On the trend recognition and forecasting ability of professional traders

Markus Glaser, Thomas Langer, and Martin Weber*

Final Version, *Decision Analysis*, forthcoming

November 15, 2007

Abstract

Empirical research documents that temporary trends in stock price movements exist so that riding a trend can be a profitable investment strategy. In this paper, we provide a thorough test of the trend recognition and forecasting ability of financial professionals who work in the trading room of a large bank and novices (students). In an experimental study using a within-subject design, we analyze two ways of trend prediction that have analogues in the real world: probability estimates and confidence intervals (quantile estimates). We find that depending on the type of task *either* underconfidence (in probability estimates) *or* overconfidence (in confidence intervals) can be observed in the same trend prediction setting based on the same information. Furthermore, we find that the degree of overconfidence in both tasks is significantly positively correlated for all experimental subjects. These findings do not only contribute to the literature on judgmental forecasting but also have important implications for financial modeling. This paper demonstrates that a theorist has to be careful when deriving assumptions about the behavior of agents in financial markets from psychological findings.

Keywords: trend recognition, forecasting, conservatism, overconfidence, professionals, financial modeling

JEL Classification Code: C9, G1

*Markus Glaser is from the Lehrstuhl für Bankbetriebslehre at the Business School, Universität Mannheim, L 13, 15, 68131 Mannheim. E-Mail: glaser@bank.BWL.uni-mannheim.de. Thomas Langer is from the Finance Center, University of Münster, Universitätsstrasse 14-16, 48143 Münster. E-Mail: Thomas.Langer@wiwi.uni-muenster.de. Martin Weber is from the Lehrstuhl für Bankbetriebslehre at the Business School, Universität Mannheim, L 5, 2, 68131 Mannheim and CEPR, London. E-Mail: weber@bank.BWL.uni-mannheim.de. We would like to thank anonymous referees, David Budescu, Ning Du, Craig Fox, David Hirshleifer, Gur Huberman, Karl Ludwig Keiber, and seminar participants at the University of Mannheim, the University of Lausanne, the University of Vienna, the University of Münster, the German Finance Association Meetings in Mainz, the SPUDM Research Conference on Subjective Probability, Utility and Decision Making in Zürich, and the 8th BDRM Conference in Chicago for valuable comments and insights. Financial Support from the Deutsche Forschungsgemeinschaft (DFG) is gratefully acknowledged.

1 Introduction

Direct probability estimates and quantile estimates are integral parts of investment decision making. Individual and professional investors have to make up their minds about the probability that bull markets will continue or reverse to decide about the investment in stocks or bonds. Furthermore, investors make, often implicitly, judgments about a plausible interval for the value of a specific stock or an index some time in the future.

Both tasks - direct probability estimates as well as quantile estimates - have extensively been studied in the literature. Studies on direct probability estimates can be subdivided into the revision-of-opinion literature and the calibration literature (see also Subsection 2.1). In the revision-of-opinion literature the main finding is underconfidence (conservatism; (Edwards 1968)) whereas the calibration literature mainly documents overconfidence (see the survey of the calibration literature by (Lichtenstein, Fischhoff, and Phillips 1982)). However, recent evidence suggests that the amount of overconfidence documented varies significantly with the elicitation mode. Furthermore, (Erev, Wallsten, and Budescu 1994) find that revision-of-opinion tasks and direct probability estimates in calibration tasks are often hard to distinguish and that the same data set can produce either overconfidence or underconfidence depending on the way the data is analyzed. In contrast, studies analyzing quantile estimates almost universally document overconfidence although the degree of overconfidence varies with the specific elicitation method (see Subsection 2.1 and in particular (Klayman, Soll, Gonzáles-Vallejo, and Barlas 1999), (Soll and Klayman 2004), (Juslin, Winman, and Hansson 2007)).

The main goal of our study is to compare the two response modes mentioned above in one experiment using a within-subject design as earlier studies usually analyze the two response modes separately. Furthermore, we analyze the role of practical expertise as we are especially interested in the relevance of our findings for financial markets. Gaining insights about how professional traders form beliefs is particularly important as financial market outcomes are mainly driven by transactions of institutional investors.[1] Our research is thus related to recent field experiments such as (Haigh and List 2005) or (Alevy, Haigh, and List 2007). They replicate standard experiments that have been formerly run with a student subject pool (such as the (Gneezy and Potters 1997) myopic loss aversion experiment or the (Anderson and Holt 1997) information cascade experiment) with professional traders and motivate this line of research by the debate about the relevance of experimental findings building on student subjects for understanding phenomena in the field. They argue that professional behavior in the field might differ from student behavior in laboratory experiments due to training or regulatory considerations, which may affect the development of decision heuristics (see (Harrison and List 2004)). (Locke and Mann 2005) argue that if professional traders' discipline minimizes behavioral costs, then models of trader irrationality describe only small numbers of investors or lightly capitalized investors whose behavior has little impact on price formation. As (Haigh and List 2005) and (Alevy, Haigh, and List 2007) do, we combine the most attractive aspects of studies analyzing expert behavior and experiments, that is, we observe professionals in a controlled environment.

In our experiment, subjects observe artificially generated stock price charts which have either a positive or a negative trend in the long run. This two state scenario is similar to the situation professionals face when judging whether the market is currently bullish or bearish. We use two natural ways of trend prediction: estimates of the probability for

[1] See (Nofsinger and Sias 1999), (Lakonishok, Shleifer, and Vishny 1992), and (Gompers and Metrick 2001).

either trend and forecasts of the future price development via confidence intervals.

Our paper is related to the recent study by (Budescu and Du 2007). They investigate the quality of direct probability judgments and quantile estimates with a focus on calibration and consistency. However, in contrast to our study, they only analyze a group of 63 graduate accounting students and 75 undergraduate business students. Our study allows us to test whether professional traders are as prone to these biases as students or whether expertise moderates deviations from rationality. We can also analyze intrapersonal differences in the strength of possible biases as we embed the two different tasks in the same informational setting.

Our main results can be summarized as follows. Depending on the type of task *either* underconfidence (in direct probability estimates) *or* overconfidence (in confidence interval estimates) can be observed in the same trend prediction setting based on the same information. Underconfidence in the trend recognition task is more pronounced the longer the price history observed by subjects and the higher the discriminability of the price path generating processes. Furthermore, we find that the degree of overconfidence in both tasks is significantly positively correlated for all experimental subjects whereas performance measures are not.[2] This result suggests robust individual differences in the degree of overconfidence. In general, professional traders are more overconfident than students in trend prediction tasks. Our study has important implications for financial modeling. We argue that the question of which psychological bias should be incorporated into a model does not depend on a specific informational setting (in our experiment both trend prediction tasks were embedded in the same informational setting) but solely on the specific task

[2] Underconfidence is defined to be a negative degree of overconfidence. Performance is the absolute value of over- or underconfidence and thus states how far subjects are away from the well calibrated benchmark.

considered. It demonstrates that a theorist has to be careful when deriving assumptions about the behavior of agents in financial markets from psychological findings.

