University of Heidelberg

Department of Economics

Discussion Paper Series | No. 669

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

Robert J. Schmidt

September 2019

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

Robert J. Schmidt

September 17, 2019

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

Highlights:

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

Keywords: private information, false consensus effect, surveys, crowd wisdom, beliefs
JEL Classifications: C83, C91, D81

Corresponding author: Robert Johann Schmidt, Alfred-Weber-Institute for Economics, University of Heidelberg, Bergheimer Str. 58, 69115 Heidelberg, Germany. Phone: +49 152 34181300, email: rojoschmidt@gmail.com. I thank Aurelien Baillon, Christian König, Christiane Schwieren, Christoph Vanberg, Jens Witkowski as well as seminar audiences in Heidelberg, the HeiKaMaX in Heidelberg, the HeiKaMaXY in Heidelberg, and the Bayesian Crowd Conference in Rotterdam for very valuable comments and suggestions.

1. Introduction

The elicitation of private information, such as preferences, beliefs, feelings or opinions, is key for social sciences (Turner and Martin, 1985), policymakers (Veenhoven, 2002), corporations (Monroe, 1973) and for public opinion research (Price and Neijens, 1997).[1] The value of such data requires that subjects exert effortful thinking when the question at hand is non-trivial and that subjects do not bias their answer towards social desirability (Li, 2007; Manski, 2004; Zizzo, 2010). Therefore, various methods have been proposed that condition a respondent´s answer on some observable fact and monetarily reward the subject for accuracy.[2] If the monetary incentive is sufficient, the mechanism is incentive compatible, as the subject is induced to honestly report her *type* (Smith, 1976).[3]

Incentivizing accurate reporting by conditioning on facts, however, limits a mechanism to the elicitation of private information about *verifiable* questions.[4] By contrast, the elicitation of *unverifiable* questions lies beyond the scope of this approach. This comprises questions that are hypothetical, counter-factual, technically unverifiable, or which refer to subjective tastes.[5] Therefore, so-called truth serums have been developed, which aim at increasing the quality of the elicitation of private data compared to non-incentivized procedures (Prelec, 2004).[6] The core assumption of these methods is that individuals are subject to the false consensus effect (Ross et al., 1977), a well-documented phenomenon that follows Bayesian reasoning (Dawes, 1989).[7] That is, individuals use their own private information when predicting private information of others.[8]

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

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

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

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

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

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

[7] In the title, we put the word "false" in parentheses, in order to remark that it is *not necessarily false* to derive beliefs about others using the signal that stems from one´s own type (Dawes, 1989). Instead, rational belief formation requires that (to some degree) subjects use their own type as a valid source when predicting the type of others (see also Prelec, 2004). Section 3.1.2 elaborates on that.

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

Given that a set of behavioral assumptions holds, truthfully answering the question is a Bayesian-Nash equilibrium and maximizes the recipient´s payoff.

However, the application of this group of mechanisms comes with two difficulties. First, they require several behavioral assumptions, such as common knowledge about a shared common prior belief, common knowledge about respondents updating their belief in an impersonally informative manner, subjects being able to identify that truth-telling represents a Bayesian Nash equilibrium and trusting others to play according to that Bayesian Nash equilibrium, too.[9] Second, the scoring rules applied in Bayesian revelation mechanisms are complicated. This makes it hard to implement the mechanisms transparently by informing subjects about the exact scoring system. As a result of that, most empirical applications have relied on faith-based implementation, without explaining the actual scoring rules in detail, but by assuring participants that truth-telling would be optimal (e.g., Barrage and Lee, 2010; Howie et al., 2011; John et al., 2012; Prelec and Weaver, 2006; Shaw et al., 2011).

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

In the first mechanism, *Benchmark*, a two-step approach is applied. First, a subject is asked about her private information in an non-incentivized manner. That statement is then used to serve as a benchmark for the elicitation of second-order beliefs (Perner and Wimmer, 1985) from another respondent, who has to guess the private information of the previously asked subject. The existence of a benchmark allows incentivization using ordinary scoring rules, such that the second respondent

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

is induced to engage in second-order reasoning in an effortful manner. The stronger the relationship between a subject´s own thought and second-order belief, the better of a proxy the elicited statement will be for her type.

In the second mechanism, *Coordination*, respondents are provided with a question and various answer alternatives. The subjects´ task is to coordinate on an answer, and each subject is paid based on the precision with which she anticipates the coordination outcome. The mechanism is inspired by the concept of focal points (Schelling, 1960), an approach to predict behavior in coordination settings with multiple equivalent equilibria. The concept postulates that participants in pure coordination settings exhibit shared perceptions about salient alternatives. Thereby, focal points emerge absent from payoffs and provide an implicit coordination device (Sudgen, 1995). Since the recognition of focal points requires subjects to form beliefs about the perception of saliences in other individuals, it involves higher-order reasoning (Camerer et al., 2004). Consequently, an individual´s coordination choice reflects her belief about the perception of others and, therefore, is informative about her own perception about the question at hand.[10]

The main advantage of the two methods is that they are tractable. The scoring and the payout function are easy to understand, such that participants are provided with a clear task that they have to solve. Therefore, the mechanisms are easy to implement for experimenters. By contrast to Bayesian revelation mechanisms, there is no theoretical necessity that subjects reveal their private information. However, it is reasonable to expect valid signals about the respondent´s thoughts. In section 4, we discuss how the mechanisms need to be implemented to maximize the discriminatory power of the two measures.

In an experiment, we mimic the elicitation of beliefs about unverifiable probabilities.[11] Subjects are provided with instructions about an ultimatum game (Güth et al., 1982) conducted by Trautmann and van de Kuilen (2014), and we elicit beliefs about proposer and responder behavior.

