of a particular subject in a particular item. We then multiply a dummy variable that indicates if option 1 is the efficient/equal/profit maximizing option (dummy = 1) or not (dummy = -1) with the choice score of the subject (i.e., 1, 1/3, -1/3 or -1) in that item. For each distributive motive, this is done for each item. The resulting scores are then averaged within the five items of a module regarding a specific distributive motive. This procedure is done for each subject, both for injunctive and for descriptive norms. The resulting scores then reflect the importance of a particular distributive motive in module 1 and module 2 on the subject level. In Appendix A5.2 the reader finds an example of how we calculated the scores.

Table 5.5. Regression Analysis on Distributive Motives

	Panel A. Injunctive Social Norms			**Panel B.** Descriptive Social Norms		
	Efficiency	Egalitarianism	Profit maximization	Efficiency	Egalitarianism	Profit maximization
Experienced	0.081**	-0.194***	0.152***	0.075**	-0.272***	0.210***
	(0.039)	(0.063)	(0.057)	(0.034)	(0.062)	(0.064)
Constant	-0.113	0.376**	-0.095	-0.137*	0.038	0.173
	(0.092)	(0.146)	(0.133)	(0.079)	(0.146)	(0.150)
# Obs.	141	141	141	141	141	141

Notes: OLS regressions. *, **, *** indicates significance at the 10%, 5%, and 1% level without correcting for multiple testing. Standard errors are clustered on the individual level and reported in parentheses. Controls are gender, age, and field of study (economics or not). As a robustness check, we conduct Tobit regressions which yield the same results.

5.4.2. Part 2: Evaluation of Unsocial Behaviors

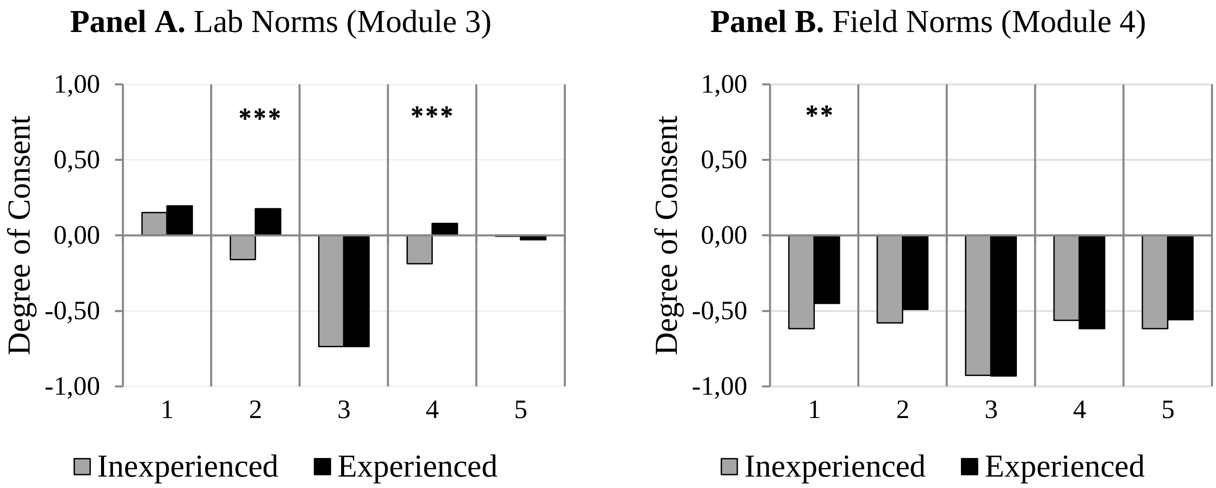
To analyze module 3 (module 4), we again first present descriptive results in Figure 5.2 and then run regression analyses on the degree of consent that the described behaviors are considered to be appropriate in the laboratory (the field).[98] Comparing perceptions about the laboratory context, we find that for experienced subjects, it is significantly more appropriate to exploit and deceive other participants within the lab context (see the regressions in Panel A of Table 5.6). The results are highly significant and robust to the correction procedure. The observed differences further support hypothesis 1.

By contrast to lab norms, real-world norms are homogenous with respect to experience (see Panel B of Table 5.6). Item 1, which refers to selfishness, is marginally significantly different between the groups, indicating that selfishness in daily life is considered more appropriate by experienced participants. This difference, however, is not robust to the correction procedure. The results thus do not support hypothesis 2.

Result 2. Lab norms differ between the two groups, as it is significantly more appropriate to exploit and deceive other participants within the lab context for experienced subjects. By contrast, real-world norms are homogenous with respect to experience.

[98] Figure 5.2 shows that behaving unsocial in the lab is considered significantly more appropriate than behaving unsocial in the context of daily life (we elaborate on that finding in section 5.5.1).

Figure 5.2. Unsocial Behaviors in the Lab and the Field



Notes: *, **, *** indicates significance at the 10%, 5%, and 1% without correcting for multiple testing; Mann-Whitney-U tests. Positive (negative) values indicate that the behavior described in an item is considered more (less) appropriate.

Table 5.6. Regression Analysis on Appropriateness of Unsocial Behaviors

	Panel A. Unsocial Behavior in the Lab (Module 3)				
	Selfishness	Exploitation	Spitefulness	Deception	Ignorance
Experienced	0.134 (0.193)	0.639***/### (0.200)	-0.019 (0.231)	0.640***/### (0.198)	-0.041 (0.193)
# Obs.	141	141	141	141	141
	Panel B. Unsocial Behavior in the Field (Module 4)				
	Selfishness	Exploitation	Spitefulness	Deception	Ignorance
Experienced	0.384* (0.204)	0.216 (0.204)	0.109 (0.327)	-0.166 (0.209)	0.222 (0.210)
# Obs.	141	141	141	141	141

Notes: Ordered probit regressions. *, **, *** indicates levels of significance at the 10%, 5%, and 1% level before correcting for multiple testing. #,##, ### indicates levels of significance at the 10%, 5%, and 1% level after correcting for multiple testing according to the Bonferroni method. Regressions in Panel A (Panel B) are ran on the degree of consent that the respective behavior is considered to be an appropriate behavior in the lab (field) context. Standard errors are clustered on the individual level and reported in parentheses. Controls are gender, age, and field of study (economics or not).

5.4.3. Part 3: Perceptions about Generalizability

In module 5, subjects are asked to evaluate the generalizability of lab behavior. Table 5.7 shows regression analyses on the degree of consent with the items used in module 5. Item 1 indicates that experienced subjects less strongly agree to the statement that, in the lab, they have the same moral standards regarding their own behavior as in daily life. That is, experienced subjects less strongly agree to a statement that suggests that the two contexts are similar. However, the difference is not robust to the correction procedure. We therefore do not find that the two groups differ in how they evaluate the generalizability of lab behavior.

Result 3. Inexperienced and experienced subjects do not differ in how they perceive the relation between behavior in the lab and behavior in the field.

Table 5.7. Regression Analysis on Perceptions about Generalizability

	Item 1 (+)	Item 2 (+)	Item 3 (-)	Item 4 (-)	Item 5 (+)
Experienced	-0.398** (0.200)	-0.022 (0.196)	-0.285 (0.198)	0.002 (0.197)	-0.191 (0.199)
# Obs.	141	141	141	141	141

Notes: Ordered probit regressions. *, **, *** indicates uncorrected levels of significance at the 10%, 5%, and 1% level. #,##, ### indicates levels of significance at the 10%, 5%, and 1% level after correcting for multiple testing according to the Bonferroni method. Regressions are ran on the degree of consent with the respective item. The (+) and (-) indicate whether the statements represent affirmations (implying similarity between the lab and field) or negations (implying dissimilarity between the lab and field). Standard errors are clustered on the individual level and reported in parentheses. Controls are gender, age, and field of study (economics or not).

5.5. Exploratory Analyses

5.5.1. Generalizability of Lab Norms

Although it is not the focus of this study, the data allows us to study generalizability by comparing modules 3 and 4 and testing whether lab and field norms correspond. As we hold the content of the items fixed and vary the framing of the context (lab vs. field), we can isolate the effect of the context. For that sake, we again quantify the answers such that the resulting scores are normalized between -1 and 1. The more positive (negative) the score, the stronger subjects agree (disagree) with the statement that a particular behavior is considered to be appropriate in the respective context.

We first conduct correlation analyses between lab norms and field norms. Table 5.8 shows that these are positive and mainly significant. We next compare the absolute values of the degree of consent between the two modules. Table 5.9 shows the same data as presented in Figure 5.2, but now the results regarding the two contexts are compared. The results indicate that perceptions about the appropriateness of the unsocial behaviors described in the items differ substantially between the two contexts, as each behavior is considered significantly *less appropriate in the field than in the lab*. This applies independently from the degree of experience. All differences remain highly statistically significant after the correction procedure. The finding stands in contrast with the hypothesis, that social preferences measured in the lab are inflated (e.g., Levitt and List 2007) and the results indicate a distinction between revealed social preferences as elicited commonly and the elicitation of normatively appropriate behavior using coordination games.

Result 3. Unsocial behavior in the lab context is considered substantially more appropriate than in the context of daily life.

Table 5.8. Correlations between Laboratory and Field Norms

	Selfishness	Exploitation	Spitefulness	Deception	Ignorance
Inexperienced	0.204*	0.302***/##	0.291**/#	0.204*	0.260**
Experienced	0.103	0.091	0.526***/###	0.268**	0.041
All subjects	0.157*	0.237***/##	0.404***/###	0.220***/##	0.157*

Notes: Spearman rank correlation. *, **, *** indicates significance at the 10%, 5%, and 1% level before correcting for multiple testing. #,##, ### indicates levels of significance at the 10%, 5%, and 1% level after correcting for multiple testing according to the Bonferroni method.

Table 5.9. Comparison Between Laboratory and Field Norms

		Selfishness	Exploitation	Spitefulness	Deception	Ignorance
Inexperienced participants	Lab context	0.15	-0.16	-0.73	-0.19	0.00
	Field context	-0.62	-0.58	-0.93	-0.56	-0.61
	Difference	-0.77***/###	-0.42***/###	-0.20***/###	-0.37***/###	-0.61***/###
Experienced participants	Lab context	0.20	0.18	-0.73	0.08	-0.03
	Field context	-0.45	-0.49	-0.93	-0.62	-0.56
	Difference	-0.65***/###	-0.67***/###	-0.20***/###	-0.70***/###	-0.53***/###

Notes: Two-sided t-tests. *, **, *** indicates significance at the 10%, 5%, and 1% level before correcting for multiple testing. #,##, ### indicates levels of significance at the 10%, 5%, and 1% level after correcting for multiple testing according to the Bonferroni method. The numbers represent the degree of consent that the behavior stated in an item is considered to be an appropriate behavior in the respective context.

5.5.2. Socio-Demographics: Age, Gender, and Field of Study

Throughout all analyses, we control for age, gender, and economic study. In Appendix A5.1, we report complete regression analyses including coefficients of control variables (age, gender, and field of study). In this section, we report those findings that are significant at the 5%-level without correction for multiple testing.

We find that all differences between inexperienced and experienced participants are fully independent of differences in age after controlling for experience. However, we identify some interesting patterns regarding gender and field of study. First, consistent with previous studies on gender effects in dictator game giving (e.g., Andreoni and Vesterlund, 2001; Eckel and

Grossman, 1998), female subjects are significantly more guided by egalitarianism and significantly less guided by efficiency in experimental allocation tasks than male subjects (see Table 5.11). However, females consider it to be more appropriate to ignore the consequences that their own decisions have on other people in a lab experiment. Second, economics students consider several unsocial behaviors (exploitation, spitefulness, and deception) in the laboratory context as more appropriate than non-economics students (see Table 5.12). These differences, however, only refer to norm perception in the laboratory context. Regarding norm perception in the field context, all sub-group differences vanish.

Result 4. Injunctive and descriptive social norms of female subjects are guided more strongly by egalitarianism and less strongly by efficiency than social norms of male subjects.

Result 5. In the lab context, economics students consider exploitation, spitefulness, and deception as more appropriate than non-economics students.

5.6. Summary and Conclusion

We compare social norm perception of inexperienced and experienced participants in economic laboratory experiments using the Krupka and Weber (2013) approach. We find that the two groups differ both concerning injunctive norms and descriptive norms in allocation decisions in the lab, with the latter diverging more strongly. Consistent with Benndorf et al. (2017) and Matthey and Regner (2013), egalitarianism is more pronounced in norm perception of inexperienced subjects, while efficiency and profit maximization dominate in experienced subjects. We complement these results with the finding that experienced subjects consider exploitation and deception of other participants in the lab as more appropriate than inexperienced subjects. The results demonstrate that not only revealed social preferences (Matthey and Regner, 2013) are related to the number of participations, but that also social norm perception, which potentially mediates the differences in behavior, is different between subjects with varying degrees of experience.

We also compare norm perception for the context of daily life and find that these do not differ between the two groups. We thus do not find support for the hypothesis that selection effects through drop-out lead to an over proportionally large share of selfish individuals in the subject pool (Casari et al. 2007, Guillén and Veszteg 2012). We therefore conclude that learning is more important than selection effects for explaining differences that are linked to experience. For a conclusive analysis of the relative importance of learning and selection, however, further research is necessary. In particular, it might be interesting to compare field behavior more comprehensively, as the scope of analyses of that context was smaller than the scope on the analysis of the lab context in our study.

Finally, we conduct exploratory analyses in order to contribute to the generalizability debate by comparing social norm perception between the lab and the field. We find that norm perception between the two contexts is correlated. However, independent from the degree of experience, behaving unsocially in the lab is considered significantly more appropriate than in the real-world. This finding stands in contrast with the hypothesis that social preferences measured in the lab are inflated (e.g., Levitt and List 2007) and indicates a distinction between revealed social preferences and the elicitation of normatively appropriate behavior using coordination games.

Our results corroborate the idea that, when conducting economic laboratory experiments, the degree of "lab experience" of participants needs to be taken care of. We therefore conclude that it is essential to make sure that this characteristic is properly randomized between treatments, and that this should be monitored in the invitation phase of the recruitment process. To ensure this, the recruitment bias identified by Benndorf et al. (2017), i.e., that the share of inexperienced subjects tends to be lower in early recruitment waves, needs to be considered.

Appendix 5

A5.1. Complete Regression Analyses reporting Coefficients of Control Variables

A5.1.1. Complete Regression Analysis on Injunctive and Descriptive Social Norms

Table 5.10. Regression Analysis on Injunctive and Descriptive Social Norms

	Panel A. Injunctive Social Norms (Module 1)				
	1	2	3	4	5
Experienced	0.302 (0.194)	-0.404* (0.211)	0.608*** (0.195)	0.123 (0.213)	0.447** (0.197)
Age	0.024 (0.018)	0.000 (0.018)	0.020 (0.018)	0.003 (0.019)	0.023 (0.019)
Female	0.141 (0.191)	0.749*** (0.208)	0.035 (0.191)	0.119 (0.211)	-0.131 (0.192)
Economics	0.223 (0.191)	-0.133 (0.207)	0.175 (0.191)	-0.274 (0.209)	0.056 (0.193)
# Obs.	141	141	141	141	141
	Panel B. Descriptive Social Norms (Module 2)				
	1	2	3	4	5
Experienced	0.555*** (0.196)	-0.545*** (0.203)	0.668*** (0.202)	0.145 (0.254)	0.524*** (0.205)
Age	-0.011 (0.017)	0.020 (0.019)	0.022 (0.020)	0.077* (0.043)	0.045* (0.027)
Female	-0.244 (0.192)	0.565*** (0.201)	-0.154 (0.197)	0.256 (0.249)	-0.208 (0.202)
Economics	-0.009 (0.191)	-0.074 (0.201)	0.136 (0.197)	-0.085 (0.248)	0.093 (0.203)
# Obs.	141	141	141	141	141

Notes: Ordered probit regressions. *, **, *** indicates significance at the 10%, 5%, and 1% level without correcting for multiple testing. Standard errors are clustered on the individual level and reported in parentheses.