The rest of the paper is organized as follows. Section 2 surveys related literature on calibration, the role of expertise, and related finance models. Section 3 describes the design of the experiment. Sections 4 and 5 present the results. Section 6 discusses the results and concludes.

2 Literature Review and our Contribution to the Literature

2.1 The Calibration Literature

Calibration studies use two types of response modes: Direct probability estimates for precisely specified events or quantile estimates (confidence interval estimates) of probability functions of continuous variables (see, for example, the survey of the literature in (Budescu and Du 2007)).

Direct probability judgments of individuals are often poorly calibrated in the sense that they do not match appropriate probabilities that are determined empirically or derived from an appropriate model. In the typical studies of direct probability judgments, subjects predict outcomes of events or answer multiple choice questions that test general knowledge. Each item has two possible answers. For each question, participants choose the answer they think is more likely to be right and indicate, on a scale from 50% to 100%, how sure they are that they have chosen correctly.[3]

[3] This response mode is called the half range method. There is also a second related response mode, the full range method (see, e.g., (Juslin, Wennerholm, and Olsson 1999) or (Juslin, Winman, and Hansson 2007)).

Some early research, especially in sequential updating or revision-of-opinion settings documented conservatism or underconfidence, i.e. probability estimates are lower than properly calculated probabilities of the favored hypothesis (see, for example, the reviews in (Edwards 1968), (Griffin and Tversky 1992), (Rapoport and Wallsten 1972), and (Slovic and Lichtenstein 1971)).

However, most studies typically find that the assessments are more extreme than the respective correct relative frequencies, suggesting that subjects are overconfident in their judgments (see, for example, (Lichtenstein, Fischhoff, and Phillips 1982) or (Keren 1991)). For example, when people express 90% confidence, they are often correct in only about 70% of the cases.

(Erev, Wallsten, and Budescu 1994) argue that revision-of-opinion tasks and direct probability estimates in calibration tasks are often hard to distinguish. Overall, the conclusion can be drawn that overconfidence in probability estimation tasks is not as general as often assumed. Some studies show that overconfidence reverses to underconfidence depending on task difficulty. For easy tasks, underconfidence is observed more often (the so-called hard-easy-effect, see, e.g., (Lichtenstein, Fischhoff, and Phillips 1982) or (Keren 1991)).[4] Furthermore, (Erev, Wallsten, and Budescu 1994) show that the same data set can produce either overconfidence or underconfidence depending on the way the data is analyzed and stress the effects of random error which may perturb true judgments. Moreover, other authors stress the importance of unrepresentative questions chosen by the experimenter and that asking for frequencies reduces or eliminates overconfidence in direct probability estimates (see, for example, (Gigerenzer, Hoffrage, and Kleinbölting 1991) or

[4] However, (Juslin, Winman, and Olsson 2000) argue that the hard-easy effect has been interpreted with insufficient attention to the scale-end effects, the linear dependency, and the regression effects in data.

(Juslin 1994)).

To elicit judgments on uncertain continuous quantities, estimates of the quantiles of probability distributions are usually used. Judges are required to provide intervals (values) that correspond to pre-stated probabilities (see (Juslin, Wennerholm, and Olsson 1999), (Keren 1991), (Klayman, Soll, Gonzáles-Vallejo, and Barlas 1999), or (Soll and Klayman 2004)). Overconfidence is usually measured by the rate of surprises, i.e., the percentage of true values falling outside the confidence intervals. For example, consider an investor who is asked to provide 90% confidence intervals for a variety of stocks at the end of the year. If a subject is perfectly calibrated, the bounds he provided should include the actual values in 90% of the cases (and 10% of the values should fall outside the stated intervals). If the percentage of surprises is higher than 10%, and the proportion of values in the intervals is lower than the pre-stated probability (e.g., only 50% of true values fall within the 90% intervals), it is inferred that the judge is overconfident. The common finding is that the empirical intervals are far too narrow. Hit rates in many studies using 90% confidence interval are less than 50%, leading to surprise rates of 50% or higher, instead of the 10% expected from well-calibrated judges (see, for example, (Hilton 2001), (Klayman, Soll, Gonzáles-Vallejo, and Barlas 1999) or (Russo and Schoemaker 1992)). However, overconfidence in interval estimates is not universal. (Budescu and Du 2007) find in a within-subject experiment that 70% intervals are well calibrated whereas overconfidence was observed when subjects were asked for 90% confidence intervals and underconfidence when subjects were asked for 50% confidence intervals. (Glaser, Langer, and Weber 2005) show that overconfidence is stronger the longer the forecast horizon in stock market predictions. Slight underconfidence was observed for short forecast horizons of one week. (Glaser, Langer, Reynders, and Weber 2007b) show that the strength of the overconfidence effect in stock market forecasts is highly significantly affected by the fact whether subjects provide price or return forecasts. Volatility estimates are lower (and the overconfidence bias is thus stronger) when subjects are asked for returns compared to price forecasts.

To summarize the discussion of the literature so far, both direct probability estimates as well as the quantile method tend to find miscalibration. However, direct probability judgments induce only a modest bias as compared to the fractile method. Some studies using direct probability judgments even found modest underconfidence (e.g. (Erev, Wallsten, and Budescu 1994)). (Juslin, Wennerholm, and Olsson 1999) referred to the pattern of extreme overconfidence with the fractile estimates and the better calibration with probability estimate as format dependence of overconfidence.

Recent research analyzes whether different judgment biases are related and whether stable individual differences in reasoning or decision making competence (see (Parker and Fischhoff 2005), (Schunk and Betsch 2006), (Stanovich and West 1998), and (Stanovich and West 2000)) or miscalibration in general (see (Glaser, Langer, and Weber 2005), (Glaser and Weber 2007), (Jonsson and Allwood 2003), (Klayman, Soll, Gonzáles-Vallejo, and Barlas 1999) or, (Budescu and Du 2007)) exist. Furthermore, there is some evidence that people show different levels of overconfidence depending on the task or domain but remain in the same rank-order over tasks or domains (see (Jonsson and Allwood 2003), p. 561, and (Glaser, Langer, and Weber 2005)).