[10] As we will show in section 3.2.2., the mechanism represents a generalization of the Krupka and Weber (2013) approach to identify social norms using coordination games. They propose to use coordination games to identify *social norms* on the *aggregate level*, while we argue that coordination games are suited to identify *any kind of private information* on the *individual level*. Indeed, there is evidence that individual coordination choices about social norm perception are related to a subject´s preferences. Schmidt (2019b) finds that injunctive and descriptive social norms elicited using coordination games predict revealed social preferences in a series of dictator games.

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

Applying both a between-subject and a within-subject design, we elicit beliefs using the mechanisms *Benchmark* and *Coordination* as well as actual first-order beliefs by conditioning payments on factual probabilities. In the between-subject comparison, we find that both mechanisms accurately reveal mean first-order beliefs of the population. In the within-subject comparison, we find that the modal difference between probabilities elicited in either mechanism and actual beliefs is zero. We therefore conclude that, in the given setting, subjects strongly anchor their statements in *Benchmark* and *Coordination* on their first-order beliefs.

The remainder of the paper is organized as follows. Section 2 reviews the literature on methods to elicit private information. Section 3 explains *Benchmark* and *Coordination*, and also provides a theoretical background for each mechanism. Section 4 illustrates how the mechanisms need to be implemented to maximize the discriminatory power. In section 5, we present the experiment to test the mechanisms in the area of probabilistic beliefs, and in section 6 we formulate testable hypotheses. Section 7 presents the results. Section 8 discusses advantages and disadvantages compared to Bayesian revelation mechanisms. Section 9 summarizes and concludes.

2. Related Literature

The seminal contribution to eliciting unverifiable subjective information is Prelec (2004), who introduces a truth-inducing scoring system that includes two additive parts. First, an information report that refers to information privately owned by a respondent (her type). The information report is scored based on the degree to which it is surprisingly common in the population.[12] Second, subjects make a prediction report. This report refers to the subject´s belief about the distribution of types in the population, and it is scored based on accuracy. Given a set of behavioral assumptions, such as common knowledge about a shared prior belief, impersonal informativeness, Bayesian reasoning and a sufficiently large sample of participants, truthfully reporting the own type represents a Bayesian Nash equilibrium.

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

Since Prelec´s (2004) innovation, various refinements have been proposed. Prelec and Seung (2006) demonstrate how to use the mechanism even when the majority of respondents are wrong. Witkowski and Parkes (2012a) propose the Robust Bayesian Truth Serum, which corrects Prelec´s (2004) drawback to operate properly only on large samples. In the Robust Bayesian Truth Serum, three subjects are sufficient to establish Bayesian Nash incentive-compatibility, but the mechanism is restricted to the elicitation of binary information. The modifications of Radanovic and Faltings (2013, 2014) allow the elicitation of non-binary signals, while still being incentive-compatible for small populations. Baillon (2017) proposes Bayesian markets, a method that simplifies previous mechanisms. Subjects make only one decision, namely whether or not to trade an asset whose value represents the share of affirmative answers to a question. Bayesian markets are thus more transparent and tractable for participants, but they are suited for binary questions only.

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

Finally, our paper is related to Carvalho et al. (2017) who discuss mechanisms that are based on the assumption that respondents exhibit social projection, a strong form of the false consensus effect.[13] They theoretically analyze payment structures that reward agreements and demonstrate that risk-neutral agents maximize their expected reward by honestly reporting their private information. In an online experiment involving text content-analysis, the subjects´ task is to review short texts under the criteria of grammar, clarity, and relevance. Their results support the hypothesis that agents report more accurate answers than when there are no incentives for honest reporting.

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

3. Benchmark and Coordination

Two methods are proposed to elicit private information in case of unverifiability: *Benchmark* and *Coordination*. In both methods, subjects are incentivized to make statements that depend on perceptions about private information of others. In section 3.1, we explain *Benchmark*, and in section 3.2, we explain *Coordination*. In both subsections, we provide theoretical backgrounds that illustrate why the methods are suited to predict a respondent´s type.

3.1. Benchmark

3.1.1. The Mechanism

Benchmark consists of two steps and requires at least two subjects, a *benchmarker*, and a *respondent*. In step 1, the experimenter asks the benchmarker about some private information in a non-incentivized manner, and her answer is then considered the *benchmark*. In step 2, the respondent is asked to guess the answer of the benchmarker, and she receives a payment that depends on the accuracy of her second-order belief.[14] Creating a benchmark in the first place circumvents the problem that scoring against an objective criterion is not feasible when an answer is unverifiable. Using the benchmark to condition the respondent´s guess allows the application of ordinary scoring rules, thereby inducing her to engage in second-order reasoning in an effortful manner. The closer the relationship between the respondent´s first-order and second-order belief, the better the prediction about her private information.

3.1.2. Theoretical Background: The False Consensus Effect

Benchmark capitalizes on the false consensus effect, a well-documented phenomenon that describes the tendency to perceive the own traits, such as preferences, habits, behaviors, choices, or opinions to be correlated with the corresponding traits of peers (Ross et al., 1977). As a result, subjects of a particular type expect over proportionally large shares of subjects similar to them in the population. Indeed, there is ample evidence that individuals overestimate the prevalence of their own characteristics (Bellemare et al., 2011; Bennett, 1999; Blanco et al., 2014; Charness and

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

Grosskopf, 2001; Marks and Miller, 1987; Mullen et al., 1985; Toussaert, 2018).[15] Also, there is experimental evidence that the false consensus effect is surprisingly robust to information provision (e.g., Ambuehl et al., 2019; Engelman and Strobel, 2012).

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

3.2. Coordination

3.2.1. The Mechanism

In *Coordination*, several subjects receive the same question, and they have to coordinate on the answer. Subjects are compensated based on their ability to anticipate the *coordination outcome*, which is determined as a function of all coordination choices. In case of verbal answers, this is usually the modal answer (e.g., Mehta et al., 1994a, 1994b; Krupka and Weber, 2013). In case of coordination with numbers, this could be the average, the median or the mode.[16] The higher a subject´s accuracy in anticipating the coordination outcome, the higher her payment.

