No. 113

THURGAU INSTITUTE
OF ECONOMICS
at the University of Konstanz

Irenaeus Wolff

If I Don't Trust Your Preferences, I Won't Follow Mine: Preference Stability, Beliefs, and Strate

Research Paper Series
Thurgau Institute of Economics
and Department of Economics
at the University of Konstanz

Member of

thurgauwissenschaft

www.thurgau-wissenschaft.ch

If I Don't Trust Your Preferences, I Won't Follow Mine: Preference Stability, Beliefs, and Strategic Choice[§]

Irenaeus Wolff

Thurgau Institute of Economics (TWI) / University of Konstanz
Hauptstrasse 90, 8280 Kreuzlingen, Switzerland
wolff@twi-kreuzlingen.ch

Abstract:
In contrast to standard theory, experimental participants often do not best-respond to their stated beliefs. Potential reasons are inaccurate belief reports or unstable preferences. Focusing on games in which participants can observe the revealed preferences of their opponents, this paper points out an additional reason for the lack of belief-action consistency. Whether a participant's best-response—or a Nash-equilibrium—predicts her behaviour depends heavily on the participant believing in others' preference stability. Believing in others' preference stability fosters predictability because it is associated with a lower variance in the participant's belief about her opponents' actions, and low-variance beliefs entail more best-responding.

Keywords: Preference stability, best-response, Nash-equilibrium, rational beliefs, public good, social dilemma, conditional cooperation, social preferences.
JEL: C72, C92, D83, H41

1 Introduction

We probably all know people who's behaviour can be categorised as 'inconsistent' or even 'erratic' (and to come up with a few names, we probably do not even have to think about certain politicians). But how much inconsistent behaviour will we expect from the next few people we get to know? As this paper shows, this is an

[§]I am grateful to Urs Fischbacher, Alexander K. Wagner, and the research group at the Thurgau Institute of Economics (TWI) for helpful comments all along the way. I thank also Gary Charness, Simon Gächter, Wieland Müller, Dirk Sliwka, Marie Claire Villeval, and Roberto Weber for very fruitful discussions on the project, as well as Jordi Brandts, Sebastian Fehrler, Botond Köszegi, Rachel Kranton, Heinrich Nax, Louis Putterman, Jean-Robert Tyran, and Ro'i Zultan for their comments on earlier drafts. Thanks go further to seminar audiences at the Universities of Vienna, Innsbruck, and Nottingham, the participants of the ESA European Meetings 2012 and 2015, the 15th International Conference on Social Dilemmas, and the EEA-ESEM meeting 2017 for the many fruitful discussions. Funding by SNF grant 100018_152788 is gratefully acknowledged.

important question: the degree of inconsistency we expect from others plays an important role when we interact with them, as it determines whether we will be acting on our own beliefs.

In economic theory, agents form beliefs and best-respond to them. Experiments have shown, however, that people do not always best-respond to their beliefs. This paper shows in two different settings—a standard public-good setting and an asset-market setting—that believing in others' preference stability plays an important role in playing a best-response to one's beliefs. *A priori*, this relationship is surprising: once I have formed a belief about my opponents' actions, my behaviour should no longer depend on my beliefs about anybody's preference stability. Best-responding to my belief about everybody else's strategies is the best I can do, no matter how I came up with my belief. However, if I believe others have stable preferences, my uncertainty decreases about what these others will do. And if I put more faith in what I think the others will do, I will best-respond to my belief more often.[1]

Interestingly, the relationship between the belief in others' preference stability and best-responding holds also when we ask for people's belief about the preference stability of *unrelated* individuals. This holds true because the belief about others' preference stability seems to be a stable personal characteristic. Thus, a person's belief about the preference stability of unrelated individuals is correlated with her belief about her opponents' preference stability.

The relationship also carries over to the question of whether people play equilibrium actions. What we have, in the end, is that whether you believe in the preference stability of person A or B predicts whether you will play an equilibrium of the game when interacting with person C. This means that we can identify *ex ante* whether someone's behaviour is modelled well by a Nash equilibrium: we only have to measure whether this person believes in others' preference stability (in addition to measuring whether the person displays stable preference herself).

This result is important because it informs our understanding of people's behaviour. Often, we do not know whether to model people's behaviour using an equilibrium concept or by some other approach. Take the example of public-good experiments: Ambrus and Pathak (2011) suggest participants are playing an equilibrium. In contrast, Fischbacher and Gächter's (2010) results suggest that participants generally best-respond to their beliefs, but that their beliefs are overly optimistic (and participants update the beliefs suboptimally). So, even in an environment that is studied as intensely as public goods, we do not know by what kind of a model we should understand behaviour.

[1] This is in line with Wolff and Bauer (2018). There, we show in a completely different context that more belief uncertainty leads to a lower rate of belief-action consistency even when holding constant the costs of an error.

This paper provides an answer. The paper shows that the behaviour of some fractions of participants can be modelled well by an equilibrium and that of others cannot—and that we can identify these participants *ex ante*. For this purpose, I will categorise participants into four categories, according to whether they (i) display stable preferences themselves, and (ii) believe in others' preference stability.

Some of the participants who display stable preferences also believe in others' preference stability. These people show behaviour that is well predicted by an equilibrium based on their elicited preferences, and should be modelled that way.[2] However, they make up only for roughly one third of the participants for our main application. Participants who display stable preferences but do not believe in others' preference stability make up for another quarter of the population. These participants do not play a best-response to their beliefs even half of the time. To model their behaviour by standard game-theoretic approaches, we thus would have to accept that both their actions and their belief reports are very noisy.

But what about the remaining 44% whose elicited preferences are unstable? For those 13% who believe in others' preference stability, the best-response concept predicts remarkably well, even though we have to base it on preferences that are empirically unstable. However, for the remaining third of the population we will definitely have to search for a new description, unless we characterise them as players with extremely high noise parameters in models like heterogeneous quantal-response equilibrium or noisy introspection.

The results are important also from an applied perspective: for example, think of workers who used to face a boss who before his recent retirement got upset at every team meeting in which one of his workers would voice an objection to his new directives. Now, a new boss comes in, and suppose the workers have the feeling that the new boss is interested in their thoughts and opinions. The results of this paper suggest that if the workers generally do not believe in preference stability, they will not necessary act on their beliefs. Hence, the new boss may be having a hard time establishing a more cooperative atmosphere despite the favourable beliefs of her workers. As a potential remedy, she may first want to identify those who believe in stability the most, before encouraging these workers in particular to voice their opinions.

[2]Note that equilibria that are based on participants' elicited preferences need not coincide with the standard textbook solution for public-good protocols: for example, two strongly inequality-averse participants in a two-player setting would face a coordination rather than a public-good game. Thus, people need not have a dominant strategy and beliefs about one's opponents' strategies may matter also theoretically. For more details, see, *e.g.*, Fehr and Schmidt (1999), Weibull (2004, who also coined the term 'protocol'), or Wolff (2017, for a systematic elicitation of equilibrium sets for public-good protocols).

The experiments. In the main part of the paper, I use a linear public-good situation as a work-horse, because it is a well-studied situation and a prime example for a situation in which we know the material structure but we do not know people's preferences. As many other papers before, I focus on conditional-contribution preferences: how much people want to contribute to a public good depending on what others contribute.[3]

To assess preference stability, I elicit participants' conditional-contribution preferences three times throughout the experimental session.[4] To measure participants' beliefs about others' preference stability, I show participants other players' choices from the first preference-elicitation experiment and elicit their beliefs about how these others will act in the second run of the preference-elicitation experiment.

Finally, I want to relate participants' strategic behaviour to their preference stability and their beliefs about others' preference stability. To analyse their strategic behaviour, I let participants play the simultaneous public-good protocol, showing them the other player's choices from the first run of the preference-elicitation experiment, and ask them for their beliefs on the other player's action in current interaction. The idea is that we should be able to predict each participant's choice in a simultaneous interaction with a new other player, based on the participant's elicited preferences and her elicited (probabilistic) belief about the other player's action. Note that all belief-elicitation tasks ask for subjective probability distributions, not point beliefs.

In a second part of the paper, I examine whether the findings generalise to a different context, using an analogous experimental setup. Here, I use the investment task introduced by Gneezy and Potters (1997) to elicit risk preferences, and add a market context where it pays to invest into a risky option security only if you believe others have sufficiently stable risk preferences to buy similar amounts of market shares as on a first (observed) occasion. By and large, the results from the main part carry over to the investment-task environment.

As a precondition to the main contributions of the paper, I also study how others' observed choices affect what people expect these others to do on the next occasion.[5] By experimental design, I abstract from reputation concerns and rule

[3]Cf. the many references provided in Fischbacher and Gächter (2010), or Fischbacher et al. (2012). Conditional-cooperation preferences may be a type of social preferences in their own right, or a manifestation of underlying preferences, e.g., for reciprocity.

[4]Note how this is different to a repeated public-good setting even under random re-matching between rounds: in repeated public-good experiments, participants virtually always get feedback about other people's contributions that may induce them to change their own behaviour. In contrast, participants do not get feedback between the different parts in my experiment.

[5]This relates to the literature on social learning. However, in this literature, the focus is usually on learning about a common state of nature, rather than about the future behaviour of others.

out both learning about the game and strategic reasoning when eliciting participants' preferences. This approach enables me to isolate the effects of knowing about the other person's preferences on a person's choices and beliefs. Thereby, I contribute to a huge literature studying reactions to observations in repeated interactions when the behaviour participants observe is determined by strategic reasoning and reputation concerns (*e.g.*, Bohnet and Huck, 2004, Dal Bó, 2005, or Duffy and Ochs, 2009).

2 Related literature.

Traditionally, economists have talked about preferences in the sense of *revealed* preferences: observing a person choosing A when B is also available meant the person "prefers A to B". However at least since Luce (1958), ideas of randomness in choice have been introduced, in particular in studies of decision-making under risk. This randomness could come in one of several ways. First, a person might have a fixed utility function but make errors while evaluating options or choosing between them (*e.g.*, Harless and Camerer, 1994, Hey and Orme, 1994). In this case, the "*underlying* preferences" are stable but choices (revealed preferences) are often not. A second concept is that of random-utility models in which the parameters of the utility function used in a decision are realisations of an unobserved random process. In these models, even the underlying preferences are stable only insofar as the random process does not change over time (*e.g.*, Luce, 1958, Loomes and Sugden, 1995).

Third, in models of deliberate randomisation an agent's non-expected-utility preferences make her want to hedge between options (*e.g.*, Machina, 1985, Fudenberg et al., 2015). Agranov and Ortoleva (2017) argue that deliberate-randomisation models predict choices to be even more variable than random-utility models. The higher variability in choices comes from agents constantly mixing between hard-to-compare options even when they know they are facing the same situation several times in a row.[6] Fourth, in our public-good setup ideas like moral licensing and moral cleansing would be applicable (*e.g.*, Sachdeva et al., 2009). Under moral licensing, an agent would feel entitled to behaving opportunistically after performing a number of 'good acts'. Under moral cleansing, the agent feels the need to perform a 'good act' after behaving opportunistically. Thus, both moral licensing

[6]The auxiliary assumption is that in situations that are known to be repeated immediately, random-utility agents do not sample their parameters more than once and evaluation-error agents do not accummulate additional information. Looking at choice errors à la Harless and Camerer (1994), of course, this argument seems much weaker. However, such choice errors predict the same amount of variance for easy-to-compare options as for hard-to-compare options, in contradiction to the experimental evidence.

and moral cleansing would also predict unstable choice behaviour.

Fifth, instability in public-good behaviour has been attributed to confusion (*e.g.*, Andreoni, 1995, Bayer et al., 2013). In particular, Muller et al. (2008) as the only paper on short-term instability of conditional-contribution preferences, attribute instability to confusion. While most of the literature has assumed—implicitly or explicitly—that eliminating confusion would lead to lower contributions, Bayer et al. (2013) show that less confusion could also leave average contributions unaffected or even increase them. Finally, and closely related to confusion, studies like Plott (1996) and Cubitt et al. (2001) have suggested that people do not have direct access to their preferences in unknown situations. Instead, people have to discover the preferences by sampling actions and observing the resulting utilities.

The main point of this paper is that people's generalised *belief* in others' behavioural stability determines whether we can predict their behaviour. I want to argue that for this point, it does not matter which of the above concepts underlies behavioural instability, which is why I do not intend to take a strong stance on the issue. However, let me briefly discuss each approach in light of the present study.

The present paper wants to measure instability that is not due to confusion. To this end, I limit the scope for confusion by inviting only experienced participants, by choosing a simplified version of the game, and by using an interface geared to making the choices-payoffs relation as transparent as possible. In terms of the data, I observe that almost as many people switch preference types between measurements 2 and 3 as do between measurements 1 and 2 (40% *vs.* 43%).[7] This constant switching rate lends support to the idea that the instability is due to something more than just confusion or learning about one's preferences.

Consider next a preference for randomisation. Agranov and Ortoleva (2017) argue that deliberately randomising participants will treat all decisions separately, in contrast to, for example, random-utility agents. By this argument, we should expect a relatively high prevalence of non-monotonicity in conditional-contribution vectors because each conditional contribution schedule consists of six decisions. Given the six decisions, a randomising agent is very likely to display a non-monotonic vector even though randomisation will be unlikely for responses to very low first-mover contributions. By the third measurement, we see 22% non-monotonic vectors, two fifths of which are due to "triangle contributors" (who thus display a very systematic non-monotonicity; this systematic non-monotonicity, however, could also at least partially be the result of randomisation). In my view, the low frequency of non-monotonic vectors in the final elicitation speaks against deliberate randomisation driving preference-type switching rates of 40% and above. The

[7]Similarly, testing sums of squared differences between measurements 1 and 2 *vs.* measurements 2 and 3 by a Wilcoxon matched-pairs signed-ranks test yields $p = 0.547$.

same argument even more forcefully applies to choice errors, and potentially also to evaluation errors.

Finally, the observed behavioural instability could be due to moral licensing or to compliance with a random-utility model. Since switches between the selfish and the conditional-cooperator categories are very rare (and virtually fully restricted to imperfect conditional cooperators), the data seem to favour the random-utility model. Alternatively, several or even all of the suggested sources of instability could contribute to the overall degree of instability.

Recent years have brought forward a handful of studies measuring (revealed-) preference stability over weeks, months, and even years also in other domains. For the case of conditional-contribution preferences, Volk et al. (2012) find stable aggregate behaviour over five months, whereas Brosig et al. (2007) and Sass and Weimann (2012) find only selfish participants to exhibit preference stability.[8] Carlsson et al. (2014) find temporal and contextual stability of behaviour in public-good situations over six years in a non-student sample in rural Vietnam. Given the mixed findings, it could be that the time intervals in the above studies are too large. So far, Muller et al. (2008) is the only paper on short-term instability of conditional-contribution preferences. They find substantial instability, which they attribute to confusion.

Next to the question of preference stability, this study's focus on beliefs in stability brings to mind three concepts from game theory. First, Selten (1975) introduced the trembling-hand perfect equilibrium, followed by Myerson's (1978) proper equilibrium. While in both concepts, beliefs in other players' potential deviations from their equilibrium strategy play a role, both are predominantly equilibrium refinements. In equilibrium, beliefs will have collapsed to equilibrium beliefs.[9]

The other two concepts are McKelvey and Palfrey's (1995) quantal-response equilibrium and Goeree and Holt's (2004) noisy-introspection model. In both models, players play noisy best-replies to their beliefs, and players will take into account that others choices are noisy. To bring my results into accordance with these models, we need to assume (i) heterogeneity in players' noise parameters, and (ii) a correlation between the level of noise governing their own choice and the level of noise they expect in others' choices.[10] Both assumptions seem plausible, but to the best of my knowledge, neither has been discussed. Given that without the two assumptions, none of the models can make sense of the data I present, my paper

[8]Closely related, Bruhin et al. (2016) find stability of preferences for reciprocity and distributional concerns over three months.

[9]Note also that Selten (1975) saw the trembles as a technical device rather than as a suitable model for errors.

