Urbig, D.; Stauf, J.; Weitzel, Utz

Working Paper

What is your level of overconfidence? A strictly incentive compatible measurement of absolute and relative overconfidence

Discussion Papers Series, No. 09-20

Provided in Cooperation with:
Utrecht University School of Economics (U.S.E.), Utrecht University

Suggested Citation: Urbig, D.; Stauf, J.; Weitzel, Utz (2009) : What is your level of overconfidence? A strictly incentive compatible measurement of absolute and relative overconfidence, Discussion Papers Series, No. 09-20, Utrecht University, Utrecht School of Economics, Tjalling C. Koopmans Research Institute, Utrecht

This Version is available at:
https://hdl.handle.net/10419/322803

Tjalling C. Koopmans Research Institute

Discussion Paper Series nr: 09-20

What is your level of overconfidence? A strictly incentive compatible measurement of absolute and relative overconfidence

Diemo Urbig
Julia Stauf
Utz Weitzel

Universiteit Utrecht

Utrecht School of Economics

Tjalling C. Koopmans Research Institute
Utrecht School of Economics
Utrecht University

Janskerkhof 12
3512 BL Utrecht
The Netherlands
telephone +31 30 253 9800
fax +31 30 253 7373
website www.koopmansinstitute.uu.nl

The Tjalling C. Koopmans Institute is the research institute and research school of Utrecht School of Economics.
It was founded in 2003, and named after Professor Tjalling C. Koopmans, Dutch-born Nobel Prize laureate in economics of 1975.

In the discussion papers series the Koopmans Institute publishes results of ongoing research for early dissemination of research results, and to enhance discussion with colleagues.

Please send any comments and suggestions on the Koopmans institute, or this series to J.M.vanDort@uu.nl

ontwerp voorblad: WRIK Utrecht

How to reach the authors

Please direct all correspondence to the first author.

Diemo Urbig
Utz Weitzel
Max Planck Institute for Economics
Entrepreneurship, Growth, and Public Policy Group
Kahlaische Straße 10, 07745
Jena, Germany
E-mail: urbig@econ.mpg.de
Utrecht University
Utrecht School of Economics
Janskerkhof 12
3512 BL Utrecht
The Netherlands.
E-mail: u.weitzel@econ.uu.nl
Julia Stauf
University of Cologne,
Cologne Graduate School in Management, Economics, and Social Sciences
Albertus-Magnus-Platz, 50923 Cologne,
Germany
E-mail: stauf@wiso.uni-koeln.de

Utrecht School of Economics
Tjalling C. Koopmans Research Institute
Discussion Paper Series 09-20

What is your level of overconfidence? A strictly incentive compatible measurement of absolute and relative overconfidence.

Diemo Urbig[ab]
Julia Stauf[c]
Utz Weitzel[ab]

[a]Max Planck Institute for Economics
Jena, Germany

[b]Utrecht School of Economics
Utrecht University

[c]Cologne Graduate School in Management, Economics, and Social Sciences
University of Cologne

August 2009

Abstract

This study contributes to the ongoing discussion on the appropriate measurement of overconfidence, in particular, its strictly incentive compatible measurement in experiments. Despite a number of significant advances in recent research, several important issues remain to be solved. These relate to the strictness of incentive compatibility, the identification of well-calibrated participants, the trichotomous classification into over- or underconfident and well-calibrated participants, and the generalization to measuring beliefs about the performance relative to other people. This paper develops a measurement of overconfidence that is improved regarding all four of these issues. We theoretically prove that our method is strictly incentive compatible and robust to risk attitudes within the framework of Cumulative Prospect Theory. Furthermore, our method allows the measurement of various levels of overconfidence and the direct comparison of absolute and relative confidence. We tested our method, and the results meet our expectations, replicate recent results, and show that a population can be simultaneously overconfident, well-calibrated, and underconfident. In our specific case, we find that more than ninety-five percent of the population believe to be better than twenty-five percent; about fifty percent believe to be better than fifty percent; and only seven percent believe to be better than seventy-five percent.

Keywords: Belief elicitation, Overconfidence, Better than average, Incentive compatibility

JEL classification: C91, D8, D83, D84

1 INTRODUCTION

Overconfidence is a frequently observed, real-life phenomenon. Individuals exaggerate the precision of their knowledge, their chances for success, for being better than others, or the precision of specific types of information. Although all these exaggerations are related to the term overconfidence (for an overview, see Griffin and Varey, 1996; Larrick et al., 2007; Moore and Healy 2008), in this paper, we particularly focus on the overestimation of own performance in a knowledge-based task. This overestimation can relate to achieving an objective standard of performance or to be better than others, which we will refer to as absolute and relative overconfidence, respectively.

Empirically, it has been shown that overconfidence in own performance can affect an entrepreneur's or manager's decision to enter a market (Camerer and Lovallo, 1999; Wu and Knott 2006) or to invest in projects (Malmendier and Tate, 2005), a stock trader's decision to buy specific stocks (Daniel et al., 1998; Stotz and Nitzsch, 2005; Cheng 2007), or an acquirer's decision to take over a target firm (Malmendier and Tate, 2008). Lawyers' and applicants' probabilities of success are likely to depend on confidence (Compte and Postlewaite, 2004), and physicists have been shown to be overconfident in their choices of medical treatment (Baumann et al., 1991). Especially the last example illustrates that the consequences of overconfidence do not only affect the decision maker, but can also have significant ramifications for third parties (e.g., patients, clients, investors, employees) as well as the economy and our society as a whole.

One stream in overconfidence research attempts to identify mechanisms that lead to overconfidence, like misestimating the difficulty of tasks based on biased experiences (Juslin and Olson, 1997; Soll, 1996). A further research stream studies how overconfidence affects evaluations of risky decision options and subsequent decisions (Simon et al., 2000; Keh et al., 2002; Cheng, 2007). Another stream of research, in which our study is embedded, is concerned with the correct definition and measurement of overconfidence. In an early study, Fischhoff et al. (1977) consider incentives within overconfidence measurements as a potential source of measurement errors. Erev et

al. (1994), Dawes and Mulford (1996), and Juslin et al. (2000) argue that measurement errors make it rather difficult to distinguish between 'real' overconfidence and 'measured' overconfidence, the latter of which may be confounded with measurement errors.

In fact, when reporting their beliefs, individuals may not be motivated to make estimations that are as precise as in situations, where a significant amount of money is at stake (Fischhoff et al., 1977). Several studies therefore incentivize the revelation of true beliefs by paying participants according to the precision of their judgments (Budescu et al., 1997; Moore and Healy, 2008), or by letting participants invest into bids on their own performance, which indirectly elicits their probability judgments (Moore and Kim, 2003). Cesarini et al. (2006) and Hoelzl and Rustichini (2005) test the effect of monetary incentives and indeed find significant differences between treatments in which participants are incentivized and those in which they are not.

However, the incentive mechanisms currently available to experimentally elicit individuals' overconfidence have their respective weaknesses. First, recently proposed incentive-based mechanisms do not elicit degrees of overconfidence, but only whether an individual or a group of individuals is overconfident or not (e.g., Hoelzl and Rustichini, 2005; Blavatskyy, 2009). Measuring levels of overconfidence may, however, be important, because more extreme levels of overconfidence are less likely to be affected or driven by measurement errors. Furthermore, believing to be better than average, as for instance investigated in Hoelzl and Rustichini (2005), does not imply overconfidence in being the best or one of the top ten percent of a population, which is likely to be more relevant for highly competitive environments as, for example, in patent races, takeover contests, or other performance-related rivalry, where the winner(s) take(s) it all. Second, most elicitation methods are distorted by risk attitudes, as in the case of uncalibrated proper scoring rules (e.g., Moore and Kim, 2003; Moore and Healy, 2008; Budescu et al., 1997). Given that risk attitudes differ within populations (e.g., Bonin et al., in press), differences in levels of overconfidence could possibly result from a confound of measuring overconfidence by variations in risk attitudes,. Third, in some cases the incentive compatibility of the measurement is based on rather limiting assumptions (e.g., Blavatskyy, 2009), such as hidden chaining, epsilon truthfulness,

or the requirement of perfectly eliciting indifference. These assumptions are to some extent inconsistent with the aim of providing clear incentives for reporting true confidence levels and, as shown in this paper, may cause serious biases in measurements, for instance by classifying too few individuals as well-calibrated.

In this paper, we therefore develop an alternative method of overconfidence elicitation that identifies levels of overconfidence, is robust to risk attitudes, and is strictly incentive compatible within the framework of rank-dependent utility theories. Moreover, our method enables a direct within-subject comparison of absolute and relative overconfidence, as both types are measured with the same methodology.

In the course of developing our method, we formally show strict incentive compatibility of the proposed design which elicits beliefs indirectly without asking about probabilities. We then apply our method by testing it experimentally. The results support our method by showing a higher detection rate of well-calibrated participants than earlier studies. Further, the experiment provides first evidence for the importance of measuring different levels of overconfidence. We find that participants are simultaneously over- and underconfident at the population level, depending on the thresholds of relative performance. For instance, although ninety-five percent of participants believe to be better than twenty-five percent of the population (implying overconfidence for small thresholds), only seven percent believe to be better than seventy-five percent (implying underconfidence for high thresholds). Although this has not been measured before, we argue that the application of our method can provide valuable new insights into, for instance, over- or underinvestment in competitive situations, where the winner takes it all, or where it is important to be among a relatively small group of top performers. Prominent examples of such situations are patent races, where investment in research and development depends on the firm's confidence in its relative performance, or takeover auctions, where the highest bid depends on the acquirer's confidence in realizing enough synergies to refinance the deal.