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

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

Coordination is different from *Benchmark* in three aspects. First, by contrast to *Benchmark*, it does not require two steps since the elicitation of private information and the creation of the coordination outcome happen simultaneously. Second, while in *Benchmark* two subjects are needed to make the mechanism work, this is not necessarily the case in *Coordination*. Specifically, the mechanism requires that subjects perceive the coordination outcome to be exogenous, i.e., a single participant is not able to influence the coordination outcome.[17] This requires that the number of participants is sufficiently large. Third, *Benchmark* and *Coordination* differ in the potential depth of reasoning required in the settings (Camerer et al., 2004; Nagel, 1995; Stahl and Wilson, 1994, 1995;). *Benchmark* only requires the formation of second-order beliefs (beliefs about the thoughts of others). By contrast, coordination games are complex, and subjects might engage in the formation of even higher-order beliefs, in order to anticipate the coordination outcome in a more sophisticated manner. This, however, does not pose a threat to the proposed mechanism, as long as beliefs of higher-order depend on a subject´s first-order belief, i.e., her own thought about the question at hand. The mechanism *Coordination* thus relies on the assumption that a subject´s first-order belief tends to be correlated with beliefs of all orders (Dawes, 1989, 1990; Engelman and Strobel, 2012; Vanberg, 2019).

3.2.2. Theoretical Background: Focal Points in Coordination Games

Coordination capitalizes on the theory of focal points, a concept proposed by Schelling (1960) to understand behavior in pure coordination settings. Schelling (1960) argues that in pure coordination games with multiple equivalent equilibria, subjects perceive varying degrees of saliences of the available alternatives, and they assume that their perception is shared by the remaining participants (Sudgen, 1995). As a result, subjects use their own perception about salient choices to make predictions about how saliencies are perceived by other participants.[18] This creates

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

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

focal points that emerge absent from payoffs, thereby constituting an implicit coordination device.[19]

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

4. Maximizing Discriminatory Power in Benchmark and Coordination

4.1. Discriminatory Power

We argue that, based on the phenomenon of false consensus, a subject´s choice in *Coordination* and *Benchmark* yields an informative signal about the respondent´s type. In order to maximize the informativeness stemming from the two mechanisms, the task should be constructed such that subjects which are of different types respond to the task in different ways. In test theory, this feature is referred to as *discriminatory power* and it has been extensively studied in that domain (e.g., Birnbaum et al., 1968; Ferrando, 2012; Hankins, 2007; Loevinger, 1954). Discriminatory power measures the degree to which a test score varies with the level of the measured trait and thus reflects the effectiveness of a test detect differences between participants concerning the respective trait.[20] To illustrate this, imagine a test that is either extremely easy or extremely hard. In both cases, the

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

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

variance will be low, as the average performance will be either close to the minimum or close to the maximum number of points. To render the distribution of scores informative, the test shall be likely to yield *higher* scores for more capable subjects and *lower* scores for less capable subjects. That is, the test shall yield variable results, *given* that the test-takers are *indeed different*. Therefore, the difficulty of the test needs to be calibrated such that average performances correspond to an expected number of solved tasks lying in the middle of the total number of items. If the difficulty of the task is appropriate, it becomes likely that heterogenous test-takers receive varying scores.

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

4.2. Example: Eliciting Minority Opinions using Numerical Questions

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

That approach, however, requires the experimenter to have a reliable prior about the distribution of first-order beliefs in the population. Another possibility is to provide subjects with numerical answers in a more nuanced way. In the above-described example, the experimenter could have subjects state percentage points for the probability that Team A wins. In the case of

Benchmark, the experimenter would first elicit the first-order belief of the benchmarker, who states an integer between 0 and 100 that shall represent the probability in percent that the event occurs. The respondent is then asked to guess the integer stated by the benchmarker and is then paid based on the accuracy of her second-order belief. Equivalently, in *Coordination* subjects could coordinate on an integer between 0 and 100. The coordination outcome is determined as a function of the coordination choices, e.g., the mean, the median, or the mode. Each participant is then paid based on the distance between her coordination choice and the coordination outcome.[21] The significant advantage of using numerical scales is that, by design, it is likely to receive variation in the respondents´ answers, since many numbers are a potentially promising bet in the settings of *Benchmark* and *Coordination*.

5. Experimental Design and Procedure

We mimic the elicitation of unverifiable beliefs and examine whether the proposed mechanisms are suited to reveal subjects´ first-order beliefs.[22] The participants´ task is to estimate empirical probabilities of events in an ultimatum game conducted by Trautmann and van de Kuilen (2014), hereafter TK.[23] At the beginning of our experiment, subjects learn that it is their task to estimate probabilities about behavior in an ultimatum game that has already been conducted. For that sake, subjects receive detailed information about TK´s ultimatum game. They are then instructed about their tasks in the respective treatments and the scoring mechanisms. In section 5.1, we elaborate on the rules of TK´s ultimatum game. In section 5.2, we present the design of our treatments and in section 5.3 the procedure of our experiment.

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

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

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

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

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

We pay significant attention to make sure that subjects understand the rules of TK´s ultimatum game and to make clear that it is not their task to play the game themselves but to assess observed behavior rates of others in that game. Subjects are provided with the original wording of TK´s instructions and answer a series of comprehension questions.[24] Also, participants in our experiment receive information about the general setting of TK´s experimental procedure (computerized laboratory experiment, show-up fee of 5€, random assignment of roles, anonymous interaction, etc.). Table 1 shows the available allocations as well as empirical probabilities of proposer choices and responder acceptance rates in TK.