2.2 The Effect of Expertise on Judgment

The analysis of the effects of expertise on judgmental forecasting and behavior in financial markets has attracted lots of attention. In this subsection, we briefly sketch the main findings that are relevant for our study. We first discuss the role of expertise in the calibration literature with a special focus on the role of expertise in forecasting. After that, we will present studies analyzing differences between the trading behavior of individual and professional investors and recent field experiments. At the end of this subsection, we discuss potential reasons for the conflicting results found in the literature on the role of expert judgment. Extensive literature reviews can also be found in (Keren 1991), (Andersson, Edman, and Ekman 2005), (Koehler, Brenner, and Griffin 2002), or (Lawrence, Goodwin, O'Connor, and Önkal 2006).

In the last subsection, we presented studies documenting that people are miscalibrated in general. An overwhelming body of research shows that experts in most domains also tend to be miscalibrated (e.g. (Koehler, Brenner, and Griffin 2002)), although some exceptions exist such as the calibration of weather forecasters ((Murphy and Brown 1984) and (Murphy and Winkler 1984)).

Experts' prediction intervals are also too tight indicating overconfidence (see, for example, (Russo and Schoemaker 1992), (Graham and Harvey 2003), or (Deaves, Lüders, and Schröder 2007)). Overconfidence is also reported with software professionals' effort prediction intervals, i.e. judgmental prediction intervals that reveal the uncertainties in software development effort ((Jorgensen, Teigen, and Molokken 2004)).

In field experiments, in which well known studies building on student subject pools (such as the (Gneezy and Potters 1997) myopic loss aversion experiment or the (Anderson and

Holt 1997) information cascade experiment) are replicated with professional traders, the results are also mixed. (Haigh and List 2005) find that professional traders exhibit behavior consistent with myopic loss aversion to a greater extent than undergraduate students. In contrast, based on 1,500 individual decisions of Chicago Board of Trade traders, (Alevy, Haigh, and List 2007) find that professionals playing an information cascade experiment are better able to discern the quality of public signals when compared to students. (Sarin and Weber 1993) evaluate the effect of ambiguity on individual decisions and the resulting market price in market settings. They examine the issue of whether ambiguity effects persist in the face of market incentives and feedback. Their subject pool consists of graduate business students and bank executives which hardly behave differently in the experiment.

In another strand of literature, studies analyze how experience affects decisions and whether biases are eliminated by trading experience and learning. Consider, for example, one of the most extensively studied biases of individual investors, the disposition effect. (Feng and Seasholes 2005) analyze the disposition effect, the investor's reluctance to realize losses and his propensity to realize gains, and find that experience eliminates the reluctance to realize losses. (Dhar and Zhu 2006) use demographic and socioeconomic variables as proxies for investor literacy, and find empirical evidence that wealthier individuals exhibit a lower disposition effect. (Shapira and Venezia 2001) compare the disposition effect of independent individual investors and investors whose accounts were managed by brokerage professionals and find that both groups are biased, but the professionals to a lesser degree. (Locke and Mann 2005) also show that full-time traders in their sample hold onto losses significantly longer than gains, but they do not find evidence of costs associated with this behavior.

To sum up so far, the question of how the strength of professionals' biases compare to that of non-professionals is difficult to answer. The results presented so far suggest the following interpretation. Financial education and financial knowledge (also called "financial literacy"), acquired by trading experience or other means of learning might help improve behavior and reduce biases in tasks in which such knowledge should actually be helpful. Further support for this conjecture is also provided by (Agnew and Szykman 2005), (Elliott, Hodge, and Jackson 2007), and (Glaser, Langer, Reynders, and Weber 2007a).

2.3 Related Finance Models

In this section, we briefly review how judgment biases are incorporated into theoretical models of financial markets.

Overconfidence is most often modeled as underestimation of the variance of information signals and, as a consequence, the uncertain value of a risky asset. Assume there is a risky asset with liquidation value v which is a realization of $\tilde{v} \sim N(0, \sigma_{\tilde{v}}^2)$. Investors receive signals $\tilde{s} = \tilde{v} + c \cdot \tilde{e}$ (assuming that $\tilde{v}$ and $\tilde{e}$ are independent) with $\tilde{e} \sim N(0, \sigma_{\tilde{e}}^2) \Rightarrow \tilde{s} \sim N(0, \sigma_{\tilde{v}}^2 + c^2 \cdot \sigma_{\tilde{e}}^2)$. If $c = 1$, investors are well calibrated, if $0 \leq c < 1$, investors are overconfident. Conditional expectation and conditional variance of $\tilde{v}$, given the realization s are

$$E[\tilde{v} \mid \tilde{s} = s] = E[\tilde{v}] + \frac{Cov[\tilde{v}, \tilde{s}]}{Var[\tilde{s}]}(s - E[\tilde{s}]) = \frac{\sigma_{\tilde{v}}^2}{\sigma_{\tilde{v}}^2 + c^2 \cdot \sigma_{\tilde{e}}^2} \cdot s \tag{1}$$

$$Var[\tilde{v} \mid \tilde{s} = s] = Var(\tilde{v}) - \frac{(Cov[\tilde{v}, \tilde{s}])^2}{Var[\tilde{s}]} = \sigma_{\tilde{v}}^2 - \frac{\sigma_{\tilde{v}}^4}{\sigma_{\tilde{v}}^2 + c^2 \cdot \sigma_{\tilde{e}}^2} \tag{2}$$

These equations show that overconfident investors underestimate the variance of the risky asset. In the extreme case $c = 0$ the conditional variance is zero. (Benos 1998), (Caballé and Sákovics 2003), (Kyle and Wang 1997), (Odean 1998), and (Wang 1998) incorporate overconfidence into different types of models such as those of (Diamond and Verrecchia 1981), (Hellwig 1980), (Grossman and Stiglitz 1980), (Kyle 1985), and (Kyle 1989). These models are useful to explain high levels of trading volume in financial markets. There are other overconfidence models that address questions like the dynamics of overconfidence, the survival of overconfident investors in markets, and the cross-section of expected returns. Examples are (Daniel, Hirshleifer, and Subrahmanyam 1998), (Daniel, Hirshleifer, and Subrahmanyam 2001), (Hirshleifer and Luo 2001), (Gervais and Odean 2001), and (Wang 2001).