[10]Of course, in either case the model must be based on participants' true preferences rather than the monetary payoffs alone.

adds an important qualification to the existing models.

This paper brings together the instability aspect with a 'behavioural-validity' aspect. I elicit conditional-contribution preferences in a sequential public-good experiment to study strategic interactions in a simultaneous public-good experiment played afterwards. This rests on the assumption that the sequential experiment measures the preferences that are relevant for the simultaneous experiment. Fischbacher et al. (2012) establish the behavioural validity of elicited conditional-contribution preferences for the simultaneous protocol. Also, note that my main finding is that I can predict the behaviour of participants very well using their conditional-contribution preferences, as long as the participants have stable preferences and generally believe in others' preference stability. This finding lends additional support to the relevance of conditional-contribution preferences for the simultaneous public-good situation.

To the best of my knowledge, this is the first paper to link public-good behaviour to equilibria computed using participants' own elicited preferences.[11] Using participants' elicited preferences is important in this context because there is a large heterogeneity in the sets of equilibria for different participant matchings. I am also not aware of any studies explaining the match between actual and predicted behaviour by situation-unspecific variables other than confusion, cognitive abilities, or strategic sophistication. Therefore, documenting the link between participants' general beliefs in others' preference stability and game-theoretic behaviour is novel and important for our picture of the world.

Last but not least, there is a huge literature in psychology on attribution theories, relating to the inferences people draw about others from observing their behaviour.[12] This research seems to focus predominantly on the correspondence bias (which is closely related to the well-known fundamental attribution error), according to which people tend to underweight situational influences and overweight personality dispositions when evaluating others' behaviour. Interestingly, researchers do not seem to have identified proneness to a correspondence bias as a personality trait. This may be explained by the idea that—taken to the extreme—believing in true character traits as a researcher might be giving evidence of being prone to the fundamental attribution error oneself.[13] At the same time, the fact

[11]See the literature on the closely related prisoners'-dilemma protocols, though (*e.g.*, Hayashi et al., 1999 or Rubinstein and Salant, 2016). In this literature, of course, behaviour by construction cannot be as rich as in public-good games. For example, in a prisoners' dilemma, there cannot be any imperfect conditional cooperators or triangle contributors, two types that have been identified robustly in the public-good literature. In particular, Fischbacher and Gächter (2010) point to imperfect conditional cooperators as an important ingredient of the explanation of behaviour in repeated public-good experiments.

[12]Cf., e.g., Gilbert (2002) or Gawronski (2004).

[13]Indeed, there is such a debate, in particular with respect to virtues, cf. Harman (2009).

that people from different cultures display the bias to different degrees in a predictable way would suggest there may also be rather stable individual differences between people with the same cultural background.[14] This paper differs from the literature on the correspondence bias in three ways: in this paper, (i) there are no situational influences that could explain differences in beliefs in stability, (ii) the paper identifies the belief in others' preference stability as a general characteristic that is correlated within-individual over different situations, and (iii) the paper examines also the consequences this characteristic has, in this case for strategic behaviour.

3 Experimental Design

The focus of this paper is on how predictability of behaviour depends on whether people have stable preferences in the short run and on whether they generally believe in stable preferences of others. I use a public-good context as the main application because it is a well-studied environment, and because it is rich enough to allow for heterogeneous behavioural patterns. To measure preference stability in this context, I elicit participants' preferences for conditional cooperation three times within a session: twice at the beginning, and a third time as the final part of the session (PREFS1, PREFS2, and PREFS3). For each of the PREFS-experiments, I use the method introduced by Fischbacher et al. (2001, see Section 3.1 for details). For none of the experimental parts do participants get any feedback before the end of the session.

To examine what potential instability means for strategic behaviour, I look at a simultaneous public-good protocol (SIMPG, played after PREFS2). As a benchmark for strategic behaviour, I chose the Nash-equilibria that result when interpreting the elicited preferences from the PREFS-experiments as participants' best-reply correspondences in the SIMPG-experiment. I call the set of all mutual best-replies resulting from the elicited preferences the set of *revealed-preference Nash equilibria* (RPNE; cf. Wolff, 2017). I show participants their interaction partner's behaviour from the PREFS1-experiment prior to the SIMPG-experiment, for two reasons (see below for a discussion of the potential signalling issue). First, Nash-equilibrium concepts typically assume common knowledge about the game being played. Showing participants the PREFS1-behaviour of their interaction partner approximates at least mutual knowledge of the game. And second, I argue that having seen how others react on different occasions is a pervasive feature of everyday life. Revealing others' preferences is an interesting benchmark case to study.

If we want to study how knowing others' reactions to past situations affects their strategic behaviour, we need to elicit their beliefs about the others' actions

[14] E.g., Choi and Nisbett (1998) or Miyamoto and Kitayama (2002).

as well as about others' preference stability. In the part after the simPG, I elicit probabilistic beliefs about their interaction partner's simPG action, that is, I ask them how likely their partner will choose each available option. In the penultimate part of each session, I then show participants the prefs1-behaviour of four other participants and ask them to report probabilistic beliefs on the participants' prefs2-behaviour. One of the four other participants is the interaction partner from the simPG-experiment, the three others are randomly chosen others. Here is a full overview of all seven parts of a session, only one of which is paid out:

FillerTask. A social-value orientation task similar to the one presented in Murphy et al. (2011). Irrelevant for this paper but for the potential signalling issue.

prefs1 The prefs-experiment as detailed in Section 3.1.

prefs2 + beliefs. Repetition of the prefs-experiment with a new interaction partner. Then, I elicited beliefs on the expected first-mover contribution, to train participants in the elicitation method used in simPGbeliefs: probabilistic beliefs elicited by a binarised scoring rule (McKelvey and Page, 1990, Hossain and Okui, 2013, probability of receiving an additional payment of 2 Euros determined by a quadratic scoring rule; I do not analyse the beliefs from this part).[15]

simPG. The simPG-experiment as detailed in Section 3.1.

simPGbeliefs. Elicitation of beliefs on the likelihood of the interaction partner choosing each possible action in the simPG-experiment. Payment by a binarised scoring rule with payoffs of 20 Euros (if successful) and 4 Euros (if not successful).

stabilityBeliefs. Elicitation of beliefs on the stability of preferences of the interaction partner in the simPG-experiment and three randomly-chosen others. Participants were shown the response vector of the other participant from the prefs1-experiment. Then, they had to state a probabilistic belief on the response of the same other participant as a second-mover in the prefs2-experiment, for each possible contribution level of the first-mover. For each of the four participants whose response stability participants had to assess, one first-mover contribution was randomly chosen. Then, participants were paid by a binarised scoring rule for their belief accuracy in the four randomly-selected cases, with a payment of 6 Euros per lottery.

[15] Note that by the transformation of payoffs into lottery tickets, the binarised scoring rule is proper under any expected-utility risk preferences, and even for non-expected-utility agents whose preferences satisfy a mild monotonicity condition.

PREFS3. Second repetition of the PREFS-experiment with a new interaction partner.[16]

Note that if participants know their behaviour in one experiment may be revealed to others in the next experiment, they may have potential signalling incentives in the first of the two.[17] My experimental design allows to counter this problem through a number of design choices, discussed in full detail in Wolff (2015) on a very similar earlier design. Most importantly, participants make decisions in seven distinct experimental parts with new interaction partners in each of them, being paid for only one randomly chosen experiment (which should make signalling prohibitively costly). They do not get any feedback about others' behaviour before the SIMPG-experiment, and each experiment is explained only as soon as it begins. While it is impossible-in-principle to show there have been no signalling attempts by participants, I could not find any evidence of signalling in the data. Also, if I were to base my analysis on the data from the final PREFS3-part in which there could not be any signalling incentives anymore, my main Result 1 would get even stronger, not weaker.

3.1 The SIMPG- and the PREFS-experiments

The SIMPG-experiment consists of a simultaneous two-player linear public-good situation with an MPCR $= \frac{2}{3}$ and an endowment of 15 Euros. Each player has to choose a contribution to the public good from the set $\{0, 3, 6, 9, 12, 15\}$ Euros, which is multiplied by $\frac{4}{3}$ and divided equally among the two players, regardless of each player's own contribution. In addition, players see the elicited PREFS1-preferences of their opponent before making their choice.

In the PREFS-experiment, participants face the same two-player linear public-good payoff structure with an MPCR $= \frac{2}{3}$ and an endowment of 15 Euros as in the SIMPG-experiment. However, the PREFS-experiment differs from the SIMPG in that there is no information on the other player, and in that the PREFS-experiments are sequential games: one participant moves first and the other moves second, being informed of the first participant's choice. Participants have to decide in either role. First, they specify their first-mover contribution to the public good that is implemented if they are not (randomly) chosen to be the second-moving player. Then, I elicit their second-mover choices using the strategy method: they are presented with all possible first-mover contributions and asked to specify their

[16]In contrast to the first two PREFS tasks, the first-mover in PREFS3 was shown the response-vector of the second-mover from the PREFS1-experiment before deciding on her (unconditional) contribution. However, the situation of the second-mover was exactly the same as in the PREFS1- and PREFS2-experiments. For the purpose of this paper, I therefore regard the PREFS3-experiment simply as a second repeat-measurement of participants' preferences.

[17]To avoid deceiving participants, the instructions included the sentence that "your behaviour from one of the earlier parts will possibly be displayed to other participants in a later part."

'conditional' contributions.[18] I hold the conditional-contribution schedules from the PREFS-experiment to be a direct expression of participants' (proximate) preferences. Therefore, I equate schedules and best-response correspondences for the remainder of this article.

To limit the scope for confusion as a major source of (measured) preference instability, I took three measures. First, I restricted the simultaneous game to a two-player six-action game rather than the usual three- or four-player games with 11-21 actions. While the MPCR may look a little complicated, all game payoffs were integer amounts. Second, I always displayed the full payoff matrix in the relevant parts. Moreover, I highlighted the relevant part of the matrix in the preference-elicitation parts of the PREFS-experiments, so that participants would know exactly what payoff profile each of their actions meant. As a third measure, I recruited *experienced* participants.[19] Participants in the experiment had participated in at least one public-good experiment and at least four additional other experiments, with no upper limits.

3.2 Procedures

On the day of the experiment, participants were welcomed and asked to draw lots in order to assign them to a cabin. There, they would find some explanation on the general structure of the experiment and on the selection of the payoff-relevant experiment (and role, if applicable). The instructions for each experiment were displayed directly on their screen during the corresponding part. The (translated) general and on-screen instructions are gathered in the Appendix.

Participants earned on average 19.33 Euros (USD 22) for about 90 minutes; this includes a 2-Euro flat payment for the completion of a post-experimental questionnaire. Altogether, seven sessions with a total of 152 participants were conducted at the LakeLab of the University of Konstanz, between June 2015 and June 2016.

4 Results

4.1 Strategic behaviour and preference instability

In the beginning of Section 3, I referred to the idea of calculating the set of equilibria that results for the preferences elicited in a PREFS-experiment. But what

[18]The order of the combinations was randomised individually for each player. Responses were elicited one-by-one for two reasons: (i) to make each decision as salient as possible, (ii) to elicit 'smooth' response-patterns only in case preferences gave rise to them.

[19]I nonetheless asked the usual comprehension questions; participants could only proceed to the experiment after answering all questions correctly.

can we expect from a *revealed-preference Nash-equilibrium* (RPNE) prediction if there is preference instability? Before I answer this question, I categorise participants into four groups, depending on whether they have stable preferences themselves and depending on whether they believe other participants have stable preferences: *stability-believers with stable preferences* (30%), *stability-believers with instable preferences* (13%), *instability-believers with stable preferences* (26%), and *instability-believers with instable preferences* (31%).

A participant is categorised into one of the *stable-preference* groups if she has *approximately stable* preferences, and into an *instable-preference* group, otherwise. I categorise as *approximately stable* all those who have an average squared difference of at most two from the mean response to each first-mover contribution. This criterion would be fulfilled with equality if a participant replies to each first-mover contribution the same way twice, deviating on the third occasion by one increment of 3 Euros in all contingencies.[20]

A participant is categorised into one of the *(in-)stability-believer* groups based on her STABILITYBELIEFS. In the STABILITYBELIEFS-part, each participant sees the choices of four other participants from the PREFS1-part and has to state a probablistic belief what each of the four others would have done in the PREFS2-part. One of the four others was the participant's interaction partner in the SIMPG-part. For the *(in-)stability-believer*-group categorisation, I focus on the participant's beliefs about the three players the participant did *not* interact with in the SIMPG-part. I then categorise a participant into one of the *stability-believer* groups if she places at least 80% probability mass on these three other players responding to all possible first-mover contributions the same way in the PREFS1- and the PREFS2-experiments, and into an *instability-believer* group, otherwise.[21] Not looking at the belief about the SIMPG-interaction partner focuses on 'stability believing' as a characteristic of the person, which promises to be more helpful for prediction than interaction-specific measures.[22]

This categorisation is useful: it informs us whether we can predict behaviour in the SIMPG-experiment or not. As a first step, I use the preferences elicited in the PREFS1-experiment together with the SIMPGBELIEFS, to predict choices in the SIMPG-experiment.[23] Then, I go one step further and predict choices in the SIMPG-

[20]Using this criterion, there are 66 *approximately stable* participants. If we were to use a median split instead, the threshold would increase to 11/3. Only eight additional participants have an average squared difference from the mean response of less than 11/3.

[21]Changing the threshold to, *e.g.*, 70% does not change the results in any meaningful way.

[22]Similar results obtain also when focusing on the interaction instead.

[23]Strictly speaking, the research design does not allow to make a clear prediction from the conditional-cooperation preferences when a participant's belief is non-degenerate. To be able to make a prediction, I use the best-response to the belief's mode. In case of multiplicity, I take the best-response to the highest other-player action that is modal in the participant's SIMPGBELIEF.

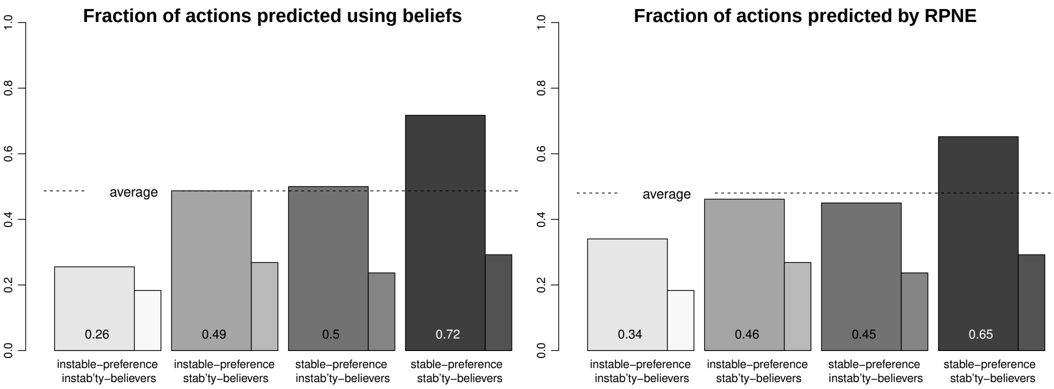


Figure 1: Fraction of SIMPG-choices correctly predicted using either preferences and beliefs (left) or only preferences (right), by whether a participant has *approximately stable* preferences and by whether the participant believes in others having stable preferences. The wide columns represent the data, the narrow columns the corresponding random benchmarks.

experiment by participants' (highest) RPNE-action.[24] Figure 1 shows the results.

Result 1. Whether a participant has stable preferences and whether the participant *in general* believes in others' preference stability determines whether we can predict reliably the participant's behaviour *in a specific* new game.