The paper is organized as follows. The next section clarifies important definitions in the measurement of overconfidence (section 2), followed by a discussion of existing methods of

incentivized overconfidence elicitation and possible further improvements (section 3). Here, the latest method, proposed by Blavatskyy (2009) in this journal, will receive special attention. In section 4, we advance a new experimental design for measuring absolute and relative overconfidence and formally show its incentive compatibility. In section 5, we report the experimental results, compare them with the findings of previously used methods and present the unique characteristics of the new method. The final sections, 6 and 7, conclude with a discussion of the limitations, future research, and possible implications of measuring levels of overconfidence.

2 DEFINITIONS

To ensure precision in the development of our method, we need to define four aspects of overconfidence measurement: the distinction between overconfidence and overprecision; the constructs of confidence in performance and actual performance; the distinction between absolute and relative performance; and the definition of incentive compatibility of belief elicitation.

2.1 Overconfidence versus overprecision

Considering the diverse contexts in which overconfidence has been investigated, e.g., concerning entrepreneurial or managerial decisions (Camerer and Lovallo, 1999; Wu and Knott, 2006; Elston et al., 2006; Köllinger et al., 2006; Malmendier and Tate, 2005, 2008), stock trading (Daniel et al., 1998; Stotz and Nitzsch, 2005; Cheng 2007), legal practice (Compte and Postlewaite, 2004), and medical practice (Baumann et al., 1991), it is not surprising that various definitions of overconfidence have been used. Griffin and Varey (1996) distinguish *optimistic overconfidence*, which refers to overestimating the likelihood that an individual's favored outcome will occur, from overestimating the validity of an individual's judgments when there is no personally favored outcome. The latter type of overconfidence is also called *overprecision* (e.g., Moore and Healy, 2008) and is typically investigated in so-called *calibration studies*. While Griffin and Varey (1996) regard overprecision as the less general and less stable bias in human decision making, Moore and Healy (2008) consider it to be more persistent than optimistic overconfidence. We focus on optimistic overconfidence because – as also discussed in the final section of this paper –

measurements of overprecision are often susceptible to confounds by optimistic overconfidence. For simplicity, we henceforth refer to optimistic overconfidence simply as 'overconfidence.'

2.2 Confidence in performance versus actual performance

Because overconfidence is by definition the difference between an individual's beliefs about the chances of her favored outcome and the actual likelihood of its happening, its measurement requires a measure of confidence in performance and a measure of actual performance. Further, the definitions of both constructs have to be clarified.

Confidence refers to the probability judgment of a distinct favored outcome of a task, e.g., solving a specific number of quiz questions or estimation tasks. A person is considered overconfident if the probability judgment is higher than that person's objective probability to achieve that outcome. In many cases, this objective probability is approximated by the actually realized outcome of this task. The fact that a person's actual outcome is an imperfect reflection of her objective probability to achieve a specific outcome has rarely been considered, but needs to be kept in mind when interpreting overconfidence measurements (Soll, 1996).

These definitions leave open whether one wants to elicit participants' abilities to predict their maximal performance, their minimal performance, a random performance, or the easiest-to-predict performance. While most overconfidence measurements implicitly assume that participants maximize their performance, e.g., the number of correct answers, optimizing the predictability by giving false answers could also be an option, especially when participants are paid for their precision in prediction. Overconfidence measurements therefore have to ensure that they measure the type of performance they claim to measure. We focus on participants' ability to judge their maximal performance and thus have to make sure that they actually maximize their performance in a task.

The measurement of confidence and performance is subject to measurement errors. To minimize these, many overconfidence studies have utilized repeated measurements. Here participants solve several tasks or quiz questions and indicate either, for each single task, how certain they are of a correct answer (e.g., Fischhoff et al., 1977; Fischhoff and MacGregor, 1982), or, subsequent to having solved all tasks, how many tasks they expect to have solved correctly. In the

former case, the responses to all tasks are averages. The latter measure represents a frequency response, which has been shown to be more appropriate (Cesarini et al., 2006). Indeed, the actual performance in solving quiz questions is usually measured with respect to the whole set of tasks. For reasons of comparability it thus makes sense to elicit the participants' confidence in their performance at the same level as a frequency response.

2.3 Absolute and relative performance

While traditional calibration studies have focused on performance judgments related to an absolute criterion, e.g., the number of correctly solved trivia questions, an increasing number of studies investigate confidence with respect to the performance of others, e.g., Hoelzl and Rustichini (2005) or Larrick et al. (2007), who call the belief to be better than average 'overplacement.' In these studies, a participant's favored outcome is to outperform other participants. While fundamentally and in technical terms, there is not a huge difference between tasks that aim at maximizing a criterion related or not related to other participants, the mental processes that lead to judgments related to these two types of tasks might exhibit vital differences. Vautier and Bonnefon (2008) show that self-related judgments can psychometrically be distinguished from comparative judgments. Building on the technical equivalence, i.e., both types of confidence are probability judgments that can be elicited in a comparable way, one can potentially use the same measurement instrument to elicit both types of overconfidence, which makes comparisons of both types, e.g., Larrick et al. (2007), less susceptible to confounds by differences in elicitation methods. The technical equivalence also allows drawing on both streams of literature to evaluate methods for eliciting confidence and performances.

2.4 Incentive compatibility of belief elicitation

The experimental setting for measuring confidence and overconfidence could be considered as a principal-agent scenario, where the experimenter as a principal wants participants as agents to behave in a certain way, i.e., to tell the truth or, more generally, to behave in a way that reveals some hidden information. Accordingly, the development of an experimental design is the equivalent of defining a payment scheme between a principal and an agent, guaranteeing that if the agent maximizes his or her own preferences, the agent behaves according to the principal's objectives

(Rasmusen, 1989). In such a case, the payment scheme is considered incentive compatible. Applied to belief elicitation methods, incentive compatibility describes the fact that a participant is confronted with incentives that make her reveal the true belief.

To theoretically prove incentive compatibility, one needs to make assumptions about the decision making of the participants in the experiment. Incentive compatibility is therefore relative to a descriptive decision theory. Hence, to develop an incentive compatible belief elicitation mechanism, we have to explicitly choose a descriptive decision theory. We use the Cumulative Prospect Theory (CPT), a rank-dependent utility theory which is currently the most widely established class of descriptive decision theories to explain deviations from expected utility theory (Tversky and Kahneman, 1992). Wakker (2004) has shown how to decompose probability weights in CPT into risk preference and a belief. We focus on the latter and furthermore include the reduction axiom, which assumes that participants are able to reduce compound lotteries to their simple representation. We need this axiom because the experimental method we propose utilizes such a compound lottery. The appropriateness of the reduction axiom, which is subject to an ongoing discussion, is briefly discussed at the end of this paper.

We can distinguish two types of incentive compatibility. Applied to belief elicitation, *strict incentive compatibility* implies that revealing the truth is always strictly preferred such that any deviation results in lowering the overall value associated with an individual's decisions. In contrast, *weak incentive compatibility* implies that she cannot improve her situation by not revealing the truth (Rasmusen, 1989). Thus, asking individuals for their beliefs without providing any incentives against lying is weakly incentive compatible, but not strictly incentive compatible. To rule out any lying, we aim to develop a strictly incentive compatible mechanism for eliciting overconfidence.

3 OVERCONFIDENCE MEASUREMENTS AND POTENTIAL IMPROVEMENTS

3.1 Need for incentives

Already in 1977 Fischhoff et al. (1977) raised doubts on whether participants in overconfidence studies are sufficiently motivated to reveal their true beliefs and therefore introduce

monetary stakes. In recent studies, Cesarini et al. (2006) as well as Hoelzl and Rustichini (2005) report significant differences depending on whether participants received additional incentives for predicting their performance correctly. These findings are in line with more general studies on monetary incentives, like those by Holt and Laury (2002, 2005), who find that elicitation of risk attitudes differs with respect to whether or not monetary incentives are provided.

3.2 Existing methods involving incentives

To deter participants from optimizing the predictability of their performance and to incentivize them to maximize their performance instead, Budescu et al. (1997b) and Moore and Healy (2008) provide positive monetary payoffs for correctly solved quiz questions. Of course, these incentives should not distract participants from predicting their performance as precisely as possible.[1] Thus, the most direct way to incentivize the reporting of a belief is to give them an additional amount of money if they come sufficiently close (below a prespecified threshold) to the actually realized performance (see, e.g., Moore and Kim, 2003). A theoretically sound generalization of this idea are so-called *proper scoring rules,* with the *quadratic scoring rule* being the most widely used instantiation (see Selten, 1998). Given this rule, which was, for instance, utilized in overconfidence studies by Budescu et al. (1997b) and by Moore and Healy (2008), participants receive a fixed amount of money if they perfectly predict the outcome, while the payoff is reduced by the square of the deviation if the prediction is not correct. This provides a strong incentive to come as close as possible to the true value, independent of how certain one is (Selten, 1998).

However, the proper scoring rules also have some disadvantages. First, they are rather complex to explain, especially if subjects do not have a sound mathematical background.[2] Second, as mentioned above, one needs to make sure that the incentive for a precise prediction does not cause adjustments in the performance in order to increase its predictability. This requirement is only

[1] As mentioned above, there is a trade-off between maximizing performance and maximizing predictability. This trade-off has not received much attention in the literature on overconfidence measurements as yet, but recently Blavatskyy (2009) proposed an elegant way to avoid it. We explain his method below together with our own approach.

[2] The complexity is reason enough for Moore and Healy (2008) to address participants' trust rather than their understanding of the mechanism. They instruct them as follows: "This formula may appear complicated, but what it means for you is very simple: You get paid the most when you honestly report your best guesses about the likelihood of each of the different possible outcomes."

satisfied in specific measurement settings, such as 'true/false' quiz questions with only two possible answers. Third, a further weakness of proper scoring rules is that they are not robust to variations in risk attitudes (Offerman et al., 2007). Participants with different risk attitudes will provide different responses even if they hold the same belief. To overcome this weakness, Offerman et al. (2007) suggest calibrating the quadratic scoring rule for every single participant. However, this has not been done in overconfidence studies yet and would make the experimental design even more complex.