Empirical research documents that temporary trends in stock price movements exist.[5] Moreover, riding a trend can be a profitable investment strategy.[6] Thus, the ability to recognize trends in stock markets influences the quality of investment decisions. This argument is in line with (Shiller 2001) who states that "ultimately, people who choose asset allocations must use their subjective judgment about the probability that stock trends will continue".[7] (Shleifer and Summers 1990) state in their survey of the noise trader approach in finance that "one of the strongest investor tendencies documented in both experimental and survey evidence is the tendency to extrapolate or to chase the

[5] See, for example, (Jones and Litzenberger 1970) or (Bernard and Thomas 1989).

[6] One example is the profitability of momentum strategies in which stocks with high (low) returns over the last three to 12 months are bought (sold short). (Jegadeesh and Titman 1993) and (Jegadeesh and Titman 2001) showed for US stocks that this strategy results in significant positive profits. This strategy has been successful in other stock markets as well (see (Rouwenhorst 1998), (Rouwenhorst 1999), and (Glaser and Weber 2003) for international evidence on the profitability of momentum strategies).

[7] (Shiller 2001), p. 3.

trend".[8] Betting on trends is closely related to noise trading, positive feedback trading and extrapolative expectations.[9] One explanation for positive feedback trading is the way investors form expectations. They naively extrapolate trends.[10] Furthermore, noise trading is related to overconfidence. (Hirshleifer 2001) argues that pure noise trading is a limiting case of overconfidence.[11]

Other models incorporate the second behavioral bias mentioned in the introduction, conservatism or underconfidence, into models of financial markets. One example is (Barberis, Shleifer, and Vishny 1998) who model investors who try to infer information from earnings series.[12] Earnings follow a random walk. Because the firm pays out all earnings as dividends the true value of the firm follows a random walk, too. Investors do not understand that dividends and firm value follow a random walk and thus do not use the random walk model to forecast future earnings. They believe that the earnings process stochastically fluctuates between either a mean-reverting or a trending regime. Using past realizations of the earnings series investors try to find out which of the two regimes is currently generating earnings. This way of modeling beliefs or updating of probabilities is able to capture the conservatism bias (underconfidence, underreaction) and the representativeness heuristic. An investor using the mean-reverting model to forecast earnings reacts too little to an earnings announcement which leads to underreaction or conservatism. However, when the trending model is used to forecast earnings, strings of good or bad news are extrapolated too far into the future. This captures the idea of representativeness. Investors see patterns

[8] (Shleifer and Summers 1990), p. 28.

[9] See, for example, (Hirshleifer 2001), p. 1545.

[10] See, for example, (Hirshleifer 2001), p. 1567, or (Daniel, Hirshleifer, and Teoh 2002), p. 145.

[11] (Hirshleifer 2001), p. 1568.

[12] To be more precise, (Barberis, Shleifer, and Vishny 1998) model a representative investor.

in random series and they thus extrapolate trends and overreact. One implication of this way of modeling beliefs is that investors erroneously use the frequency of recent earnings reversals to predict the likelihood that the current earnings change will be reversed in the future.[13]

A recent, closely related model is (Hong, Stein, and Yu 2007). They study the asset pricing implications of learning in an environment in which the true model of the world is a multivariate one, but agents update only over the class of simple univariate models. Thus, if a particular simple model does a poor job of forecasting over a period of time, it is discarded in favor of an alternative simple model.

This overview shows that there are many different ways of modeling judgment biases: Some models assume an underestimation of the variance of signals (or overestimation of their precision), which means too tight confidence intervals. Other theories model how investors form beliefs and predict future outcomes when they observe past realizations of a time series. Furthermore, there is the mechanistic modeling approach in positive feedback trading models in which investors form expectations of future prices by extrapolating trends or in which differences of past prices are directly incorporated into asset demand functions. Our study helps to evaluate these various modeling approaches and competing assumptions.

2.4 Contribution of our Paper to the Literature: Summary

The main contributions of our paper to the literature can be summarized as follows:

Overconfidence is far less uniform and universal than is often assumed. In particular, it

[13] See (Bloomfield and Hales 2002) for an experimental investigation of this assumption.

seems to vary dramatically across elicitation methods such as direct probability estimates or quantile estimates. The first goal of this study is to compare directly the quality of the judgments obtained from the two methods. Apart from (Budescu and Du 2007), there is little work on this issue.

The second goal is to contribute to the literature on field experiments in the spirit of (Harrison and List 2004), (Haigh and List 2005), and (Alevy, Haigh, and List 2007). (Harrison and List 2004) distinguish (1) a conventional lab experiment (one that employs a standard subject pool of students, an abstract framing, and an imposed set of rules); (2) an artefactual field experiment (the same as a conventional lab experiment but with a nonstandard subject pool); (3) a framed field experiment (the same as an artefactual field experiment but with field context in either the commodity, task, or information set that the subjects can use); and (4) a natural field experiment (the same as a framed field experiment but where the environment is one where the subjects naturally undertake these tasks and where the subjects do not know that they are in an experiment). Our present study can be regarded as a framed field experiment according to (Harrison and List 2004).

The third goal of our study is to test modeling assumptions of recent behavioral models in finance. In particular, we analyze how the two main groups of judgment biases modeled in finance are related.

3 Experimental Design

The experiment we present in this paper was part of a larger project that consisted of four phases.[14] In the first phase, a questionnaire was presented that contained knowledge related overconfidence tasks that required a controlled environment (see (Glaser, Langer, and Weber 2005)). In this phase, we further assigned secret IDs to subjects to guarantee anonymity and collected demographic data. The three other phases of the project were internet based. In prespecified time windows subjects had to access a web page and log in to the experimental software. Each phase took about 30 minutes to complete. Overall, 31 professionals of a large German bank participated in the project, 29 of them completed all parts of the study, the remaining two subjects dropped out or missed a phase due to vacation or other reasons. Based on a self report in the first phase, 11 professionals assigned their job to the area 'Derivatives', 10 to the area 'Proprietary Trading', 12 to the area 'Market Making' and 6 to 'Other Area'.[15] The age of the professionals ranged from 23 to 55 with a median age of 33 years. 14 subjects had an university degree and the median subject had worked for the bank for five years (range 0.5 to 37 years). In addition to the professionals, we had a comparison group of 64 advanced students, all specializing in Banking and Finance at the University of Mannheim. Their median age was 24 and ranged from 22 to 30. While there were no women in the group of professionals, the comparison group consisted of 6 female and 58 male students. The comparison group was faced with exactly the same procedure as the professionals with the one exception that for organizational reasons the questionnaire of phase 1 was filled out at the end rather

[14] In addition, there was a preexperimental meeting in which we interviewed the professional subjects to better understand their decision scope and goals.

[15] Subjects could assign themselves to more than one area.

than the beginning of the project.