Table 1. Ultimatum Game of Trautmann and van de Kuilen (2014)

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

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

5.2. Treatments and Scoring

Treatments. Four main treatments are conducted: *SURVEY*, *BELIEF*, *BENCHMARK*, and *COORDINATION*. Additionally, we conduct *CONTROL*, a control treatment that is intended to capture the degree of noise inherent in the elicitation of beliefs in the given setting. Subjects are instructed about the task in the respective treatment, i.e., whether their task is to state first-order beliefs, second-order beliefs, or whether their task is to coordinate. Probabilities are separately elicited for the 12 possible events in TK´s ultimatum game. Precisely, subjects state for each of the six possible allocations (i) how probable it was that a proposer chose a particular allocation and (ii) how probable it was that a responder accepted a particular allocation. Our design is intended to compare first-order beliefs, second-order beliefs, and coordination choices both in a between subject-manner and in a within-subject manner (see table 2).

• ***SURVEY:*** In treatment *SURVEY*, first-order beliefs are elicited in an non-incentivized manner. Subjects asses the probabilities of the 12 events of TK´s ultimatum game and receive a fixed payment of 12.50€ for their participation in the experiment. Treatment *SURVEY* is intended to yield non-incentivized beliefs that are then used to score second-order beliefs elicited in *BENCHMARK*.[25]

• ***BELIEF:*** In treatment *BELIEF*, first-order beliefs are elicited in an incentivized manner. Subjects are instructed that their payment depends on the precision of their first-order beliefs about the factual probabilities in TK. At the end of the experiment, the computer randomly draws one item, and the performance in that item determines a respondent´s payoff.

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

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

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

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

Scoring. In each treatment (except *SURVEY*) subjects are paid based on accuracy regarding the respective task, and performance is evaluated relative to the other subjects within a session. Subjects within a session are ranked from highest to lowest accuracy regarding the respective task. In *BELIEF*, subjects are ranked according to the accuracy of their first-order belief in one randomly drawn item. In *BENCHMARK*, subjects are ranked according to the accuracy of their second-order belief. In *COORDINATION*, subjects are ranked according to their ability to anticipate the coordination outcome. The subject with the highest accuracy earns 15€. Payment then linearly diminishes by 0.75€ by each rank. That is, the subject with the second-highest performance earns 14.25€, the subject with the third-highest performance earns 13.50€, and so forth.[26] Since all sessions were conducted with 20 participants, the incentives created through the relative payment scheme are identical. In addition to that payment, subjects receive a show-up fee of 5€. By contrast to these treatments, in treatment *SURVEY*, subjects receive a show-up fee of 12.50€.[27]

After instructing subjects about their specific task and the scoring mechanisms, they answer a series of control questions. This way, we make sure that they understand their task, and how their compensation would be determined in the respective treatments. Table 2 summarizes the structure

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

[27] In expectancy, the payment between the five treatments is (almost) identical. The expected payoff in the treatments with relative payments is 12.88€.

of treatments and illustrates the between-subject and the within-subject comparisons that the experiment allows.

Table 2. Treatment Overview and Content

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

5.3. Procedure

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

6. Hypotheses

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

second-order beliefs are $\overline{SB_j} = (\sum_{i=1}^{n} SB_{ij})/n$ and average coordination choices are $\overline{C_j} = (\sum_{i=1}^{n} C_{ij})/n$.

We model second-order beliefs and coordination choices as a function of first-order beliefs: $SB_{ij} = FB_{ij} + \varepsilon_{ij}^{SB}$ and $C_{ij} = FB_{ij} + \varepsilon_{ij}^{C}$. The error terms ε_{ij}^{SB} and ε_{ij}^{C} capture the difference between a respondent´s statement in the respective mechanism and her actual first-order belief. One way to interpret these error terms is that they result from an anchoring and adjustment procedure (Tversky and Kahnemann, 1974).[28] That is, subjects *anchor* their statements in *Benchmark* and *Coordination* on their first-order belief, and they then *adjust* it deepening on their perception about the coherence between their own perception and their best guess about the perception of others (Epley et al., 2004).[29] Accordingly, averages of the population can be formulated as $\overline{SB_j} = \overline{FB_j} + \overline{\varepsilon_j^{SB}}$ and $\overline{C_j} = \overline{FB_j} + \overline{\varepsilon_j^{C}}$.[30] The model illustrates when the mechanisms Benchmark and Coordination work best, namely when ε_{ij}^{SB} and ε_{ij}^{C} are small. The following hypotheses formulate how ε_{ij}^{SB} and ε_{ij}^{C} as well as $\overline{\varepsilon_j^{SB}}$ and $\overline{\varepsilon_j^{C}}$ look like.