Recent research on individuals' beliefs to be better than others has suggested alternative methods to elicit (relative) overconfidence; some of which can also be applied to elicit confidence in (absolute) performance. Moore and Kim (2003) provide participants with a fixed amount of money and allow them to wager as much as they want on their performance. If they succeed in solving a task, i.e., solving it correctly, the wagered amount is doubled. The basic idea is that the more likely one is to win, the more one would invest into own performance. While this method has the advantage of avoiding explicit probability judgments and could be perceived as simpler than proper scoring rules, it is not robust to risk attitudes. The more risk averse a participant is, the less she wagers, which confounds the measurement of the participant's belief with his or her risk attitude.

While the investment approach is principally a trade-off between a safe income and a risky income, where the risk is rooted in own performance, Hoelzl and Rustichini (2005) and Blavatskyy (2009) suggest eliciting overconfidence by implementing a trade-off between the performance risks and a lottery with predetermined odds. Hoelzl and Rustichini utilize this idea for measuring better-than-average beliefs by letting participants choose between playing a fifty-fifty lottery and being paid if they are better than fifty percent of all participants. Blavatskyy (2009) measures absolute overconfidence in answering trivia questions and lets participants choose between being paid according to their performance and playing a lottery. The first option results in a fixed payoff M if a randomly drawn question has been answered correctly. The second option yields the same payoff M with a probability that equals the fraction of questions that have been answered correctly so that the expected value of both options is equivalent. However, participants are not told how many questions they have answered correctly, nor do they know that the lottery's odds are based on this number.

Participants choosing the lottery are considered underconfident. Those that choose to be paid according to their performance are considered overconfident. If they are indifferent between both alternatives (indicated by an explicit response leading to a random choice between both), they are considered well-calibrated. This method has empirically been found to be robust to risk attitudes (Blavatskyy, 2009).

Blavatskyy's (2009) method to elicit beliefs about own performance is in fact a special application of what Offerman et al. (2007) call *measuring canonical probabilities,* and what Abdellaoui et al. (2005) label the *elicitation of choice-based probabilities*. These methods aim at eliciting an individual's belief about the probability of a binary random process with payoffs H and L. This is done by determining a probability for a binary lottery with the same payoffs H and L such that individuals are indifferent between the random process and the lottery. Since this lottery represents an equivalent to the random process, the probability of the lottery is considered the (canonical) probability associated with the random process. These methods have theoretically been shown to be robust to risk attitudes (Wakker, 2004), supporting Blavatskyy's (2009) empirical finding.

3.3 Potential improvements

Although some of the points discussed in this section also apply to other studies, we focus on Blavatskyy's (2009) method, because it represents the state of the art in overconfidence research and, more importantly, offers some desirable characteristics: it manages to integrate the incentives for performance maximization and for the elicitation of true beliefs about this performance, it is robust to risk attitudes, and it does not rely on verbally stated probability judgments. Furthermore, it is incentive compatible, although, as we will show below, not strictly incentive compatible.

Despite the method's elegance and potential, however, it also has some weaknesses that call for improvements. These relate to its incentive compatibility, its classification of well-calibrated individuals, and its limitation in eliciting only three levels of confidence, i.e., overconfidence, underconfidence, and well-calibrated confidence. Furthermore, it is restricted to eliciting

overconfidence in absolute abilities and does not generalize easily to eliciting relative overconfidence.

3.3.1 Incentive compatibility

Blavatskyy's (2009) method makes use of *hidden chaining*, where participants do not know that the lotteries they are confronted with in a subsequent step of the experiment are actually based on the choices they made in earlier steps, i.e., when solving their quiz questions. If participants knew about the chaining, they would also know that the options to choose from in the second step represent identical winning chances. As a consequence, they would always be indifferent. For eliciting participants' confidence, Blavatskyy's (2009) method therefore requires that participants do not know about this chaining and, more precisely, erroneously believe that there is no chaining. Accordingly, the design requires the experimenter to convince participants of a false belief about how their outcomes are affected by their choices. This assumption is in conflict with the *salience requirement* in experimental economics, which requires that participants correctly know and understand all the rules, how their payoffs are determined.[3] A similar argument was established by Harrison (1986) with respect to the experimental design presented by Becker et al. (1964), as well as by Harrison and Rutström (2008) with respect to the trade-off method introduced by Wakker and Deneffe (1996) and extended by Abdellaoui (2000). For an improved method of overconfidence measurement, we therefore suggest an elicitation procedure that does not utilize hidden chaining.

3.3.2 Classification of well-calibrated confidence

Another characteristic of Blavatskyy's (2009) design is the assumption of *epsilon truthfulness*, which states that participants tell the truth when there is no incentive to lie (Rasmusen, 1989; Cummings et al., 1997). If they are indifferent between choosing to be paid according to their performance and being paid according to a lottery, then Blavatskyy's design expects participants to explicitly indicate that indifference. Only if they do this, are they classified as well-calibrated.

[3] Salience requires that "the reward received by the subject depends on her actions (and those of other agents) as defined by institutional rules that she understands. That is, the relation between actions and the reward implements the desired institution, and subjects understand the relation" (Friedman and Sunder, 1994, p. 13). It is thus commonly understood that participants in economic experiments have perfect knowledge about how their choices affect their outcomes.

Without the assumption of epsilon truthfulness, any distribution of overconfident, underconfident, and well-calibrated measurements can be explained by a population of well-calibrated participants who choose randomly in case of indifference. Only under the assumption of epsilon truthfulness, self-reporting beliefs without incentives is sufficient to let participants reveal their true beliefs (the indifference) because there is no (monetary) incentive either to lie or to tell the truth. However, as Blavatskyy's method generally incentivizes belief elicitation with regard to over- and underconfidence, it is contradictory to rely on (non-incentivized) epsilon truthfulness when it comes to the identification of well-calibrated behavior. Our method therefore aims at strict incentive compatibility without the need to assume epsilon truthfulness.

Related to the identification of well-calibrated participants is the fact that in Blavatskyy's (2009) design the probability that a well-calibrated participant is indifferent (and is thus correctly classified as such) approaches zero.[4] Indeed, with six and a quarter percent (three participants) the share of well-calibrated participants in Blavatskyy's (2009) study is rather small. In the experimental test of our own method of overconfidence measurement, we therefore expect a significantly higher percentage of well-calibrated confidence than reported in Blavatskyy (2009).

3.3.3 Precision and comparability of confidence measurements

Despite being rooted in well-established belief elicitation methods, the methods suggested both by Blavatskyy (2009) and Hoelzl and Rustichini (2005) only reveal whether or not a belief exceeds a certain threshold. In Blavatskyy (2009), subjects are classified as overconfident, underconfident or well-calibrated, but it is impossible to state whether one is more or less overconfident than another. Similarly, Hoelzl and Rustichini (2005) can only show that a participant

[4] To illustrate this, assume that a participant is well-calibrated, which means that her confidence is equal to the *expected* performance. According to Blavatskyy's (2009) method, a participant is classified as well-calibrated if her confidence is equal to the *realized* performance (not the expected performance). Note that as long as the tasks involve a stochastic component, i.e., include imperfect knowledge, the realized performance is a random variable represented by a distribution with a mean mirroring the expected performance of a participant. The values that the realized performance can take are determined by the number of tasks solved and do not need to contain the mean. If performances are assumed to be drawn from a continuous distribution, then the probability that a participant's true performance, whether she is well-calibrated or not, perfectly matches the realized performance, approaches zero. Because Blavatskyy's (2009) method classifies only those participants as well-calibrated whose confidence perfectly matches the realized performance, the probability to classify a well-calibrated participant as such approaches zero.

believes to be worse or better than fifty percent of all participants, but not whether she believes to be better than thirty or seventy percent.

Hoelzl and Rustichini consider a population as overconfident if more than fifty percent believe to be better than fifty percent. While this definition of overconfidence does indeed describe a population's relative overconfidence in being better than average, we argue that this classification does not generalize to other levels of performance, because levels of relative overconfidence have been found to depend on the level of task difficulty (e.g., Larrick et al., 2007; Moore and Healy, 2008). Thus, a population could be overconfident with respect to being better than average, but simultaneously underconfident with respect to belonging to, e.g., the top twenty-five percent of a population. It is therefore important to measure beliefs more precisely. By determining the percentage of a population which believes to be better than the rest of this population, our method can measure different degrees of a population's overconfidence. Based on the experimental test of our method, we will be able to plot altogether ten levels of a population's relative confidence in a range of from five to ninety-five percent.

In addition to measuring overconfidence and underconfidence at more levels, our method also allows a direct comparison between (absolute) overconfidence and relative overconfidence. This enables new empirical tests in an ongoing theoretical debate. Moore and Healy (2008) propose a theory, which they have tested, based on Bayesian updating that explains why individuals who are overconfident also believe that they perform below average, and those who are underconfident believe that they perform above average. Larrick et al. (2007) show under which conditions Moore and Healy's results can be expected and argue that relative confidence and (absolute) confidence, both being part of overconfidence, essentially represent a common underlying factor: subjective ability. Our improvements measure (absolute) confidence and relative confidence with the same incentive compatible method and therefore enable a more robust comparison of both constructs. Our experimental test, which is based on the new method, thus represents a first step toward such a comparison of (absolute) confidence and relative confidence.