A5.1.2. Complete Regression Analysis on Distributive Motives

Table 5.11. Regression Analysis on Distributive Motives

	Injunctive Social Norms			Descriptive Social Norms		
	Efficiency	Egalitarianism	Profit maximization	Efficiency	Egalitarianism	Profit maximization
Experienced	0.081**	-0.194***	0.152***	0.075**	-0.272***	0.210***
	(0.039)	(0.063)	(0.057)	(0.034)	(0.062)	(0.064)
Age	0.000	-0.007	0.006	-0.002	0.001	0.005
	(0.004)	(0.006)	(0.005)	(0.003)	(0.006)	(0.006)
Female	-0.127***	0.112*	0.010	-0.078**	0.156**	-0.058
	(0.039)	(0.062)	(0.056)	(0.034)	(0.062)	(0.064)
Economics	0.028	-0.082	0.028	0.026	-0.034	0.018
	(0.039)	(0.062)	(0.056)	(0.034)	(0.062)	(0.064)
Constant	-0.113	0.376**	-0.095	-0.137*	0.038	0.173
	(0.092)	(0.146)	(0.133)	(0.079)	(0.146)	(0.150)
# Obs.	141	141	141	141	141	141

Notes: OLS regressions. *, **, *** indicates significance at the 10%, 5%, and 1% level; standard errors are clustered on the individual level and reported in parentheses. As a robustness check, we conduct Tobit regressions which yield the same results.

A5.1.3. Complete Regression Analysis on Evaluation of Unsocial Behaviors

Table 5.12. Regression Analysis on Appropriateness of Unsocial Behaviors

	Panel A. Unsocial Behavior in the Lab (Module 3)				
	Selfishness	Exploitation	Spitefulness	Deception	Ignorance
Experienced	0.134 (0.193)	0.639*** (0.200)	-0.019 (0.231)	0.640*** (0.198)	-0.041 (0.193)
Age	-0.014 (0.017)	0.012 (0.018)	-0.001 (0.024)	-0.028 (0.019)	0.018 (0.018)
Female	-0.014 (0.191)	0.230 (0.195)	0.367 (0.230)	0.249 (0.193)	0.561*** (0.195)
Economics	0.037 (0.191)	0.453** (0.196)	0.636*** (0.221)	0.628*** (0.197)	0.154 (0.192)
# Obs.	141	141	141	141	141
	Panel B. Unsocial Behavior in the Field (Module 4)				
	Selfishness	Exploitation	Spitefulness	Deception	Ignorance
Experienced	0.384* (0.204)	0.216 (0.204)	0.109 (0.327)	-0.166 (0.209)	0.222 (0.210)
Age	-0.012 (0.019)	0.000 (0.018)	-0.008 (0.036)	-0.002 (0.019)	-0.003 (0.019)
Female	-0.248 (0.200)	-0.221 (0.203)	0.416 (0.336)	-0.082 (0.206)	0.260 (0.208)
Economics	0.175 (0.200)	-0.074 (0.203)	-0.021 (0.320)	-0.087 (0.207)	-0.066 (0.207)
# Obs.	141	141	141	141	141

Notes: Ordered probit regressions. *, **, *** indicates significance at the 10%, 5%, and 1% level without correcting for multiple testing. Standard errors are clustered on the individual level and reported in parentheses. Regressions in Panel A (Panel B) are ran on the degree of consent that the respective behavior is considered to be an appropriate behavior in the lab (field) context.

A5.1.4. Complete Regression Analysis on Perceptions about Generalizability

Table 5.13. Regression Analysis on Perceptions about Generalizability

	Agreement with Statements				
	Item 1 (+)	Item 2 (+)	Item 3 (-)	Item 4 (-)	Item 5 (+)
Experienced	-0.398** (0.200)	-0.022 (0.196)	-0.285 (0.198)	0.002 (0.197)	-0.191 (0.199)
Age	-0.003 (0.018)	-0.001 (0.018)	0.007 (0.018)	-0.001 (0.018)	0.016 (0.018)
Female	-0.115 (0.196)	0.098 (0.194)	0.123 (0.196)	0.001 (0.195)	0.065 (0.196)
Economics	0.133 (0.196)	0.133 (0.195)	0.048 (0.196)	0.313 (0.197)	0.302 (0.198)
# Obs.	141	141	141	141	141

Notes: Ordered probit regressions. *, **, *** indicates significance at the 10%, 5%, and 1% level without correcting for multiple testing. Standard errors are clustered on the individual level and reported in parentheses. Regressions are ran on the degree of consent with the respective item. The (+) and (-) indicate whether the statements represent affirmations (implying similarity between the lab and field) or negations (implying dissimilarity between the lab and field).

A5.2. Example Calculation of Scores for Distributive Motives

We illustrate how we calculate the scores that reflect the relative importance of efficiency, egalitarianism, and profit maximization on the individual level. The following table depicts choices of an exemplary subject in module 1 or module 2. The columns “Efficiency-Dummy” / “Egalitarianism-Dummy” / “Profit-Dummy” indicate if Option 1 in the respective item is the efficient / egalitarian / profit-maximizing choice.

Item	Choices in Module 1	Choices in Module 2	Efficiency-Dummy	Egalitarianism-Dummy	Profit-Dummy
1	Somewhat inappropriate (-1/3)	Option 2 somewhat more often (-1/3)	-1	-1	1
2	Very inappropriate (-1)	Option 2 much more often (-1)	-1	1	- - -
3	Somewhat inappropriate (-1/3)	Option 1 somewhat more often (+1/3)	1	-1	1
4	Very appropriate (+1)	Option 1 much more often (+1)	-1	1	1
5	Somewhat appropriate (+1/3)	Option 1 much more often (+1)	- - -	-1	1

The score representing the importance of the distributive motives in the two modules are then calculated as follows:

- Efficiency (injunctive): [(-1/3)*(-1)+(-1)*(-1)+(-1/3)*1+1*(-1)+(1/3)*0]/5=**0.00**
- Egalitarianism (injunctive): [(-1/3)*(-1)+(-1)*1+(-1/3)*(-1)+1*1+(1/3)*(-1)]/5=**0.07**
- Profit max. (injunctive): [(-1/3)*1+(-1)*0+(-1/3)*1+1*1+(1/3)*1]/5=**0.13**

- Efficiency (descriptive): [(-1/3)*(-1)+(-1)*(-1)+(1/3)*1+1*(-1)+1*0]/5=**0.13**
- Egalitarianism (descriptive): [(-1/3)*(-1)+(-1)*1+(1/3)*(-1)+1*1+1*(-1)]/5=**-0.20**
- Profit max. (descriptive): [(-1/3)*1+(-1)*0+(1/3)*1+1*1+1*1]/5=**0.40**

In Table 5.5 and Table 5.11, the level of experience is then regressed on these values.

Chapter 6

Point Beauty Contest: Measuring the Distribution of Focal Points on the Individual Level

Robert J. Schmidt

Abstract: We propose the Point Beauty Contest, a mechanism to identify the distribution of focal points on the individual level. By contrast to conventional coordination, subjects coordinate by the distribution of points. This allows for nuanced coordination strategies, as subjects can invest in multiple alternatives at the same time and weigh their choice. A subject´s strategy choice then reveals her perception of the distribution of focal points. In an experiment on the elicitation of social norms, we compare the mechanism with conventional coordination. The data confirms the theoretical predictions regarding coordination behavior and demonstrates that the proposed technique is suited to identify the distribution of focal points on the individual level. Using Monte Carlo simulations, we find that the proposed mechanism identifies focal points on the population level more efficiently than conventional coordination. We point to the possibility of using the mechanism as a simple method to directly measure strategic uncertainty.

Highlights:

- A method to identify the distribution of focal points on the individual level is proposed
- In an experiment, the proposed method is compared to conventional coordination
- The data demonstrates that the technique reveals focal points on the individual level
- On the aggregate level, the proposed technique identifies focal points more efficiently

Acknowledgments: I thank Dietmar Fehr, Florian Kauffeldt, Marco Lambrecht, Gerhard Minnameier, Illia Pasichnichenko, Hannes Rau, Christiane Schwieren, Stefan Trautmann, Marie Claire Villeval, Adam Zylbersztejn as well as seminar audiences in Frankfurt, Heidelberg, GATE Lyon Saint Etienne, the 9th International Conference of the ASFEE in Nizza, the ESA World Meeting in Berlin, and the ESA Asia-Pacific Meeting in Abu Dhabi for very valuable comments and suggestions.

6.1. Introduction

Schelling (1960) argues that in coordination games with multiple equilibria, subjects perceive varying degrees of saliences regarding the available alternatives. This renders some of the equilibria more or less "focal" and constitutes an implicit coordination device.[99] Focal points are interesting not only because they help subjects coordinate, but because of their potential to reveal shared perceptions. For example, in the original version of the Keynesian Beauty Contest (Keynes, 1936), respondents are provided with pictures of women, and their task is to coordinate on the most attractive pictures. According to Schelling´s concept, focal points might be induced by prevalent beauty ideals within the guessers´ population. Capitalizing on the same mechanism, Krupka and Weber (2013) propose using coordination games to elicit social norm perception.[100]

In the described settings, however, subjects choose only *one* alternative. As a result, the coordination choice of a single participant only reveals which alternative she considers *most* focal. For example, a subject´s coordination choice in the Keynesian Beauty Contest reveals which picture the respondent considers most salient, but it is not identified which picture is ranked second or third. In order to analyze how one alternative relates to other alternatives in terms of focality and to determine a ranking, it is necessary to combine the choices of many participants.[101] Yet, such a ranking would only emerge on the *population* level, i.e., based on the choices of *many* participants. By contrast, the ranking of focal points on the *individual* level, i.e., regarding a single respondent, remains unidentified. This results from the nature of the technique since subjects can only bet on one alternative.

We propose the *Point Beauty Contest*, a method that allows eliciting the ranking of focal points on the individual level. The Point Beauty Contest allows participants to bet on multiple outcomes and to weigh their choices. In contrast to conventional coordination, where subjects

[99] Schelling (1960) himself conducted a series of informal experiments to illustrate this effect. For example, he asked subjects whether they would pick either "heads" or "tails" in a coordination game. Of the 42 respondents, 36 chose heads. As no formal differences between the strategies or the respective equilibria were present in that setting, he concluded that the obvious presence of a coordination device could only be attributed to shared perceptions and that, apparently, "heads" appeared to be more focal than "tails". Since then, both experimental and theoretical work has corroborated the relevance of focal points in a variety of coordination settings (e.g., Binmore and Samuelson, 2006; Casajus, 2000; Crawford et al., 2008; Fehr et al., 2019; Isoni et al., 2013, 2014, 2019; Janssen, 2001, 2006; Metha et al. 1992, 1994a, 1994b; Pope et al., 2015; Sudgen, 1995; Sugden and Zamarrón, 2006).

[100] In that approach, subjects are confronted with the description of a particular behavior and they have to coordinate on appropriateness ratings. The method assumes that social norms are constituted through shared perceptions (Crawford and Ostrom, 1995), which thereby determine the focality of alternatives. Consequently, subjects´ coordination choices reveal perceptions about prevailing social norms.

[101] The term *focality* is meant to represent the *degree* to which an alternative appears to be focal to a player.

coordinate by choosing one alternative, subjects are equipped with a budget of points that they can distribute among multiple alternatives. Like in conventional coordination games, subjects are incentivized to reveal their beliefs about the other participants´ behavior, as they are paid according to the precision with which they anticipate the other participants´ choices. While coordination with a single choice reveals the most focal alternative, this approach allows the elicitation of the ranking of focal points on the level of a *single participant*.

Fine-grained coordination is present in many real-world coordination settings. In a bank run, for example, depositors might not only think about withdrawing *none* of their money or *all* of their money from a bank. Instead, due to a conflict of pecuniary incentives (play withdrawal) and social preferences (play no withdrawal), a subject might want to engage in both strategies simultaneously. The proposed mechanism captures two aspects of such a setting. First, subjects can invest in multiple alternatives in a coordination setting. Second, accordingly, the coordination outcome not only depends on the number of subjects choosing a particular alternative, but also on the *weights* that are put on the alternatives. Thus, the Point Beauty Contest provides a framework that reflects the interaction between subjects when nuanced strategy choices are feasible.[102]

We analyze the Point Beauty Contest both theoretically and experimentally. In the theoretical part, we derive predictions for coordination behavior that depend on risk preferences and strategic uncertainty. In an experiment on the elicitation of social norms (Krupka and Weber, 2013), we compare the proposed mechanism with conventional coordination. On the aggregate level, we find that the coordination outcomes correspond, i.e., the average ranking produced by the Point Beauty Contest matches the ranking elicited using the conventional approach. Looking at the choices on the subject level confirms the theoretical predictions and demonstrates that the proposed technique is suited to identify the ranking of focal points on the individual level. Moreover, using Monte Carlo simulations, we find that the Point Beauty Contest identifies focal points on the population level more efficiently, as it yields a given level of precision about the underlying distribution with significantly fewer subjects.

We see several fields of application for the Point Beauty Contest. First, the mechanism is suited to reflect coordination settings where fine-grained coordination is feasible. This allows to study coordination behavior when subjects opt for nuanced coordination strategies that involve engagement in multiple alternatives. Second, the mechanism allows to uncover the

[102] Note that such kind of nuanced strategies conceptually differ from mixed strategies, where subjects assign probabilities to every pure strategy. Instead, in the Point Beauty Contest, subjects engage in multiple strategies at the same point in time.

distribution of focal points in coordination games on the individual level. That is, subjects not only reveal the most salient alternative in a coordination setting but reveal their ranking of saliences. Third, the mechanism is useful when an experimenter is interested in the identification of focal points on the population level with fewer resources since the Point Beauty Contest yields results that are as precise as the results from conventional coordination with substantially fewer subjects.

The Point Beauty Contest also contributes to the elicitation of social norms using coordination games. As Krupka and Weber (2013) state, their results show that "a social norm is not always a single action that should or should not be taken, but rather a profile of varying degrees of social appropriateness for different available actions".[103] Social norms as such a profile can only be detected on the population level with conventional coordination, while the Point Beauty Contests elicits such profile of social norm perception on the subject level.

Finally, the Point Beauty Contest serves as a simple and direct tool to measure strategic uncertainty in coordination games, as the assignment of points depends on the risk preferences and the degree of strategic uncertainty that the subjects perceive. Controlling for risk preferences thus allows isolating the degree of strategic uncertainty on the individual level. For example, Heinemann et al. (2009) propose to measure strategic uncertainty by eliciting certainty equivalents and identify the payment that renders a subject indifferent between the certain payoff and an uncertain payoff that is subject to strategic uncertainty. Our approach would facilitate the elicitation of uncertainty in strategic settings, as the subject´s behavior (i.e., the distribution of points) reflects a direct measure for that kind of uncertainty.

The remainder of the paper is organized as follows. Section 6.2 contains the theoretical framework to derive predictions for coordination behavior. Section 6.3 presents the experiment and section 6.4 the experimental results. Section 6.5 contains simulations results on efficiency measurement. Section 6.6 summarizes and concludes.

6.2. Theoretical Framework

6.2.1. The Game

Consider a one-shot coordination game where subjects $i = 1, \dots, n$ see alternatives $j = 1, \dots, m$. Each subject receives a budget of X points and distributes the points between alternatives. The

[103] See the abstract of Krupka and Weber (2013).

number of points that individual i assigns to j is denoted as x_{ij}. All points must be used, i.e., $\sum_{j=1}^{m} x_{ij} = X$. We refer to the vector $X_i = (x_{i1}, \dots, x_{im})$ as a subject´s coordination choice. After all subjects decided about X_i, the average number of points $\bar{x}_j$ assigned to alternative j is calculated as $\bar{x}_j = (\sum_{i=1}^{n} x_{ij})/n$. The alternative that received most points on average is considered the *winning alternative* j^*. If more than one alternative received the maximum number of points, j^* is determined randomly among these alternatives.[104] Finally, each participant receives a payoff π_i that is proportional to the number of points x_{ij^*} that she assigned to the winning alternative, i.e., $\pi_i \sim x_{ij^*}$.

6.2.2. Belief Formation, Preferences, and Strategic Uncertainty

Focal Points and Belief Formation. For each alternative j, a subject perceives focality $\varphi_{ij} \geq 0$ and the vector $\Phi_i = (\varphi_{i1}, \dots, \varphi_{im})$ determines a subject´s ranking of focalities. A subject´s Φ_i is induced by the framing of the game, i.e., the question at hand. By definition, subjects assume that perceptions about focalities are correlated among participants and that the remaining subjects use it as a coordination device (Sudgen, 1995). Based on Φ_i, a subject derives beliefs p_{ij} which reflect the probability that alternative j becomes the winning alternative j^*. Specifically, stronger focality renders the respective alternative as a more promising bet for the investment of points: $\varphi_{ik} > \varphi_{il} \Leftrightarrow p_{ik} > p_{il}$ for two alternatives k and l. The vector $P_i = (p_{i1}, \dots, p_{im})$, with $\sum_{j=1}^{m} p_{ij} = 1$, is the perceived probability distribution about the coordination outcome of an individual i.[105] The translation of focalities into actual probabilities allows viewing the agent´s problem as a game against nature (Luce and Raiffa, 1957).