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

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

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

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

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

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

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

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

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

7. Results

7.1. Between-Subject Analysis of Averages

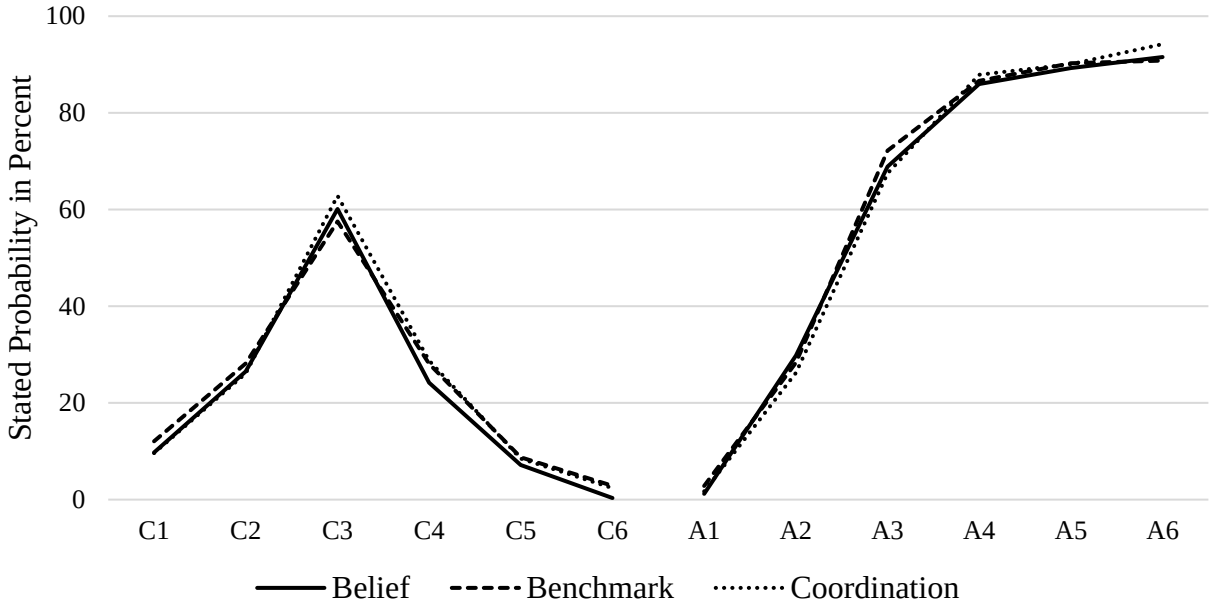
We start with aggregate level analysis by comparing average first-order beliefs with (i) average second-order beliefs and (ii) average coordination choices. Figure 1 shows average first-order beliefs elicited in *BELIEF*, average second-order beliefs elicited in stage 1 of *BENCHMARK*, and average coordination choices elicited in stage 1 of *COORDINATION*. Items C1 to C6 refer to probabilities of proposers-choices and A1 to A6 refer to acceptance-rates of responders. Mann-Whitney-U tests are conducted to test for equality of averages. Before correcting for multiple testing, item C6 differs between *BENCHMARK* and *BELIEF* ($p < 0.05$) and the same item differs between *COORDINATION* and *BELIEF* ($p < 0.01$).[31] Both significances vanish when correcting for multiple testing using the Bonferroni procedure.[32] We therefore cannot reject hypotheses 1A and 1B, which state that the average probabilities elicited in *BENCHMARK* and *COORDINATION* are identical to average first-order beliefs elicited in *BELIEF*.

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

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

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

Figure 1. Between-Subject Comparison of Elicited Probabilities



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

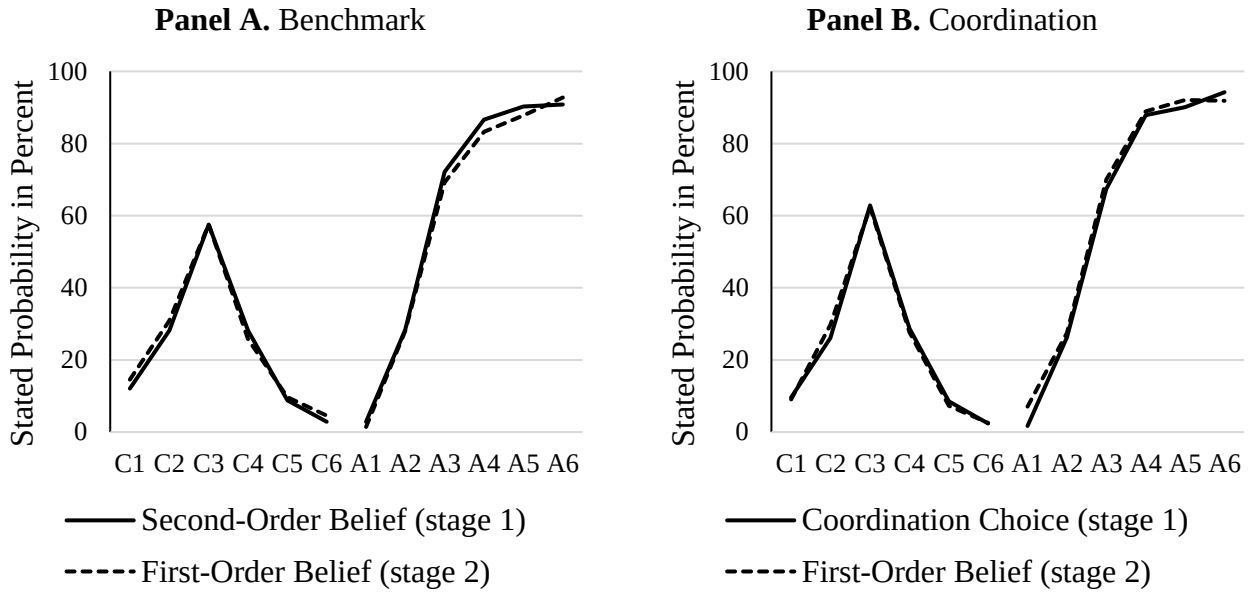
7.2. Within-Subject Analysis of Averages

To examine differences in *BENCHMARK* and *COORDINATION* to first-order beliefs on the individual level, we start by comparing average outcomes between stage 1 and stage 2 in these treatments. Remember that in stage 1, the respective mechanisms are applied, i.e., subjects state their second-beliefs in stage 1 of *BENCHMARK*, and they coordinate in stage 1 of *COORDINATION*. In stage 2, first-order beliefs are elicited in the same manner as in *BELIEF*. By contrast to the above section, we now compare the outcomes of the mechanisms with first-order beliefs in a within-subject manner. Panel A of Figure 2 compares averages of stage 1 and stage 2 in *BENCHMARK*, and Panel B compares averages of stage 1 and stage 2 in *COORDINATION*. Wilcoxon Signed-Rank tests are conducted to detect differences between averages. Before correcting for multiple testing, item C1 ($p < 0.1$) and item C4 ($p < 0.05$) differs between stage 1 and stage 2 in *BENCHMARK*. In *COORDINATION*, item C2 ($p < 0.05$), item A1 ($p < 0.05$) and

item A4 ($p < 0.1$) differ between stage 1 and stage 2. Again, these significances vanish after the correction procedure. The results are thus consistent with those reported in the previous section, i.e., average second-order beliefs elicited in *BENCHMARK* and average coordination choices elicited in *COORDINATION* do not differ from average first-order beliefs of subjects.