4 DEVELOPMENT OF A NEW METHOD

In an attempt to improve the measurement of overconfidence along the lines discussed above, we propose a method that elicits canonical probabilities based on binary choices. We thereby build on methods suggested for belief elicitation with respect to probabilities that do not depend on own activities (Abdellaoui et al., 2005). The choice is between being paid according to own performance (success or failure) and participating in a lottery with a given winning probability. The new method combines the strengths of Blavatskyy's (2009) method with improvements based on our discussion in the previous section. Instead of utilizing hidden chaining, we ask participants for multiple binary choices, one of which is randomly selected to determine the payoff (*random lottery design*). We utilize choices between being paid according to performance in general and being paid according to a lottery's outcome (as in Blavatskyy, 2009) to approximate the indifference probability. By selecting an appropriate set of choices, our measurement method classifies participants as well-calibrated if their confidence level is closer to the actually realized performance than to any other possible performance level. This removes the need for well-calibrated participants to explicitly indicate their indifference.

The new method can be applied to various definitions of performance. We exemplify this by eliciting performance beliefs with respect to two different types of performance, absolute and relative. Both are based on participants' answers to ten quiz questions without feedback. We selected these questions from a larger set of questions with an equivalent level of difficulty.

(1) Absolute performance: a participant succeeds if she answered one quiz question correctly and failed otherwise. The question is determined randomly.

(2) Relative performance: a participant succeeds if she answered more questions correctly than another randomly assigned participant who answered the same questions. She fails if she answered fewer questions. If both answered the same number of questions correctly, one is considered to have succeeded and the other to have failed. If this happens, who succeeds and who fails is determined randomly.

We do not directly translate the absolute performance measurement (1) into the relative measurement (2), because this requires participants to elicit their belief about the probability that they have a higher probability to be correct compared to other participants. We believe that this would be rather complicated to communicate to participants. We therefore ask them to compare the number of correct questions, which is easier for participants to understand. As the number of correct questions is the best estimate of the probability to be correct, the direct and the indirect measure we use for the absolute and relative performance are, in fact, equivalent.

4.1 Experimental design

The experiment consists of five stages. Although these are explained below, more details can be found in the instructions in the appendix.

Step 1: Test for understanding the instructions: To ensure that participants correctly understand the instructions, see Appendix B, where they had to answer ten yes/no questions regarding the experimental design. They could only proceed if all answers were correct.

Step 2: Solving quiz questions: As usual in overconfidence experiments, participants solve ten quiz questions without feedback. For this experiment we used multiple choice questions with four possible answers. These ten questions are randomly drawn from a set of 28 questions. We selected these questions from a larger set used by Eberlein et al. (2006) that were correctly answered by forty to fifty percent of the participants.

Step 3: Select card stack and relevant quiz question: The experimenter presents ten stacks of 20 cards each, containing 1, 3, 5, ..., 17, 19 cards with a green cross (wins) and a complementary number of white cards (blanks). Participants do not see the number of green cards and do not (yet) know the distribution of green cards. One randomly drawn participant can inspect the stacks, and after mixing them again another participant randomly chooses one stack. All other stacks are removed. The same procedure is repeated for a second set of 10 stacks of cards. The experimenter also lets one participant draw one card out of a third stack of 10 (numbered from 1 to 10) that determines the question that counts for the absolute performances of all participants.

Step 4: Strategy-based choice: This part of the experiment is split into two steps. In a first step, participants choose between being paid according to their *absolute performance* and drawing a card from stack one. In a second step, they choose between being paid according to their *relative performance* and a card from stack two. Participants thus choose twice between two payoff schemes. At that time, they do not know the number of green cards in the stack that was previously selected in Step 3. However, we allow participants to condition their choice, as shown on the screenshot in the appendix and in the following example of their response: "*If there are 5 green and 15 white cards in the stack and I have the choice between 'cards' and 'quiz - own results,' I choose ...,"* followed by a choice between 'cards' and 'quiz - own results.' This mechanism mirrors the so-called strategy method introduced by Selten (1967), which is usually applied to game theoretical experiments (e.g., Selten et al., 1997), but is applied here to a situation without strategic interaction. One half of the participants (randomly determined) complete the two steps in reverse order, i.e., relative performance first and absolute performance second.[5]

Step 5: Disclosure of cards and application of participants' strategies: In a last step, the number of green cards in the two stacks and the number of the relevant question are disclosed. Participants who chose to draw a card from any of the two stacks can individually draw a card.[6] Payoffs are calculated and individually paid to participants.

4.2 Measurements

Our experimental design provides us with the following individual measurements:

(1) *Absolute performance p* as the number of correctly answered questions, divided by 10. The value is an integer value between 0 and 1.

(2) *Relative performance rp* is 1 if one participant was better than the other randomly assigned participant, 0 if one was worse, and 0.5 if one solved as many question as the other.

---

[5] Rationally, for an increasing number of green cards participants should never choose performance-based payoff once they have chosen cards for less green cards. If participants violated this assumption, the software displayed a popup with the text "Please check your input. Are you really sure? Yes, continue / No, back." For a single person, a sequence ended with "performance," "cards," "performance." This "cards" choice was considered as "performance."

[6] The individual random draw is used because we wanted to avoid that by accident all lose, which would be bad for the reputation of the lab, or all win, which would be bad for our budget.

(3) *Confidence c* in own absolute performance is the mean of (i) the highest probability for *cards* for which a participant would choose the absolute performance-based payoff rule and (ii) the lowest probability for *cards* for which a participant would choose the draw of a card from the stack of cards. If one participant always, respectively never, chooses the cards, then confidence is set to 0.0 respectively 1.0 (both cases did not occur).

(4) *Relative confidence rc* in relative performance is the mean of (i) the highest probability for which a participant would choose the relative performance-based payoff rule, and (ii) the lowest probability for which a participant would choose the draw of a card from the stack of cards. If a participant always, respectively never, chooses the cards, then confidence is set to 0.0 respectively 1.0 (both cases did not occur).

(5) *Absolute overconfidence oc* is the difference between absolute confidence and absolute performance, $oc=c-p$. We consider participants as well-calibrated, when overconfidence *oc* equals zero. Note that *c* is an approximation of a participant's confidence, and the exact value of *c* lies in the closed interval between $c=-0.05$ and $c=+0.05$. As shown below, participants are well-calibrated when their confidence is closer to their performance than to any other possible performance.

(6) *Relative overconfidence roc* is computed analogously to *oc*.

4.3 Formal proof of strict incentive compatibility

Before we report the results of the experimental test, we formally show that our improved method is strictly incentive compatible and that it has the claimed properties. First, participants prefer a higher performance over a lower one, i.e., they maximize their performance. Second, participants choose the lottery if the winning probability of the lottery is at least as high as their believed performance.[7] Third, a participant is considered well-calibrated if her true performance expectation is closer to the actually realized performance than to any other possible performance. Fourth, elicited probability judgments are theoretically robust to risk attitudes.

[7] This also covers the case of exact equality, which does not need any further requirements.

In order to formally show the incentive compatibility (formally), it is necessary to make assumptions about the participants' behavior in the form of a descriptive decision theory. In this paper, we apply the cumulative prospect theory (CPT) by Tversky and Kahneman (1992), although the following proof holds for expected utility theory, too, and can be extended to a variety of related decision theories. Since CPT does not explicitly consider compound lotteries, i.e., lotteries over lotteries (used in our experiment), we need to include the reduction axiom (Starmer and Sudgen, 1991) as an additional assumption. It states that participants can reduce compound lotteries to their simple representation. Within the framework of CPT and the reduction axiom the claimed properties of our experimental design can be proven. The proof will be applied to the specific design we use in our experiment, including the double elicitation of (absolute) confidence and relative confidence. With marginal adjustments, the proof also holds if the two types of confidence are considered individually.

Participants are confronted with N choices for each of the two elicited confidences; N therefore determines the precision. In our specific case, N equals 10. Without loss of generality, let $c_{a,i} \in \{0,1\}$ with $1 \leq i \leq N$ be the participant's choice between being paid according to own performance and the lottery i with winning probability p_{Li} . In our case, the lottery i is characterized by $2i$-1 winning cards among the total of $2N$ cards (in our case, it results in 20 cards); thus, $p_{LI}=(i-0.5)/N$. If the task is chosen (and not the lottery), $c_{a,i}$ equals 1, otherwise 0. Let $c_{r,i} \in \{0,1\}$ be the same for the choices between relative performance and a lottery. Vectors $c_a=(c_{a,1}, c_{a,2}, \ldots , c_{a,N})$ and $c_r=(c_{r,1}, c_{r,2}, \ldots , c_{r,N})$ represent vectors of these decisions. Furthermore, let q be the performance expectation by the participant. Let us assume that the ex ante performance of the participant varies between q_{min} and q_{max}, i.e $q_{min} \leq q \leq q_{max}$ (depending on the participant's choice). Let us further assume that the expectation of the relative performance rq is a strictly monotonic function of the performance expectation, i.e., the first derivative of $rq'(q)$ is strictly larger than 0. Let H be the amount of money that can be won in the lottery or earned when the task (absolute or relative) has been performed successfully. If the lottery is lost or the task has not been performed successfully then a participant earns nothing. As participants are assumed to follow cumulative prospect theory, the preference value V for a given set of decisions (c_a, c_r, q) is given by (1), with p being the belief

about the occurrence of payoff H. The function $v(x)$ represents the CPT value function applied to payoffs with $v(0)=0$. For simplicity, we also assume that $v(x)>0$ for $x>0$. The function $\pi(p)$ represents the CPT probability weighting function with $\pi(p)\in[0,1]$. Both functions are assumed to be monotonically increasing in the payoff x respectively the probability p.

$$V(c_a, c_r, q, H) = v(H)\pi(p(c_a, c_r, q)) \tag{1}$$

Applying the reduction axiom, a participant is assumed to form a belief about the occurrence of H depending on her decisions, and she is able to reduce compound lotteries to their simple representation. This is necessary as our method implements a random choice between alternatives that are themselves uncertain Since each of the $2N$ (in our case 20) choices between the absolute respectively relative performance and a lottery can become relevant with equal probability, the probability for H is the average of the probabilities of all single decisions ($c_{a,1}$ to $c_{a,N}$ and $c_{r,1}$ to $c_{r,N}$). As shown in (2), for a single decision (between absolute performance and lottery with winning probability p_{Li}) the probability is determined by $c_{a,i}\, q$ + (1-$c_{a,i}$) $p_{Li,}$ which is q if the performance is chosen and p_{Li} if the lottery is chosen. For the choice between relative performance and a lottery the probability of a payoff H is determined correspondingly.