Preferences. Subjects exhibit von-Neumann-Morgenstern utility functions. For convenience, we normalize $\pi_i = x_{ij^*}$, so that profit will simply equal the number of points assigned to j^*. As a result, utility simplifies to $U_i = \sum_{j=1}^{m} p_{ij}\, u(x_{ij})$. The utility function u is continuous and twice differentiable with $u'(x) > 0$. Subjects can be risk-averse ($u''(x) < 0$), risk-neutral ($u''(x) = 0$), or risk-seeking ($u''(x) > 0$).

[104] It is necessary that only one j becomes the winning alternative. This ensures that subjects are not incentivized to equalize points among all alternatives which would maximize the profit of all participants, but render the outcome uninformative.

[105] We assume that subjects perceive the probabilities to be exogenous, i.e., they do not strategically assign points in an attempt to influence the probability distribution. This assumption is adequate when the number of participants is sufficiently large.

Strategic Uncertainty and Coordination Behavior. A subject is *certain* if she is sure about the outcome of the game: $p_{ij} = 1$ for some j. A subject is *partially uncertain* if she considers at least one alternative k to be more promising than another alternative l, without being fully confident: $0 < p_{il} < p_{ik} < 1$ for some k and l. A subject is *fully uncertain* if she is clueless concerning the outcome of the game: $p_{ij} = 1/m$ for all j.

Accordingly, we say that a subject is *gambling* if she assigns all points to one alternative: $x_{ij} = X$ for some j. A subject is *ranking* if she assigns more points to one alternative k than to another alternative l: $0 < x_{il} < x_{ik} < X$. A subject is *hedging* if she fully hedges her profit by assigning equally many points to all alternatives: $x_{ij} = X/m$ for all j.

6.2.3. Predictions for Coordination Behavior and Revelation of Focalities

For simplicity, predictions refer to a game with two alternatives k and l, without loss of generality. Table 6.1 shows predictions for coordination behavior depending on risk preference and strategic uncertainty. If subjects are either risk-averse or if subjects are certain about the coordination outcome, then a subject´s coordination choice X_i reflects her perception of underlying focalities Φ_i. That is, subjects are *gambling* in case of *certainty*, they are *ranking* in case of *partial uncertainty*, and they are *hedging* in case of *full uncertainty*. In these cases, subjects reveal their ranking of focal points, as they assign more points to alternatives that are considered more promising: $p_{ik} > p_{il} \Leftrightarrow x_{ik} > x_{il}$. Since we assume that subjects derive success probabilities of alternatives based on their degree of focalitity, i.e., $\varphi_{ik} > \varphi_{il} \Leftrightarrow p_{ik} > p_{il}$, a subject´s ranking of points will correspond to her ranking of focalities, i.e., $\varphi_{ik} > \varphi_{il} \Leftrightarrow x_{ik} > x_{il}$ in these cases.[106]

Proposition 1. If an individual is risk-averse or certain about the coordination outcome, then she fully reveals her ranking of focalities by assigning more points to alternatives that she considers more focal.

Proof. See Appendix A6.1.

[106] This also holds in a game with more than two alternatives, because the predictions apply to all pairwise comparisons of alternatives.

Table 6.1. Predictions for Coordination Behavior

	Certainty: $p_k = 1$	Partial uncertainty: $0 < p_l < p_k < 1$	Full Uncertainty: $p_k = p_l$
Risk-averse	$\boldsymbol{x_k = X}$	$\boldsymbol{0 < x_l < x_k < X}$	$\boldsymbol{x_k = x_l}$
Risk-neutral	$\boldsymbol{x_k = X}$	$x_k = X$	Indifferent
Risk-seeking	$\boldsymbol{x_k = X}$	$x_k = X$	$x_k = X$ or $x_l = X$

Notes: The bold printing refers to those cases where a subject's coordination choice fully reflects her beliefs.

6.3. Experiment

6.3.1. Design

We experimentally test our predictions by applying the Point Beauty Contest to elicit social norm perception. The idea to use coordination games to measure social norm perception has been proposed by Krupka and Weber (2013). In their approach, subjects are asked to evaluate a particular behavior (e.g., "how appropriate is it to do X?"), and they are provided with different answer alternatives to evaluate that behavior (e.g., "very appropriate", "somewhat appropriate", "somewhat inappropriate", "very inappropriate"). The subjects´ task is to choose the answer of which they think the majority of participants would choose it. That approach is equivalent to the classical Keynesian Beauty Contest, in which it is not optimal for subjects to state their own opinion, but to anticipate the modal choice of the group. We compare their method, where subjects can only bet on one alternative, with our approach, where subjects can bet on multiple alternatives and weigh their choices. Note that the elicitation of social norm perception is just one context to test the proposed mechanism. Any experimental setting, in which participants coordinate, would be suited for an experimental test.

We conduct two treatments: *Classical Beauty Contest (CBC)* and *Point Beauty Contest (PBC)*. In both treatments, we elicit injunctive social norms (part 1) and descriptive social norms (part 2) for five daily life behaviors. Injunctive social norms refer to perceptions of *normatively appropriate* behavior while descriptive social norms refer to perceptions of *common behavior,* i.e., the behavior practiced by most people (Cialdini et al., 1990). Table 6.2 shows the five behaviors that we use for the elicitation of injunctive and descriptive social norms.

Table 6.2. Items Used for the Elicitation of Social Norms

1.	Taking some money out of a found wallet before bringing it to the lost-property office.
2.	Lying for reasons of courtesy.
3.	Treating unfairly a person of which one has been treated unfairly before.
4.	Keeping the money when the cashier accidentally returned too much change.
5.	Mainly paying attention to the own well-being in daily life.

For the elicitation of injunctive social norms, subjects are confronted with a particular item and they are asked, how they evaluate the respective behavior regarding its appropriateness. Subjects then have to coordinate on the answer options: “very appropriate”, “somewhat appropriate”, “somewhat inappropriate”, “very inappropriate”. For the elicitation of descriptive social norms, subjects are confronted with a particular item, and they are then asked how many people would engage in the described behavior. Subjects then coordinate on the answer options: “a large majority”, “a majority”, “a minority”, “a small minority”.

In *CBC*, we employ conventional coordination, as done by Krupka and Weber (2013). That is, for each item, a subject receives 10€ if she manages to pick the answer alternative that is chosen by the majority of the respondents in the session (and zero otherwise). In the *PBC*, subjects are endowed with 100 points in each item, and their task is to distribute the 100 points between the available alternatives. In each item, subjects gain 0.10€ for each point that they assign to the *winning alternative*, i.e., the alternative that receives most points on average. Therefore, the payoff profile of *CBC* is also feasible in *PBC*, since assigning all 100 points to one alternative in *PBC* is equivalent to *CBC* in payoff terms.

In both treatments, subjects receive detailed instructions on the coordination mechanisms in parts 1 and 2 and about how their payment is determined. Specifically, subjects are provided with several examples to illustrate how their payment is calculated depending on their behavior and the behavior of others. Subjects answer several control questions in which they compute profits in a series of hypothetical scenarios. In particular, we pay attention to make clear that subjects are not asked about their own opinion. To make sure that subjects consider this feature, we remind them on each screen, on which they enter a coordination choice, that their task is not to state their own opinion but to coordinate with the remaining participants in the room.

Finally, in part 3, we elicit risk preferences using the Eckel and Grossmann (2008) approach, in order to test whether risk preferences affect coordination behavior in *PBC* as predicted by our theory. In part 3, subjects have to choose one of the lotteries from the menu shown in Table 6.3.

Table 6.3. Lotteries Choices Used to Elicit Risk Preferences

Lottery	50%	50%	EV	Risk Preference
1	4.00	4.00	4.00	RA
2	3.50	5.00	4.25	RA
3	3.00	6.00	4.50	RA
4	2.50	7.00	4.75	RA
5	2.00	8.00	5.00	RA
6	1.50	9.00	5.25	RA
7	1.00	10.00	5.50	RA / RN
8	0.50	10.50	5.50	RN / RS

Notes: EV = expected value; RA = risk-averse; RN = risk-neutral; RS = risk-seeking. In the experiment, subjects only see the first three columns.

At the end of a session, one of the three parts is drawn by chance to determine the payment. If part 1 or part 2 are drawn, then one item within that part is drawn by chance, and it determined the payment of a subject. If part 3 is drawn, then subjects play the lottery that they previously chose.

6.3.2. Procedure

The experiment was programmed in z-Tree (Fischbacher, 2007), and recruitment was done via hroot (Bock et al., 2014). In total, 158 subjects participated, and the sessions were conducted at the experimental lab of Heidelberg University in January and February 2018. We conducted 8 sessions, each with 20 participants (except for one session with 18 participants in *PBC*). 80 subjects participated in the *CBC* and 78 participated in the *PBC*. Participation in either treatment took about 35 minutes, and subjects earned on average 9.40€ (including a show-up fee of 5€).[107]

6.4. Results

6.4.1. Comparison of Coordination Outcomes

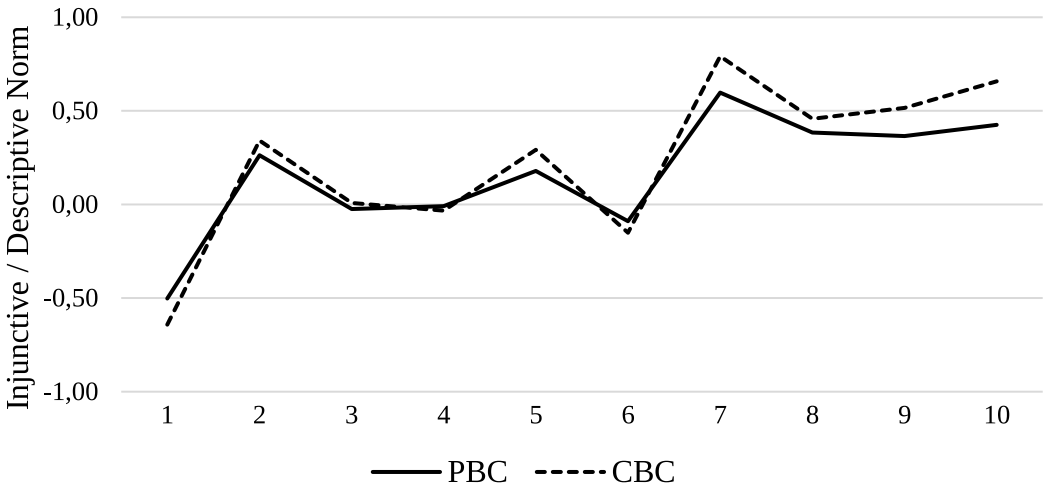
To analyze coordination choices, we quantify the answers such that the resulting scores are normalized between -1 and 1. Injunctive social norms are quantified as: 1 = " very appropriate", 1/3 = "somewhat appropriate", -1/3 = "somewhat inappropriate", -1 = "very inappropriate". Descriptive social norms are quantified as: 1 = "a large majority", 1/3 = "a majority", -1/3 = "a minority", -1 = "a small minority". Thus, the more positive (negative) the score in part 1, the

[107] A replication package, including instructions in German and English language, raw data, and data analysis files is available at the repository for research data of Heidelberg University: https://heidata.uni-heidelberg.de.

more appropriate (inappropriate) it is considered to engage in the described behavior. The more positive (negative) the score in module 2, the more common the described behavior is considered to be.

Figure 6.1 shows a comparison of mean norms elicited in *CBC* and *PBC*. Mann-Whitney-U tests are conducted to test whether mean outcomes differ between the treatments. We find that four items differ on the 5%-level (items 1, 7, 9, and 10). After correcting for multiple testing using the Bonferroni procedure, three items remain significant on the 5%-level (items 1, 7, and 10).[108] Generally, the results in *PBC* tend to be somewhat flatter than the results in *CBC*.

Figure 6.1. Comparison of Mean Coordination Results



Notes: Items 1 to 5 are injunctive norms from part 1, and items 6-10 are descriptive norms from part 2. The more positive (negative) the score in items 1-5, the more appropriate (inappropriate) it is considered to engage in the described behavior. The more positive (negative) the score in items 6-10, the more common the described behavior is considered to be.

We next compare ordinal rankings (Table 6.4). In *CBC*, alternatives in each item are ranked with respect to the share of subjects that chose a particular alternative. In *PBC*, alternatives in each item are ranked with respect to the average number of points assigned to the alternatives. We do not find that the rankings systematically differ. Precisely, the rankings produced by *CBC* and *PBC* correspond in eight of the ten items. In two items (4 and 10), we find that the rankings do marginally differ, as the order of two of the four alternatives is switched. These differences,

[108] We account for the fact that multiple items are used to detect treatment differences. In order to take care of the inflation of the overall type-I-error rate, we multiply the *p*-values by the number of items (i.e., by ten).

however, seem to result from noise, as the alternatives that do not correspond are extremely close to one another.[109]

Result 1. One the aggregate level, the coordination outcomes of *PBC* and *CBC* do not differ.

Table 6.4. Comparison of Rankings of Alternatives

Item	Point Beauty Contest					Classical Beauty Contest					Ranks identical
	+ +	+	-	- -	Mean	+ +	+	-	- -	Mean	
1	6	13	29	51	-0,50	1	5	40	54	-0,64	✓
2	27	42	22	8	0,26	15	73	11	1	0,34	✓
3	14	32	40	14	-0,02	11	33	53	4	0,01	✓
4	(19)	28	35	(17)	-0,01	(13)	34	40	(14)	-0,03	x
5	22	42	27	9	0,18	25	45	29	1	0,29	✓
6	13	31	35	21	-0,09	4	36	44	16	-0,15	✓
7	58	28	10	4	0,60	75	20	4	1	0,79	✓
8	35	43	16	6	0,38	26	68	5	1	0,46	✓
9	37	38	18	7	0,37	43	44	13	1	0,52	✓
10	(39)	(41)	14	6	0,42	(51)	(46)	3	0	0,66	x

Notes: Items 1-5 are injunctive social norms, and items 6-10 are descriptive social norms. Responses are: "very appropriate" (+ +), "somewhat appropriate" (+), "somewhat inappropriate" (-), "very inappropriate" (- -) in items 1-5 and "large majority" (+ +), "majority" (+), "minority" (-), "small minority" (- -) in items 6-10. For *PBC*, the numbers represent the average numbers of points that have been assigned to the respective alternatives. For *CBC*, the numbers represent the share (in percent) of subjects that chose the respective alternative. The modal response is shaded. Means are calculated using the above-described scoring. The numbers in parentheses in items 4 and 10 indicate those numbers, where the ranking of alternatives is not identical between the two treatments.

6.4.2. Coordination Behavior and the Role of Risk Preferences in the PBC

We look at all 780 decisions made in *PBC* (78 participants times 10 items per subject) and classify whether subjects apply gambling, ranking, or hedging. We observe almost no hedging (less than 0.1%), but some gambling (9.1%). In most of the decisions, subjects apply ranking, i.e., they assign varying numbers of points to the available alternatives (90.8%). More precisely, in 34.2% of cases, subjects fully rank their choices by assigning varying numbers of points to all four alternatives. In 53.4% of cases, subjects assign three different numbers to the four alternatives, and in 3.3% of cases, subjects assign two different numbers to the four alternatives.

[109] For example, in item 4, in the *CBC* alternative 1 is chosen by 12.5% and alternative 4 by 13.8%. By contrast, in the *PBC*, alternative 1 received 18.9 points on average and alternative 4 received 17.4 points on average. That is, in the *CBC*, alternative 4 is more popular, while in the *PBC*, alternative 1 is more popular. From a qualitative point of view, however, the two alternatives seem to be equally popular in both treatments. We therefore conclude that the differences concerning their ranking are not systematic, but result from noise.