The left-hand part of Figure 1 shows that the SIMPG-choices of *stability-believers with stable preferences* can be predicted using the PREFS1- and SIMPGBELIEFS-data in 72% of the cases. The corresponding figure for *instability-believers with instable preferences* equals 26% which even cannot be distinguished statistically from random sampling from all SIMPG choices (binomial test, $p = 0.191$). The choices of *stability-believers with instable preferences* and *instability-believers with stable preferences* can be predicted in roughly half of the cases (49% and 50%, $p \leq 0.014$).[25,26]

Note that Result 1 is not a consequence of easy-to-predict *Defectors* (definition follows in Section 4.4 below) being overrepresented among the *stability-believers*

[24] In case of multiplicity, I use the highest RPNE-contribution. Using the lowest, the mean, or the median RPNE-contributions (rounded to the nearest possible value) does not change the results in any substantial way. I use the highest RPNE-contribution because this yields the overall highest fraction of predicted choices.

[25] The above figures also tend to be statistically different: a χ^2-test for all four figures yields $p < 0.001$, Boschloo tests for pairwise comparisons yield $p \in [10^{-5}, 0.080]$ except for the comparison of the 'intermediate' groups ($p = 1$).

[26] If we define a best-response as the reaction to any of: the belief mode, the average belief rounded to the next-possible value or rounded down to the next-possible value (to allow for some pessimism), best-responses account for 43%, 64%, 50%, and 78% of choices (same order as in Fig. 1). Also, the finding is not a consequence of certain types having degenerate beliefs and others not: if we exclude the 21 people with degenerate beliefs, the figures change to: 24%, 50%, 35%, and 66%.

with stable preferences. Not including *Defectors* would yield the following predictability figures (in the same order as in Figure 1): 24%, 45%, 23%, 66%. Result 1 also cannot be explained by *instable-preference* participants learning or revealing their true preferences only by the end: using the preferences elicited in PREFS3 does not decrease the differences, with predictability figures of 26%, 49%, 55%, and 74%. Rather, the unexpected difference in best-response rates between those who generally believe in others' preference stability and those who do not is related to the uncertainty expressed in the participants' beliefs over actions. The standard deviation of the other player's expected action in *instability believers*' beliefs for the simultaneous game is clearly higher compared to that in *stability believers*' beliefs (3.9 vs 2.9, $p = 0.002$, Wilcoxon-Mann-Whitney test). In turn, a higher standard deviation of the other's expected action leads to a lower best-response-to-belief rate.[27] There is no such statistical difference with respect to the player's own preference stability.[28]

The right-hand part of Figure 1 shows what fraction of SIMPG-choices can be predicted using only data from the PREFS1-experiment. The choices of *stability-believers with stable preferences* are predicted in about two thirds of the cases (65%), whereas those of *instability believers with instable preferences* are predicted in only one third (34%). Choices by participants categorised in the 'intermediate' groups are predicted in slightly less than half of the cases (46% and 45%; $p \leq 0.034$ for binomial tests against chance for all groups).[29] The frequencies of predictable choices go along with the corresponding beliefs: *stability believers with stable preferences* on average put a probability mass of 56% on the event that the other player chooses the (highest) RPNE action, whereas *instability believers* assign only 26% (those in the 'intermediate' groups assign 45% and 47%).[30] In summary, we can predict *stability believers* and participants with *approximately stable* preferences more easily because they deem it more likely that others will choose a RPNE action, they are more positive about it, and hence, they are more likely to best-respond to it.

[27] Spearman's $\rho = -0.14, p = 0.091$; in a probit regression of best-responses on only the expected action's standard deviation (and a constant), the coefficient has $p = 0.045$.

[28] A Wilcoxon-Mann-Whitney test for *instability believers*' expectations *vs stability believers*' expectations yields $p = 0.002$ (standard deviations 3.91 *vs* 2.88). The same test within the *instability believers* yields $p = 0.848$ (3.89 *vs* 3.95) and $p = 0.670$ within the *stability believers* (2.99 *vs* 2.79).

[29] The given figures also tend to be statistically different: the χ^2-test for over all groups yields $p = 0.026$, but Boschloo tests indicate differences only for *stability believers with stable preferences vs. instability-believers with instable preferences* ($p = 0.003$) and potentially *vs.* the 'intermediate' groups ($p = 0.142$ and $p = 0.080$).

[30] Wilcoxon Mann-Whitney tests yield $p < 0.01$ for *instability believers with instable preferences vs* each of the other groups, but only $p = 0.171$ for *stability believers with stable preferences vs* the 'intermediate' groups.

	Model 1	Model 2	Model 3	Model 4	Model 5
(Intercept)	0.386 (0.104)***	0.643 (0.052)***	0.438 (0.189)*	0.687 (0.150)***	0.270 (0.216)
Prob(others' stability)	0.003 (0.002)*		0.003 (0.002)·		0.006 (0.002)**
Preference instability		−0.007 (0.005)		−0.004 (0.005)	0.016 (0.010)
Male			0.212 (0.083)*	0.210 (0.085)*	0.210 (0.083)*
A-Levels (average grade)			−0.072 (0.069)	−0.090 (0.069)	−0.062 (0.068)
Choice of 0 in acquire-a-company			−0.529 (0.481)	−0.584 (0.488)	−0.594 (0.483)
Economist			0.111 (0.091)	0.108 (0.092)	0.121 (0.091)
Prob(others'stability)×preference instability					−0.000 (0.000)·
R^2	0.032	0.012	0.107	0.088	0.131
Adj. R^2	0.025	0.005	0.077	0.057	0.089
Num. obs.	152	152	152	152	152
RMSE	0.486	0.491	0.473	0.478	0.469

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{\cdot}p < 0.1$

Table 1: Linear-probability-model regressions of best response play on a participant's own preference (in-)stability, the average probability placed on others' stability of preferences ("Prob(others' stability)"), and a number of controls capturing cognitive abilities, gender, and field of study.

4.2 Explaining behaviour

Section 4.1 shows that classifying participants by their preference stability and by their belief in others stability goes a long way to identify those people whose behaviour we can predict *ex ante*. This section asks whose SIMPG contributions are consistent with best-response behaviour, analysing the roles of preference stability and of beliefs in others' stability in more detail. The analysis in this part relaxes the requirements of prediction in an important way. For the analysis in this part, I define a best-response as the optimal reaction to any of: the belief mode, the average belief rounded to the next-possible value or rounded down to the next-possible value (*cf.* ftns. 23 and 26).

Table 1 provides a number of regressions of best-response play on a participant's: average subjective probability of others reacting in the same way in PREFS1 and PREFS2 ("Prob(others' stability)"); "Preference instability" as measured by the average squared difference between her three preference measurements; the interaction between the participant's preference instability and her subjective probability of others' preference stability; gender; cognitive abilities as proxied by her average A-Levels grade and by whether the participant chose the optimal offer in a hypothetic acquire-a-company game included in the post-experimental questionnaire; field of study (economics or otherwise).[31]

[31] Using the chosen number in the acquire-a-company game does not affect the results; including the A-levels math grade, or participants' CRT score which is possible for only 82 of the participants does not affect the results meaningfully, either. Notably, in all three cases, the corresponding coefficient is not significantly different from 0, corroborating that cognitive abilities do not seem to play an important role here. Further, including age or religiousness does not affect the results.

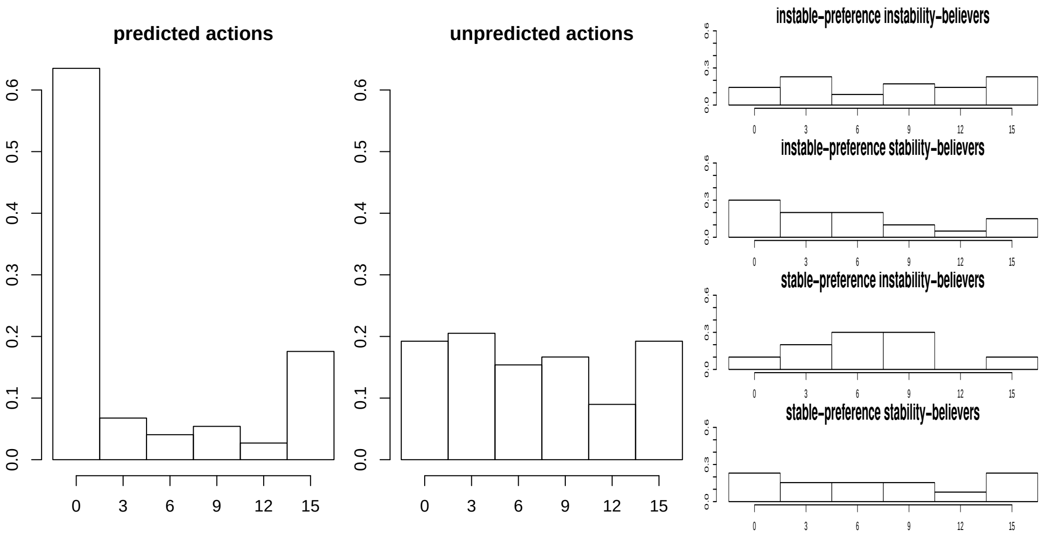


Figure 2: Relative frequencies of successfully predicted actions (left) and unpredicted actions, overall (middle) and by the participant's preference stability and general belief in preference stability (right).

The second row of Table 1 provides further support for the idea that a participants' general belief in others' preference stability is an important determinant of whether she will play according to her stated belief. Surprisingly, the influence of participants' degree of preference stability that is apparent from Figure 1 does not shine through in the third row of Table 1. While the coefficient at least has the expected sign in Models 2 and 4, it has the 'wrong' sign in Model 5 (which is, however, counteracted by the weakly significant interaction effect in the final row). At the same time, we note that participants' cognitive abilities do not seem to play an important role, either. In other words, it is not the case cognitively able participants are more capable of playing in accordance with their beliefs.

4.3 Behaviour of the unpredictable participants

The middle panel of Figure 2 shows the behaviour of those participants whose action could not be predicted by their preferences together with their beliefs, next to the behaviour of 'predictable' participants (on the left). The distribution of unpredicted actions looks close to uniform (also by a χ^2-test, $p = 0.528$).

In the right-most panel of Figure 2, we see the distribution of unpredicted actions split up by the participant's preference stability and general belief in preference stability. While there are too few observations in each category for a meaningful statistical analysis, eyeballing suggests that behaviour may not be completely random, except for the *instability-believers with instable preferences.*

Stability-believers with instable preferences seem to prefer to contribute rather little, whereas *instability-believers with stable preferences* seem to prefer to choose some intermediate contribution as a consequence of their uncertainty over others' behaviour. *Stability-believers with stable preferences*, finally, tend towards the extremes, thereby showing the pattern that is closest to the choices of 'predictable' participants.

4.4 Preference stability in the PREFS-experiments

Result 2. There is substantial preference instability in our sample. Slightly more than half of the participants show preferences that are not even approximately stable over three measurements within the same session.

In total, only 44 out of 152 participants (29%) respond exactly the same way in all three PREFS-repetitions. The *approximate stability* criterion (defined in Section 4.1) which allows for some variation is fulfilled by 66 out of 152 participants (43%). Conversely, the elicited preferences of 57% of the participants are not even approximately stable.

I now examine whether this instability also translates into instability of preference types and how instability and preference type are related. To this end, I categorise the participants into the usual conditional-contribution types according to their preferences as elicited in the PREFS-experiment. I use the categories suggested by Fischbacher et al. (2001) but divide the conditional cooperators into three subcategories. To be precise, I use the following categories:

Defectors Always respond by a contribution of 0 Euros.

PerfCCs Perfect conditional cooperators always respond by mirroring exactly the contribution of the first-mover.

ImpCCs Imperfect conditional cooperators have a monotonically-increasing response vector. Respond by less than the first-mover contribution at least some of the time.

NmImpCCs Non-monotonic imperfect conditional cooperators have a non-monotonic response vector, for which the Spearman correlation coefficient with first-mover contributions is positive with $p \leq 0.05$ (one-sided).

Triangles Have a hump-shaped response-vector.

Altruists Always contribute fully with at most one deviation to 12 Euros.

Unclassifiables Cannot be classified into any of the above categories.

Table 2 displays the type distributions, depending on the time of measurement, that is, depending on whether I use the response vector from the PREFS1-, the PREFS2-, or the PREFS3-experiment. Overall, the numbers of conditional cooperators (including all three categories, 41-53%, or roughly half of our sample) and of *Defectors* (roughly one quarter of our sample) are similar to those in Fischbacher et al. (2001).

	PREFS1	PREFS2	PREFS3
Defectors	35	42	41
PerfCCs	29	38	41
ImpCCs	25	24	30
NmImpCCs	8	16	9
Triangles	17	14	14
Altruists	6	2	2
Unclassifiables	32	16	15
Total	152	152	152

Table 2: Distribution of player types, by time of measurement.

Table 2 confirms Result 2 on the level of preference types. The frequencies of players classified as specific types fluctuate considerably: the number of participants classified as *Defectors* and *PerfCCs* or *ImpCCs* increases, and the number of participants classified as *Triangles*, *Altruists*, or *Unclassifiable* participants decreases. At the same time, Table 2 might suggest that stability increases over time, as the numbers change less from column PREFS2 to PREFS3 than they do from PREFS1 to PREFS2. However, this aggregate trend is misleading. The percentage of participants who switch preference-type categories only falls minimally (and insignificantly), from 43% to 40%.[32]

Result 3. There is considerable heterogeneity in terms of preference stability: among *Defectors* and perfect conditional cooperators (*PerfCCs*), two thirds have completely stable preferences, but only 1 out of the remaining 88 participants does.

Specifically, 23 out of 35 *Defectors* and 20 out of 29 *PerfCCs* (as classified in the PREFS-experiment) respond exactly the same way in all three PREFS-repetitions. Out of the remaining 88 participants, this holds true only for a single *Triangle*. Approximate preference stability is fulfilled for 24 out of 35 *Defectors*, 21 out of 29 *PerfCCs*, 13 out of 25 *ImpCCs*, half of the *NmImpCCs*, 1 *Triangle*, 1 *Altruist* and 2 *Unclassifiables*. Again, we see that some 70% of the participants classified as

[32] An analysis on the contribution level yields the same result: the sum of squared differences in a participant's responses does not differ significantly between the comparison PREFS1–PREFS2 and the comparison PREFS2–PREFS3 ($p = 0.547$, Wilcoxon matched-pairs signed-ranks test).

	Coefficient	Std. Error	p-value
(Intercept)	71.1	(4.01)	$1 \cdot 10^{-56}$
appr. stable preference	12.9	(3.02)	$2 \cdot 10^{-5}$
same type	7.1	(4.29)	0.0968
appr. stable preference × same type	−5.5	(5.94)	0.3573
PerfCC	2.7	(3.26)	0.4086
ImpCC	−6.2	(3.62)	0.0895
NmImpCC	−1.3	(5.14)	0.8081
Triangle	1.0	(4.38)	0.8253
Altruist	−15.3	(6.74)	0.0235
Unclassifiable	−15.9	(3.88)	0.0001
other player is a *PerfCC*	1.4	(3.67)	0.7033
other player is an *ImpCC*	−12.8	(3.65)	0.0005
other player is a *NmImpCC*	−23.3	(4.62)	$6 \cdot 10^{-7}$
other player is a *Triangle*	−19.4	(4.23)	$5 \cdot 10^{-6}$
other player is an *Altruist*	−26.1	(8.79)	0.0031
other player is *Unclassifiable*	−28.4	(3.60)	$1 \cdot 10^{-14}$

Table 3: Average probability mass a player's belief placed on the other player choosing the same reaction in PREFS2 as in PREFS1, regressed on own and other's preference type and on whether the player has approximately stable preferences. Ordinary least squares model with standard errors clustered on participants.

Defectors or *PerfCCs* are approximately stable, compared to less than a quarter of the remaining participants.

4.5 Beliefs about preference stability

Do participants themselves believe in preference stability? As a measure for the degree of belief in preference stability, I use the probability mass a participant places on the event that the other participant will respond to the first-mover contribution exactly the same way in the PREFS2- as in the PREFS1-experiment, averaged over all possible first-mover contributions. Table 3 shows the results of a regression of this 'average belief in stability' on the participant's preference type, the other player's preference type, on whether the two types are the same and on whether the participant has approximately stable preferences.