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

Figure 2. Within-Subject Comparison of Elicited Probabilities



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

7.3. Correlations

We now analyze to what degree probability statements elicited in *BENCHMARK* and *COORDINATION* are related to first-order beliefs of individuals by conducting correlation analyses. Looking at the combined data of all items, we find that the statements in stage 1 and stage 2 are strongly and statistically significantly correlated in both treatments ($r = 0.87$ in *BENCHMARK*;

r = 0.90 in *COORDINATION*; $p < 0.00001$ in both treatments; Pearson correlation).[33] That result is consistent with the idea promoted in section 4, i.e., that the statements extracted in *BENCHMARK* and *COORDINATION* vary with the underlying first-order belief of individuals. Likewise, the correlation between stage 1 and stage 2 is strongly positive and statistically significant in treatment *CONTROL* (r = 0.89; $p < 0.00001$; Pearson correlation), but the correlation is not higher than in *BENCHMARK* and *COORDINATION*.

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

7.4. Analysis of Error Terms

We proceed by analyzing the congruence between numbers stated in stage 1 and stage 2 of *BENCHMARK* and *COORDINATION*. For that sake, we examine the distribution of error terms ε_{ij}^{SB} and ε_{ij}^{C} defined in section 6, which emerge when a subject states different numbers in stage 1 and stage 2 for the same item.[34] It is reasonable to expect that subjects will exhibit noise when stating their beliefs for 12 items two times in a row. To have a baseline to compare the distribution of error terms with, we use the error terms observed in treatment *CONTROL*, which provide a measure for the degree of noise that occurs when subjects state first-order beliefs.

In order to get an impression about that measure, Figure 3 shows the distribution of individual error terms based on the combined data of all items.[35] The distribution is centered around zero, and the modal error term, as well as the median error term in each treatment, equal zero (see Table 3). Two-sided t-tests are conducted to test if mean error terms on the item level differ from zero.[36] We do not find that error terms in any item differ from zero, neither in *BENCHMARK* nor *COORDINATION*.[37] The fact that error terms do not differ from zero is consistent with Hypotheses 1A and 1B.

---

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

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

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

[36] In Appendix A.1, Table 5, Panel A, we report mean error terms on the item level.

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

To investigate Hypotheses 2A and 2B, we analyze means of *absolute* error terms: $|\varepsilon_{ij}^{SB}|$ and $|\varepsilon_{ij}^{C}|$.[38] Two-sided t-tests are conducted to test if mean absolute error terms on the item level differ from zero.[39] We find that in all three treatments, in most items mean absolute error terms are significantly different from zero on the 5%-level.[40] After the correction procedure, still, most items remain significantly different from zero on the 5%-level. This result is not consistent with Hypotheses 2A and 2B. In order to identify what part of these differences is due to noise, we compare mean absolute error terms in *BENCHMARK* and *COORDINATION* with mean absolute error terms observed in treatment *CONTROL*. Conducting Mann-Whitney-U tests to identify differences between treatments, we do not find that *BENCHMARK* or *COORDINATION* differs from *CONTROL* in terms of absolute error terms.

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

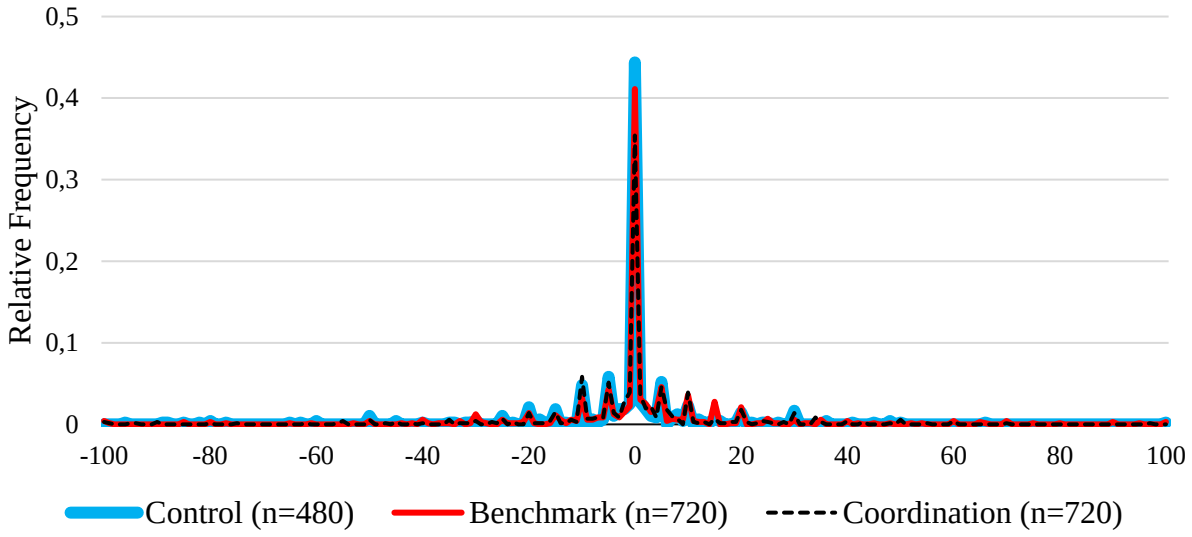
Result 5. Mean absolute error terms significantly differ from zero in each treatment. However, mean absolute error terms observed in *BENCHMARK* and *COORDINATION* do not differ from mean absolute error terms in *CONTROL*.