Our theoretical framework predicts that subjects “manage” the degree of payoff risk, such that it suits their risk preference. Indeed, we find that the proportion of gambling is over-proportionally high in participants with low or negative risk aversion. While the share of gambling decisions is 22.4% for subjects that chose lottery 7 or lottery 8 (i.e., subjects that are potentially risk-neutral or risk-seeking as measured by the lottery task), it is only 3.0% for participants that chose lottery 1-6 (subjects that are clearly risk-averse as measured by the lottery task). Moreover, we examine risk induced in coordination choice. Table 6.5 reports regression results on the standard deviation of the assignment of points. We find that behavior in the risk elicitation task is significantly related to the standard deviation of the distribution of points. The more risk-averse subjects are in the lottery choice, the more balanced is the distribution of points, i.e., the lower the standard deviation implied in the coordination choice X_i. Also, older subjects are more prone to coordinate in a risky manner in the *PBC*. By contrast, gender and economics study remain insignificant, once we control for risk-attitude.

Result 2. In the *PBC*, most of the subjects rank their alternatives by assigning different numbers to the available alternatives. The more risk-averse subjects are, the less dispersed is the assignment of points.

Table 6.5. Risk Induced in Coordination Choice X_i

	Standard deviation of points assigned to alternatives		
Risk attitude	1.830*** (0.480)		1.792*** (0.521)
Female		-5.384*** (2.007)	-1.953 (1.985)
Age		0.569** (0.231)	0.658*** (0.186)
Economics		0.179 (0.460)	0.420 (0.429)
Constant	18.609*** (1.979)	15.343*** (5.728)	2.484 (5.780)
N	780	780	780

Notes: Tobit regressions. *, **, *** indicates significance at the 10%, 5%, and 1% level. The variable “risk attitude” indicates which of the lotteries (coded as number between 1 and 8) a subject chose. The higher the number, the less risk-averse is a subject. Robust standard errors are clustered on the individual level and reported in parentheses. As a robustness check, OLS regressions are conducted that yield the same results.

6.5. Simulation

We run Monte Carlo simulations in order to test which of the techniques uncovers the underlying ranking more efficiently, i.e., with fewer observations. We consider the realized coordination outcomes from the 78 subjects in *PBC* and the 80 subjects in *CBC* as benchmark (i.e., the results described in section 6.4.1.). We then run Monte Carlo simulations and mimic our original experiment with varying numbers of *n* participants, with $n = 1, \ldots, 100$. For each of the ten items, each *n* is simulated 10.000 times both for the *CBC* and the *PBC*. We then use the simulated data to study how fast the simulated results converge to the benchmark when *n* grows larger. The degree of convergence is measured using convergence of the mean and convergence of the ordinal ranking of the alternatives.[110] Convergence of the mean is measured as realized confidence intervals (50% and 90%) of the simulated means. Convergence of ordinal rankings is measured as the share of simulated items, in which the ordinal ranking corresponds to the benchmark. The more efficient the mechanism, the smaller should the confidence intervals of means become when *n* grows larger. Equivalently, the more efficient the mechanism, the higher should be the share of simulated items in which the ordinal ranking produced by the simulation is identical with the benchmark when *n* increases. Holding a particular *n* constant thus allows us to compare the degree of efficiency between *PBC* and *CBC*.

The simulation results show that in the *PBC*, the examined confidence intervals are lower for each *n* in either of the 10 items (see Figure 6.2).[111] That is, the precision with which the mean is approached when the number of participants increases is higher for the *PBC* for each size of *n*. Regarding convergence to the ordinal rankings, the *PBC* converges faster to the underlying ranking in 9 of the 10 items, while in one item the *CBC* converges faster.

The efficiency gains are particularly strong for the usual numbers of participants used in economic experiments. For example, both the 90%-confidence and the 50%-confidence intervals for the mean that are realized in the *CBC* with $n = 50$ participants are reached in the *PBC* with $n = 30$ participants already. The share of ordinal rankings that corresponds to the benchmark that is produced in the *CBC* with $n = 50$ is reached in the *PBC* already with $n = 16$. This indicates that the *PBC* is more efficient as an experimental method, in particular regarding the elicitation of ordinal rankings of focal points.

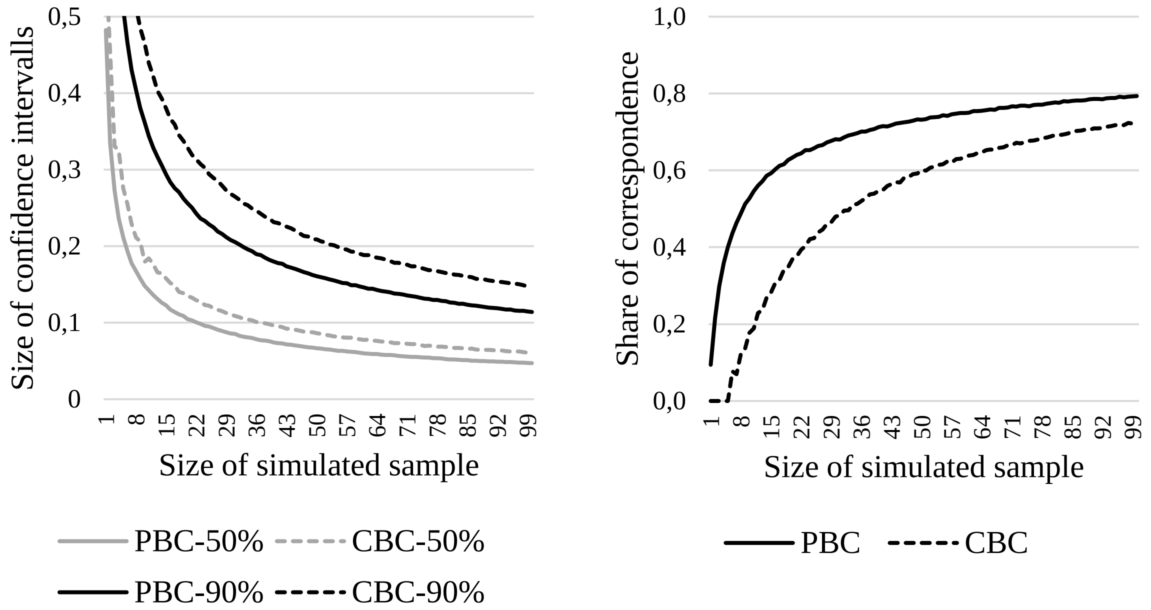
[110] To derive the mean, we use the same scoring system as in the results section. That is, the ratings are normalized between -1 and 1.
[111] Figure 6.2 shows the average of all 10 items. In Appendix A6.2 the reader finds figures of simulations results separately for each item.

Result 3. The *PBC* is more efficient than the *CBC* in identifying the means and the ordinal rankings of coordination choices on the population level.

Figure 6.2. Simulation Results

Panel A. Convergence to the Mean

Panel B. Convergence to the Ordinal Ranking



Notes: The x-axis of both graphs indicates the *n*, i.e., the number of participants that is being simulated. Panel A shows confidence intervals of means. Panel B shows the share of simulation runs in which the ordinal ranking of a simulation run corresponds with the ranking of the benchmark. Both graphs contain the data of 100.000 simulation runs (10.000 simulation runs for each of the ten items). Appendix A.2 contains graphs of simulations separately for each item.

6.6. Summary and Conclusion

We propose a Point Beauty Contest to identify the ranking of focal points in coordination games on the individual level. By contrast to conventional coordination where subjects can only bet on one alternative, subjects are endowed with points, which they assign to the available alternatives. This enables subjects to bet on multiple outcomes and to weigh their choices. We examine the proposed method both theoretically and experimentally. In the theoretical part, we derive that the assignment of points depends on strategic uncertainty and risk preferences. In an experiment, we find that the mechanism is suited to identify the heterogeneity of focal points on the individual level, as most of the subjects assign varying numbers of points to the different alternatives. Using Monte Carlo simulations, we find the mechanism to be more efficient regarding the identification of focal points on the population level.

We see four contributions. First, the Point Beauty Contest provides a framework to formally represent coordination settings in which subjects do not coordinate by exclusively choosing one alternative, but in which subjects coordinate in a fine-grained manner by choosing multiple alternatives at the same time. For example, the framework is useful to model a bank-run if one assumes that depositors do not only think about withdrawing none of their money or all of their money from a bank, but want to engage in both strategies simultaneously. Second, the mechanism allows to uncover the distribution of focal points in coordination games on the individual level. This allows, for example, to measure social norms on the individual level as a profile, i.e., varying degrees of social appropriateness for different available actions (cf. Krupka and Weber, 2013). Third, the Point Beauty Contest provides a possibility to measure focal points on the population level with significantly fewer participants compared to conventional coordination. Fourth, the Point Beauty Contest serves as a simple and direct tool to measure strategic uncertainty in coordination settings (Heinemann et al., 2009), as the assignment of points in the Point Beauty Contest yields a measure that is directly related to that kind of uncertainty.

Appendix 6

A6.1. Proof of Proposition 1

We suppress the script i from now on and analyze a representative individual. If $\pi = x_j^*$, the utility function is $U = \sum_{j=1}^{m} p_j * u(x_j)$.

Risk aversion. Take two arbitrary alternatives k and l. Assume $\varphi_k > \varphi_l \Leftrightarrow p_k > p_l$. First-order conditions require that $p_k * u'(x_k) \equiv p_l * u'(x_l)$. If $p_k > p_l$, then $u'(x_k) < u'(x_l)$. Since utility is marginally decreasing in case of risk aversion, it needs to be that $x_k > x_l$. As a result, for each comparison of two arbitrary alternatives and independent from strategic uncertainty, utility maximization requires to assign more points to alternatives that are more focal: $\varphi_{ik} > \varphi_{il} \Leftrightarrow x_{ik} > x_{il}$.

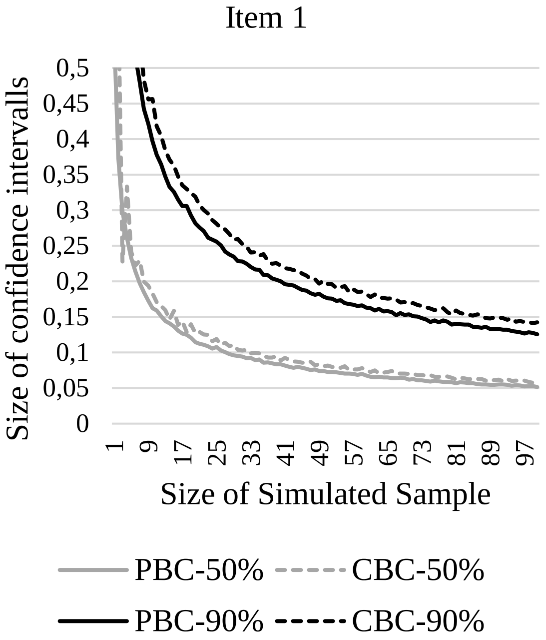
Risk-neutrality. Take two arbitrary alternatives k and l. Assume $\varphi_k > \varphi_l \Leftrightarrow p_k > p_l$. In case of "certainty", it is the dominant strategy to assign all points to the alternative that is expected to become the winning alternative with certainty. The same reasoning applies to a subject that perceives "partial uncertainty", i.e., assigning all points to the more promising alternative is dominant. Risk-neutral subjects that face "full uncertainty" are indifferent between all possible distributions of points, since they can neither control expected profit nor can they affect the expected payoff. Therefore, in case of risk-neutrality, $\varphi_{ik} > \varphi_{il} \Leftrightarrow x_{ik} > x_{il}$ applies if subjects are certain, while it does not when subjects are partially uncertain or fully uncertain.

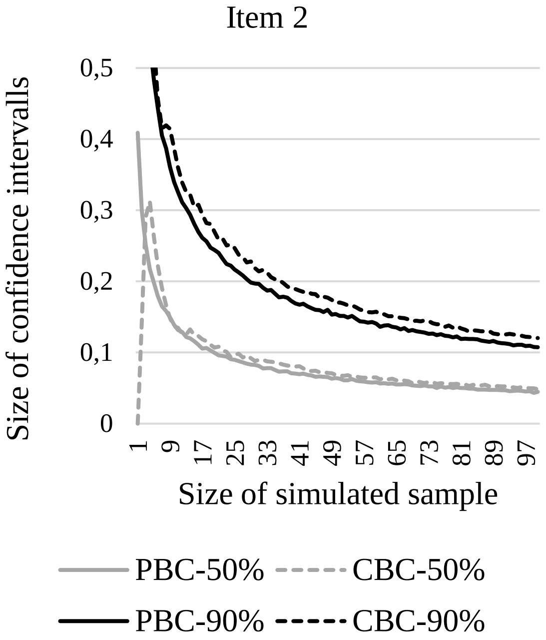
Risk-seeking. Take two arbitrary alternatives k and l. Assume $\varphi_k > \varphi_l \Leftrightarrow p_k > p_l$. In case of "certainty", it is the dominant strategy to assign all points to the alternative that is expected to become the winning alternative with certainty. This maximizes the expected payoff, while risk cannot be affected in case of certainty. The same reasoning applies to a subject that is "partially uncertain". In case of "full uncertainty", risk-seeking subjects will invest all points into a random alternative, because, although they cannot affect expected payoff in case of full uncertainty, gambling will maximize risk, thus it maximizes utility. Therefore, in case of risk-seeking, $\varphi_{ik} > \varphi_{il} \Leftrightarrow x_{ik} > x_{il}$ applies if subjects are certain, while it does not when subjects are partially uncertain or fully uncertain. ■

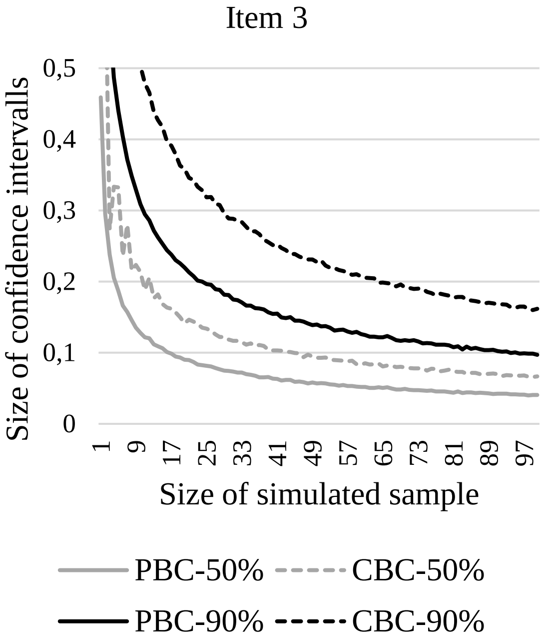
A6.2. Figures of Simulation Results for each Item

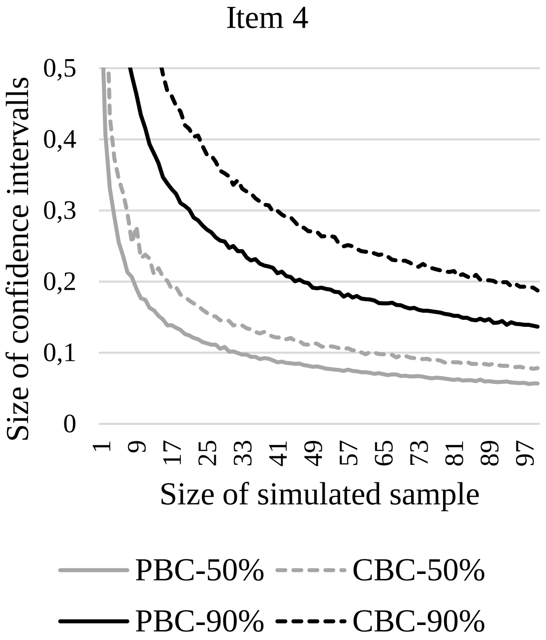
The graphs show simulation results separately for each of the ten items. Each simulation contains 10.000 simulation runs.