Result 4. Participants' beliefs in preference stability on aggregate are 'rational': participants correctly believe *Defectors* and *PerfCCs* to display a high degree of response-stability, whereas they correctly expect other types to display more variable responses.

As we see in the lower part of Table 3, only *Defectors* (baseline category) and *PerfCCs* are expected to display a high degree of preference stability. The 'average belief in stability' drops sharply when the other player is of other types (by 13-28 percentage points). This corresponds very well to the finding that *Defectors* and *PerfCCs* truly are the most stable types.

Result 5. Having stable preferences strongly boosts participants' beliefs in others' preference stability, and participants' own type also has an influence. Hence, there seems to be a personal element in believing in others' preference stability.

As the second row in the main part of Table 3 shows, a participant assigns on average 13 percentage points more probability mass to another participant showing the same preferences again if the participant has *approximately stable* preferences than when she does not. Rows five to ten of Table 3 show that *Unclassifiables*, *Altruists*, and potentially also *ImpCCs* place substantially less probability weight on others showing stable preferences, compared to the baseline category of the *Defectors*. Interestingly, we see only weak evidence that beliefs in others' preference stability are boosted by both players having the same type of preferences, and no evidence at all if the focal participant has stable preferences.

4.6 Beliefs about others' simPG-choices

How good are people at predicting others' simPG-choices? This question is addressed in Figure 3. Figure 3 plots how much probability mass participants placed on the action chosen by their interaction partners, again depending on whether a participant has *approximately stable* preferences and on whether the participant believes in others having stable preferences.

Result 6. Participants are not able to predict others' behaviour in a simultaneous public-good situation very well even when they are shown the preferences elicited in a sequential public-good situation like in Fischbacher et al. (2001). (Only) Part of the reason might be that participants underestimate the predictive power of this information.

As we can see in Figure 3, the overall average probability mass participants place on the actual choice of their interaction partner is only one third. Interestingly, participants who believe in others' preference stability have the (insignificantly) more accurate beliefs, with a probability mass of up to 40% placed on the actually chosen contribution level. This figure increases to 44% if we define 'stability believing' in terms of the belief on the simPG-interaction partner herself (in

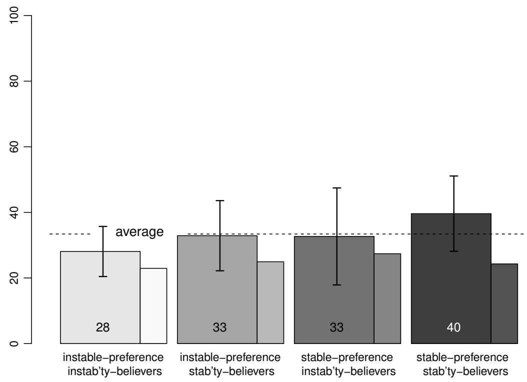


Figure 3: Empirical accuracy of beliefs: probability mass placed on the action chosen by the interaction partner, depending on whether a participant has *approximately stable* preferences and on whether the participant believes in others having stable preferences. The wide columns represent the data, the narrow columns the corresponding random benchmarks. Error bars indicate 95%-confidence intervals.

that case, irrespective of preference stability). The fact that *stability-believers* have the more accurate beliefs on the behaviour of the other player might suggest that *instability-believers* place too little weight on the information they are given. On the other hand, the fact that *stability-believers* also have rather inaccurate beliefs clearly shows that believing in the information can be at most part of the story.[33]

5 Beliefs in preference stability in a risk context

So far, we have seen that the degree to which a participant believes in others' preference stability in general predicts whether we as researchers can predict the participant's behaviour in a different interaction with a new partner. This section briefly summarises the results of a second experiment that tests whether we find similar results in a completely different context, this time on risky decisions.

5.1 Experimental setup

Experiment 2 had five parts. Again, each part was described only as it started, and the payoff-relevant part was determined randomly. In parts 1, 2, and 5, I measured participant's risk attitudes by the method of Gneezy and Potters (1997): in

[33]Comparing the distribution of probability masses placed on the other player's SIMPG-action to that of the corresponding probability masses from beliefs randomly sampled from the same participant category (category-mean probability mass depicted in the narrow bars in Figure 3), by means of bootstrapped Kolmogorov-Smirnov tests, yields p-values of 0.016, 0.121, 0.572, and 0.016 (ordered as in Figure 3).

each of these parts, participants had to choose an integer amount x to invest into a risky asset out of an endowment of 10 Euros. With a probability of two thirds, the asset turned worthless, yielding a payoff of $(10 - x)$ Euros, but with a probability of one third, the asset went up in price, yielding a payoff of $(10 + 2.5x)$. In part 3, I showed participants the part-1 behaviour $x_j^{(1)}$ of three other participants and told them they could buy an option security on a market index, whose value was determined by the three other participants' part-2 behaviour $x_j^{(2)}$: if the three participants had invested at least one Euro less than their part-1 sum (i.e., if $\sum x_j^{(2)} \geq \sum x_j^{(1)} - 1$), the option would pay off, otherwise it would be worthless. The price of the option was set to the focal participant's part-1 investment $x_i^{(1)}$ plus 2, the payoff in case of success was $(10 + 1.5[x_i^{(1)} + 2])$ Euros. I chose this payoff to make it a difficult decision for participants, not in order to keep the payoff structure similar to the other parts.[34] In part 4, I asked for participants' beliefs by the same method as in the simPGbeliefs- and stabilityBeliefs-parts of the main experiment. I asked them about their belief concerning the "market" in part 3 first (about $\sum x_j^{(2)}$), followed by their beliefs about 6 individual other participants' part-2 behaviour $x_j^{(2)}$ (the second, fourth, and fifth being the participants determining the market). For each of the seven beliefs, they got 3 Euros in case their earned lottery tickets were successful. 168 participants participated in 6 sessions in January and February 2017.

5.2 Results

As in the main experiment, I classified participants into those with *approximately stable* vs *instable* preferences and into *stability-believers* vs *instability-believers*. As before, a participant was categorised as having *approximately stable* preferences if her average squared deviation was less than two thirds of the smallest increment (in this case, of 1) from the mean choice in parts 1, 2, and 5, yielding 85/168 participants in the *stable* category. A participant was categorised as a *stability-believer* if she placed an average probability of 60% on the other participants repeating their part-1 investments in part 2, again evaluating only the beliefs on those participants that were irrelevant for the focal individual's part-3 decision.[35] Figure 4 shows the

[34]Note that, had participants known part 3 in advance, they would have had an incentive to change their behaviour in part 1. To avoid any issues of deception, I used the technique introduced by Bardsley (2000): participants knew that 1 out of the 5 experimental parts (part 3) could not become payoff-relevant, without knowing which part this would be. They did not get to know the connection between their part-1 choice and the part-3 situation at any time of the experiment.

[35]I use a cutoff of 60% rather than of 80% as in the main experiment, for two reasons: first, the smallest increment in Experiment 2 was 1 rather than 3, so erroneous deviations should be expected more often in Experiment 2 (if participants expect others to make Fechner-type errors). And second, this leaves us with 33/168 *stability-believers* rather than with only 16, giving us more statistical

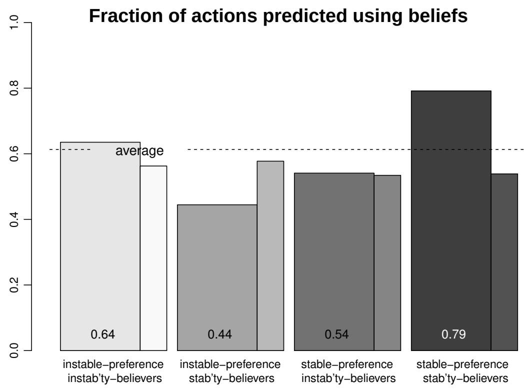


Figure 4: Fraction of part-3 choices correctly predicted using part-1 preferences and market beliefs, by whether a participant has *approximately stable* preferences and by whether the participant believes in others having stable preferences. The wide columns represent the data, the narrow columns the corresponding random benchmarks.

fractions of part-3 choices correctly predicted by participants' risk attitudes (as measured in part 1) together with their market beliefs (on $\sum x_j^{(2)}$).[36]

Result 7. The general belief in others' risk-preference stability determines how reliably we can predict the participant's behaviour in a market game, in particular for participants with stable preferences.

Result 7 is immediately obvious from Figure 4 for participants with *approximately stable* preferences as well as for participants with *instable* preferences; the only surprise is that this time, the belief or non-belief in others preference stability seems to predict belief-action correspondence *negatively* amongst the participants with *instable* preferences.[37] To understand this surprising effect, we have to take

power. Even with the cutoff at 60%, we only have 24 *stability-believers with stable preferences* (and, as in the public-good context, very few—9—*stability-believers with instable preferences*).

[36] I use the CRRA utility function $u(x) = \frac{1}{1-\rho}x^{(1-\rho)}$, in which part-1 investments $x_i^{(1)}$ directly translate into an estimate of ρ. For choices $x_i^{(1)} = 10$, I postulate $\rho = 0$, whereas for $x_i^{(1)} = 0$, I use an arbitrary $\rho = 1.5$ (which roughly would be the next component of the sequence of ρs when going from $x_i^{(1)} = 9$ to $x_i^{(1)} = 0$). Both of these extreme cases together make up for 12% of the observations (split roughly evenly). Nothing substantial changes if we exclude these observations from the analysis. Using the estimate of ρ, I then compare the expected utility of buying the option security in part 3 given the probability assigned to the event that the option yields a positive payoff (i.e., that $\sum x_j^{(2)} \geq \sum x_j^{(1)} - 1$) to the utility of 10 Euros, to predict the participant's part-3 choice.

[37] A χ^2-test is inconclusive on whether correctly predicted choices differ between the four categories ($p = 0.122$), while Boschloo-tests indicate a significant difference between *stability-* and *instability-believers with stable preferences* ($p = 0.041$) but not between *stability-* and *instability-believers with instable preferences* ($p = 0.261$; note, however, that there are only 9 observations in the former group).

a closer look at the participants in the respective categories. For the participants with *instable preferences*, 91-100% of the predicted actions are to invest in the option security in part 3 (compared to 72-75% for the groups with *stable preferences*). Moreover, the *instability-believers with instable preferences* are the only group who clearly increase their investments from investment decision 1 to investment decision 2 (by 0.70 on average, compared to negative average changes of up to -0.13 in the other groups). If the *instability-believers with instable preferences* then project their own increase in investments from decision 1 to decision 2 onto others (which they do, predicting an average increase of 0.42 amongst those others who do not affect the outcome of the market decision and of 0.34 amongst those who do), their disbelief in others' preference stability will make many of them optimistic enough to invest in the option security. Which means they will tend to choose the predicted option relatively often. Consider now the *stability-believers with instable preferences.* First of all, note that the share of participants for whom the investment in part 1 (which is used to predict the part-3 decision) is larger than their average investment from parts 1, 2, and 5 is largest in this group (44%, compared to 20-25% in the other groups). This means that the prediction will tend to overestimate these participants' willingness to take risks. Moreover, their belief in stable preferences makes them much less optimistic than their *instability-believing* counterparts, both of which leads them to choose the option security much less often—which in turn leads to a lower share of correct predictions.

Let us turn to the *stable-preference* groups. Here, there is no systematic shift in risk taking over the different investment decisions, and no systematic expectation that others' behaviour will shift in a certain direction. This means that the story from the public-good game carries over: *stability-believers* have a clearer idea of what they think others will do, and thus act accordingly, while *instability-believers* are much more uncertain about what will happen.[38] This difference then translates into more noise in the *instability-believers*' decision-making.

6 Conclusions

In economics, we traditionally thought of preferences as of something stable. In fact, the usefulness of the concept of preferences hinges on them being sufficiently stable. Studies suggesting that the temporal stability of preferences may be limited have opened the door to a completely new world. This paper looks at what is behind this door. However, I have not addressed the question of what it means when I classify somebody as playing a best-response to her beliefs when her elicited

[38]Standard deviations in the market beliefs are 1.3 vs 2.7, $p = 0.002$, Wilcoxon Mann-Whitney test. Note that the same holds true for the *instable-preference* groups, with 1.0 vs 3.0, $p < 0.001$. However, in this case, the effects described above dominate the uncertainty effect.

preferences are instable. I have not addressed how we should think about a 'best-response', an 'equilibrium', or a 'game' without preference stability, either.[39] For this paper, I have defined all the concepts empirically, to analyse whether these benchmarks help in understanding behaviour. And I argue they do, at least in the public-good experiment: we can predict about half of all choices even of those participants with instable preferences—and even if we use no more than their elicited (instable!) preferences—provided they believe in others' preference stability.[40]

With one exception, participants who generally believe in others' preference stability are 25 percentage points more likely to play a best-response to their belief about their interaction partner's action, compared to participants who do not believe in others' preference stability. This result might seem odd from a game-theoretic perspective: I should do the best I can do given my belief about the other's action, no matter where that belief comes from. However, if I have a general tendency to be uncertain about what others will do, my belief may be highly volatile also in a specific situation. Then, by the time I am asked to report it, my belief may have changed enough to make my action no longer a best-response to the reported belief. At the same time, my expectation of the other player's contribution will have a higher variance, which is exactly what we observe.[41] Importantly, none of these results seems to be due to a preference-discovery process: the results continue to hold (or even get stronger) when we use the preference elicitation at the end of the experiment for the analysis.

The findings shed new light on the debate of how we should model people's behaviour in strategic interactions. To take the main example of this paper, public-good protocols, there are strong proponents of an equilibrium approach like Ambrus and Pathak (2011). On the other hand, the findings of Fischbacher and Gächter (2010) make a strong case that beliefs are not in equlibrium. My paper suggests a heterogeneous approach. One third acts well in line with an equilibrium based on their elicited preferences (a 'revealed-preference Nash equilibrium'). One third potentially could be modelled as participants playing a noisy best-response to a similarly noisy belief. But the remaining third can only be modelled as (quantal-

[39]For a specific form of instability, decision-making with evaluation errors, the questions have been answered, of course. McKelvey and Palfrey's (1995) quantal-response equilibrium provides the necessary framework in this case.

[40]For the risk-context experiment, the benchmark of about 50% best-responses under random play unfortunately prevents us from seeing similar effects.

[41]The only exception are participants in the risk-context experiment who have instable preferences; for the instability-believers among them, the initial preference measurement tends to be biased towards too much risk-aversion, so that they have a greater tendency to buying the option security in part 3 of the experiment than their initial preference measurement suggests. Given the experiment was set up in a way as to make buying the market-security the optimal choice for a majority of the participants, we have a(n insignificantly) higher fraction of best-responses amongst these participants, compared to their stability-believing counterparts.

response-)equilibrium players if we assume their noise parameters to be extremely high.

The focus of this paper has been on unstable preferences and on people's belief in others' preference stability. These aspects provide a plausible 'micro-foundation' for participants' heterogeneity in terms of their behavioural 'distance' to equilibrium. While the paper remains largely agnostic about where the instability of revealed preferences comes from, the noise parameter in people's behaviour and belief on actions seems to be grounded in how positive people are about being able to predict others. Whether this extends beyond strategic uncertainty to other features of the environment is an important question for further research.

Technical acknowledgements

The experiments were computerised using z-Tree (Fischbacher, 2007), participants were recruited using ORSEE (Greiner, 2004) with Mozilla Firefox. The equilibria of the game were calculated during the experiment using R (R Development Core Team, 2001, 2012, Ihaka, 1998), which was also used to analyse the data in combination with RKWard (Rödiger et al., 2012) and RStudio (RStudio Team, 2015). For the statistical testing, R packages Exact (Calhoun, 2015, Boschloo test), dgof (Arnold and Emerson, 2011, bootstrapped Kolmogorov-Smirnov test), plm (Croissant and Millo, 2008) and lmtest (Zeileis and Hothorn, 2002, both for the regression with cluster-robust standard errors), texreg (Leifeld, 2013, conversion of regression output to LaTeX), and doBy (Højsgaard and Halekoh, 2016, calculating groupwise summary statistics) were of particular value. Most of this was done on a computer running on KDE-based (KDE e.V., 2012) Kubuntu, which required the use of wine for the programming of the experiment. The article was written using Kile.