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

[39] In Appendix A.1, Table 5, Panel B, we report mean absolute error terms on the item level.

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

Figure 3. Distribution of Individual Error Terms (all Items)



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

Table 3. Error Terms and Absolute Error Terms (all Items)

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

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

7.5. External Validity

Figure 4 (Appendix A.1) indicates a high external validity of all mechanisms, as subjects are accurately assessing the patterns of proposer choices and responder acceptance rates in each treatment. To evaluate external validity, mean Brier scores (Brier, 1950), i.e., average squared deviations between factual data and elicited beliefs, are calculated and reported in Table 4. Before the correction procedure, Brier scores of item C6 differ between *BENCHMARK* and *BELIEF* ($p < 0.05$), and the same item differs between *COORDINATION* and *BELIEF* ($p < 0.01$). None of these differences survive the correction procedure. That is, external validity in *BENCHMARK* and

COORDINATION does not differ from the degree of external validity that is obtained when first-order beliefs are elicited in an ordinary manner.

Table 4. Mean Brier Scores

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

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

7.6. Discussion of Results and Evaluation of Hypotheses

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

However, two considerations put the results on absolute error terms into perspective. First, the mean of absolute error terms is significantly larger than the median of absolute error terms in both treatments (see Table 3). Almost half of the estimations extracted in *BENCHMARK* and *COORDINATION* are identical with first-order beliefs, and also the median differences indicate negligible deviations between statements extracted in the two mechanisms and actual first-order beliefs. In the given setting, the median difference is more informative, since the mean is strongly affected by few subjects that enter strongly diverging numbers in the two stages (thereby strongly increasing the mean of absolute error terms). Second, as seen in treatment *CONTROL*, the degree of noise inherent in the setting equals the extent of error terms in *BENCHMARK* and

COORDINATION. The deviations on the individual level thus seem to be driven by subjects being ambiguous about their actual first-order belief, thus creating noise.

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

8. Advantages Compared to Bayesian Revelation Mechanisms

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

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

lack of trust regarding Bayesian Nash play of other subjects might refrain a subject from playing Bayesian Nash herself (Knez and Camerer, 1994).

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

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

9. Summary and Conclusion

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

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

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

subjects strongly anchor their statements in *Benchmark* and *Coordination* on their first-order beliefs.

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

Appendix

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

A.1.1. Mean Error Terms

Panel A of Table 5 shows means of error terms ε_{ij} on the item level. The average error term $\overline{\varepsilon_j}$ of item j is calculated as $\overline{\varepsilon_j} = (\sum_{i=1}^{n} \varepsilon_{ij})/n$. Two-sided t-tests are conducted to test if mean error terms on the item level differ from zero. We do not find that error terms in any item differs from zero neither in *BENCHMARK* or *COORDINATION* nor in *CONTROL*.

A.1.2. Mean Absolute Error Terms

Panel B of Table 5 shows means of absolute error terms $|\varepsilon_{ij}|$ on the item level. The average absolute error term $\overline{|\varepsilon_j|}$of item j is calculated as $\overline{|\varepsilon_j|} = (\sum_{i=1}^{n} |\varepsilon_{ij}|)/n$. Two-sided t-tests are conducted to test if mean absolute error terms on the item level differ from zero. We find that in all three treatments, in most items mean absolute error terms are significantly different from zero on the 5%-level. Precisely, in *BENCHMARK* and *CONTROL*, in 11 of the 12 items mean absolute error terms differ from zero with $p < 0.05$; in *COORDINATION*, in 10 items mean absolute error terms differ from zero with $p < 0.05$. Mann-Whitney-U tests are conducted to test for differences between treatments. Before the correction procedure, item C6 differs between *BENCHMARK* and *CONTROL* ($p < 0.05$)

and items C3 ($p < 0.1$), C5 ($p < 0.01$) and A1 ($p < 0.05$) differ between *COORDINATION* and *CONTROL*. None of these differences survives the correction procedure.

Table 5. Analysis of Error Terms on the Item Level

Panel A. Mean Error Terms $\overline{\varepsilon_j}$

	C1	C2	C3	C4	C5	C6	A1	A2	A3	A4	A5	A6
Benchmark	-2,5	-2,7	0,0	2,4	-0,9	-1,7	1,4	0,3	3,0	3,3	2,5	-1,9
Coordination	0,5	-3,6	0,5	1,0	1,2	-0,2	-5,4	-1,2	-2,6	-1,0	-2,0	2,3
Control	-1,6	-0,6	-1,5	1,4	0,1	-0,2	-0,2	-4,5	-3,1	-3,3	-5,2	-7,9