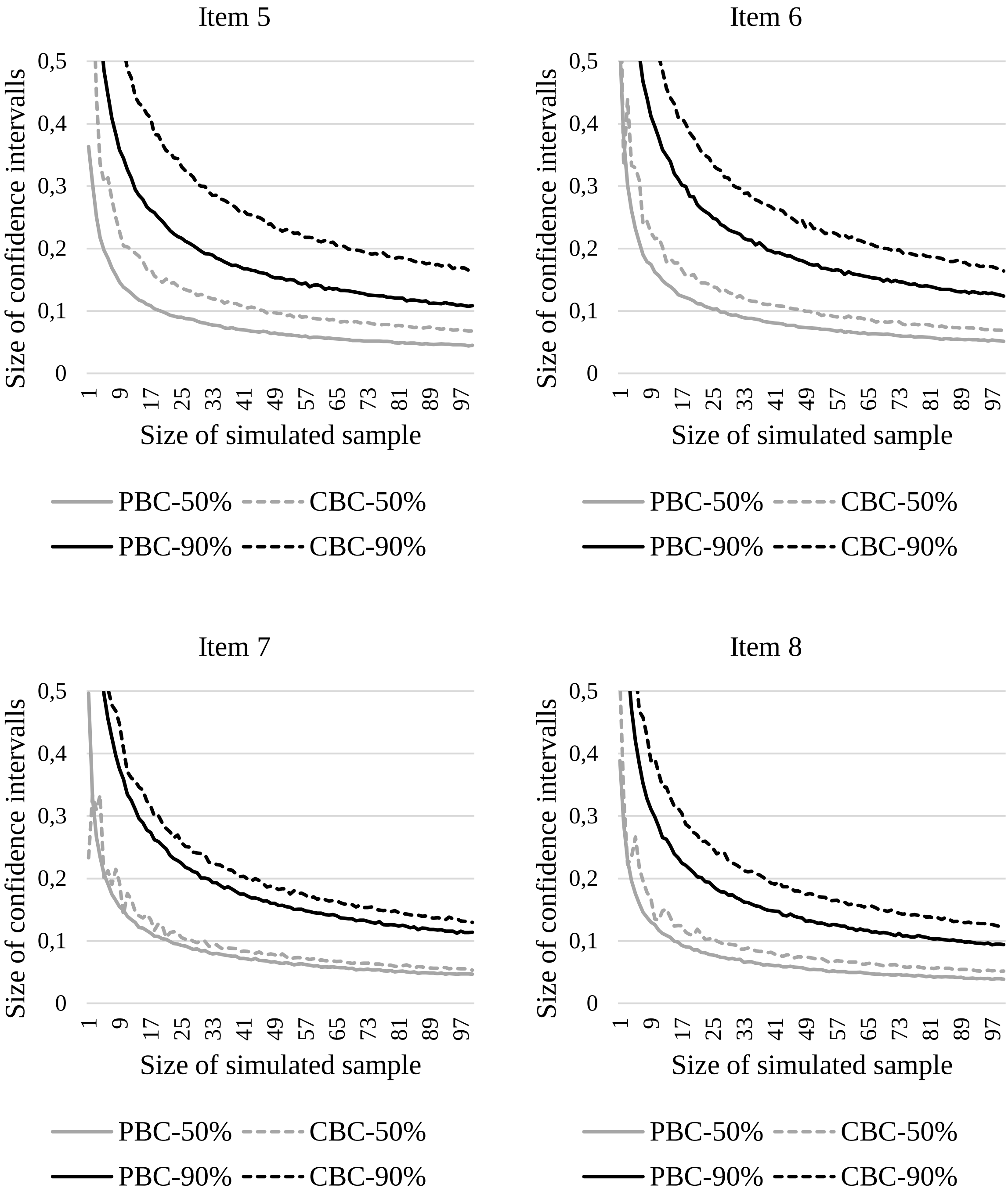
A6.2.1. Convergence to the Mean

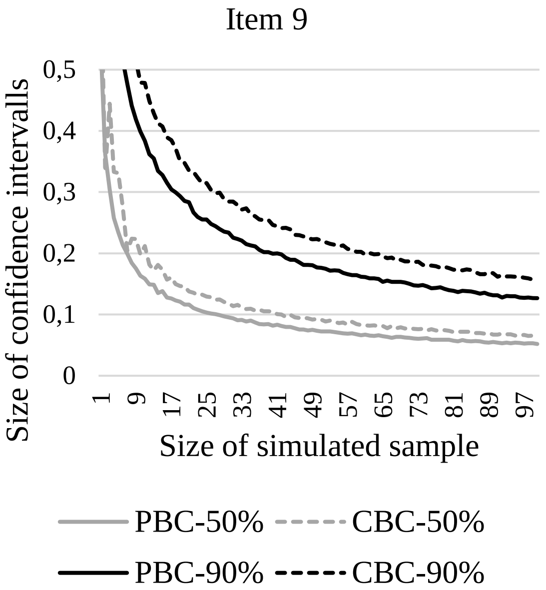


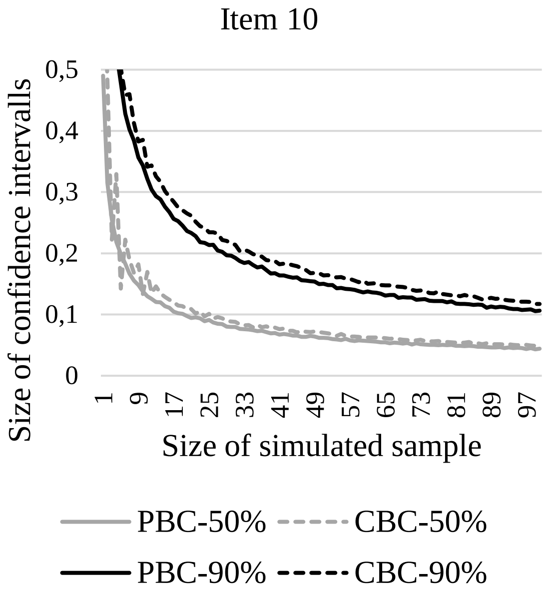




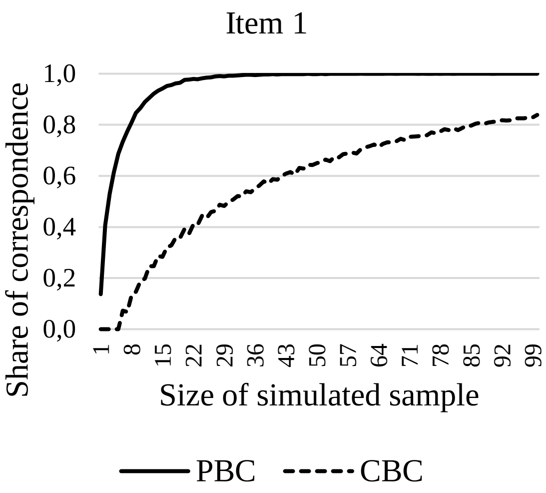


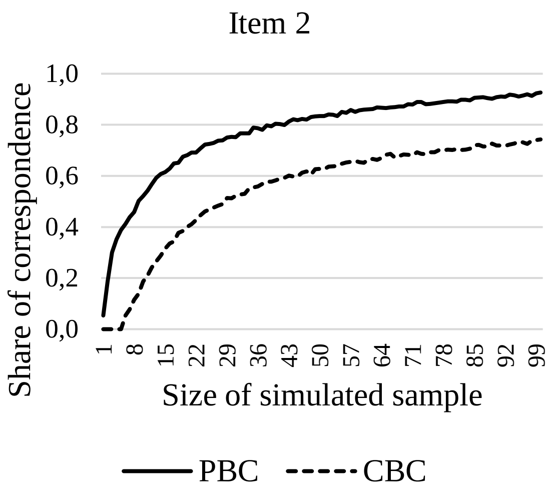

Item 5
Size of confidence intervalls
0,5
0,4
0,3
0,2
0,1
0
1 9 17 25 33 41 49 57 65 73 81 89 97
Size of simulated sample
PBC-50% CBC-50%
PBC-90% CBC-90%
Item 6
Size of confidence intervalls
Size of simulated sample
PBC-50% CBC-50%
PBC-90% CBC-90%
Item 7
Size of confidence intervalls
Size of simulated sample
PBC-50% CBC-50%
PBC-90% CBC-90%
Item 8
Size of confidence intervalls
Size of simulated sample
PBC-50% CBC-50%
PBC-90% CBC-90%

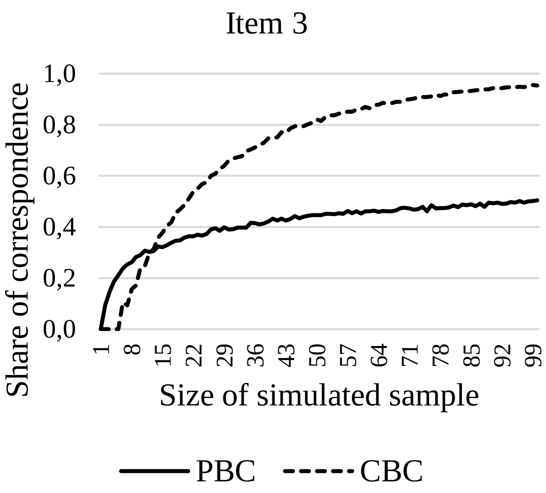


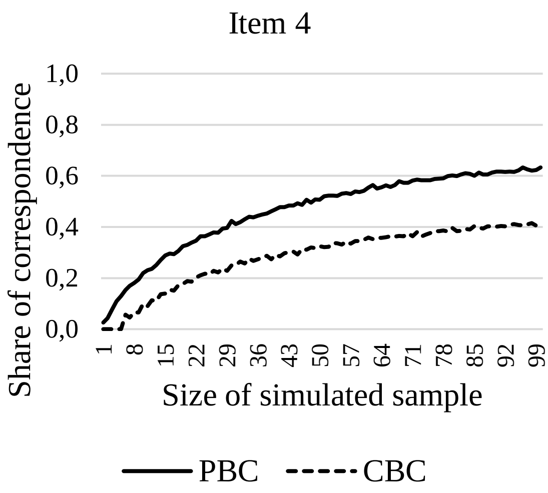


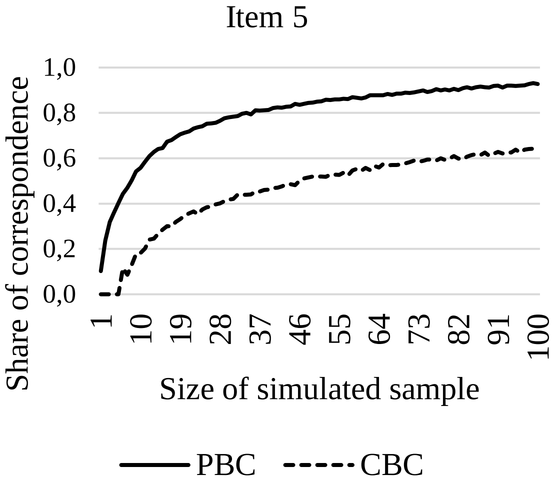
A6.2.2. Convergence to the Ordinal Ranking

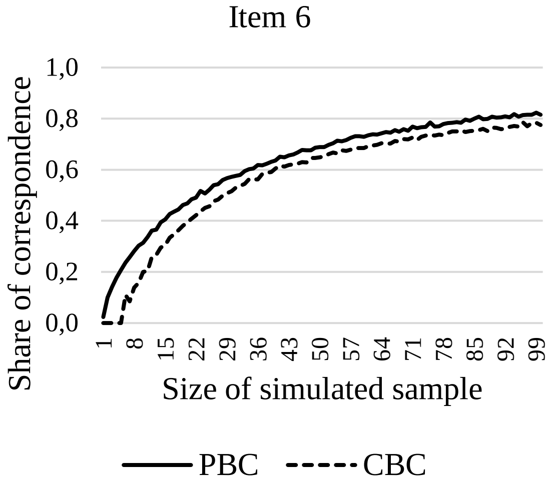


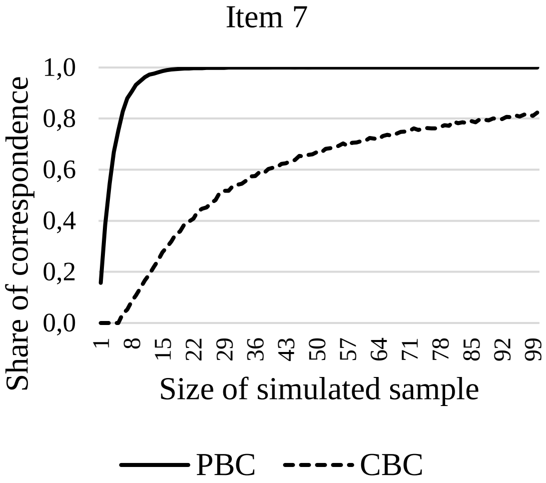


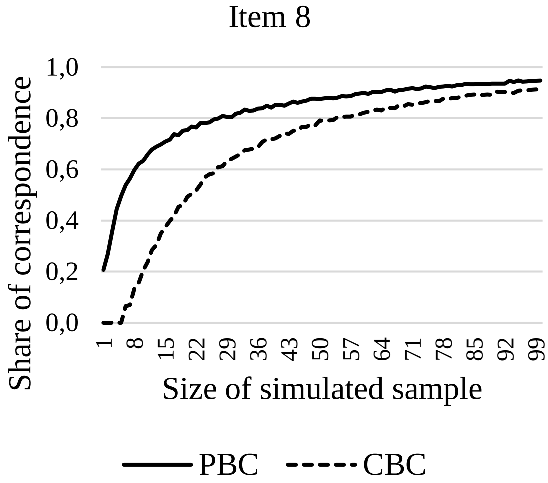


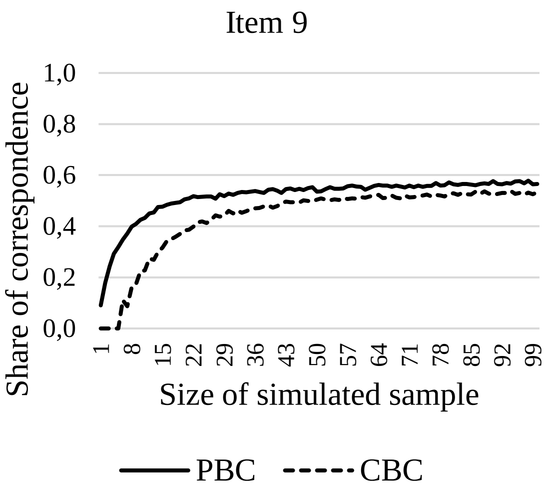


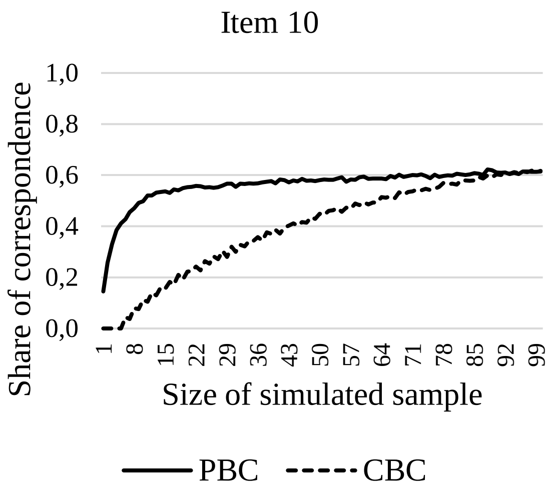

Item 5
Share of correspondence
1,0
0,8
0,6
0,4
0,2
0,0
1 10 19 28 37 46 55 64 73 82 91 100
Size of simulated sample
PBC
CBC


Item 6
Share of correspondence
1,0
0,8
0,6
0,4
0,2
0,0
1 8 15 22 29 36 43 50 57 64 71 78 85 92 99
Size of simulated sample
PBC
CBC


Item 7
Share of correspondence
1,0
0,8
0,6
0,4
0,2
0,0
1 8 15 22 29 36 43 50 57 64 71 78 85 92 99
Size of simulated sample
PBC
CBC


Item 8
Share of correspondence
1,0
0,8
0,6
0,4
0,2
0,0
1 8 15 22 29 36 43 50 57 64 71 78 85 92 99
Size of simulated sample
PBC
CBC


Item 9
Share of correspondence
1,0
0,8
0,6
0,4
0,2
0,0
1 8 15 22 29 36 43 50 57 64 71 78 85 92 99
Size of simulated sample
PBC
CBC


Item 10
Share of correspondence
1,0
0,8
0,6
0,4
0,2
0,0
1 8 15 22 29 36 43 50 57 64 71 78 85 92 99
Size of simulated sample
PBC
CBC

Chapter 7

7. Discussion and Conclusion

There are multiple conclusions that can be drawn from the results presented in this dissertation, most of which have already been stated in the conclusion sections of the respective chapters. This chapter aims to provide conclusions that consider the implications of the research results from a broader perspective, while also taking into account the limitations and shortcomings of the respective studies.