References

Agranov, M. and Ortoleva, P. (2017). Stochastic choice and preferences for randomization. *Journal of Political Economy*, 125(1):40–68.

Ambrus, A. and Pathak, P. A. (2011). Cooperation over finite horizons: A theory and experiments. *Journal of Public Economics*, 95(7-8):500–512.

Andreoni, J. (1995). Cooperation in public-goods experiments: Kindness or confusion? *The American Economic Review*, 85(4):891–904.

Arnold, T. A. and Emerson, J. W. (2011). Nonparametric goodness-of-fit tests for discrete null distributions. *The R Journal*, 3(2):34–39.

Bardsley, N. (2000). Control without deception: Individual behaviour in free-riding experiments revisited. *Experimental Economics*, 3(3):215–240.

Bayer, R.-C., Renner, E., and Sausgruber, R. (2013). Confusion and learning in the voluntary contributions game. *Experimental Economics*, 16(4):478–496.

Bohnet, I. and Huck, S. (2004). Repetition and reputation: Implications for trust and trustworthiness when institutions change. *American Economic Review*, 94(2):p362 – 366.

Brosig, J., Riechmann, T., and Weimann, J. (2007). Selfish in the end? An investigation of consistency and stability of individual behavior. Working paper.

Bruhin, A., Fehr, E., and Schunk, D. (2016). The many faces of human sociality: Uncovering the distribution and stability of social preferences. Working paper.

Calhoun, P. (2015). *Exact: Unconditional Exact Test.* R package version 1.6.

Carlsson, F., Johansson-Stenman, O., and Nam, P. K. (2014). Social preferences are stable over long periods of time. *Journal of Public Economics*, 117:104–114.

Choi, I. and Nisbett, R. E. (1998). Situational salience and cultural differences in the correspondence bias and the actor-observer bias. *Personality and Social Psychology Bulletin*, 24:949–960.

Croissant, Y. and Millo, G. (2008). Panel data econometrics in R: The plm package. *Journal of Statistical Software*, 27(2).

Cubitt, R. P., Starmer, C., and Sugden, R. (2001). Discovered preferences and the experimental evidence of violations of expected utility theory. *Journal of Economic Methodology*, 8(3):385–414.

Dal Bó, P. (2005). Cooperation under the shadow of the future: Experimental evidence from infinitely repeated games. *American Economic Review*, 95(5):1591–1604.

Duffy, J. and Ochs, J. (2009). Cooperative behavior and the frequency of social interaction. *Games and Economic Behavior*, 66(2):785–812.

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

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

Fischbacher, U. and Gächter, S. (2010). Social preferences, beliefs and the dynamics of free riding in public good experiments. *American Economic Review*, 100(1):541–556.

Fischbacher, U., Gächter, S., and Fehr, E. (2001). Are people conditionally cooperative? Evidence from a public goods experiment. *Economics Letters*, 71(3):397–404.

Fischbacher, U., Gächter, S., and Quercia, S. (2012). The behavioral validity of the strategy method in public good experiments. *Journal of Economic Psychology*, 33(4):897 – 913.

Fudenberg, D., Iijima, R., and Strzalecki, T. (2015). Stochastic choice and revealed perturbed utility. *Econometrica*, 83(6):2371–2409.

Gawronski, B. (2004). Theory-based bias correction in dispositional inference: The fundamental attribution error is dead, long live the correspondence bias. *European Review of Social Psychology*, 15(1):183–217.

Gilbert, D. T. (2002). *Heuristics and Biases: The Psychology of Intuitive Judgment*, chapter Inferential Correction, pages 167–184. Cambridge University Press.

Gneezy, U. and Potters, J. (1997). An experiment on risk taking and evaluation periods. *Quarterly Journal of Economics*, 112(2):631–645.

Goeree, J. K. and Holt, C. A. (2004). A model of noisy introspection. *Games and Economic Behavior*, 46(2):365–382.

Greiner, B. (2004). An online recruitment system for economic experiments. In Kremer, K. and Macho, V., editors, *Forschung und wissenschaftliches Rechnen 2003*, volume 63, pages 79–93. Ges. für Wiss. Datenverarbeitung, Göttingen.

Harless, D. W. and Camerer, C. F. (1994). The predictive utility of generalized expected utility theories. *Econometrica*, 62(6):1251–1289.

Harman, G. (2009). Skepticism about character traits. *Journal of Ethics*, 13:235–242.

Hayashi, N., Ostrom, E., Walker, J., and Yamagishi, T. (1999). Reciprocity, trust, and the sense of control: A cross-societal study. *Rationality and Society*, 11(1):27–46.

Hey, J. D. and Orme, C. (1994). Investigating generalizations of expected utility theory using experimental data. *Econometrica*, 62(6):1291–1326.

Højsgaard, S. and Halekoh, U. (2016). *doBy: Groupwise Statistics, LSmeans, Linear Contrasts, Utilities*. R package version 4.5-15.

Hossain, T. and Okui, R. (2013). The binarized scoring rule. *Review of Economic Studies*, 80(3):984–1001.

Ihaka, R. (1998). R: Past and future history. In Weisberg, S., editor, *Proceedings of the 30th Symposium on the Interface*, pages 392–396.

KDE e.V. (2012). About kde. http://www.kde.org/community/whatiskde/. Last accessed on 31st January, 2013.

Leifeld, P. (2013). texreg: Conversion of statistical model output in R to LaTeX and HTML tables. *Journal of Statistical Software*, 55(8):1–24.

Loomes, G. and Sugden, R. (1995). Incorporating a stochastic element into decision theories. *European Economic Review*, 39(3–4):641–648.

Luce, R. D. (1958). A probabilistic theory of utility. *Econometrica*, 26(2):193–224.

Machina, M. J. (1985). Stochastic choice functions generated from deterministic preferences over lotteries. *Economic Journal*, 95(379):575–594.

McKelvey, R. D. and Page, T. (1990). Public and private information: An experimental study of information pooling. *Econometrica*, 58(6):1321–1339.

McKelvey, R. D. and Palfrey, T. R. (1995). Quantal response equilibria for normal form games. *Games and Economic Behavior*, 10(1):6 – 38.

Miyamoto, Y. and Kitayama, S. (2002). Cultural variation in correspondence bias: The critical role of attitude diagnosticity of sosocial constrained behavior. *Journal of Personality and Social Psychology*, 83:1239–1248.

Muller, L., Sefton, M., Steinberg, R., and Vesterlund, L. (2008). Strategic behavior and learning in repeated voluntary contribution experiments. *Journal of Economic Behavior & Organization*, 67(3–4):782–793.

Murphy, R. O., Ackermann, K. A., and Handgraaf, M. J. J. (2011). Measuring social value orientation. *Judgment and Decision Making*, 6(8):771–781.

Myerson, R. (1978). Refinements of the nash equilibrium concept. *International Journal of Game Theory*, 7(2):73–80.

Plott, C. R. (1996). *The Rational Foundations of Economic Behaviour*, chapter Rational individual behaviour in markets and social choice processes: the discovered preference hypothesis, pages 225–250. Palgrave MacMillan, Basingstoke.

R Development Core Team (2001). What is R? *R News*, 1:2–3.

R Development Core Team (2012). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing.

Rödiger, S., Friedrichsmeier, T., Kapat, P., and Michalke, M. (2012). Rkward: A comprehensive graphical user interface and integrated development environment for statistical analysis with r. *Journal of Statistical Software*, 49(9):1–34.

RStudio Team (2015). *RStudio: Integrated Development Environment for R*. RStudio, Inc., Boston, MA.

Rubinstein, A. and Salant, Y. (2016). "isn't everyone like me?": On the presence of self-similarity in strategic interactions. *Judgment and Decision Making*, 11(2):168–173.

Sachdeva, S., Iliev, R., and Medin, D. L. (2009). Sinning saints and saintly sinners. the paradox of moral self-regulation. *Psychological Science*, 20(4):523–528.

Sass, M. and Weimann, J. (2012). The dynamics of individual preferences in repeated public good experiments. Working Paper 2/2012, Otto-von-Guericke-Universität Magdeburg.

Selten, R. (1975). Reexamination of the perfectness concept for equilibrium points in extensive games. *International Journal of Game Theory*, 4(1):25–55.

Volk, S., Thöni, C., and Ruigrok, W. (2012). Temporal stability and psychological foundations of cooperation preferences. *Journal of Economic Behavior & Organization*, 81(2):664–676.

Weibull, J. W. (2004). Testing game theory. In Huck, S., editor, *Advances in Understanding Strategic Behaviour: Game Theory, Experiments and Bounded Rationality. Essays in Honour of Werner Güth*, pages 85–104. Palgrave Macmillan.

Wolff, I. (2015). When best-replies are not in equilibrium: Understanding cooperative behaviour. Research Paper Series 97, Thurgau Institute of Economics.

Wolff, I. (2017). What are the equilibria in public-good experiments? *Economics Letters*, 150:83–85.

Wolff, I. and Bauer, D. (2018). Elusive beliefs: How uncertainty leads to stochastic choice and errors. Research Paper Series 111, Thurgau Institute of Economics.

Zeileis, A. and Hothorn, T. (2002). Diagnostic checking in regression relationships. *R News*, 2(3):7–10.

Appendix

For the Appendix, please refer to
<http://www.wiwi.uni-konstanz.de/fischbacher/home/staff/dr-irenaeus-wolff/>.

Appendix

The Appendix is meant for online publication only.

Appendix Instructions (translated)[42]

I General instructions

General information

You will now participate in an economic experiment. If you read the following explications thoroughly, you can—depending on your choices—earn money. Therefore, it is very important that you read these explications thoroughly.

The instructions you receive from us are for your personal information only. During the experiment, communication is absolutely prohibited. Non-compliance with this rule will lead to exclusion from the experiment and all payments. If you have questions, please raise your hand. We then answer your question at your cubicle.

In this experiment, you will receive money. The amount you receive depends on your decisions and on the decisions of the other experiment participants. Additionally, you receive a compensation of 2 Euros for completing the ensuing questionnaire.

The experiment

The experiment you are participating in today consists of six independent parts. In each of these parts, you will be matched with a different participant. In any case, the participants matched to you will be different people. You will not get to know the identities of the participants you are matched with, neither during nor after the experiment. In the same vein, the participants you are matched with will not get to know your identity.

In some of the parts, there are several participant roles. The role you will take on in actual fact in the different parts will be announced only at the end of the experiment. Therefore, you will make all potentially relevant decisions. Similarly, we will announce only at the very end which of the six parts is relevant for payment. Therefore, you have to determine for all parts what you decide in the

[42] The German original is available from the author upon request.

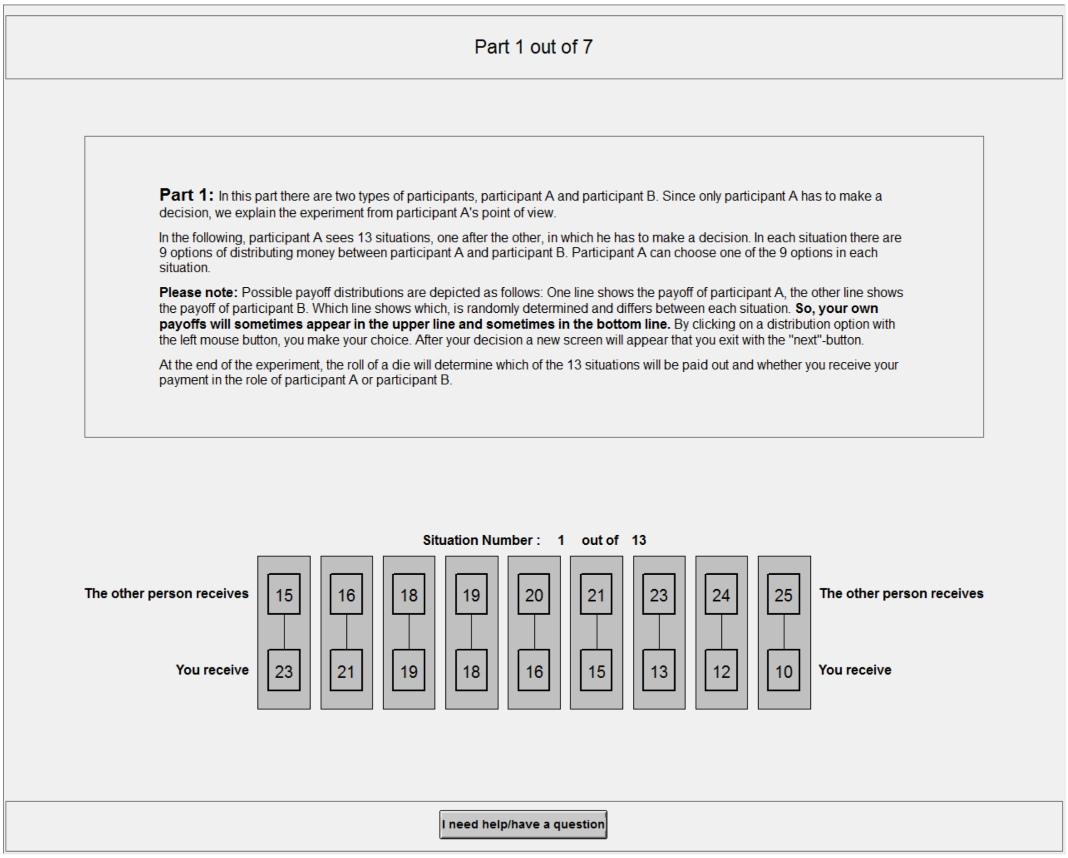
according roles. At the end, you will be paid according to the decision you have taken in the relevant role of the randomly-drawn part of the experiment.

Your role and the relevant part are determined by the roll of a die by the participant we have randomly chosen to be the person making the lucky draw at the beginning of the experiment.[43] However, we will announce the realisations of the die rolls only at the end of the experiment. Hence you will know only then which of your decisions will be relevant for your payment.

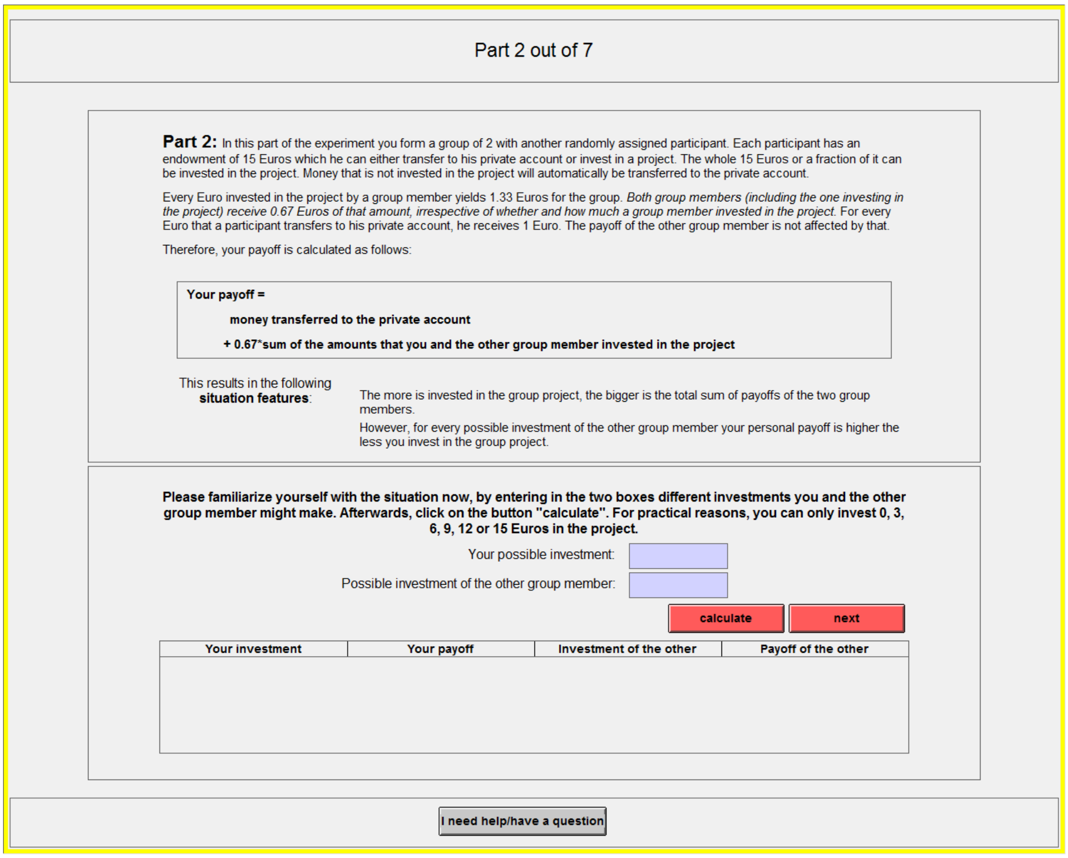
We describe the individual parts directly on the screen. At each point of the experiment, you only receive the description of the according part. We point out to you that your behaviour from one of the earlier parts will possibly be displayed to other participants in a later part. Further, we would like to inform you that the average payoff to be expected from each of the parts is the same.