Panel B. Mean Absolute Error Terms $\overline{|\varepsilon_j|}$

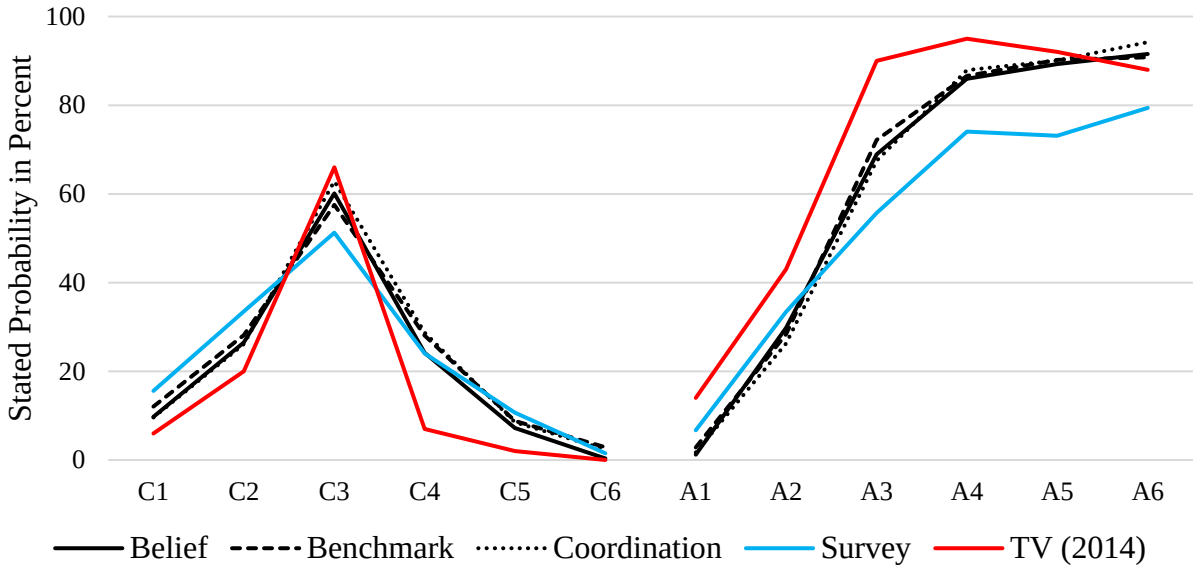
	C1	C2	C3	C4	C5	C6	A1	A2	A3	A4	A5	A6
Benchmark	10,0	12,8	10,8	9,0	6,6	3,7	1,8	13,3	14,6	12,6	9,7	10,0
Coordination	7,9	15,0	12,7	15,5	8,1	3,7	6,1	12,0	11,0	8,1	5,3	3,0
Control	5,4	13,8	9,7	8,2	3,6	0,5	0,5	10,6	12,6	11,4	11,1	13,9

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

A.2. Results of Treatment Survey

Graphical analysis (figure 4) as well as mean Brier scores (table 6) indicate a lower external validity of *SURVEY*, compared to *BELIEF*, *BENCHMARK*, and *COORDINATION*. The number of participants, however, is not sufficient to draw statistical inferences on that question.

Figure 4. Extracted Beliefs and Factual Data of TK (2014)



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

Table 6. Mean Brier Scores

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

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

References

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

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

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

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

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

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

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

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

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

Brier, Glenn W. (1950). Verification of Forecasts Expressed in Terms of Probability. *Monthly*

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Schelling, Thomas C. 1960. The Strategy of Conflict. Cambridge: Harvard University Press.

Schlag, K. H., Tremewan, J., & Van der Weele, J. J. (2015). A penny for your thoughts: A survey of methods for eliciting beliefs. *Experimental Economics*, *18*(3), 457-490.

Schmidt, R. J. (2019a). Point Beauty Contest: Measuring the Distribution of Focal Points on the Individual Level, University of Heidelberg: AWI Discussion Paper Series.

Schmidt, R. J. (2019b). Do Injunctive or Descriptive Social Norms Elicited Using Coordination Games Better Explain Social Preferences?, University of Heidelberg: AWI Discussion Paper Series.

Schoenbaum, M. (1997). Do smokers understand the mortality effects of smoking? Evidence from the Health and Retirement Survey. *American Journal of Public Health*, *87*(5), 755-759.

Schotter, A., & Trevino, I. (2014). Belief elicitation in the laboratory. *Annu. Rev. Econ.*, *6*(1), 103-128.

Selten, R. (1967). Die Strategiemethode zur Erforschung des eingeschränkt rationalen Verhaltens im Rahmen eines Oligopolexperiments, S. 136–168. *Tübingen: JCB Mohr (Paul Siebeck)*.

Shaw, A. D., Horton, J. J., & Chen, D. L. (2011, March). Designing incentives for inexpert human raters. In *Proceedings of the ACM 2011 conference on Computer supported cooperative work* (pp. 275-284). ACM.

Slovic, P., Fischhoff, B., & Lichtenstein, S. (1980). Facts and fears: Understanding perceived risk. In *Societal Risk Assessment* (pp. 181-216). Springer, Boston, MA.

Smith, V. L. (1976). Experimental economics: Induced value theory. *The American Economic Review*, *66*(2), 274-279.

Stahl II, D. O., & Wilson, P. W. (1994). Experimental evidence on players' models of other players. *Journal of Economic Behavior & Organization*, *25*(3), 309-327.

Stahl, D. O., & Wilson, P. W. (1995). On players′ models of other players: Theory and experimental evidence. *Games and Economic Behavior*, *10*(1), 218-254.

Sugden, R. (1995). A theory of focal points. *The Economic Journal*, *105*(430), 533-550.

Sugden, R., & Zamarrón, I. E. (2006). Finding the key: the riddle of focal points. *Journal of Economic Psychology*, *27*(5), 609-621.

Toussaert, S. (2018). Eliciting Temptation and Self-Control Through Menu Choices: A Lab Experiment. *Econometrica*, *86*(3), 859-889.

Trautmann, S. T., & van de Kuilen, G. (2014). Belief elicitation: A horse race among truth serums. *The Economic Journal*, *125*(589), 2116-2135.

Turner, C., & Martin, E. (1985). *Surveying subjective phenomena* (Vol. 2). Russell Sage Foundation.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*(4157), 1124-1131.

Vanberg, C. (2019). A short note on the rationality of the false consensus effect, University of Heidelberg: AWI Discussion Paper Series No. 662.

Veenhoven, R. (2002). Why social policy needs subjective indicators. *Social Indicators Research*, *58*(1-3), 33-46.

Witkowski, J., & Parkes, D. C. (2012a). A robust bayesian truth serum for small populations. In *Twenty-Sixth AAAI Conference on Artificial Intelligence*.

Witkowski, J., & Parkes, D. C. (2012b). Peer prediction without a common prior. In *Proceedings of the 13th ACM Conference on Electronic Commerce* (pp. 964-981). ACM.

Zizzo, D. J. (2010). Experimenter demand effects in economic experiments. *Experimental Economics*, *13*(1), 75-98.