For that sake, I divide the dissertation into two parts: Chapter 2 on the one hand and Chapters 3 to 6 on the other hand. This split is guided by the methodology used in the respective parts. While subjects make choices about payoffs in the experiment presented in Chapter 2, the experiments in Chapters 3 to 6 are devoted to examining choices in coordination settings. Section 7.1 provides concluding remarks about Chapter 2 by discussing a shortcoming of our study and by providing avenues for future research on process fairness. Sections 7.2 and 7.3 provide concluding remarks about Chapters 3 to 6, by discussing the methodology of using pure coordination games to answer research questions and by evaluating the limitations of my dissertation to draw inferences about that methodology.

7.1. Futures Avenues for Research on Process Fairness

The main aim of chapter 2 is to examine how institutions that are in charge of allocating indivisible goods need to be designed so that decision-makers adhere to process fairness. This is motivated by the idea that process fairness, ceteris paribus, is valuable and should be aimed for from a prescriptive perspective (e.g., Andreozzi et al., 2013; Bolton et al., 2005; Krawczyk, 2011; Saito, 2013; Trautmann, 2009; Trautmann and van de Kuilen, 2016; Trautmann and Wakker, 2010). In accordance with the existing literature, our results demonstrate the detrimental effects of lack of perceived process fairness: subjects´ perceptions about outcome fairness are negatively affected by low perceptions of process fairness (although the level of outcome inequality is constant across treatments) and recipients engage more in costly punishing the allocator. Thus, our results are consistent with previous research results demonstrating that perceptions of process fairness are crucial for the success of social and economic interactions (e.g., Kollock, 1998; Ostrom 1990).

A significant part of previously conducted research on process fairness deals with understanding the factors that shape fairness perceptions, the interaction of different concepts of fairness, and the consequences that are associated with low levels of perceived fairness. However, I think that the importance of process fairness, the detrimental effects of lack of process fairness, and its necessity in social interactions are well understood (e.g., Adler et al., 1983; Alm et al., 1995; Andreoni et al., 1998; Bowles and Gintis, 2000; Brockner, 2002; Falk et al., 2017; Fehr, 2018; Garonzik et al., 2000; Kessler and Leider, 2016; Skarlicki and Folger, 1997; Trautmann and van de Kuilen, 2016). Thus, the direction of research that we step into is innovative in that we conduct empirical research pursuing to understand *how* it can be made sure (or at least more probable) that process fairness is *actually implemented and maintained* in situations where a fair process is *not easily feasible*. In particular, what our study demonstrates is that it is not granted that subjects are always *able* to implement process fairness, even though they might have a preference for doing so. Based on our results, I believe that it is worth to follow that path and to try to understand more thoroughly how the implementation of process fairness can be fostered in real-world allocation settings.

Likewise, I think it is worthwhile to continue researching to better understand how perceptions of process fairness can be maximized in settings where outcome inequality is unavoidable. Since process fairness serves as a cushion for outcome inequality, it is essential to learn more about how groups, organizations, and the society can mitigate the detrimental effects of distributional inequality through proper implementation of fair allocation processes. This seems to be particularly necessary, given that both income inequality and wealth inequality will tend to rise in our society in the future (e.g., Piketty, 2014). Under this premise, outcome fairness progressively becomes less feasible, so that society has to make an effort wherever it can to help allocators to apply fair procedures and to maximize perceptions of process fairness on the side of recipients.

Before providing ideas for how future research could extend the study presented in Chapter 2, I want to state a shortcoming that is associated with our experiment. In our setting, we present two subjects as being equally deserving, but we do not induce a situation in which one participant is clearly *more* entitled than another recipient, based on some objective measure (e.g., a real-effort task or through different endowments). In order to corroborate the claim that subjects actually exhibit favoritism, it would have been interesting to confront the decision-maker with such treatments. Specifically, having varying constellations of *relative entitlements*, combined with varying measures of *relative similarity* (i.e., less/more deserving participant has higher/lower relative similarity with the allocator) in a 2×2 factorial design, would allow to examine the trade-off between preferences for process fairness and preferences for favoritism in a more comprehensive manner. As our design is not suited to explore this trade-off, I hope that further studies will devote to examining this aspect.

In addition to considering this limitation, I want to sketch three further aspects in which future research could extend our experimental allocation setting. First, it might be worthwhile to examine how subjects make use of the possibility to communicate in the given setting, and how communication affects fairness perceptions. Here, different treatments are thinkable. For example, one could provide the allocator with the possibility to explain her decision to recipients. This is a feature that is present in a series of real-world allocation settings, e.g., when assigning jobs or promotions among candidates. One hypothesis could be that decision-makers make use of their possibility to explain their decision and to help recipients understand the dilemma in which the allocator is situated. It is not implausible to expect that this affects the recipients´ interpretation of the allocator´s behavior, in particular if the allocator uses the opportunity to verbally signal that she is seeking to implement a fair procedure. As a result, it might positively affect perceptions of process fairness on the side of recipients. Further treatment variations could allow for communication between all involved players, or allow for asymmetric communication, where only a part of the recipients has the possibility to communicate with the allocator.

Second, it might be interesting to vary the kind of information that is transmitted between participants. While we use political profiles to create relative distance between the allocator and the recipients, there are other interesting ways to create such differences. One straight-forward example would be to examine the effect of revealing the gender of subjects. This might be interesting because it potentially helps understanding gender concentrations in different areas if one assumes that selection decisions have similar characteristics as our allocation setting (i.e., decisions are one-shot, and multiple *independent* allocators determine allocations). Further possibilities include socio-economic background or the level of income. Also, one could think about using a catalog of questions, as we did, but with a different emphasis.

Third, it might be interesting to study the given setting in a dynamic context. In our experiment, we apply a one-shot interaction. However, one shot-interactions conceptually differ from dynamic settings where subjects repeatedly interact over multiple periods. In the real-world, allocation decisions sometimes appear one shot, and the recipients do not have the possibility to further interact with the allocator. However, there are also settings where the participants interact in a repeated manner (take, for example, the assignment of attractive tasks or projects within a company). Therefore, it would be worth studying whether in repeated interactions, the allocator simply repeats the actions that she takes in the one-shot setting, or whether she varies her behavior. Another limitation of studying one-shot settings is that it is not possible to detect whether an allocator would have implemented process fairness in the long run by alternating which recipient receives the favorable outcome. Additionally, it would be interesting to examine how recipients´ retaliation behavior, as well as fairness perceptions of all involved individuals look like in dynamic environments.

Finally, one could think about various further interventions, since the interventions that we examine are only a small section of possible mechanisms. I thus hope that further research uses our design to continue exploring the underlying research question while taking care of our limitations and by enriching our experimental setting in sensible ways.

7.2. The Informativeness of Coordination Choices on the Individual Level

Regarding Chapters 3 to 6, I want to draw joint conclusions about the methodology used in these studies. In all experiments contained in these chapters, we use coordination games to draw inferences about subjects based on their coordination choices. The implicit assumption that underlies this approach is that it is appropriate to use such decisions to draw inferences about beliefs (Chapter 3) or social norm perception (Chapters 4 to 6) of participants. In that regard, we go one step further than Krupka and Weber (2013), who propose that coordination games are suited to predict changes across contexts and treatment variations on the *aggregate level* (cf. abstract of Krupka and Weber, 2013). Looking at the combined data of my projects, my conclusion is that coordination choices are also indicative of subjects´ traits on the *individual level*.

To make that point, I first recapitulate the results from Chapters 3 and 4. Remember that in these experiments, we not only elicit coordination choices from subjects, but we additionally elicit beliefs (Chapter 3) and social preferences (Chapter 4) in an incentive-compatible manner using standard procedures from experimental economics. Beliefs are elicited using a scoring rule that pays subjects based on precision, and social preferences are measured via incentivized allocation settings (i.e., dictator games, which are a workhorse in experimental economics to measure social preferences). In both studies, we find that coordination choices are statistically significantly related to the respective trait of interest on the individual level (i.e., either to subjects´ beliefs or to their revealed social preferences). More precisely, in Chapter 3, coordination choices do not differ from first-order beliefs, neither on the aggregate nor on the individual level. In Chapter 4, we find that when subjects coordinate on social norms, their individual coordination choices are strongly related to their actual decisions in the dictator games.

The results, however, need to be put into perspective, given one fundamental methodological limitation that is shared by the studies contained in Chapter 3 and Chapter 4. Since we are eliciting individual level traits, we cannot rule out the existence of spillover effects between the two kinds of decisions that subjects take. In Chapter 3, in treatment *COORDINATION*, it might be that spillover effects emerge between elicited *coordination choices* (elicited in stage 1 of the treatment) and *first-order beliefs* (elicited in stage 2). Equivalently, in Chapter 4, it might be that spillover effects emerge between elicited *social preferences* (in the allocation stage) and *coordination choices* (in the norm elicitation stage). Put differently, there are two competing explanations for the close relationship: A subject could choose to coordinate in a particular manner either (i) because it corresponds to her actual trait (which is our claim when applying that methodology), or (ii) because of spillover effects that occur between the two kinds of decisions that individuals make. The *fundamental* methodological problem with this approach is that it is difficult to refute the hypothesis of spillover effects since this would require to *independently* elicit *two* individual decisions from *one* subject at the *same* point in time. Therefore, since the decisions that subjects make in Chapter 3 and Chapter 4 are *not independent*, spillover effects between the two choices cannot be ruled out.

Arguing with the data provided in these chapters therefore suffers from that fundamental limitation, as the results are only *consistent* with the hypothesis that coordination choices are informative, while the data is not suited to refute the hypothesis of spillover effects. In order to get around that problem, one needs to design an experiment where subjects play coordination games, and the coordination choices would need to be compared to another measure on the individual level that is either (i) *not elicited* in the same experiment (so that a direct spillover effect is avoided), or that is (ii) *fixed* (so that it is not affected by spillover effects, even if spillover effects do exist). If the relationship between the subject´s type and her coordination choice remains tight, this would not be explained by spillover effects.

However, the experiment from Chapter 3 provides additional support for the hypothesis that spillover effects do not explain the relationship between coordination choices and first-order beliefs. The fact that first-order beliefs in treatment *BELIEF* and first-order beliefs elicited in stage 2 of treatment *COORDINATION* do not differ on average indicates that having subjects coordinate beforehand (in stage 1 of treatment *COORDINATION*), does on average *not affect* their first-order beliefs. Although this data does not refute the hypothesis that spillover effects do exist, it provides support for the hypothesis that they do not play a significant role.

The data in Chapter 5 also supports that conclusion, when looking at the coordination choices of females and males. These indicate that female subjects are significantly more guided by egalitarianism and significantly less guided by efficiency in experimental allocation tasks than male subjects. This is in accordance with previous studies on gender effects in dictator game giving (e.g., Andreoni and Vesterlund, 2001; Eckel and Grossman, 1998) and supports the idea that individual characteristics shape how subjects choose to coordinate. The relationship between a subject´s coordination choice and her gender cannot be explained by spillover effects.

However, one limitation in that regard is that the experiment in Chapter 5 was not designed to test for such (gender) differences. Therefore, one needs to be cautious when using this data to evaluate hypotheses (otherwise, one would fall short regarding the issue of multiple testing, i.e., to use the very same data to shed light on more than one hypothesis). I therefore document that the gender differences identified in Chapter 5 are consistent with the idea that coordination choices of males and females are indicative of gender-specific social preferences found by Andreoni and Vesterlund (2001) as well as Eckel and Grossman (1998). However, in order to produce more profound experimental evidence, one would need to design an experiment whose intention it is to explicitly test for gender differences.

Taken together, the data of these projects lead me to conclude that coordination games are an interesting tool to draw inferences about individual participants. Our combined results support the idea that subjects use their own type when predicting the coordination choices of others (Dawes, 1989; Vanberg, 2019). The methodological value of this insight is that coordination games are a powerful tool to identify traits of subjects that are otherwise difficult to explore. This includes beliefs about unobservable facts ("what is the probability that some hypothetical event materializes until year 2100?") or preferences that are difficult to measure ("how much would you donate if you would win a million dollar?"). More generally, it includes all kinds of traits and characteristics that are difficult to explore using the standard approaches of experimental economics, i.e., observing how subjects express their actual traits in properly incentivized decision situations (Smith, 1976). However, in order to corroborate that claim, more experiments, which fundamentally rule out spillover effects, are necessary.

Finally, the mechanism proposed in Chapter 6 does not explicitly yields further insights regarding the previous conclusion. However, given that the above conclusion would be valid, the Point Beauty Contest would allow drawing inferences about subjects on the individual level in a more fine-grained manner. A limitation of the experiment in Chapter 6 is that we examine the Point Beauty Contest only for the elicitation of social norms. In order to generate further support for the claim that focal points on the individual level are measurable using that approach, it would be desirable to test it with other coordination games.

7.3. Coordination Games as an Incentivized Crowd Wisdom Device

Finally, I want to conclude on the idea of using coordination games as a prediction or crowd wisdom device. To do so, I again briefly recapitulate the experiments conducted in Chapter 3 and Chapter 4. In Chapter 3, the subjects´ task is to estimate realized probabilities about behavior in an ultimatum game conducted by Trautmann and van de Kuilen (2014). Subjects are confronted with a particular event, and their task is to coordinate on the realized probability of that event. Each subject states an integer between 0 and 100 that is meant to represent probability in percentage terms, and the closer their number is to the average number stated by the crowd (i.e., all subjects within a session), the higher their payoff. In Chapter 4, subjects have a similar task, as they have to coordinate on descriptive social norms, which represent how prevalent a particular behavior is considered to be. Subjects are confronted with the description of a particular behavior in the dictator game, and their task is to coordinate on one of the four categories: “a large majority”, “a majority”, “a minority”, “a small minority”.

In both experiments, we find that the coordination outcomes very well correspond to the observed behavior. In Chapter 3, the average probabilities elicited using coordination games correspond to actual behavior rates in the ultimatum game conducted by Trautmann and van de Kuilen (2014), and in Chapter 4, the elicited descriptive norms yield accurate estimations about actual dictator behavior.

In particular, Chapter 3 yields interesting insights about the suitability of using coordination games as crowd wisdom or prediction device. First, the assessments of realized probabilities using the conventional elicitation of first-order beliefs are not more accurate than the elicitation of beliefs using coordination games. Second, by contrast, beliefs elicited in a non-incentivized manner seem to differ from incentivized beliefs, and their external validity is also lower. Although the second result needs to be considered with caution due to the small sample of subjects that state their beliefs in a non-incentivized manner, descriptive analysis indicates that beliefs elicited using coordination are a better indicator for actual first-order beliefs than are non-incentivized beliefs. This suggests that using coordination games to elicit beliefs or descriptive social norms are a potentially powerful tool to generate predictions about questions that are otherwise difficult to incentivize.

Of course, the evidence in this dissertation that leads me to these conclusions is limited, since the experiments comprise only a small number of possible fields of applications. In order to back up these conclusions, one needs to run experiments in further areas, such as sports matches, political events, events on the financial market, and the like. That data could then be compared to predictions generated in other settings (e.g., prediction markets). My impression is that coordination games are an interesting tool to make incentivized predictions in such settings, and I believe that it would be worthwhile to employ a “horse race” to compare that approach with other mechanisms. I hope that further experimental work will be conducted to explore that claim.

References

Abeler, J., & Nosenzo, D. (2015). Self-selection into laboratory experiments: pro-social motives versus monetary incentives. *Experimental Economics*, *18*(2), 195-214.

Adler, J. W., Hensler, D. R., Nelson, C. E., & Rest, G. J. (1983). *Simple justice*. Rand Corporation.

Allcott, H. (2011). Social norms and energy conservation. *Journal of Public Economics*, *95*(9-10), 1082-1095.

Ambuehl, S., Bernheim, D., & Ockenfels, A. (2019). Projective Paternalism. Working Paper.

Andreoni, J., & Vesterlund, L. (2001). Which is the fair sex? Gender differences in altruism. *The Quarterly Journal of Economics*, *116*(1), 293-312.

Andreozzi, L., Ploner, M., & Soraperra, I. (2013). *Justice among strangers. On altruism, inequality aversion and fairness* (No. 1304). Cognitive and Experimental Economics Laboratory, Department of Economics, University of Trento, Italia.