[43]The participant making the lucky draw did not take part in the actual experiment and did not get to know anything about it. The participant was merely asked to roll the die three times, record the results on screen as well as on a sheet of paper (the latter was later put up at the wall in the laboratory), and come to the experimenters' room directly afterwards to collect 8 Euros for the faithful completion of the task.

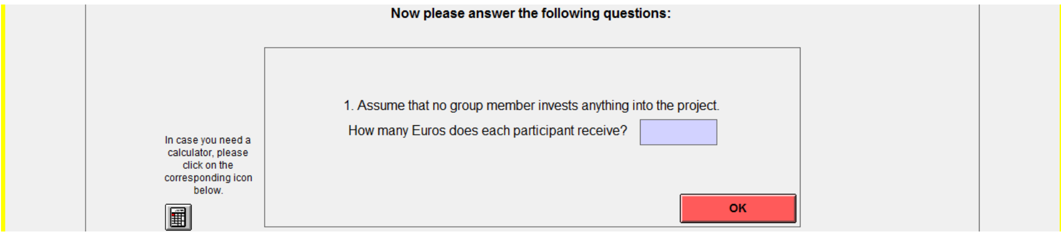
II On-screen instructions (translated)



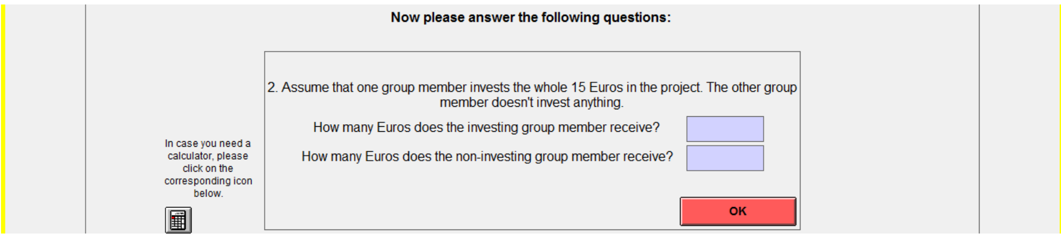
Part 1: Screenshot of the instruction stage.



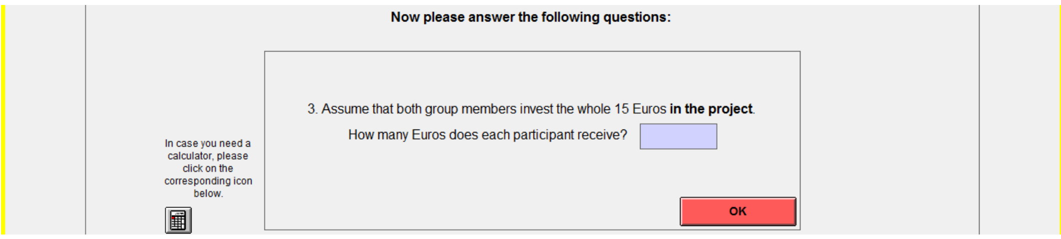
Part 2, PREFS1: instructions; screen with only text (as in upper half) omitted.



Part 2, PREFS1: instructions; comprehension question 1 (upper part as above).

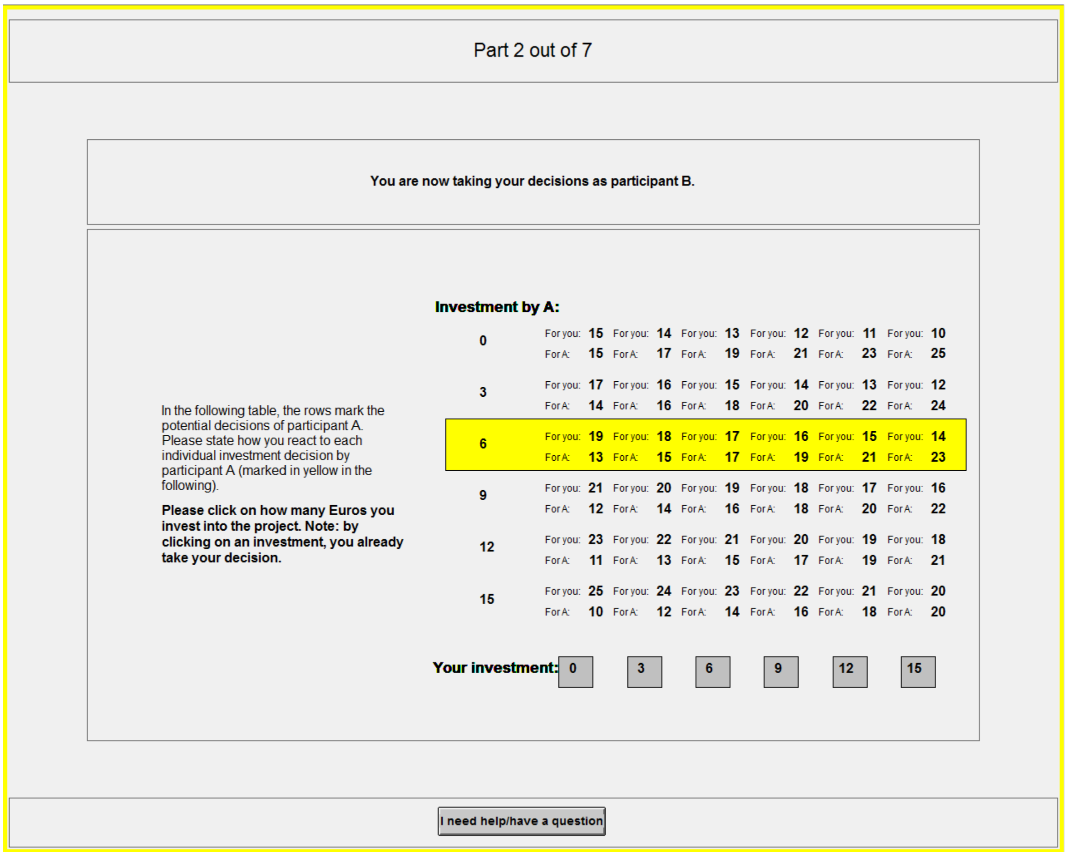


Part 2, PREFS1: instructions; comprehension question 2 (upper part as above).

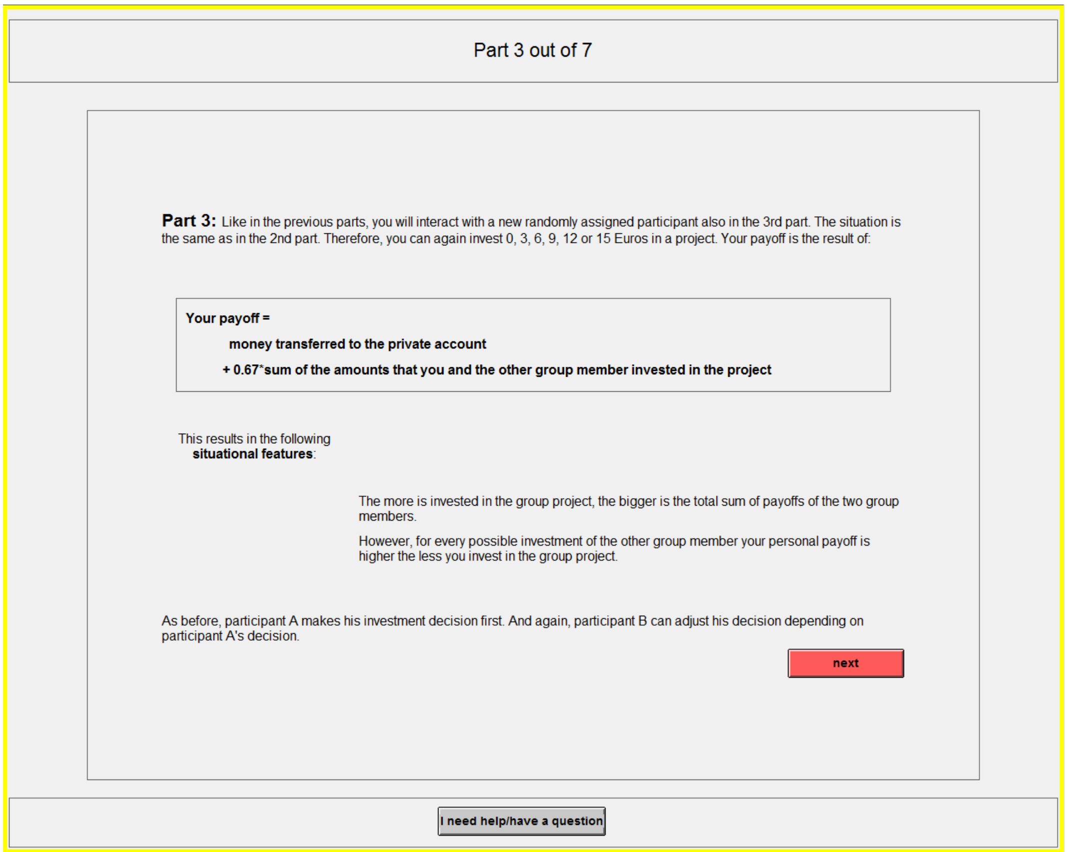


Part 2, PREFS1: instructions; comprehension question 3 (upper part as above).

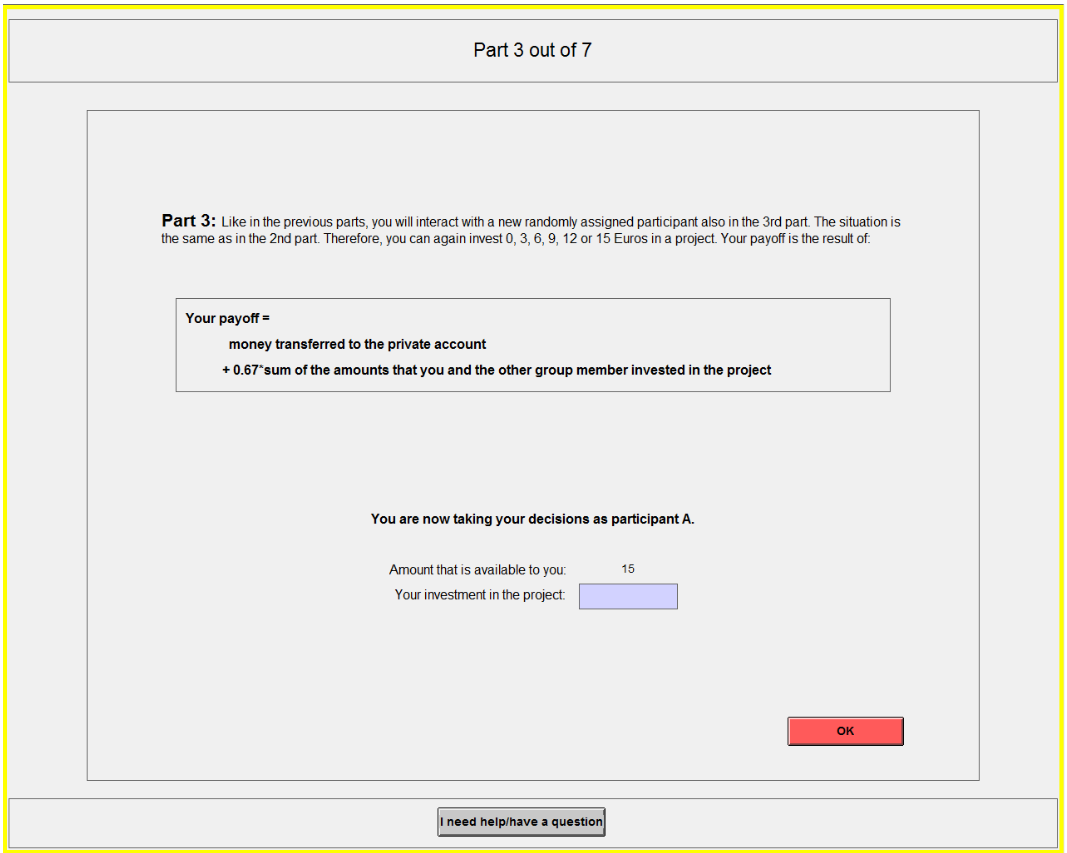
Part 2, PREFS1: unconditional-contribution choice.



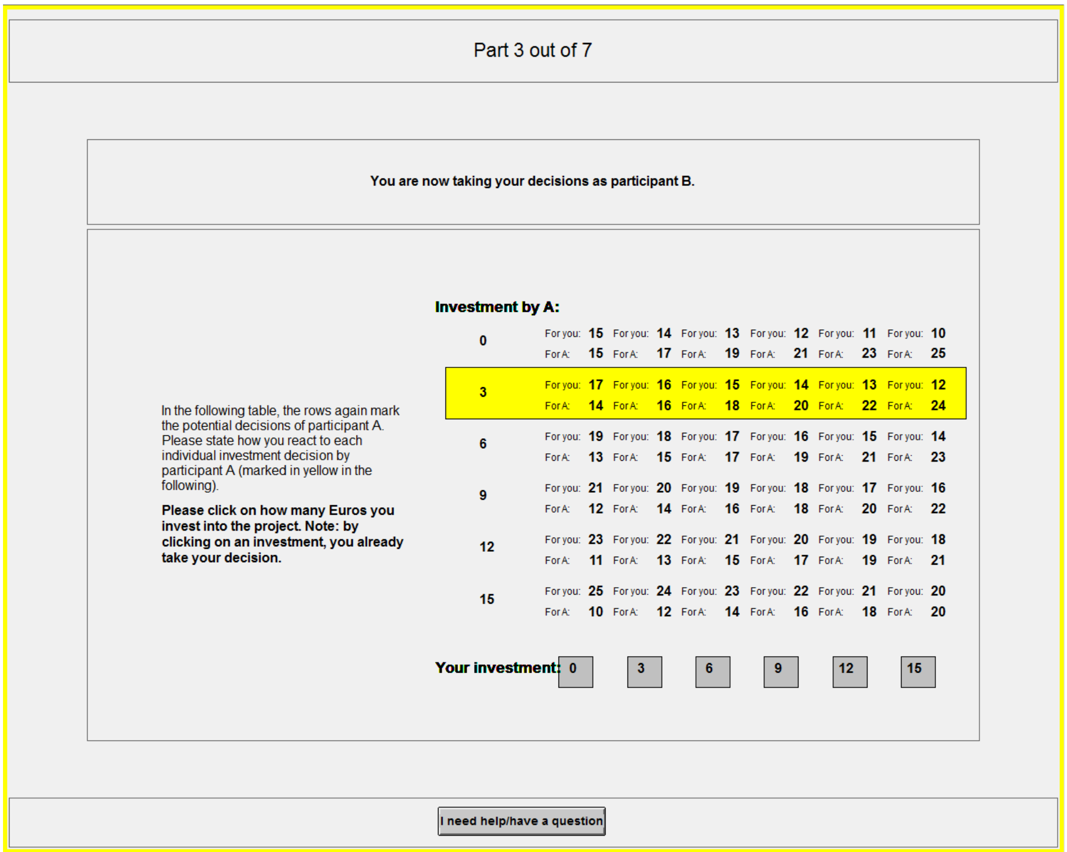
Part 2, PREFS1: conditional-contribution choice (preference elicitation).