$$V(c_a, c_r, q, H) = v(H)\pi\left(\frac{\sum_{i=1}^{N}\left(c_{a,i}q + (1-c_{a,i})p_{Li}\right) + \sum_{i=1}^{N}\left(c_{r,i}rq(q) + (1-c_{r,i})p_{Li}\right)}{2N}\right) \tag{2}$$

Note that $V(c_a,c_r,q,H)$ is always larger than or equal to zero. Equation 2 can be simplified to

$$V(c_1, c_2, q, H) = v(H)\pi(p) \tag{3}$$

$$\text{with} \qquad p = \frac{1}{2N}\sum_{i=1}^{N}\left(c_{1i}(q - p_{Li}) + c_{2i}(rq(q) - p_{Li}) + 2p_{Li}\right)$$

with the following derivatives with respect to the decision variables:

$$\frac{\partial V(c_1, c_2, q, H)}{\partial q} = v(H)\pi'(p)\frac{1}{2N}\sum_{i=1}^{N}(c_{1i} + c_{2i}rq'(q))$$
$$\frac{\partial V(c_1, c_2, q, H)}{\partial c_{1,i}} = v(H)\pi'(p)\frac{1}{2N}c_{1i}(q - p_{Li})$$
$$\frac{\partial V(c_1, c_2, q, H)}{\partial c_{2,i}} = v(H)\pi'(p)\frac{1}{2N}c_{2i}(rq(q) - p_{Li})$$

The terms v(H) and $\pi'(p)$ are by definition strictly positive. Given that $rq'(q)$ is strictly larger and c_{1i} and c_{2i} are never less than 0, we can conclude that the preference value is strictly increasing in q as long as at least one decision is made in favor of being paid according to own absolute or relative performance, i.e., at least one $c_{a,i}$ or $c_{r,i}$ equals 1. In our mechanism, this is only the case if the belief about own performance or the belief about relative performance is greater than five percent. Below this threshold, being paid according to own absolute or relative performance is never chosen, and thus there is no strict incentive to maximize this performance, i.e., the first derivative is zero. However, the minimal expected probability of success with respect to absolute performance in our experiment is twenty-five percent (random choice out of four answers per question) such that the belief about own performance always lies above five percent. Hence, a participant always maximizes her performance expectation.[8]

Furthermore, the preference value is strictly increasing in the decisions $c_{x,i}$ for x being a or r, if q respectively $rq(q)$ are greater than the winning probability of the alternative lottery. Thus, participants will always choose the task if their belief about a good performance is greater than the probability to win the lottery. Participants therefore always reveal their true beliefs through their choice behavior. If they do not choose the lottery for lottery i but for i+1, then the best estimation of the participant's belief is $(p_{Li} + p_{Li+1})/2$, which in our case is i/N.

Note that participants with a confidence between *i/N-0.05* and *i/N+0.05* are all classified to have a confidence level of *i/N*. Such participants are considered well-calibrated if they have solved *i* out of *N* tasks correctly. Therefore, even if their confidence level differs only slightly from the elicited performance, they will still be classified as well-calibrated. This holds as long as the

[8] Note that when excluding the elicitation of confidence in absolute performance such a lower threshold for performance expectations is not present.

difference between confidence and actual performance does not exceed *1/2N*, which is equivalent to the condition that confidence is closer to another level of performance that could be elicited.

All results above are based on CPT and the reduction axiom. As such, they are independent of an individual's risk attitude as long as it satisfies the axioms of CPT. Our results are thus theoretically robust to variations in risk attitudes, modeled via value and probability weighting functions with characteristics following CPT.

4.4 Experimental sessions and participants

We conducted the experiment in two sessions, on August 25 and September 1, 2008, with altogether 31 women and 29 men, who were recruited from the student body of the University of Jena, Germany. Their average age was 23.85 years, with a minimum of 18 years, a maximum of 30 years, and a standard deviation of 2.45. We recruited students from all disciplines, ranging from the natural to the social sciences, with the exception of psychology, the reason being that psychology students might have had previous experience with psychological experiments in which critically different mechanisms are frequently used. These experiences might have produced an *ex ante* bias in the students' expectations with regard to our (economic) experiment; more specifically, these students might not have trusted the experimenter. On average, the experimental session lasted 60 minutes, and participants earned 11.10 euro.

5 RESULTS

Table 1 provides some summary statistics for our experiment. The average absolute performance p is 0.497 with a median of 0.5; the average relative performance is 0.5 with a median of 0.5. On average, participants have a confidence in their performance of 0.493 with median 0.5 and a confidence in their relative performance of 0.498 with a median 0.5.[9] In this experiment, the

[9] Two out of 60 participants violated a basic principle, namely that under the condition that an individual tries to maximize her performance the probability to win in a multiple choice task with four alternatives is at least 25%. One participant switched between 15% and 25%. It is, however, possible that this person had a confidence of 25% and was therefore indifferent between the 25% lottery and her performance. Choosing the lottery in this case is still rational and consistent. The behavior of the second person who switched between 5% and 15% is, however, not captured by the

participants are thus, on average, well-calibrated in absolute and in relative terms. We did not find any order effects; the order of elicitation of absolute and relative confidence did not cause significant differences between the two treatments.

Table 1

Table 1 also reports the correlation of variables with performance p, and with confidence regarding absolute and relative performance. We find that (absolute) performance and relative performance are positively correlated, as one would have expected, because participants with a higher performance have a greater chance to be better than others. Participants are partially aware of their performance as their (absolute) confidence and relative confidence in their performance increasing with their performance (Pearson correlations are significant at the five percent level). However, (absolute) overconfidence and relative overconfidence in performance both decrease with the level of their performance. This result is consistent with prior findings in overconfidence studies. Based on their theory and empirical results, Moore and Healy (2008) argue that, with higher performance, participants tend to become less overconfident and even underconfident, but at the same time believe to be better than others. While the theory by Moore and Healy (2008) tentatively suggests a negative correlation between relative confidence *rc* and overconfidence *oc*, we find a positive relation in our data. We furthermore find a positive correlation between overconfidence *oc=c-p* and confidence *c*, which means that the more confident a participant is, the more overconfident she becomes (Pearson correlation is 0.47 with $p<0.05$). To better understand the relation between correlations involving overconfidence *oc=c-p* and relative overconfidence *roc=rc-rp,* on one side, and statistics about the constituent terms, *c*, *rc*, *p*, and *rp,* on the other, we refer the reader to Appendix A in Larrick at al. (2007), which provides a formal analysis of correlations that can be extended to many cases, where one variable in a correlation is used to calculate the second variable in that correlation.

---

theories applied here, i.e., rank-dependent utility theories. Since results do not change qualitatively, we kept this data point in the data set

5.1 Well-calibrated participants

When developing our elicitation method, we we expected that more participants would be classified as well-calibrated than in Blavatskyy's (2009) study. Figure 1 compares Blavatskyy's results with our own. The left-hand graph shows the distribution of performances within the data sets. The mean performance is slightly higher in Blavatskyy's study, i.e., fifty-five versus fifty percent in our study, and the variance of performance is also greater in his study, i.e., 0.0485 versus 0.0329. The right-hand graph in Figure 1 plots the relative frequency of participants classified as underconfident, well-calibrated, and overconfident. As expected, we find that in our study more participants are identified as well-calibrated, i.e., twenty-three versus six percent in Blavatskyy's study. Based on a Chi-square test, we find that the two binary distributions of well-calibrated versus not well-calibrated participants are different and statistically significant at the five percent level. This fully supports our expectation.

Figure 1

5.2 Simultaneous over- and underconfidence at the population level

Above we argued for a more precise measurement of several levels of over- and underconfidence instead of focusing on a binary belief to be better or worse than the average of a population (e.g., Hoelzl and Rustichini, 2005). If it is important to be among the top five percent of a population, an optimistic better-than-average belief may not generalize to an optimistic better-than-top five percent belief. As Table 1 reports, relative confidence, i.e., the perceived probability that one participant is better than another correlates positively with (absolute) performance, and, even more importantly, relative overconfidence correlates negatively with performance. This suggests that the better a participant is, the less she is unrealistically optimistic regarding her position relative to others. At the population level, it is thus possible that about fifty percent believe to be better than fifty percent ─ indicating a well-calibrated population (Hoelzl and Rustichini, 2005) ─ but that significantly less than twenty-five percent believe to be better than seventy-five percent.

Figure 2

Figure 2 addresses this question by plotting the relative frequency of participants who believe to be better than five, fifteen, twenty-five, …, and ninety-five percent. Consistent with our conclusion from the mean of relative confidence, approximately fifty percent believe to be better than fifty percent of all participants. Thus, regarding this benchmark the group of our participants is neither over- nor underconfident. However, considering other benchmarks, the conclusion differs. About ninety-five percent of all participants believe to be better than twenty-five percent, implying overconfidence for a small threshold; but only about seven percent believe to be better than seventy-five percent, implying underconfidence. Our group of participants is therefore underconfident for large and overconfident for small thresholds.