All relevant data for this paper was electronically collected in phases 2 and 3.[16] The study consists of two related tasks which we call 'trend prediction by probability estimates' or trend recognition and 'trend prediction by confidence intervals' or forecasting. In the trend prediction by probability estimates part, subjects were confronted with two simple distributions of price changes. Both processes could generate price movements of size -2, -1, 0, $+1$, and $+2$ with different probabilities. They were constructed such that one process had a positive trend, i.e. a positive expected value, and the other process had a negative trend. Both processes were graphically and numerically displayed and their meaning explained in detail. Figure 1 shows a screenshot of the experiment. Subjects were further informed that one of the two processes was randomly picked to generate a price path.[17] After observing this path it would be their task to state a subjective probability that this path is driven by a positive or a negative trend. Great care was taken to make sure subjects understood that this was a synthetically generated path and no relation to real market phenomena and price patterns must be assumed. Subjects then saw a developing price path. The path started in $t = 0$ at price 100 and moved on to $t = 5$, where subjects had to make their first judgment. The probability elicitation was implemented as a two stage procedure. First, subjects had to state if they were at least 90% sure that a specific process would underly the observed path. Then, according to their answer they could choose an explicit percentage number in the restricted domain. This procedure was perfectly symmetric. When moving the slider the numerical percentage values simultaneously increased and decreased for both processes (see Figure 1). Thus,

[16] There were further tasks in phases 2 and 3 that we do not address in this paper.

[17] Note that in this design, the critique of a potential unrepresentative sample of tasks in the spirit of (Gigerenzer, Hoffrage, and Kleinbölting 1991) or (Juslin 1994) does not apply.

the fine tuning offered probabilities between 90% and 100% for the negative trend if subjects were at least 90% sure about the negative trend in the first step. The scale was restricted to a respective percentage range [0%, 10%], if they were 90% sure about a positive trend and to [10%, 90%] otherwise. It was possible in stage 2 to go back and change the judgment of the first stage. Subjects made such judgments at dates $t = 5$, 6, 7, 8, 9, 10 and 20.[18] Three different pairs of price processes were used within the experiment. In all cases the distribution with a positive trend was a mirror image of the distribution with a negative trend. Thus, by stating the negative trend probabilities for the outcomes -2, -1, 0, $+1$, and $+2$ the complete pair is defined. The following distributions were used:

$$\begin{aligned} D_{1.5} &= (18\%, \quad 24\%, \quad 34\%, \quad 16\%, \quad 8\%) \\ D_{1.75} &= (9.8\%, \quad 28\%, \quad 43\%, \quad 16\%, \quad 3.2\%) \\ D_{2} &= (12\%, \quad 32\%, \quad 37\%, \quad 16\%, \quad 3\%) \end{aligned}$$

The index of D_k is chosen as the quotient of the probabilities for the outcome -1 and the outcome $+1$. These odds k will later turn out to be central for deriving the correct trend probabilities given some price change. In addition, k can be interpreted as a measure of the discriminability of the two trends (see also (Griffin and Tversky 1992) and (Massey and Wu 2005)). In phase 2 of the experiment, all subjects saw each process pair exactly once. So, overall they had to estimate 21 probabilities (one for each of the seven dates times three price paths). Phase 3 was identical to phase 2 with different price paths presented to subjects. Hence, overall we have 42 probability judgments for each subject in this 'trend prediction by probability estimates' task.

[18] On entering the second stage the slider was always set to the stated percentage of the previous round or the closest available percentage value (i.e. 90% or 10%). In $t = 5$ it was set to 50% or the closest available number. Furthermore, subjects did not receive feedback about the mathematically correct probabilities after each estimate.

After stating the subjective probability at time $t = 20$ subjects were further asked to predict how the price path will develop until $t = 40$. They had to provide a price interval that they expect to contain the price at $t = 40$ with 90% probability. The price at $t = 40$ should fall short of the lower boundary and pass over the higher boundary with only 5% probability each. Such an interval judgment was collected for each process pair in phases 2 and 3. Thus, overall we have six intervals for each subject in this 'trend prediction by confidence intervals' part.[19]

4 Results

4.1 Trend Prediction by Probability Estimates

In this subsection we compare the subjective probability estimate with the mathematically correct probability that a given chart was generated by the process with the negative trend (which, of course, determines the probability that this chart was generated by the process with the positive trend). It turns out that $P(\text{neg. trend}|\text{level} = x)$ neither depends on date t nor on details of the price path but only on the absolute price change from the starting price level 100. This can be seen as follows: If we observe a price step of +1 then the odds are given as

$$\frac{P(+1|\text{pos. })}{P(+1|\text{neg. })} = k. \tag{3}$$

Due to the specific construction of the distributions it is easily verified that it holds more generally

[19] Two professionals participated in only one of the two phases and thus provided only three intervals.

$$\frac{P(z|\text{pos. })}{P(z|\text{neg. })} = k^z \quad \text{for} \quad z \in \{-2, -1, 0, +1, +2\}. \tag{4}$$

From $P(\text{level} = 100 \mid \text{pos. })/P(\text{level} = 100 \mid \text{neg. }) = 1$ it then follows by induction

$$\frac{P(\text{level} = x|\text{pos.})}{P(\text{level} = x|\text{neg.})} = k^{x-100} \quad \text{for all } x. \tag{5}$$

Bayes' Law yields

$$P(\text{neg.} \mid \text{level} = x) = \frac{P(\text{level} = x|\text{neg.}) \cdot P(\text{neg. })}{P(\text{level} = x|\text{neg.}) \cdot P(\text{neg.}) + P(\text{level} = x|\text{pos.}) \cdot P(\text{pos. })} \tag{6}$$

and with $P(\text{neg.}) = P(\text{pos.})$ it follows from equation (5),

$$P(\text{neg.}|\text{level} = x) = \frac{1}{1 + k^{x-100}}. \tag{7}$$

Assume, for example, $k = 2$, that is the distribution $D_2 = (12\%, 32\%, 37\%, 16\%, 3\%)$ and a price level of 102 after 5 periods ($t = 5$). Then, the probability $P(\text{neg. trend}|\text{level} = 102)$ is $1/(1+2^{102-100}) = 1/5$. Our design is similar to the studies of (Griffin and Tversky 1992) and (Massey and Wu 2005). For example, in the study of (Griffin and Tversky 1992), there was a coin with either a 3 : 2 bias in favor of heads or tails. Subjects received outcomes of coin spinning, in other words the number of heads and the number of tails.[20] Then, subjects were asked to state their subjective confidence that the respective outcome was generated by the coin with the 3 : 2 bias in favor of heads. As in our experiment, the probability distributions were mirror-images of each other. It turned out that the only

[20] There was always a majority of the number of heads.

relevant determinant of the mathematically correct probability was the difference between the number of heads and the number of tails whereas sample size and sequence of heads and tails were irrelevant. This shows the analogy to our study if one interprets the biases in favor of heads or tails as positive or negative trend. Counterparts of sample size and sequence are time period and price path in our study.