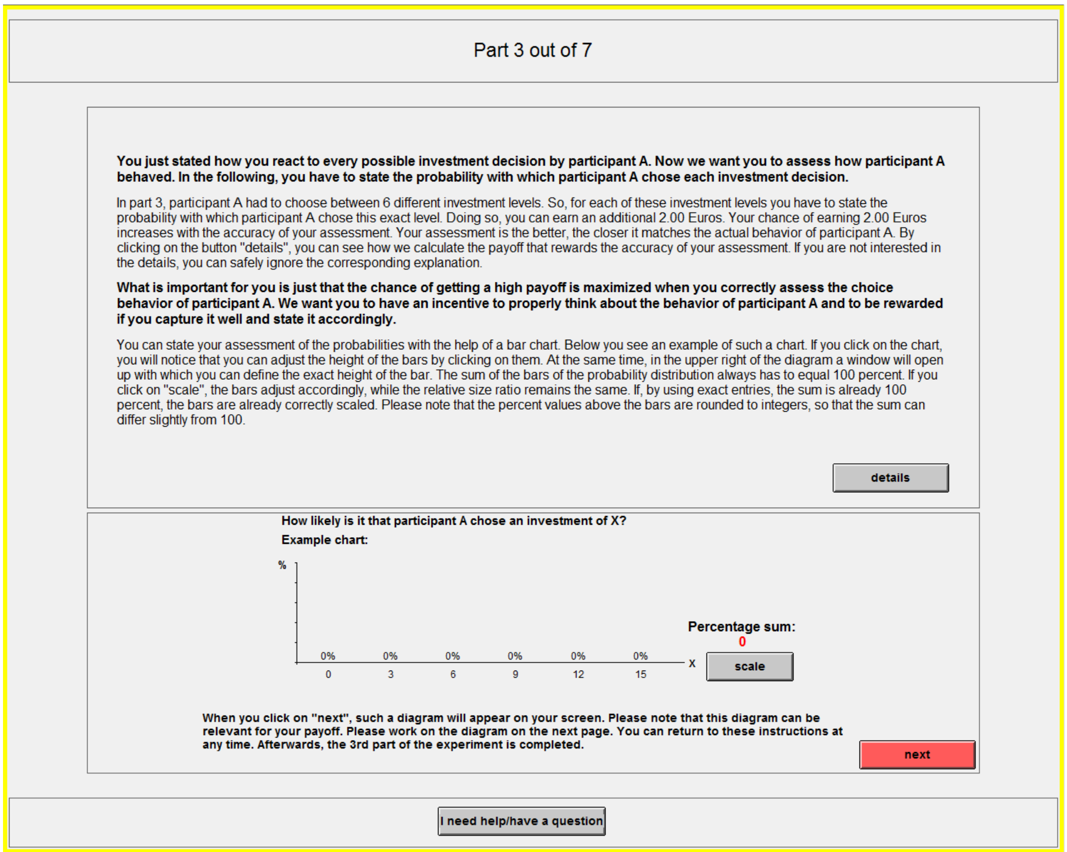
Part 3, PREFS2: instructions.



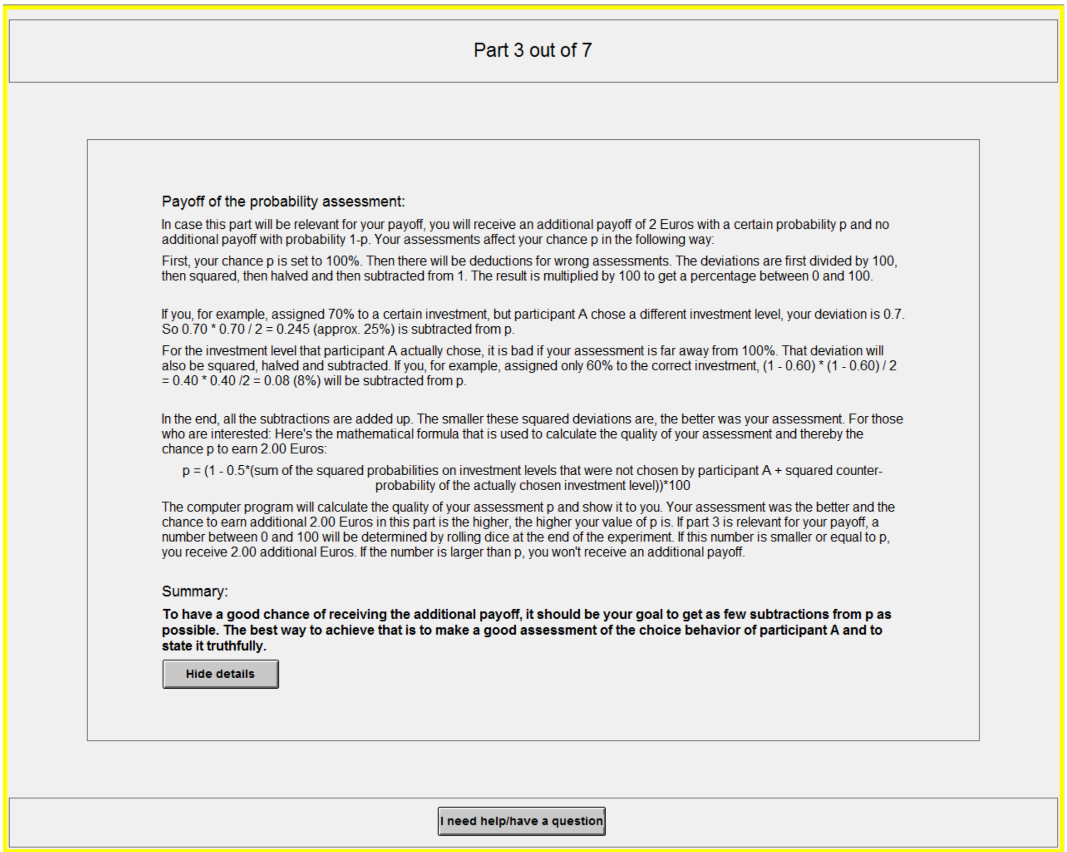
Part 3, PREFS2: unconditional-contribution choice.



Part 3, Prefs2: conditional-contribution choice (preference elicitation).



Part 3, Prefs2: instructions for the belief-elicitation on A's choice (training for the simPGbeliefs and stabilityBeliefs experiments).



Part 3 out of 7

Payoff of the probability assessment:

In case this part will be relevant for your payoff, you will receive an additional payoff of 2 Euros with a certain probability p and no additional payoff with probability 1-p. Your assessments affect your chance p in the following way:

First, your chance p is set to 100%. Then there will be deductions for wrong assessments. The deviations are first divided by 100, then squared, then halved and then subtracted from 1. The result is multiplied by 100 to get a percentage between 0 and 100.

If you, for example, assigned 70% to a certain investment, but participant A chose a different investment level, your deviation is 0.7. So 0.70 * 0.70 / 2 = 0.245 (approx. 25%) is subtracted from p.

For the investment level that participant A actually chose, it is bad if your assessment is far away from 100%. That deviation will also be squared, halved and subtracted. If you, for example, assigned only 60% to the correct investment, (1 - 0.60) * (1 - 0.60) / 2 = 0.40 * 0.40 /2 = 0.08 (8%) will be subtracted from p.

In the end, all the subtractions are added up. The smaller these squared deviations are, the better was your assessment. For those who are interested: Here's the mathematical formula that is used to calculate the quality of your assessment and thereby the chance p to earn 2.00 Euros:

p = (1 - 0.5*(sum of the squared probabilities on investment levels that were not chosen by participant A + squared counter-probability of the actually chosen investment level))*100

The computer program will calculate the quality of your assessment p and show it to you. Your assessment was the better and the chance to earn additional 2.00 Euros in this part is the higher, the higher your value of p is. If part 3 is relevant for your payoff, a number between 0 and 100 will be determined by rolling dice at the end of the experiment. If this number is smaller or equal to p, you receive 2.00 additional Euros. If the number is larger than p, you won't receive an additional payoff.

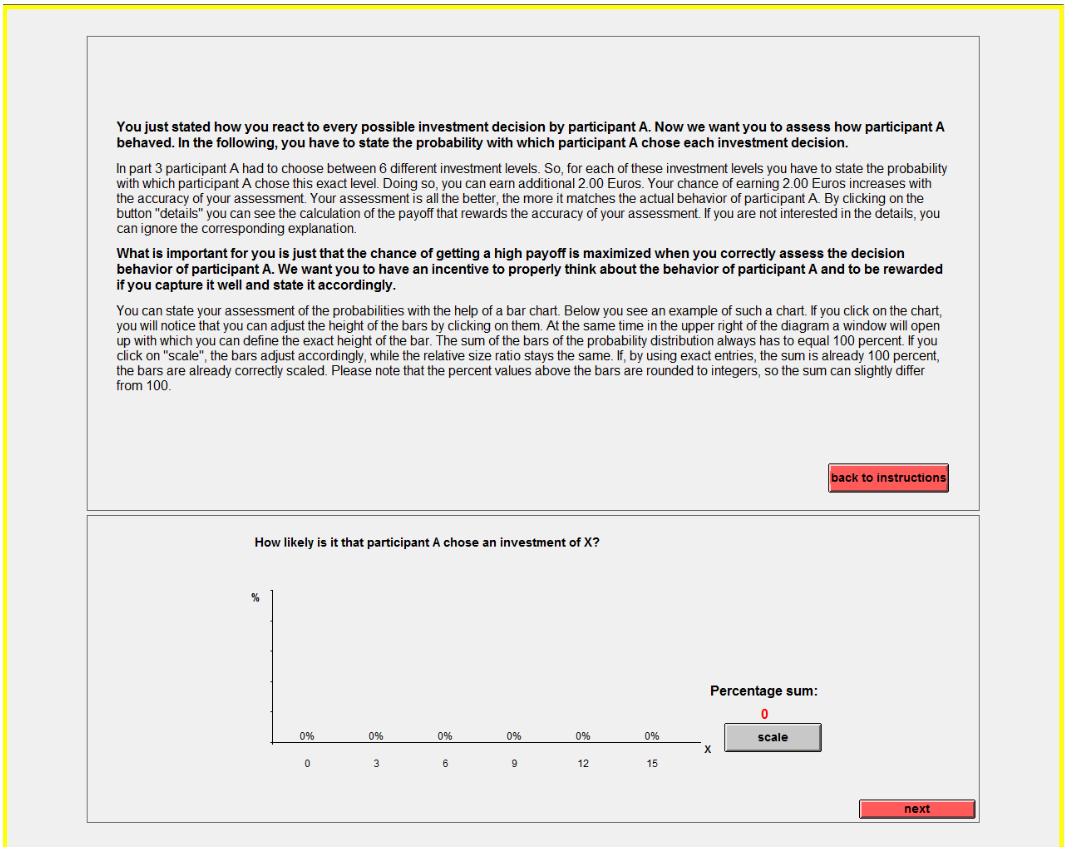
Summary:

To have a good chance of receiving the additional payoff, it should be your goal to get as few subtractions from p as possible. The best way to achieve that is to make a good assessment of the choice behavior of participant A and to state it truthfully.

Hide details

I need help/have a question

Part 3, PREFS2: instructions for the belief-elicitation on A's choice, details (training for the SIMPGBELIEFS and STABILITYBELIEFS experiments).



You just stated how you react to every possible investment decision by participant A. Now we want you to assess how participant A behaved. In the following, you have to state the probability with which participant A chose each investment decision.

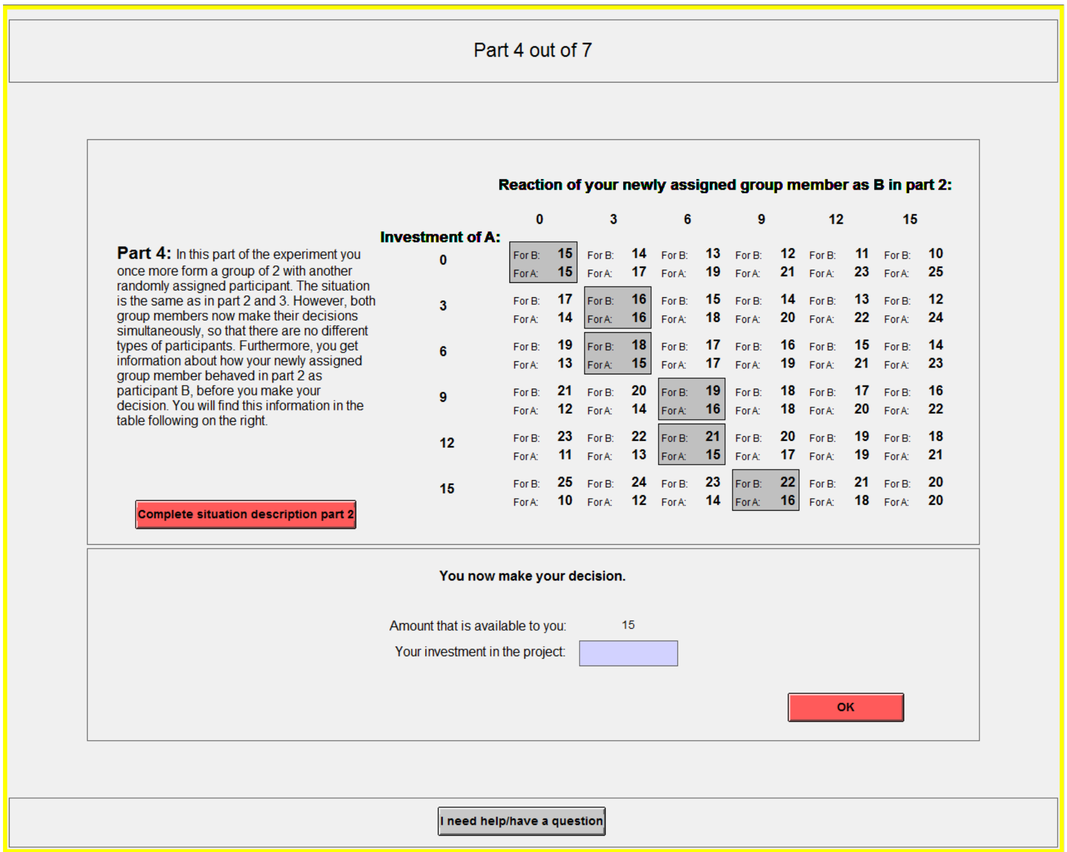
In part 3 participant A had to choose between 6 different investment levels. So, for each of these investment levels you have to state the probability with which participant A chose this exact level. Doing so, you can earn additional 2.00 Euros. Your chance of earning 2.00 Euros increases with the accuracy of your assessment. Your assessment is all the better, the more it matches the actual behavior of participant A. By clicking on the button "details" you can see the calculation of the payoff that rewards the accuracy of your assessment. If you are not interested in the details, you can ignore the corresponding explanation.

What is important for you is just that the chance of getting a high payoff is maximized when you correctly assess the decision behavior of participant A. We want you to have an incentive to properly think about the behavior of participant A and to be rewarded if you capture it well and state it accordingly.