5.3 Confidence in absolute versus relative performance

In our experiment, we used the new method to elicit confidence in own absolute performance (confidence) and confidence in own relative performance (relative confidence) with the same methodology and at the same time. This enables a direct comparison of the two types of confidence. A correlation of 0.728 (see Table 1) already indicates that both are closely related. Figure 3 visualizes the relation between both variables. Besides plotting the data points, it provides conditional means and a fitted linear approximation of the relation between both variables. Since both are subject to measurement errors, conditional means as well as simple regression analysis yield biased results; especially the slope of the fitted linear function might be attenuated.[10] However, running a direct regression (variable 1 on variable 2) and a reverse regression (variable 2 on variable 1), as illustrated in Figure 3, provides bounds on the true parameter (Wansbeek and Meijer, 2000). Despite one outlier with little relative confidence but more or less average (absolute) confidence, Figure 3 suggests a very compelling relation of confidence and relative confidence. In fact, we cannot reject the hypothesis that both are identical for our data. Although this identity might not be

[10] For an in-depth discussion of the consequences of measurement errors for overconfidence research, see Erev et al. (1994), Soll (1996), Pfeifer (2994), Brenner et al. (1996), and Juslin and Olsson (2000).

observed in experiments, where the average performance is not fifty percent, we would nevertheless expect a close relation of both constructs, although at a different level.

Figure 3

---

6 DISCUSSION

In developing our method, we have focused on measuring optimistic overconfidence in contrast to overprecision. As this is an important distinction (see section 2 about definitions), our choice of focus is discussed in more detail below. Subsequently, we discuss four limitations of this study, highlight its implications, and suggest avenues for further research.

6.1 Why we do not measure overprecision

While overconfidence and overprecision are related, we have focused on measuring overconfidence. Besides overconfidence being the more robust belief distortion (Griffin and Varey, 1996), we initially argued that measuring overprecision is likely to be confounded by overconfidence. Overprecision has often been measured by asking subjects to estimate the numerical answer to a question and subsequently to report a confidence interval around this number (e.g., Fischhoff et al., 1977; Cesarini et al., 2006). This is usually repeated for a number of questions. Overprecision refers to the phenomenon that most of these confidence intervals are found to be too narrow. Following Cesarini et al. (2006), we argue that asking for an estimate makes participants favor a narrow over a wider interval, indicating a less precise estimation. Similarly, asking participants for estimates that come sufficiently close, i.e., within a given interval around the target value, as done by Larrick et al. (2007), as well as asking them to solve trivia questions and judge their likelihood to solve them correctly (Moore and Healy, 2008), make the former prefer a correct answer. The salience of a notion of good performance in measurements of overprecision increases even more if interval elicitations are extended by asking participants for their estimation of the number of intervals that correctly capture the real number as in the case of Cesarini et al. (2006), or by asking them for the estimation and a probability of being correct, e.g., Fischhoff et al. (1977).

Most of the overprecision measurements can therefore be interpreted to be based on measuring optimistic overconfidence in a setting, where own performance equals precision of judgments. If participants suffer from optimistic overconfidence, they exaggerate their likelihood of performing well and, by the same token, of estimating the numbers or solving the questions correctly. These optimistic overconfident participants will therefore appear to exhibit overprecision. If precision directly or indirectly becomes the performance benchmark, then overprecision and optimistic overconfidence cannot be perfectly distinguished. This is the reason why we have focused purely on optimistic overconfidence and consider beliefs about outcomes that are favored.

6.2 Limitations and further research

Among the salient limitations of our method is, first, the fact that the theoretical proof of incentive compatibility requires the reduction axiom for the random lottery mechanism. This axiom has been challenged with respect to its empirical justification, particularly in connection with the use of random lottery mechanisms, which is widespread in experimental economics. However, empirical studies on whether or not the mechanism leads to distortion conclude that "experimenters can continue to use the random-lottery incentive mechanism" (Hey and Lee, 2005, p. 263). Their results are consistent with what has been found by many others, for instance Starmer and Sugden (1991) and, most recently, Lee (2008) published in this journal.[11]

A second limitation relates to the question whether or not the incentives make participants more or less well-calibrated. We replicate findings by Blavatskyy (2009) that, on average, participants are well-calibrated. While one might be tempted to attribute this to the incentivized methods, a theory recently put forward by Moore and Healy (2007, 2008) provides an alternative explanation. The authors suggest that, for intermediate levels of difficulty, participants are, on

[11] To ensure that participants take every decision as if it was the only one, we additionally adjusted the traditional random lottery method. Instead of randomly selecting the relevant decision at the end, we physically selected a decision at the beginning of the experiment, but kept it covered until the end. We then asked participants to condition their decisions such that they respond "If this choice is …, then I will choose … ." This design does not require participants to envision a future random draw. Instead, they observe the crucial element of the design before they make their decisions. This design applies the strategy method, developed for experiments on strategic interaction (Selten et al., 1997), to a decision experiment, where one person plays against nature.

average, neither over- nor underconfident. In our study the levels of difficulty, i.e., the average performance, are close to fifty percent, which has also been found in the study by Blavatskyy (2009).

Our method includes a parameter *N* describing the number of binary choices used to elicit confidence beliefs. In our method, the precision of performance elicitation is driven by this parameter. Increasing *N* also increases the precision of both confidence and performance measurements. Note, as a third limitation, that the probability to identify a well-calibrated participant subsequently decreases. Studies that refer to a dichotomous classification of well-calibrated versus ill-calibrated participants should account for this dependency on the precision of measurements. We therefore recommend to base overconfidence analyses on a range of degrees of confidence and overconfidence instead of dichotomous or trichotomous classifications based on single thresholds (such as underconfident, well-calibrated, and overconfident). Due to their complex dependency on measurement errors, the latter are generally difficult to interpret. This dependency on precision also needs to be considered when comparing studies that utilize different levels of precision.

A general, fourth, limitation that should be noted is that the incentive compatibility of our mechanism as well as the interpretation of the results are based on the assumption that risk attitudes are independent of the source of risk (which is common in belief elicitation methods and standard in rank-dependent utility theories). This assumption is critical for the inference that the elicited confidence reflects the risk associated with the participant's own performance. Empirical work seems to suggest that risk attitudes differ for both sources of risk, own performance, and lotteries (Heath and Tversky, 1991; Kilka and Weber, 2001). Because of the fact that this limitation is very rather common, this issue clearly calls for more research into belief elicitation under conditions of source-dependent risk attitudes.

7 CONCLUSIONS

This study has been motivated by the ongoing discussion about the appropriate measurement of overconfidence and, in particular, how to measure it in strictly incentive compatible experiments. We have identified some major challenges, e.g., the necessary balance between

incentives to maximize own performance and the incentives to predict own performance as accurately as possible. Most recently, in this journal, Blavatskyy (2009) has suggested a mechanism that elicits overconfidence in an arguably incentive compatible way. This mechanism has several crucial advantages: it does not require the statement of probabilities, it has been empirically shown to be robust to variations in risk attitudes, and it balances the two incentives for confidence and performance in an elegant way. While these characteristics are indeed desirable, we have nevertheless identified weaknesses. Based on existing empirical and theoretical studies, we substantially extend and adapt Blavatskyy's (2009) mechanism, mainly with regard to the strictness of its incentive compatibility and the identification of well-calibrated participants. Another significant improvement is the measurement of more than a maximum of three levels of confidence, i.e., overconfidence, underconfidence, and well-calibrated confidence. We finally develop and test a new method that retains all the advantages of Blavatskyy's (2009) mechanism but is, in addition, strictly incentive compatible (theoretically shown within the framework of CPT), identifies those participants as well-calibrated whose confidence is closer to their actual performance than to any other possible performance, and is suited to measure overconfidence with much greater precision.

Besides this methodological advance, this paper also provides more applied results. Research on relative overconfidence generally focuses on the belief to be better than the average of a population. We argue that for many social and economic situations the belief to be better than average is of less relevance than the belief to be the best or among the best. Since our mechanism elicits degrees of overconfidence, we can test whether, for instance, more than ten percent of participants believe to be better than ninety percent. In fact, our analysis (visualized in Figure 2) shows that, simultaneously, too few participants believe to be among the best while too many believe not to be among the worst. This may have significant economic implications, which would be worthwhile to investigate in more depth. For instance, a general underconfidence to be among the best could lead to pessimism in highly competitive environments, such as patent races, where the winner takes it all, possibly triggering underinvestments into research and development.

LITERATURE

Abdellaoui, M.; Vossmann, F. & Weber, M. (2005) Choice-Based Elicitation and Decomposition of Decision Weights for Gains and Losses Under Uncertainty. ***Management Science***, 51, 1384-1399.

Baumann, A. O.; Deber, R. B. & Thompson, G. G. (1991) Overconfidence among physicians and nurses: The 'micro-certainty, macro-uncertainty' phenomenon. ***Social Science & Medicine***, 32, 167-174.

Becker, G. M.; DeGroot, M. H. & Marschak, J. (1964) Measuring utility by a single response sequential method. ***Behavioral Science***, 9, 226-232.

Blavatskyy (2009) Betting on own knowledge: Experimental test of overconfidence. ***Journal of Risk and Uncertainty***, 38, 39-49.

Bonin, H.; Constant, A.; Tatsiramos, K. & Zimmermann, K. F. (in press) Native-Migrant Differences in Risk Attitudes. ***Applied Economics Letters***.

Brenner, L. A.; Liberman, V. & Tversky, A. (1996) Overconfidence in probability and frequency judgements: A critical examination. ***Organizational Behavior & Human Decision Processes***, 65, 212-219.

Budescu, D. V.; Wallsten, T. S. & Au, W. T. (1997a) On the Importance of Random Error in the Study of Probability Judgment. Part II: Applying the Stochastic Judgment Model to Detect Systematic Trends. ***Journal of Behavioral Decision Making***, 10, 173-188.

Budescu, D. V.; Wallsten, T. S. & Au, W. T. (1997b) On the Importance of Random Error in the Study of Probability Judgment. Part II: Applying the Stochastic Judgment Model to Detect Systematic Trends. ***Journal of Behavioral Decision Making***, 10, 173-188.

Camerer, C. F. & Lovallo, D. (1999) Overconfidence and Excess Entry: An Experimental Approach. ***The American Economic Review***, 89, 306-318.