(Massey and Wu 2005) transfer the (Griffin and Tversky 1992) logic into a regime shift detection setting. They conduct three experiments to test how effective individuals are at detecting such regime shifts. Specifically, they investigate when individuals are most likely to underreact to change and when they are most likely to overreact to it. They develop and confirm a system-neglect hypothesis: Individuals react primarily to the signals they observe and secondarily to the environmental system that produced the signal. This system is, for example, described by the signal diagnosticity and the transition probability. (Massey and Wu 2005) find that individuals pay inordinate attention to the signal, and neglect diagnosticity and transition probability, the aspects of the system that generated the signal. Moreover, they suggest that system neglect leads to a predictable pattern of under- and overreaction: individuals are most prone to underreaction in unstable environments with precise signals, and to overreaction in stable environments with noisy signals.

Equation (7) shows that as long as the price level is higher than 100, the probability that the chart was generated by the process with the negative trend should be lower than 0.5. If the price level is lower than 100, the probability that the chart was generated by the process with the negative trend should be higher than 0.5. If subjects predict the correct direction, we are able to compare the subjective probability with the mathematically correct probability. Assume, for example, the correct probability for the negative trend is 0.7.

If a person states a subjective probability of 0.9 this person is classified as overconfident. If the person states a probability of 0.6 the subject is regarded as underconfident. This classification is far less clear if a subject states a probability of, say, 0.2 for the negative trend. Is this person extremely overconfident? Or is she underconfident because her subjective certainty is lower than the objective certainty? We argue that these subjects cannot be reasonably classified as either overconfident or underconfident.[21] We thus omit these subjects from our analysis. In period $t = 5$, 24 of 552 probability (4.3%) estimates have to be eliminated (37 in period $t = 10$ and 28 in period $t = 20$).

In the following, we focus on the three time periods $t = 5$, $t = 10$, and $t = 20$.[22] Table 1 presents the number of probability estimates above (overconfident), equal to (well calibrated), and below the correct probability (underconfident).[23] The majority of probability estimates is below the correct probability thus indicating underconfidence. For example, in time period $t = 5$, 175 probability estimates are above the correct probability (33.14%) whereas 293 probability estimates are below the correct probability (55.50%).

These preliminary results are similar to (Griffin and Tversky 1992) who report an underestimation of the mathematically correct probability in this scenario. Furthermore, Table

[21] There are other studies that measure overconfidence in a different way. According to these studies we should classify subjects that predict the wrong trend to be true as extremely overconfident. However, it can be shown that this methodology could lead to serious biases. Furthermore, other studies restrict the probability range for the subjects' answers. (Griffin and Tversky 1992), for example, only analyze situations with a majority of heads. Other studies present knowledge questions. Subjects have to provide an answer and have to state their subjective confidence from 50% to 100%. In both cases, the probability range for the subjects' answers is restricted so that apparently illogical answers (e.g. the answer yes with a subjective certainty of 40%) are impossible.

[22] Analysis of the remaining time periods yields similar results. We omit these results to save space.

[23] Almost 90% of the correct probability estimates ("well calibrated") correspond to price levels of 100 with a correct probability of 0.5.

1 shows that underconfidence increases with time. (Griffin and Tversky 1992) also find that underconfidence is stronger when the weight of evidence (sample size) is higher.

To analyze these preliminary observations more deeply we define an overconfidence measure doc_1 which is simply the difference between the subjective and the mathematically correct probability.[24] This measure indicates overconfidence if $doc_1 > 0$ and underconfidence if $doc_1 < 0$. Table 1 presents the medians of doc_1 for dates $t = 5$, $t = 10$, and $t = 20$. The medians of doc_1 are always negative indicating underconfidence. For example, doc_1 in $t = 5$ is -4.12, which means that the median probability estimate underestimates the mathematically correct probability by -4.12 percentage points. Again, underconfidence increases with time. In $t = 10$ and $t = 20$, doc_1 is even lower.

When we analyze the medians of doc_1 separately for each distribution we find that the higher the discriminability between the two distributions (the larger k), the larger the underconfidence (lower doc_1-values). This result is also consistent with (Griffin and Tversky 1992) and (Massey and Wu 2005). (Massey and Wu 2005), for example, find that underreaction is most prevalent in high diagnosticity systems.

Our doc_1 measure reflects the deviation of the subject's answer from the rational benchmark on the probability scale. Alternatively, the bias could be measured on the price scale as each probability uniquely corresponds to some price level. The choice of scale is not irrelevant, since the transformation from probabilities to prices is not linear. For example, a doc_1 value of $+5$ corresponds to a much larger price difference if it is derived from a 95% probability estimate and a 90% correct probability than if it results from the values 65% and 60%, respectively. To make sure that our results are robust with respect to the

[24] Note, that we only analyze subjects that predict the correct direction. Thus, all probabilities are equal to or above 0.5.

chosen scale, we consider a second measure doc_2 on the price scale that is defined as the difference between the price level implied by the probability estimate and the observed price level.[25] If doc_2 is positive, subjects are overconfident, if doc_2 is negative, they are underconfident.

Table 1 presents the medians of doc_2 for dates $t = 5$, $t = 10$, and $t = 20$. Again, all doc_2 are negative indicating underconfidence.

So far, we have analyzed medians for all probability estimates although these observations are not independent because several observations come from the same subject. We now calculate doc_1 and doc_2 separately for each individual as simply the median of doc_1 and doc_2, respectively.

Table 1 shows medians of the median doc_1 and doc_2 values per subject. All values indicate underconfidence. All doc_1 and doc_2 values are significantly different from zero at the 1% level except for doc_2 in $t = 5$ which is only significant at the 5% level.[26]

When interpreting our results, we have to keep in mind that our theoretical prediction

[25] We replace stated probabilities of 100 and 0 with 99.75 and 0.25, respectively. This takes into account that a subject with a subjective probability estimates in the range [99.5, 100] has to choose the answer 100 due to the discrete scale of possible probability estimates.