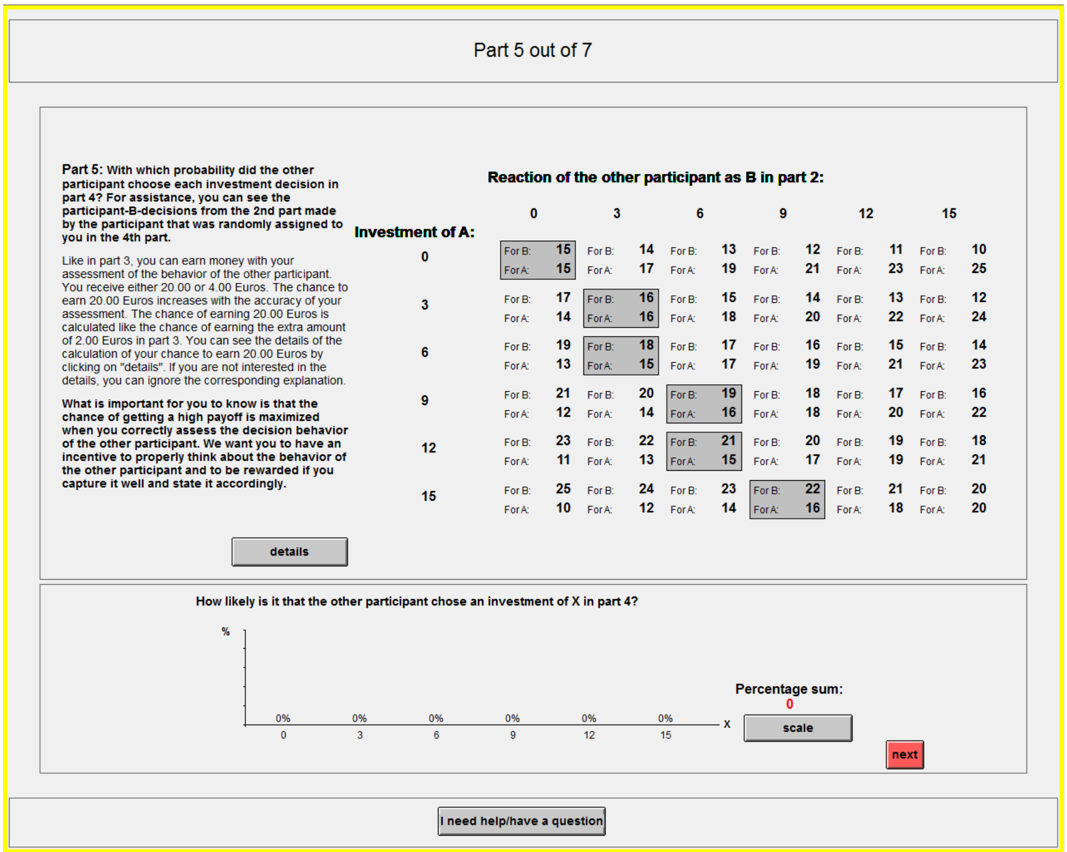
You can state your assessment of the probabilities with the help of a bar chart. Below you see an example of such a chart. If you click on the chart, you will notice that you can adjust the height of the bars by clicking on them. At the same time in the upper right of the diagram a window will open up with which you can define the exact height of the bar. The sum of the bars of the probability distribution always has to equal 100 percent. If you click on "scale", the bars adjust accordingly, while the relative size ratio stays the same. If, by using exact entries, the sum is already 100 percent, the bars are already correctly scaled. Please note that the percent values above the bars are rounded to integers, so the sum can slightly differ from 100.

back to instructions

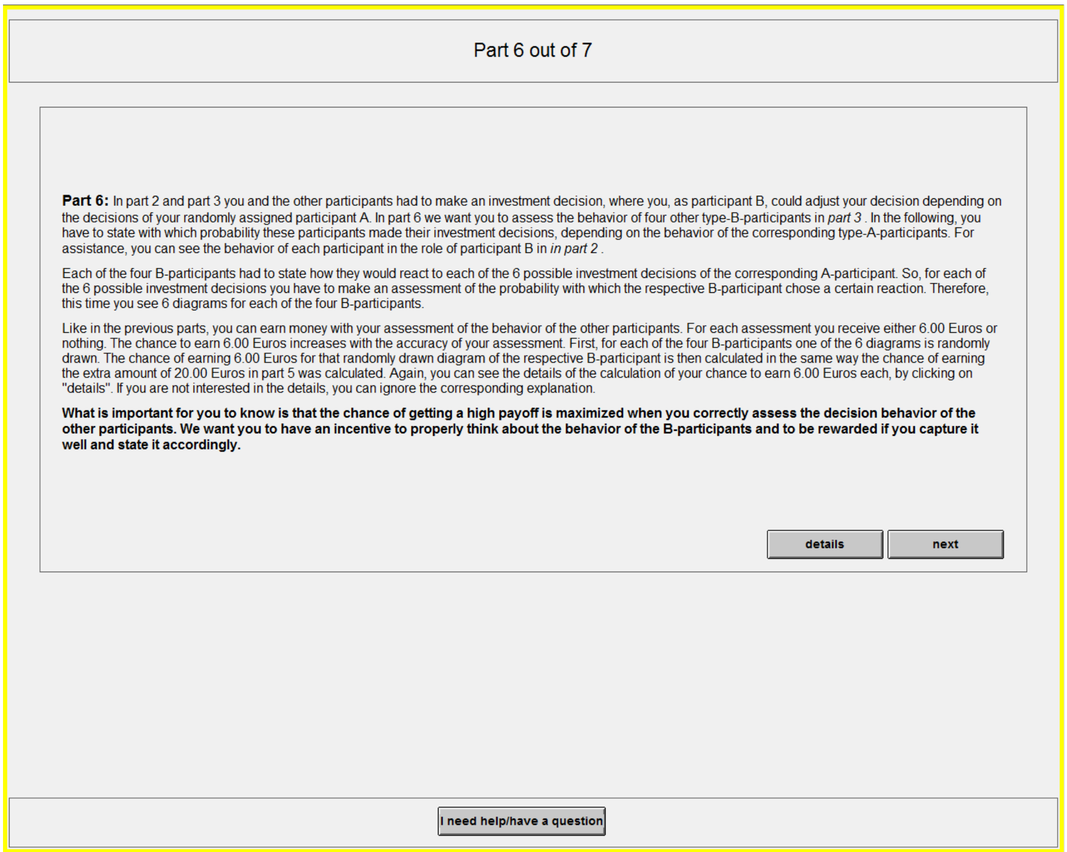
Part 3, PREFS2: belief-elicitation on A's choice (training for the SIMPGBELIEFS and STABILITYBELIEFS experiments).



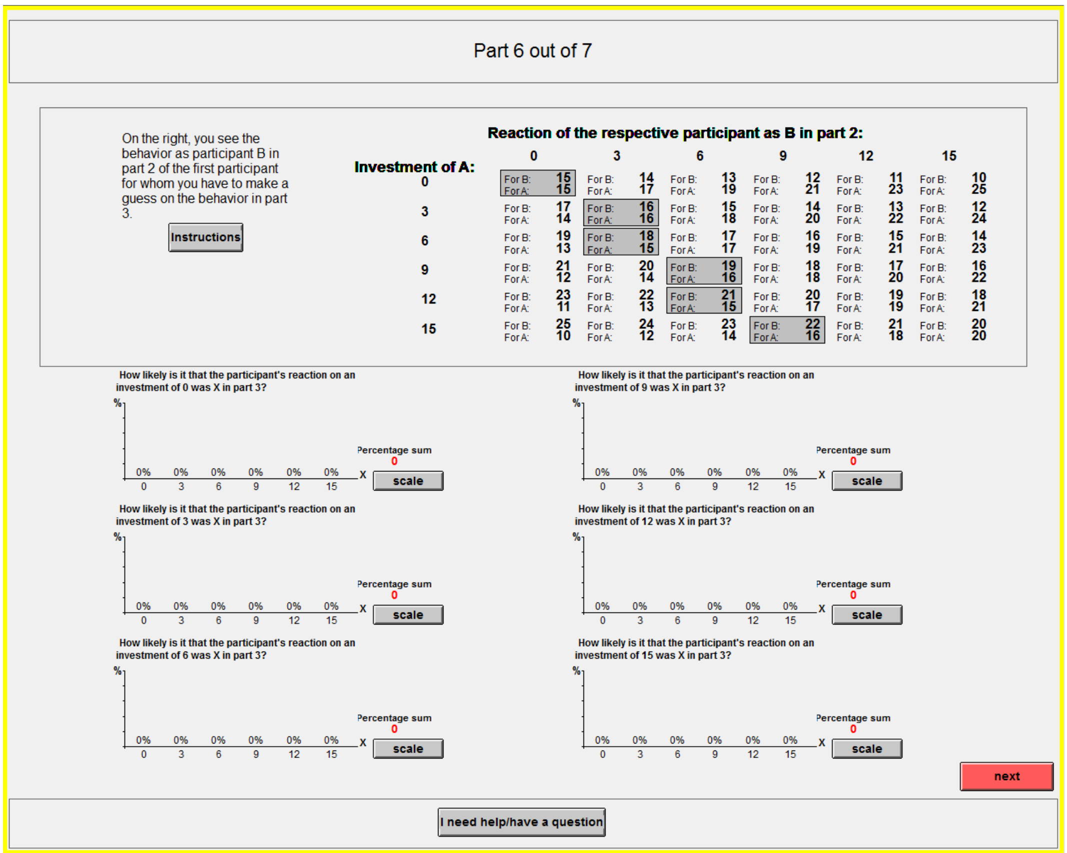
Part 4, sɪᴍPG: contribution choice; screen with only instructions (as in upper half) omitted.



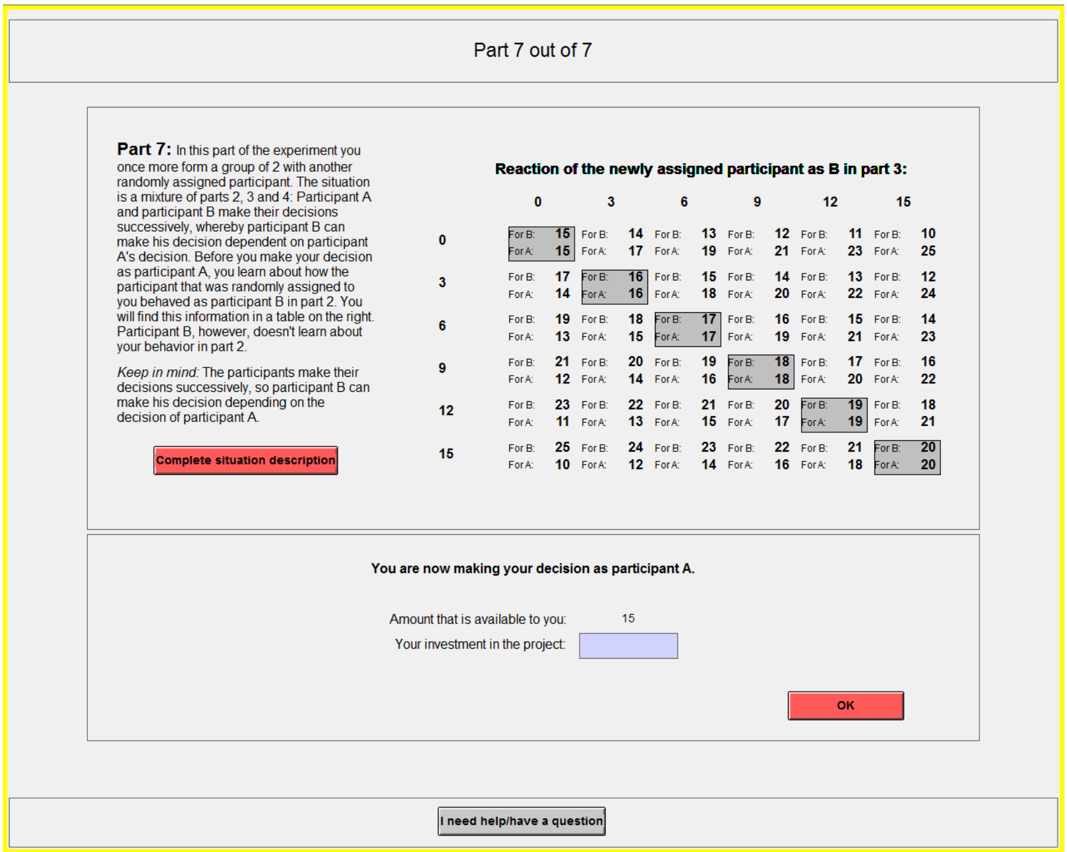
Part 5, sɪᴍPGBᴇʟɪᴇꜰꜱ: belief elitication; screen with only instructions (as in upper half) omitted. Clicking on “details” led to an analogous screen as in Part 3.



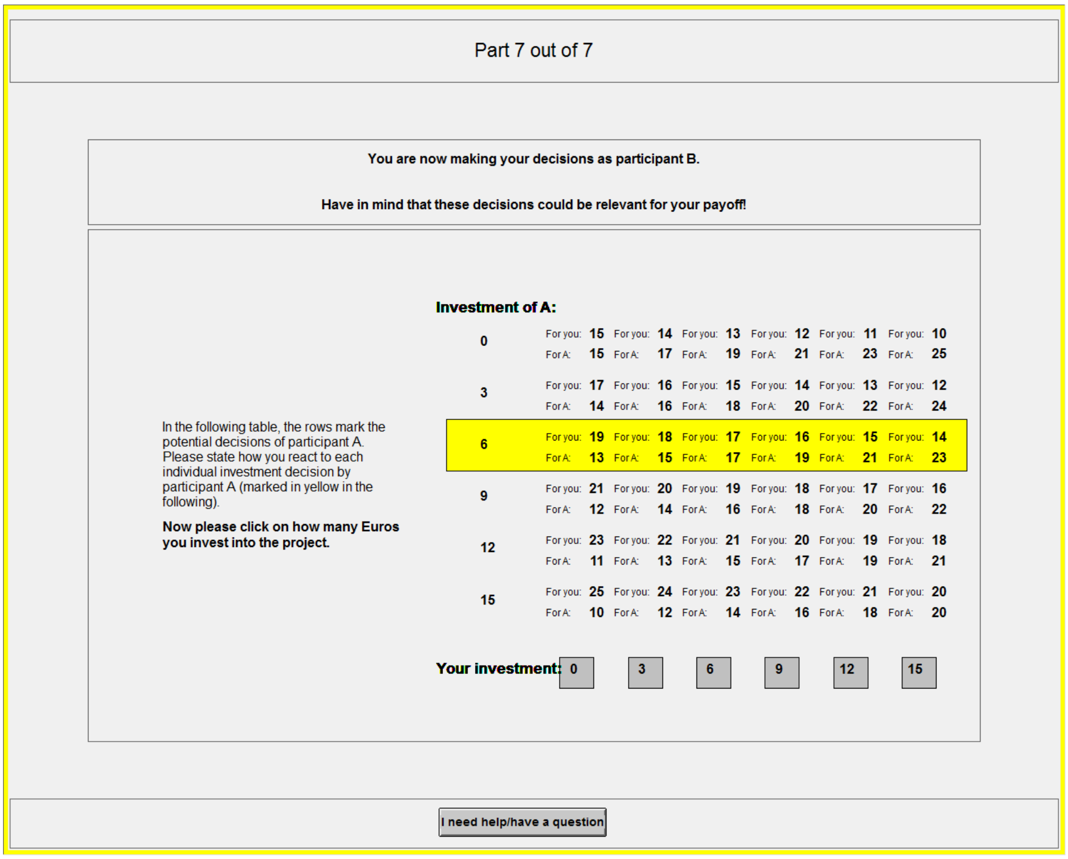
Part 6, STABILITYBELIEFS: instructions; "details" led to an analogous screen as in Part 3.



Part 6, STABILITYBELIEFS: belief elicitation.



Part 7, PREFS3: unconditional contribution; "Complete situation description" led to a screen similar to the instructions screen in Part 3.



Part 7, PREFS3: conditional contribution (preference elicitation).

THURGAU INSTITUTE
OF ECONOMICS
at the University of Konstanz

Hauptstrasse 90
Postfach
CH-8280 Kreuzlingen 2

T +41 (0)71 677 05 10
F +41 (0)71 677 05 11

info@twi-kreuzlingen.ch
www.twi-kreuzlingen.ch