Asch, S. E. (1956). Studies of independence and conformity: I. A minority of one against a unanimous majority. *Psychological Monographs: General and Applied*, *70*(9), 1.

Azrieli, Y., Chambers, C. P., & Healy, P. J. (2018). Incentives in experiments: A theoretical analysis. *Journal of Political Economy*, *126*(4), 1472-1503.

Baer, J. S., & Carney, M. M. (1993). Biases in the perceptions of the consequences of alcohol use among college students. *Journal of Studies on Alcohol*, *54*(1), 54-60.

Baer, J. S., Stacy, A., & Larimer, M. (1991). Biases in the perception of drinking norms among college students. *Journal of Studies on Alcohol*, *52*(6), 580-586.

Bagozzi, R. P., & Warshaw, P. R. (1990). Trying to consume. *Journal of Consumer Research*, *17*(2), 127-140.

Baillon, A. (2017). Bayesian markets to elicit private information. *Proceedings of the National Academy of Sciences*, *114*(30), 7958-7962.

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

Baron, J., & Hershey, J. C. (1988). Outcome bias in decision evaluation. *Journal of Personality and Social Psychology*, 54(4), 569.

Barr, A., Lane, T., & Nosenzo, D. (2018). On the social inappropriateness of discrimination. *Journal of Public Economics*, *164*, 153-164.

Barrage, L., & Lee, M. S. (2010). A penny for your thoughts: Inducing truth-telling in stated preference elicitation. *Economics Letters*, *106*(2), 140-142.

Bartling, B., & Fischbacher, U. (2012). Shifting the blame: On delegation and responsibility. *The Review of Economic Studies*, *79*(1), 67-87.

Bastek, C., Klaus, B., & Kübler, D. (2018). How lotteries in school choice help to level the playing field (No. SP II 2018-205). WZB Discussion Paper.

Bellemare, C., Sebald, A., & Strobel, M. (2011). Measuring the willingness to pay to avoid guilt: estimation using equilibrium and stated belief models. *Journal of Applied Econometrics*, *26*(3), 437-453.

Bénabou, R., & Tirole, J. (2006). Incentives and prosocial behavior. *American Economic Review*, *96*(5), 1652-1678.

Benndorf, V., Moellers, C., & Normann, H. T. (2017). Experienced vs. inexperienced participants in the lab: do they behave differently?. *Journal of the Economic Science Association*, *3*(1), 12-25.

Bennett, R. (1999). Sports sponsorship, spectator recall and false consensus. *European Journal of Marketing*, *33*(3/4), 291-313.

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

Bicchieri, C., & Dimant, E. (2019). Nudging with care: The risks and benefits of social information. *Public Choice, Forthcoming*.

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

Binmore, K., & Samuelson, L. (2006). The evolution of focal points. *Games and Economic Behavior*, *55*(1), 21-42.

Birnbaum, A., Lord, F. M., & Novick, M. R. (1968). Statistical theories of mental test scores. *Some latent trait models and their use in inferring an examinee's ability. Addison-Wesley, Reading, MA.*

Blanco, M., Engelmann, D., Koch, A. K., & Normann, H. T. (2014). Preferences and beliefs in a sequential social dilemma: a within-subjects analysis. *Games and Economic Behavior*, *87*, 122-135.

Blanton, H., Köblitz, A., & McCaul, K. D. (2008). Misperceptions about norm misperceptions: Descriptive, injunctive, and affective 'social norming'efforts to change health behaviors. *Social and Personality Psychology Compass*, *2*(3), 1379-1399.

Bock, O., Baetge, I., & Nicklisch, A. (2014). hroot: Hamburg registration and organization online tool. *European Economic Review*, 71, 117-120.

Bolton, G. E., & Ockenfels, A. (2000). ERC: A theory of equity, reciprocity, and competition. *American Economic Review*, *90*(1), 166-193.

Bolton, G. E., Brandts, J. & Ockenfels, A. (2005). Fair procedures: Evidence from games involving lotteries. *Economic Journal*, *115,* 1054-1076.

Borsari, B., & Carey, K. B. (2003). Descriptive and injunctive norms in college drinking: a meta-analytic integration. *Journal of Studies on Alcohol*, *64*(3), 331-341.

Boyle, C. (1998). Organizations selecting people: how the process could be made fairer by the appropriate use of lotteries. *Journal of the Royal Statistical Society: Series D (The Statistician)*, *47*(2), 291-321.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly weather review*, *78*(1), 1-3.

Brock, J. M., Lange, A., & Ozbay, E. Y. (2013). Dictating the risk: Experimental evidence on giving in risky environments. *American Economic Review*, *103*(1), 415-37.

Brockner, J. (2002). Making sense of procedural fairness: How high procedural fairness can reduce or heighten the influence of outcome favorability. *Academy of Management Review*, *27*(1), 58-76.

Broome, J. (1990, January). Fairness. In *Proceedings of the Aristotelian Society* (Vol. 91, pp. 87-101). Aristotelian Society, Wiley.

Burks, S. V., & Krupka, E. L. (2012). A multimethod approach to identifying norms and normative expectations within a corporate hierarchy: Evidence from the financial services industry. *Management Science*, *58*(1), 203-217.

Camerer, C. F., Ho, T. H., & Chong, J. K. (2004). A cognitive hierarchy model of games. *The Quarterly Journal of Economics*, *119*(3), 861-898.

Carey, K. B., Borsari, B., Carey, M. P., & Maisto, S. A. (2006). Patterns and importance of self-other differences in college drinking norms. *Psychology of Addictive Behaviors*, *20*(4), 385.

Carvalho, A., Dimitrov, S., & Larson, K. (2017). Inducing honest reporting of private information in the presence of social projection. *Decision*, *4*(1), 25.

Casajus, A. (2000). Focal points in framed strategic forms. *Games and Economic Behavior*, *32*(2), 263-291.

Casari, M., Ham, J. C., & Kagel, J. H. (2007). Selection bias, demographic effects, and ability effects in common value auction experiments. *American Economic Review*, *97*(4), 1278-1304.

Charness, G., & Grosskopf, B. (2001). Relative payoffs and happiness: an experimental study. *Journal of Economic Behavior & Organization*, *45*(3), 301-328.

Chen, Y. & Li, S. (2009). Group Identity and Social Preferences. *American Economic Review* 99, 431-457.

Choshen-Hillel, S., Shaw, A., & Caruso, E. M. (2015). Waste management: How reducing partiality can promote efficient resource allocation. *Journal of Personality and Social Psychology*, *109*(2), 210.

Cialdini, R. B. (2003). Crafting normative messages to protect the environment. *Current Directions in Psychological Science*, *12*(4), 105-109.

Cialdini, R. B., Demaine, L. J., Sagarin, B. J., Barrett, D. W., Rhoads, K., & Winter, P. L. (2006). Managing social norms for persuasive impact. *Social Influence*, *1*(1), 3-15.

Cialdini, R. B., Kallgren, C. A., & Reno, R. R. (1991). A focus theory of normative conduct: A theoretical refinement and reevaluation of the role of norms in human behavior. In *Advances in Experimental Social Psychology* (Vol. 24, pp. 201-234). Academic Press.

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.

Cleave, B. L., Nikiforakis, N., & Slonim, R. (2013). Is there selection bias in laboratory experiments? The case of social and risk preferences. *Experimental Economics*, *16*(3), 372-382.

Coffman, L. C., & Real, A. G. (2018). Moral Perception of Advised Actions. Working paper, Harvard University.

Cole, H. L., Mailath, G. J., & Postlewaite, A. (1992). Social norms, savings behavior, and growth. *Journal of Political Economy*, *100*(6), 1092-1125.

Coleman, J. (1990). Foundations of Social Theory. Cambridge, Mass.: Belknap Press of Harvard University Press.

Crawford, S. E., & Ostrom, E. (1995). A grammar of institutions. *American Political Science Review*, *89*(3), 582-600.

Crawford, V. P., Gneezy, U., & Rottenstreich, Y. (2008). The power of focal points is limited: Even minute payoff asymmetry may yield large coordination failures. *American Economic Review*, *98*(4), 1443-58.

Cremer, J., & McLean, R. P. (1988). Full extraction of the surplus in Bayesian and dominant strategy auctions. *Econometrica: Journal of the Econometric Society*, 1247-1257.

Dana, J., Weber, R. A., & Kuang, J. X. (2007). Exploiting moral wiggle room: experiments demonstrating an illusory preference for fairness. *Economic Theory*, *33*(1), 67-80.

d'Aspremont, C., & Gérard-Varet, L. A. (1979). Incentives and incomplete information. *Journal of Public Economics*, *11*(1), 25-45.

Dawes, R. M. (1989). Statistical criteria for establishing a truly false consensus effect. *Journal of Experimental Social Psychology*, *25*(1), 1-17.

Dawes, R. M. (1990). The potential nonfalsity of the false consensus effect.

Dickinson, D. L., Masclet, D., & Peterle, E. (2018). Discrimination as favoritism: The private benefits and social costs of in-group favoritism in an experimental labor market. *European Economic Review*, *104*, 220-236.

Dong, L. & Huang, L. (2018). Favoritism and Fairness in Teams. *Games* 9, article 65.

Dwenger, N., Kübler, D., & Weizsäcker, G. (2018). Flipping a coin: Evidence from university applications. *Journal of Public Economics*, *167*, 240-250.

Dyreng, S. D., Mayew, W. J., & Williams, C. D. (2012). Religious social norms and corporate financial reporting. *Journal of Business Finance & Accounting*, *39*(7-8), 845-875.

Eckel, C. C., & Grossman, P. J. (2000). Volunteers and pseudo-volunteers: The effect of recruitment method in dictator experiments. *Experimental Economics*, *3*(2), 107-120.

Eckel, C. C., & Grossman, P. J. (2008). Forecasting risk attitudes: An experimental study using actual and forecast gamble choices. *Journal of Economic Behavior & Organization*, *68*(1), 1-17.

Edgeworth, F. Y. (1890). The element of chance in competitive examinations. *Journal of the Royal Statistical Society*, *53*(4), 644-663.

Elek, E., Miller-Day, M., & Hecht, M. L. (2006). Influences of personal, injunctive, and descriptive norms on early adolescent substance use. *Journal of Drug Issues*, *36*(1), 147-172.

Ellingsen, T., Johannesson, M., Tjøtta, S., & Torsvik, G. (2010). Testing guilt aversion. *Games and Economic Behavior*, *68*(1), 95-107.

Elster, J. (1989). Social norms and economic theory. *Journal of Economic Perspectives*, *3*(4), 99-117.

Engelmann, D., & Strobel, M. (2012). Deconstruction and reconstruction of an anomaly. *Games and Economic Behavior*, *76*(2), 678-689.

Epley, N., Keysar, B., Van Boven, L., & Gilovich, T. (2004). Perspective taking as egocentric anchoring and adjustment. *Journal of Personality and Social Psychology*, *87*(3), 327.

Falk, A., & Heckman, J. J. (2009). Lab experiments are a major source of knowledge in the social sciences. S*cience*, *326*(5952), 535-538.

Falk, A., Meier, S., & Zehnder, C. (2013). Do lab experiments misrepresent social preferences? The case of self-selected student samples. *Journal of the European Economic Association*, *11*(4), 839-852.

Fehr, D., Heinemann, F., & Llorente-Saguer, A. (2019). The power of sunspots: An experimental analysis. *Journal of Monetary Economics*.

Fehr, E., & Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. *The Quarterly Journal of Economics*, *114*(3), 817-868.

Ferrando, P. J. (2012). Assessing the discriminating power of item and test scores in the linear factor-analysis model. *Psicológica*, *33*(1), 111-134.

Fischbacher, U. (2007). z-Tree: Zurich toolbox for ready-made economic experiments. *Experimental Economics*, *10*(2), 171-178.

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

Gächter, S., Nosenzo, D., & Sefton, M. (2013). Peer effects in pro-social behavior: Social norms or social preferences?. *Journal of the European Economic Association*, *11*(3), 548-573.

Galizzi, M. M., & Navarro-Martínez, D. (2018). On the external validity of social preference games: a systematic lab-field study. *Management Science*.

Garonzik, R., Brockner, J., & Siegel, P. A. (2000). Identifying international assignees at risk for premature departure: The interactive effect of outcome favorability and procedural fairness. *Journal of Applied Psychology*, *85*(1), 13.

Gerber, A. S., & Rogers, T. (2009). Descriptive social norms and motivation to vote: Everybody's voting and so should you. *The Journal of Politics*, *71*(1), 178-191.

Göckeritz, S., Schultz, P. W., Rendón, T., Cialdini, R. B., Goldstein, N. J., & Griskevicius, V. (2010). Descriptive normative beliefs and conservation behavior: The moderating roles of personal involvement and injunctive normative beliefs. *European Journal of Social Psychology*, *40*(3), 514-523.

Goeschl, T., Kettner, S., Lohse, J., & Schwieren, C. (2018). From Social Information to Social Norms: Evidence from Two Experiments on Donation Behaviour. *Games*, *9*(4), 91.

Goldstein, N. J., Cialdini, R. B., & Griskevicius, V. (2008). A room with a viewpoint: Using social norms to motivate environmental conservation in hotels. *Journal of consumer Research*, *35*(3), 472-482.

Greiner, B. (2015). Subject pool recruitment procedures: organizing experiments with ORSEE. *Journal of the Economic Science Association*, *1*(1), 114-125.

Grosch, K., & Rau, H. A. (2017). Do discriminatory pay regimes unleash antisocial behavior? Working paper, Göttingen.

Guillén, P., & Veszteg, R. F. (2012). On “lab rats”. *The Journal of Socio-Economics*, *41*(5), 714-720.

Guiso, L., Jappelli, T., & Terlizzese, D. (1992). Earnings uncertainty and precautionary saving. *Journal of Monetary Economics*, *30*(2), 307-337.

Guiso, L., Jappelli, T., & Terlizzese, D. (1996). Income risk, borrowing constraints, and portfolio choice. *The American Economic Review*, 158-172.

Güth, W., Schmittberger, R., & Schwarze, B. (1982). An experimental analysis of ultimatum bargaining. *Journal of Economic Behavior & Organization*, *3*(4), 367-388.

Hankins, M. (2007). Questionnaire discrimination:(re)-introducing coefficient δ. *BMC Medical Research Methodology*, *7*(1), 19.

Heinemann, F., Nagel, R., & Ockenfels, P. (2009). Measuring strategic uncertainty in coordination games. *The Review of Economic Studies*, *76*(1), 181-221.

Hertwig, R., & Engel, C. (2016). Homo ignorans: Deliberately choosing not to know. *Perspectives on Psychological Science*, *11*(3), 359-372.

Hong, F., Riyanto, Y., & Zhang, R. (2016). *Multidimensional group identity and redistributive allocation: An experimental study*. Economic Growth Centre Working Paper Series 1701, Nanyang Technological University, School of Social Sciences, Economic Growth Centre.

Howie, P. J., Wang, Y., & Tsai, J. (2011). Predicting new product adoption using Bayesian truth serum. *Journal of Medical Marketing*, *11*(1), 6-16.

Isoni, A., Poulsen, A., Sugden, R., & Tsutsui, K. (2013). Focal points in tacit bargaining problems: Experimental evidence. *European Economic Review*, *59*, 167-188.

Isoni, A., Poulsen, A., Sugden, R., & Tsutsui, K. (2014). Efficiency, equality, and labeling: An experimental investigation of focal points in explicit bargaining. *American Economic Review*, *104*(10), 3256-87.

Isoni, A., Poulsen, A., Sugden, R., & Tsutsui, K. (2019). Focal points and payoff information in tacit bargaining. *Games and Economic Behavior*.

Janssen, M. C. (2001). Rationalizing focal points. *Theory and Decision*, *50*(2), 119-148.

Janssen, M. C. (2006). On the strategic use of focal points in bargaining situations. *Journal of Economic Psychology*, *27*(5), 622-634.

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological science*, *23*(5), 524-532.

Johnson, S., Pratt, J. W., & Zeckhauser, R. J. (1990). Efficiency despite mutually payoff-relevant private information: The finite case. *Econometrica: Journal of the Econometric Society*, 873-900.

Jurca, R., & Faltings, B. (2006). Minimum payments that reward honest reputation feedback. In *Proceedings of the 7th ACM conference on Electronic commerce* (pp. 190-199). ACM.