[26] Note that by looking at median differences per subject instead of means we mostly circumvent the problem that one can have very large differences in one direction but not in the other (a fact that e.g. (Erev, Wallsten, and Budescu 1994) relied on to explain simultaneous over- and underconfidence). A participant displays underconfidence in the majority of cases if her median doc_1-difference is negative. There remains a subtle problem though due to the fact that the medians are calculated from six values and are thus an average of the two intermediate differences. Thereby, it could be argued that the above mentioned skewness results in the negative medians even though positive and negative differences are actually equally likely (i.e. they are simply noise). Such an explanation is not supported by the data though. Defining for each subject the sum of the signs of the differences as an additional and simple confidence measure, we still find a highly significant effect (for $t = 5$, 10, and 20). The strongest effect is observed for $t = 20$ where the measure is negative for 82.80% of the subjects (for 10.75% it is positive) and the median level of underconfidence is -4.

(the mathematical correct probability) is based on a particular model with a rather strong assumption, namely conditional independence. In earlier literature this was considered to be a potential problem in situations where artificial laboratory settings did not coincide with a real-world decision environment. (Winkler and Murphy 1973)), for example, have used this fact to explain conservatism. However, in a finance setting as ours, the assumption that the change of the value of a stock over time is independent from the previous ones is not unrealistic (see (Fama 1965)).

The main results of this subsection can be summarized as follows:

- Subjects are, on average, underconfident, i.e. they underestimate the mathematically correct probability.
- Underconfidence is stronger the higher the sample size, i.e. the longer the time period.
- Underconfidence is stronger the higher the discriminability of the two distributions, i.e. the higher k.

4.2 Trend Prediction by Confidence Intervals

Intertwined with the trend probability estimation, a second task was used to further examine possible judgment biases related to trend prediction. Following each final probability judgment, based on a price path from $t = 0$ to $t = 20$, subjects were asked to make a prediction about the price level at time $t = 40$. This judgment was elicited via a confidence interval, consisting of an upper and lower limit. Subjects were instructed that the price at $t = 40$ should be above the upper limit and below the lower limit with 5% probability each. Thus the stated confidence interval was supposed to contain the price at $t = 40$ with 90% probability. The experimental subjects were asked to provide such confidence intervals for three price paths in phase 2 as well as in phase 3. Thus overall we have six [low, high]-intervals, two for each of the price process pairs $D_{1.5}$, $D_{1.75}$, and D_2. The correct distribution of prices at $t = 40$ can be computed from the price at $t = 20$ and the process distributions. Thus, the stated confidence intervals can be translated into probability intervals. For a perfectly calibrated subject these induced probability intervals $[p_{low}, p_{high}]$ should be [5%, 95%].[27] We measure the degree of under-/overconfidence via the length of the induced probability interval. If a subject is too sure about the price at $t = 40$ and provides too narrow an interval, it is classified as overconfident. More explicitly, our third measure of overconfidence doc_3 is defined as: $doc_3 = 90\% - (p_{high} - p_{low})$. Positive doc_3-values correspond to overconfidence, negative doc_3-values to underconfidence.

Such estimates of the quantiles of probability distributions are often elicited for uncertain continuous quantities, usually general knowledge questions (see Subsection 2.1). Judges are required to provide intervals that correspond to pre-stated probabilities (see, for example, (Juslin, Wennerholm, and Olsson 1999), (Klayman, Soll, Gonzáles-Vallejo, and Barlas 1999), (Soll and Klayman 2004), (Cesarini, Sandewall, and Johannesson 2006), (Juslin, Winman, and Hansson 2007)). Over- or underconfidence is usually measured by the rate of surprises, i.e., the percentage of true values falling outside the confidence intervals. The common finding is that the empirical intervals are far too narrow (see, for example, Lichtenstein, Fischhoff, and Phillips (1982), (Glaser, Nöth, and Weber 2004), (Klayman, Soll, Gonzáles-Vallejo, and Barlas 1999), or (Hilton 2001)).

Such confidence intervals are also used to elicit predictions of time series such as stock price charts (see, for example, (Budescu and Du 2007), (Glaser and Weber 2005)). In such

[27] Note, however, that due to the discreteness of the distribution it was not possible in general to exactly hit this target.

studies, hit rates are also often calculated (such as in (Budescu and Du 2007)). However, especially when real financial time series are forecasted within a particular time window, such hit rates are problematic as the development of stock prices of different firms are not independent.[28] Consider for example an investor who predicted in July 2001 that a set of stocks will slightly increase until the end of the year with error bounds around median forecast that correspond to historical stock price volatility. After the terror attacks of September 11, 2001, such a person would have been classified as extremely overconfident although the ex ante prediction looked quite reasonable. This is the reason why in such tasks several studies compare the volatility expectation implied by the width of the confidence interval stated with a reasonable benchmark such as the historical volatility or the volatility implied by the option market (see, for example, (Glaser and Weber 2005) or (Graham and Harvey 2003)). However, when forecasting financial time series, financial econometricians will never agree on one single "correct" volatility estimate which is necessary to judge the appropriateness of confidence intervals stated by subjects. The "optimal" volatility forecast does hardly exist (see (Poon and Granger 2004)).

However, in our setting, we are able to exactly calculate the probability implied by the intervals stated by subjects. Our doc_3 thus enables us to precisely calculate the degree of overconfidence by the tightness of the intervals provided. We thus circumvent the problems of other studies that calculate hit rates in stock prediction tasks or use questionable volatility benchmarks.[29]

[28] See (Jorgensen, Teigen, and Molokken 2004) for an extensive discussion of the hit rate and other evaluation measures of judgmental confidence intervals.

[29] In principal, we could also calculate "usual" hit rates by producing ex post price paths for periods 21 to 40 extending the price path from 1 to 20 that the subject saw before they made their final judgment. But this would just mean that we add a layer of noise (the noise that the final price is just one draw from the distribution and the noise that the price path from 1 to 20 could have happened to be atypical for the underlying process) to a situation that we are able to analyze

Table 2 presents the results. From the 552 observations, 335 are classified as overconfident ($doc_3 > 0$) and 217 as underconfident ($doc_3 < 0$) with a median doc_3-value of +10.03.[30] In Table 2, we also present doc_3-values on a more aggregated level. From the 93 subjects in the data set 61 have a positive median doc_3-value and are thus classified as overconfident. The 32 other subjects are underconfident. The median subject has a doc_3-value of 8.01. These doc_3-values are significantly different from 0. To address a potential concern, we will next look at a slightly different overconfidence measure doc_4. In the computation of the doc_3-values we had to determine the correct price distribution at $t = 40$. This t_{40}-distribution depends on the trend process pair as well as on the current ($t = 20$) price level. The current price level enters the computation in two ways. It obviously determines the absolute location of the t_{40}-distribution, but it also influences the shape of the distribution via the correct trend probabilities that follow from the price level. By an individual misjudgment of the trend probabilities in $t = 20$, as documented in the previous section, a different "rational" t_{40}-distribution would thus be derived. In the computation of our fourth overconfidence measure doc_4 we replace the correct trend probability derived from the price level at $t = 20$ by the trend probability stated by the individual in $t = 20$. As shown in Table 2, this alteration of the rational benchmark does not have a strong impact on the results, however. From the 524 observations, 341 are classified as overconfident and 183 as underconfident with a median doc_4-value of 9.11.[31] These doc_4-values as well as the aggregated numbers (67 subjects are classified as overconfident, 26 as underconfident) are significantly different from 0.

without any noise.