Cesarini, D.; Sandewall, O. & Johannesson, M. (2006) Confidence interval estimation tasks and the economics of overconfidence. ***Journal of Economic Behavior & Organization***, 61, 453-470.

Cheng, P. Y. K. (2007) The Trader Interaction Effect on the Impact of Overconfidence on Trading Performance: An Empirical Study. ***Institutional Investor***, 8, 50-63.

Compte, O. & Postlewaite, A. (2004) Confidence-Enhanced Performance. ***The American Economic Review***, 94, 1536-1557.

Cummings, R. G.; Elliott, S.; Harrison, G. W. & Murphy, J. (1997) Are hypothetical referenda incentive compatible? ***Journal of Political Economy***, 105, 609-621.

Daniel, K.; Hirshleifer, D. & Subrahmanyam, A. (1998) Investor Psychology and Security Market Under- and Overreactions. ***Journal of Finance***, 53, 1839-1885.

Dawes, R. M. & Mulford, M. (1996) The false consensus effect and overconfidence: Flaws in judgment or flaws in how we study judgment? ***Organizational Behavior & Human Decision Processes***, 65, 201-211.

Dohmen, T.; Falk, A.; Huffman, D.; Sunde, U.; Schupp, J. & Wagner, G. (2005) Individual Risk Attitudes: New Evidence from a Large, Representative, Experimentally-Validated Survey. IZA Discussion Paper No. 1730.

Eberlein, M.; Ludwig, S. & Nafziger, J. (2006) The Effects of Feedback on Self-Assessment. University of Bonn, Department of Economics.

Elston, J. A., G. W. Harrison, & Rutström, E. E. (2006) Experimental economics, entrepreneurs and the entry decision. *University of Central Florida working paper* 06-06.

Erev, I.; Wallsten, T. S. & Budescu, D. V. (1994) Simultaneous over- and underconfidence: The role of error in judgment processes. ***Psychological Review***, 101, 519-527.

Fischhoff, B.; Slovic, P. & Lichtenstein, S. (1977) Knowing with certainty: The appropriateness of extreme confidence. ***Journal of Experimental Psychology***, 3, 552-564

Fischhoff, B. & MacGregor, D. (1982) Subjective confidence in forecasts. ***Journal of Forecasting***, 1, 155-172.

Friedman, D. & Sunder, S. Experimental Methods Cambridge University Press, 1994.

Griffin, D. W. & Varey, C. A. (1996) Commentary: Towards a consensus on overconfidence. ***Organizational Behavior & Human Decision Processes***, 65, 227-231.

Harrison, G. W. (1986) An experimental test for risk aversion. ***Economic Letters***, 21, 7-11.

Harrison, G.W.; Johnson, E.; McInnes, M. M. & Rutström, E. E. (2005) Individual Choice and Risk Aversion in the Laboratory: A Reconsideration. Working Paper 03-18. Department of Economics, College of Business Administration, University of Central Florida.

Harrison G., Rutström E. (2008) Risk Aversion in the Laboratory. In: J. C. Cox and G.W. Harrison (eds.), Risk Aversion in Experiments (Greenwich, CT: JAI Press, Research in Experimental Economics, Volume 12).

Heath, C. & Tversky, A. (1991) Preference and Belief: Ambiguity and Competence in Choice under Uncertainty. ***Journal of Risk and Uncertainty***, 4, 5-28.

Hey, John D. & Jinkwon Lee (2005) Do Subjects Separate (or Are They Sophisticated)? ***Experimental Economics*** 8, 233-265.

Hoelzl, E. & Rustichini, A. (2005) Overconfident: Do you put money on it? ***The Economic Journal***, 115, 305-318.

Holt, C. A. & Laury, S. K. (2002) Risk Aversion and Incentive Effects. ***The American Economic Review***, 92, 1644-1655.

Holt, C. A. & Laury, S. K. (2005) Risk Aversion and Incentive Effects: New Data without Order Effects. ***The American Economic Review***, 95, 902-904.

Kilka, M. & Weber, M. (2001) What Determines the Shape of the Probability Weighting Function under Uncertainty? ***Management Science***, 47, 1712-1726

Juslin, P. & Olsson, H. (1997) Thurstonian and Brunswikian Origins of Uncertainty in Judgment: A Sampling Model of Confidence in Sensory Discrimination. ***Psychological Review***, 104, 344-366.

Juslin, P.; Winman, A. & Olsson, H. (2000) Naive Empiricism and Dogmatism in Confidence Research: A Critical Examination of the Hard-Easy Effect. ***Psychological Review***, 107, 384-396.

Keh, H. T.; Foo, M. D. & Li, B. C. (2002) Opportunity Evaluation under Risky Conditions: The Cognitive Processes of Entrepreneurs. ***Entrepreneurship Theory & Practice***, 27, 125-148.

Koellinger, P.; Minniti, M. & Schade, C. (2006) "I think I can, I think I can": Overconfidence and entrepreneurial behavior. ***Journal of Economic Psychology***, 28(4), 502-527.

Larrick, R. P.; Burson, K. A. & Soll, J. B. (2007) Social comparison and confidence: When thinking you're better than average predicts overconfidence (and when it does not). ***Organizational Behavior & Human Decision Processes***, 102, 76-94.

Lee, Jinkwon (2008), The Effect of the Background Risk in a Simple Chance Improving Decision Model. ***The Journal of Risk and Uncertainty*** 36, 19-41.

Malmendier, U. & Tate, G. (2005) CEO Overconfidence and Corporate Investment. ***Journal of Finance***, LX, 2661-2700.

Malmendier, U. & Tate, G. (2008) Who makes acquisitions? CEO overconfidence and the market's reaction. ***Journal of Financial Economics***, 89, 20-43.

Moore, D. A. & Healy, P. J. (2008) The Trouble With Overconfidence. ***Psychological Review***, 115, 502-517.

Moore, D. A. & Healy, P. J. (2007) Bayesian overconfidence. ***Academy of Management Proceedings***, 1-6.

Moore, D. A. & Kim, T. G. (2003) Myopic Social Prediction and the Solo Comparison Effect. ***Journal of Personality and Social Psychology***, 85, 1121-1135.

Offerman, T.; Sonnemans, J.; van de Kuilena, G. & Wakker, P. P. (2007) A Truth-Serum for Non-Bayesians: Correcting Proper Scoring Rules for Risk Attitudes.

Pfeifer, P. E. (1994) Are we overconfident in the Belief that Probability Forecasters are Overconfident? ***Organizational Behavior & Human Decision Processes***, 58, 203-213.

Rasmusen, E. (1989) Games and Information: An Introduction to Game Theory. Blackwell.

Selten, R. (1967) Die Strategiemethode zur Erforschung des eingeschränkt rationalen Verhaltens im Rahmen eines Oligopolexperiments. Beiträge zur Experimentellen Wirtschaftsforschung, Tübingen: J.C.B. Mohr, 136-168.

Selten, R.; Mitzkewitz, M. & Uhlich, G. R. (1997) Duopoly strategies programmed by experienced players. ***Econometrica***, 65, 517-555.

Selten, R. (1998) Axiomatic Characterization of the Quadratic Scoring Rule. ***Experimental Economics***, 1, 43-62.

Simon, M.; Houghton, S. M. & Aquino, K. (2000) Cognitive biases, risk perception, and venture formation: How individuals decide to start companies. ***Journal of Business Venturing***, 15, 113-134.

Soll, J. B. (1996) Determinants of Overconfidence and Miscalibration: The Roles of Random Error and Ecological Structure. ***Organizational Behavior & Human Decision Processes***, 65, 117-137.

Starmer, C. & Sugden, R. (1991) Does the Random-Lottery Incentive System Elicit True Preferences? An Experimental Investigation. ***The American Economic Review***, 81, 971-978.

Stotz, O. & von Nitzsch, R. (2005) The Perception of Control and the Level of Overconfidence: Evidence from Analyst Earnings Estimates and Price Targets. ***Journal of Behavioral Finance***, 6, 121-128.

Tversky, A. & Kahneman, D. (1992) Advances in prospect theory: Cumulative representation of uncertainty. ***Journal of Risk and Uncertainty***, 5, 297-323.

Vautier, S.; Raufaste, E. & Cariou, M. (2003) Dimensionality of the Revised Life Orientation Test and the status of filler items. ***International Journal of Psychology***, 38, 390-400.

Wakker, P. P. (2004) On the Composition of Risk Preference and Belief. ***Psychological Review***, 111, 236-241.

Wakker, P. & Deneffe, D. (1996) Eliciting von Neumann-Morgenstern Utilities when Probabilities are Distorted or Unknown. ***Management Science***, 42, 1131-1150.

Wandsbeek, T. & Meijer, E. (2000) Measurement Error and Latent Variables in Econometrics. (Advanced Textbooks in Economics, vol. 37) Elsevier Science B V: Amsterdam.

Wu, B. & Knott, A. M. (2006) Entrepreneurial Risk and Market Entry. ***Management Science***, 52, 1315-1330.

APPENDIX A: EXPERIMENT INSTRUCTIONS FOR PARTICIPANTS

There were two versions of the instructions. Both versions differ with respect to the order of treatments. In the version reported below, the first set of decisions is related to own performance while the second set of decisions is related to own relative performance. In the second, unreported version, the order is reversed.

Welcome to our experiment!

General information

You will be participating in an experiment in the economics of decision making in which you can make money. The amount of money you will receive will depend on your general knowledge and on your decisions during the experiment. Irrespective of the result of the experiment, you will receive a participation fee of €2.50.

Please do not communicate with other participants from now on. If you have any questions, please refer to the experimenters.

All decisions are made anonymously.

You will now receive detailed instructions regarding the course of the experiment.

It is crucial for the success of our study that you fully understand the instructions. After having read them, you will therefore have to answer a number of test questions to control whether you understood them correctly. The experiment will not start until all participants have answered the test questions.