Kallgren, C. A., Reno, R. R., & Cialdini, R. B. (2000). A focus theory of normative conduct: When norms do and do not affect behavior. *Personality and Social Psychology Bulletin*, *26*(8), 1002-1012.

Keren, G., & Teigen, K. H. (2010). Decisions by coin toss: Inappropriate but fair. *Judgment and Decision Making*, *5*(2), 83.

Kessler, J. B., & Leider, S. (2016). Procedural fairness and the cost of control. *The Journal of Law, Economics, and Organization*, *32*(4), 685-718.

Keynes, J. M., 1936, The General Theory of Employment, Interest and Money, Macmillan, London.

Khwaja, A., Silverman, D., Sloan, F., & Wang, Y. (2009). Are mature smokers misinformed?. *Journal of Health Economics*, *28*(2), 385-397.

Khwaja, A., Sloan, F., & Salm, M. (2006). Evidence on preferences and subjective beliefs of risk takers: The case of smokers. *International Journal of Industrial Organization*, *24*(4), 667-682.

Kimbrough, E. O., & Vostroknutov, A. (2016). Norms make preferences social. *Journal of the European Economic Association*, *14*(3), 608-638.

Knez, M., & Camerer, C. (1994). Creating expectational assets in the laboratory: coordination in ‘weakest-link’games. *Strategic Management Journal*, *15*(S1), 101-119.

König-Kersting, C., & Trautmann, S. T. (2018). Countercyclical risk aversion: Beyond financial professionals. *Journal of Behavioral and Experimental Finance*, *18*, 94-101.

Konow, J. (2003). Which is the fairest one of all? A positive analysis of justice theories. *Journal of Economic Literature, 41*(4), 1188-1239.

Krawczyk, M. (2010). A glimpse through the veil of ignorance: Equality of opportunity and support for redistribution. *Journal of Public Economics*, *94*(1-2), 131-141.

Krawczyk, M. (2011). What brings your subjects to the lab? A field experiment. *Experimental Economics*, *14*(4), 482-489.

Krawczyk, M. W. (2011). A model of procedural and distributive fairness. *Theory and decision*, *70*(1), 111-128.

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

Larimer, M. E., Turner, A. P., Mallett, K. A., & Geisner, I. M. (2004). Predicting drinking behavior and alcohol-related problems among fraternity and sorority members: examining the role of descriptive and injunctive norms. *Psychology of Addictive Behaviors*, *18*(3), 203.

Lazear, E. P., Malmendier, U., & Weber, R. A. (2012). Sorting in experiments with application to social preferences. *American Economic Journal: Applied Economics*, *4*(1), 136-63.

Lee, C. M., Geisner, I. M., Lewis, M. A., Neighbors, C., & Larimer, M. E. (2007). Social motives and the interaction between descriptive and injunctive norms in college student drinking. *Journal of Studies on Alcohol and Drugs*, *68*(5), 714-721.

Levitt, S. D., & List, J. A. (2007). What do laboratory experiments measuring social preferences reveal about the real world?. *Journal of Economic perspectives*, *21*(2), 153-174.

Levitt, S. D., & List, J. A. (2008). Homo economicus evolves. *Science*, *319*(5865), 909-910.

Li, W. (2007). Changing one's mind when the facts change: incentives of experts and the design of reporting protocols. *The Review of Economic Studies*, *74*(4), 1175-1194.

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

Loevinger, J. (1954). The attenuation paradox in test theory. *Psychological Bulletin*, *51*(5), 493.

Loewenstein, G. (1994). The psychology of curiosity: A review and reinterpretation. *Psychological Bulletin*, *116*(1), 75.

Loughran, T. A., Paternoster, R., & Thomas, K. J. (2014). Incentivizing responses to self-report questions in perceptual deterrence studies: An investigation of the validity of deterrence theory using Bayesian truth serum. *Journal of Quantitative Criminology*, *30*(4), 677-707.

Luce, R. D., & Raiffa, H. (1957). Games & Decisions. John Willey & Sons. *Inc., New York*.

Lumsden, J. (1976). Test theory. *Annual review of psychology*, *27*(1), 251-280.

Mair, J., & Bergin-Seers, S. (2010). The effect of interventions on the environmental behaviour of Australian motel guests. *Tourism and Hospitality Research*, *10*(4), 255-268.

Manning, M. (2009). The effects of subjective norms on behaviour in the theory of planned behaviour: A meta-analysis. *British Journal of Social Psychology*, *48*(4), 649-705.

Manski, C. F. (2004). Measuring expectations. *Econometrica*, *72*(5), 1329-1376.

Marks, G., & Miller, N. (1987). Ten years of research on the false-consensus effect: An empirical and theoretical review. *Psychological Bulletin*, *102*(1), 72.

Matthey, A., & Regner, T. (2013). On the independence of history: experience spill-overs between experiments. *Theory and Decision*, *75*(3), 403-419.

McAfee, R. P., & Reny, P. J. (1992). Correlated information and mecanism design. *Econometrica: Journal of the Econometric Society*, 395-421.

McLean, R., & Postlewaite, A. (2002). Informational size and incentive compatibility. *Econometrica*, *70*(6), 2421-2453.

Mehta, J., Starmer, C., & Sugden, R. (1992). An experimental investigation of focal points in coordination and bargaining: some preliminary results. In *Decision Making under Risk and Uncertainty* (pp. 211-219). Springer, Dordrecht.

Mehta, J., Starmer, C., & Sugden, R. (1994a). Focal points in pure coordination games: An experimental investigation. *Theory and Decision*, *36*(2), 163-185.

Mehta, J., Starmer, C., & Sugden, R. (1994b). The nature of salience: An experimental investigation of pure coordination games. *The American Economic Review*, *84*(3), 658-673.

Merton, R. K. (1957). Priorities in scientific discovery: a chapter in the sociology of science. *American Sociological Review*, *22*(6), 635-659.

Miao, B., & Zhong, S. (2018). Probabilistic social preference: how Machina’s Mom randomizes her choice. *Economic Theory*, *65*(1), 1-24.

Miller, N., Resnick, P., & Zeckhauser, R. (2005). Eliciting informative feedback: The peer-prediction method. *Management Science*, *51*(9), 1359-1373.

Monroe, K. B. (1973). Buyers’ subjective perceptions of price. *Journal of Marketing Research*, *10*(1), 70-80.

Moorman, R. H. (1991). Relationship between organizational justice and organizational citizenship behaviors: Do fairness perceptions influence employee citizenship?.*Journal of Applied Psychology*, *76*(6), 845.

Mullen, B., Atkins, J. L., Champion, D. S., Edwards, C., Hardy, D., Story, J. E., & Vanderklok, M. (1985). The false consensus effect: A meta-analysis of 115 hypothesis tests. *Journal of Experimental Social Psychology*, *21*(3), 262-283.

Nagel, R. (1995). Unraveling in guessing games: An experimental study. *The American Economic Review*, *85*(5), 1313-1326.

Neighbors, C., O'Connor, R. M., Lewis, M. A., Chawla, N., Lee, C. M., & Fossos, N. (2008). The relative impact of injunctive norms on college student drinking: The role of reference group. *Psychology of Addictive Behaviors*, *22*(4), 576.

Nikiforakis, N., & Slonim, R. (2015). Editors' preface: introducing JESA. *Journal of the Economic Science Association*, *1*(1), 1-7.

Nolan, J. M., Schultz, P. W., Cialdini, R. B., Goldstein, N. J., & Griskevicius, V. (2008). Normative social influence is underdetected. *Personality and Social Psychology Bulletin*, *34*(7), 913-923.

Organ, D. W., & Ryan, K. (1995). A meta-analytic review of attitudinal and dispositional predictors of organizational citizenship behavior. *Personnel Psychology*, *48*(4), 775-802.

Ostrom, E. (2000). Collective action and the evolution of social norms. *Journal of Economic Perspectives*, *14*(3), 137-158.

Perner, J., & Wimmer, H. (1985). "John thinks that Mary thinks that…" attribution of second-order beliefs by 5-to 10-year-old children. *Journal of Experimental Child Psychology, 39*(3), 437-471.

Piketty, Thomas, Capital in the 21stCentury, Harvard University Press, 2014.

Pope, D. G., Pope, J. C., & Sydnor, J. R. (2015). Focal points and bargaining in housing markets. *Games and Economic Behavior, 93*, 89-107.

Prelec, D. (2004). A Bayesian truth serum for subjective data. *science*, *306*(5695), 462-466.

Prelec, D., & Seung, S. (2006). An algorithm that finds truth even if most people are wrong. *Unpublished manuscript*, *69*.

Prelec, D., & Weaver, R. G. (2006). Truthful answers are surprisingly common: Experimental tests of the bayesian truth serum. In *Proceedings of the Conference on Econometrics and Experimental Economics (CEEE'06)*.

Price, V., & Neijens, P. (1997). Opinion quality in public opinion research. *International Journal of Public Opinion Research*, *9*(4), 336-360.

Rabin, M. (1993). Incorporating fairness into game theory and economics. *The American Economic Review*, 1281-1302.

Radanovic, G., & Faltings, B. (2013). A robust bayesian truth serum for non-binary signals. In *Twenty-Seventh AAAI Conference on Artificial Intelligence*.

Radanovic, G., & Faltings, B. (2014). Incentives for truthful information elicitation of continuous signals. In *Twenty-Eighth AAAI Conference on Artificial Intelligence*.

Reese, G., Loew, K., & Steffgen, G. (2014). A towel less: Social norms enhance pro-environmental behavior in hotels. *The Journal of Social Psychology*, *154*(2), 97-100.

Reno, R. R., Cialdini, R. B., & Kallgren, C. A. (1993). The transsituational influence of social norms. *Journal of Personality and Social Psychology*, *64*(1), 104.

Rivis, A., & Sheeran, P. (2003). Descriptive norms as an additional predictor in the theory of planned behaviour: A meta-analysis. *Current Psychology*, *22*(3), 218-233.

Robin, S., Rusinowska, A., & Villeval, M. C. (2012). Ingratiation and favoritism: Experimental evidence.

Ross, L., Greene, D., & House, P. (1977). The “false consensus effect”: An egocentric bias in social perception and attribution processes. *Journal of Experimental Social Psychology, 13*(3), 279-301.

Saito, K. (2013). Social preferences under risk: Equality of opportunity versus equality of outcome. *American Economic Review*, *103*(7), 3084-3101.

Schelling, T. C. (1960). The Strategy of Conflict. Harvard university press.

Schlag, K. H., Tremewan, J., & Van der Weele, J. J. (2015). A penny for your thoughts: A survey of methods for eliciting beliefs. *Experimental Economics*, *18*(3), 457-490.

Schmidt, R. J. (2019a). Point Beauty Contest: Measuring the Distribution of Focal Points on the Individual Level, University of Heidelberg: AWI Discussion Paper Series Nr. 667.

Schmidt, R. J. (2019b). Do Injunctive or Descriptive Social Norms Elicited Using Coordination Games Better Explain Social Preferences?, University of Heidelberg: AWI Discussion Paper Series Nr. 668.

Schmidt, R. J. (2019c). Capitalizing on the (False) Consensus Effect: Two Tractable Methods to Elicit Private Information, University of Heidelberg: AWI Discussion Paper Series Nr. 669.

Schoenbaum, M. (1997). Do smokers understand the mortality effects of smoking? Evidence from the Health and Retirement Survey. *American Journal of Public Health*, *87*(5), 755-759.

Schotter, A., & Trevino, I. (2014). Belief elicitation in the laboratory. *Annu. Rev. Econ.*, *6*(1), 103-128.

Schultz, P. W. (1999). Changing behavior with normative feedback interventions: A field experiment on curbside recycling. *Basic and Applied Social Psychology*, *21*(1), 25-36.

Schultz, W. P., Khazian, A. M., & Zaleski, A. C. (2008). Using normative social influence to promote conservation among hotel guests. *Social Influence*, *3*(1), 4-23.

Selten, R. (1967). Die Strategiemethode zur Erforschung des eingeschränkt rationalen Verhaltens im Rahmen eines Oligopolexperiments, S. 136–168. *Tübingen: JCB Mohr (Paul Siebeck).*

Shaw, A. D., Horton, J. J., & Chen, D. L. (2011). Designing incentives for inexpert human raters. In *Proceedings of the ACM 2011 conference on Computer supported cooperative work* (pp. 275-284). ACM.

Shaw, A., & Olson, K. (2014). Fairness as partiality aversion: The development of procedural justice. *Journal of Experimental Child Psychology, 119*, 40-53.

Sherif, M. (1936). The Psychology of Social Norms. Harper and Row, New York.

Skarlicki, D. P., & Folger, R. (1997). Retaliation in the workplace: The roles of distributive, procedural, and interactional justice. *Journal of Applied Psychology, 82*(3), 434.

Slonim, R., Wang, C., Garbarino, E., & Merrett, D. (2013). Opting-in: Participation bias in economic experiments. *Journal of Economic Behavior & Organization*, *90*, 43-70.

Slovic, P., Fischhoff, B., & Lichtenstein, S. (1980). Facts and fears: Understanding perceived risk. In *Societal Risk Assessment* (pp. 181-216). Springer, Boston, MA.

Smith, J. R., Louis, W. R., Terry, D. J., Greenaway, K. H., Clarke, M. R., & Cheng, X. (2012). Congruent or conflicted? The impact of injunctive and descriptive norms on environmental intentions. *Journal of Environmental Psychology*, *32*(4), 353-361.

Smith, V. L. (1976). Experimental economics: Induced value theory. *The American Economic Review*, *66*(2), 274-279.

Stahl II, D. O., & Wilson, P. W. (1994). Experimental evidence on players' models of other players. *Journal of Economic Behavior & Organization*, *25*(3), 309-327.

Stahl, D. O., & Wilson, P. W. (1995). On players′ models of other players: Theory and experimental evidence. *Games and Economic Behavior, 10*(1), 218-254.

Stutzer, A., & Lalive, R. (2004). The role of social work norms in job searching and subjective well-being. *Journal of the European Economic Association*, *2*(4), 696-719.

Sugden, R. (1995). A theory of focal points. *The Economic Journal*, *105*(430), 533-550.

Sugden, R., & Zamarrón, I. E. (2006). Finding the key: the riddle of focal points. *Journal of Economic Psychology*, *27*(5), 609-621.

Thaler, R. H., & Sunstein, C. R. (2009). *Nudge: Improving decisions about health, wealth, and happiness*. Penguin.

Toussaert, S. (2018). Eliciting Temptation and Self‐Control Through Menu Choices: A Lab Experiment. *Econometrica*, *86*(3), 859-889.

Trautmann, S. T. (2009). A tractable model of process fairness under risk. *Journal of Economic Psychology*, *30*(5), 803-813.

Trautmann, S. T., & van de Kuilen, G. (2014). Belief elicitation: A horse race among truth serums. *The Economic Journal*, *125*(589), 2116-2135.

Trautmann, S. T., & van de Kuilen, G. (2016). Process fairness, outcome fairness, and dynamic consistency: Experimental evidence for risk and ambiguity. *Journal of Risk and Uncertainty*, *53*(2-3), 75-88.

Trautmann, S. T., & Wakker, P. P. (2010). Process fairness and dynamic consistency. *Economics Letters*, *109*(3), 187-189.

Turner, C., & Martin, E. (1985). *Surveying subjective phenomena* (Vol. 2). Russell Sage Foundation.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*(4157), 1124-1131.

Vanberg, C. (2019). A short note on the rationality of the false consensus effect, University of Heidelberg: AWI Discussion Paper Series No. 662.

Veenhoven, R. (2002). Why social policy needs subjective indicators. *Social Indicators Research*, *58*(1-3), 33-46.

Walker, D. D., Neighbors, C., Rodriguez, L. M., Stephens, R. S., & Roffman, R. A. (2011). Social norms and self-efficacy among heavy using adolescent marijuana smokers. *Psychology of Addictive Behaviors*, *25*(4), 727.

Witkowski, J., & Parkes, D. C. (2012a). A robust bayesian truth serum for small populations. In *Twenty-Sixth AAAI Conference on Artificial Intelligence*.

Witkowski, J., & Parkes, D. C. (2012b). Peer prediction without a common prior. In *Proceedings of the 13th ACM Conference on Electronic Commerce* (pp. 964-981). ACM.

Zizzo, D. J. (2010). Experimenter demand effects in economic experiments. *Experimental Economics*, *13*(1), 75-98.