Please read the instructions carefully and do not hesitate to contact the experimenters in case you have any questions.

Course of the experiment

After all participants have read the instructions and answered the test questions, we will begin with the first part of the experiment.

In this part, you will see a sequence of 10 questions, for each of which you will have to choose 1 out of 4 possible answers. One other player in this room will be randomly assigned to you and will have to solve exactly the same series of questions.

In (the following) parts 2 and 3, we will offer you the opportunity to choose a payoff mechanism. A payoff mechanism is a method that describes how your payoff will be determined. In both parts, 2 and 3, you will have to choose between two Options: **cards** and **quiz.**

1. Cards

For this mechanism, 20 playing cards will be shuffled. A certain number of these cards bear a green cross. You will draw one card from the stack. If it bears a green cross, you will receive €7. If it does not bear a green cross, you will receive €0. By the time you have to decide for or against this payoff mechanism, you will know exactly how many of the cards in the stack bear a green cross.

2. Quiz

If you choose this mechanism, your payoff depends on your answers to the quiz questions. The more questions you have answered correctly, the higher is your chance of receiving a payoff of €7. There are two variants of the payoff mechanism "quiz": **own result** and **relative result**.

a) **Own result:** One out of the 10 quiz questions will be drawn randomly. If you answered this question correctly, you will receive a payoff of €7. Otherwise, you will receive €0. With this payoff mechanism, your payoff will only depend on your own performance.

b) **Relative result:** If you answered more questions correctly than the player that has been assigned to you in the beginning and had to answer exactly the same questions, you will receive €7. If you answered fewer questions correctly, you will receive €0. In case of a draw, it will be randomly decided who will receive the €7.

In the second part of the experiment, you will be able to choose between the payoff mechanisms

(1) cards and
(2a) **quiz – own result**.

In the third part of the experiment, you will be able to choose between the payoff mechanisms

(1) **cards** and
(2b) **quiz – relative result**.

In both parts, one of your options will be to draw a card from a stack which might bear a green cross, which is a pure random mechanism. The other option will always be a payoff mechanism, which determines your payoff based on your result from answering the quiz questions. This means that, in any case, you should try to correctly answer as many questions as possible. It may happen that the number of cards with a green cross is always so small that may you prefer to be paid according to your answers. In this case, your chances are better the more questions you answered correctly.

The diagram below shows the course of the experiment schematically:

<table>
<tr><td>Part 1</td><td colspan="3">Answer quiz questions</td></tr>
<tr><td rowspan="2">Part 2</td><td rowspan="2">Choose a payoff mechanism</td><td>(1) Cards
or</td><td>• One out of 20 cards is drawn
• Green cross: €7
• No green cross: €0</td></tr>
<tr><td>(2a) Quiz – own result</td><td>• One quiz question is randomly drawn
• Correct answer: €7
• Wrong answer: €0</td></tr>
<tr><td rowspan="2">Part 3</td><td rowspan="2">Choose a payoff mechanism</td><td>(1) Cards
or</td><td>• One out of 20 cards is drawn
• Green cross: €7
• No green cross: €0</td></tr>
<tr><td>(2b) Quiz - relative result</td><td>• Another player has been randomly assigned to you
• You answered more questions correctly than he/she: €7
• You answered fewer questions correctly than he/she: €0</td></tr>
</table>

If you have understood the course of the experiment, you may now start to answer the test questions you see on your computer screen. You may always, before and during the experiment, refer to these instructions. The sole aim of the test questions is to control whether you understood the instructions. They are not the quiz questions you will see in part 1 of the experiment!

The experiment will start when all participants have answered the test questions correctly.

APPENDIX B: SCREENSHOT OF THE EXPERIMENT

TABLE 1

Descriptive statistics (mean, standard deviation, mode) and selected correlations

	Descriptive statistics			Correlations		
	μ	**σ**	**median**	***p***	***c***	***rc***
Performance *p*	0.497	0.181	0.5	-	0.472	0.475
Relative performance *rp*	0.500	0.441	0.5	0.551	0.491	0.388
Confidence *c*	0.492	0.172	0.5	0.472	-	0.728
Relative confidence *rc*	0.498	0.173	0.5	0.475	0.728	-
Overconfidence *oc*	-0.005	0.182	0.00	-0.551	0.476	0.215
Relative overconfidence *roc* [1)]	-0.002	0.407	0.05	-0.395	-0.223	0.005

Sample size n=60

[1)] There are 30 cases with *roc* less than or equal to 0.00 and 30 cases greater than or equal to *0.10*; thus, 0.05 is by definition the median, despite the fact that this value could not be chosen.

FIGURE 1

Comparison of performance and classification for Blavatskyy (2009) and this study

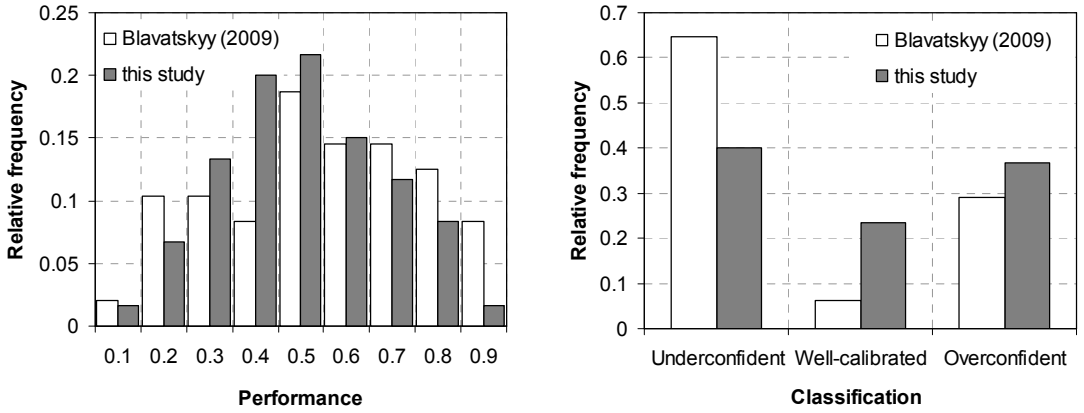


FIGURE 2

Population's better-than-others beliefs

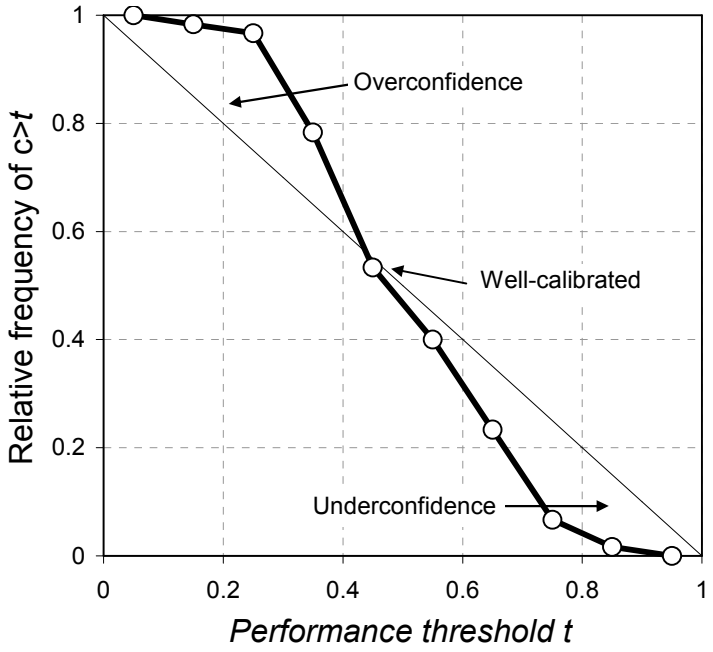
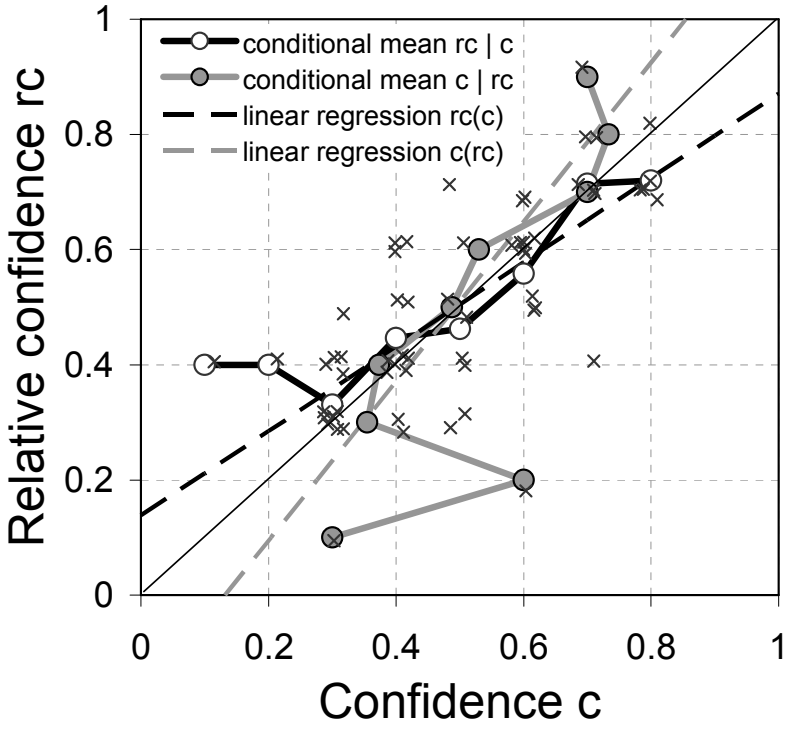


FIGURE 3

Comparison of confidence regarding absolute performance and relative performance
(including conditional means and linear regressions)



To improve the visibility of data points, we added some small white noise to single data points (but not to the data used for conditional means and regressions).