[30] We again report median instead of mean values because of the obvious asymmetry in possible doc_3-values (ranging from -10% to +90%).

[31] We restrict the doc_4-analysis to the same 524 observations that we considered in the analysis of the measures doc_1 and doc_2 in $t = 20$.

Our finding of extreme overconfidence when people state 90% confidence intervals is consistent with several other studies. However, when interpreting this result, we have to keep in mind that in general, the amount of overconfidence can vary depending on how the intervals are elicited and whether 90% confidence intervals or, say, 70% confidence intervals are elicited (see (Budescu and Du 2007)). Furthermore, it would be interesting to analyze whether subjects providing too tight 90% confidence intervals also state too tight 70% confidence intervals to analyze whether robust individual differences in the degree of miscalibration in the spirit of (Jonsson and Allwood 2003) also exist in this setting.

To summarize, the main result of this subsection is: In trend prediction by confidence intervals, subjects are overconfident on average, i.e. they provide too narrow intervals.

5 Traders versus Students: The Correlation of Overconfidence Measures

5.1 Traders versus Students

Table 3 analyzes the various overconfidence measures defined in the previous sections separately for the two subject groups: professional traders and lay people (students). The first part of Table 3 shows the medians of all observations. Accordingly, the second column shows figures that are also part of Table 1 and Table 2. For example, Table 3 shows again that the median doc_1-value among all observations in $t = 5$ is -4.12. The next two columns present the median doc_1-value in $t = 5$ as well as all the other doc_i-values, $i \in \{1, 2, 3, 4\}$, for the two groups of subjects separately. The main observation is that financial professionals have higher doc_i-values than students. Thus, the degree of overconfidence is always higher for traders. Note, again, that underconfidence is considered to be a negative degree of overconfidence.

There are two interpretations of these findings. Focusing on doc_1- and doc_2-values shows that all medians are negative (except for doc_1 and doc_2 in $t = 5$ with a median value of 0) implying that professionals perform better in this task; they are closer to the correct probability. In other words: They are less underconfident than students in this task. However, analysis of doc_3- and doc_4-values yields that professionals are not generally better than students but, again, more overconfident. The last column of Table 3 presents the p-value of a Kruskal-Wallis test. Null hypothesis is equality of populations. The results show that we can reject the null hypothesis of the equality of populations in most cases at least at the 5% level. This first interpretation would thus be that professionals are better calibrated in direct probability estimation tasks whereas students are better calibrated in quantile estimates. Our results would thus be in line with the unclear relation between expertise and calibration in the studies sketched in Subsection 2.2. A second interpretation could be as follows: There are stable individual differences in the degree of overconfidence across subject. While the amount of overconfidence measured differs from task to task the ranking of people does not (see also (Glaser, Langer, and Weber 2005) or (Jonsson and Allwood 2003)). According to this interpretation, professionals are generally more overconfident.

The second part of Table 3 shows the median doc_i-value of the medians per subject. In contrast to Table 1 we calculate the median of doc_1 and doc_2 over the three dates $t = 5$, $t = 10$, and $t = 20$ for ease of exposition. Again, all doc_i-values are higher for traders although a Kruskal-Wallis test suggests that we cannot reject the null hypothesis of equality of populations for doc_3 and doc_4.

Another question is who is the real expert in this task? One group of subjects has more practical expertise (the professionals) whereas the other group is presumably better trained in statistics (the student comparison group). Given that our tasks heavily build on statistical proficiency one has to be cautious to regard the professional traders as the real experts in our experiments. The same point also applies to the field experiments by (Haigh and List 2005) and (Alevy, Haigh, and List 2007) that also document differences between professionals and a student control group.

The main message of this subsection can be summarized as follows. Traders always have a higher degree of overconfidence. As in the 'trend prediction by probability estimates' task underconfidence is observed, with professionals performing better in this task. But traders are not better in general. The 'trend prediction by confidence intervals' task shows that professionals are again more overconfident.

5.2 Correlation of Overconfidence and Performance Measures

Table 4 presents Spearman rank correlations as well as significance levels of correlations of doc_i-measures, $i \in \{1, 2, 3, 4\}$. doc_1 and doc_2 are the medians over the three dates $t = 5, t = 10$, and $t = 20$ per subject. doc_3 and doc_4 are the medians per subject. Table 4 shows that all overconfidence measures are positively correlated. These correlations are significant at the 1% level except for $corr(doc_1, doc_4)$ with a p-value of 0.0217. When we calculate Kendall's Tau, the results are similar. Accordingly, we find interpersonal differences. A higher degree of overconfidence in the 'trend prediction by probability estimates' task implies a higher degree of overconfidence in the 'trend prediction by confidence intervals' task and vice versa. This result is remarkable as these are completely different tasks.

Similar results are obtained by (Budescu and Du 2007) who find that direct probability estimates and quantile estimates are internally consistent across subjects.

Table 5 shows an according analysis for the subjects' performance instead of their overconfidence. The variables $perf_i$ are calculated as absolute values of doc_i, $i \in \{1, 2, 3, 4\}$, and are aggregated to median values as described above. The lower the $perf_i$ values, the closer the subjects approached the rational benchmark, and the better they performed in the task. In contrast to the overconfidence measures, performance measures in the different task types are hardly correlated (i.e. correlations between $perf_1$, $perf_2$ and $perf_3$, $perf_4$). As shown in Table 5, the lowest p-value is 0.0442 for the correlation between $perf_2$ and $perf_3$. Combining the results of Table 4 and Table 5, we can conclude that overconfidence rather than performance is a robust individual characteristic: Individuals are not generally better or worse in such judgment tasks, but generally more or less overconfident.

6 Discussion and Conclusion

Our main findings can be summarized as follows.

1. Depending on the type of task *either* underconfidence *or* overconfidence can be observed in the same trend prediction setting based on the same information.

2. The degrees of overconfidence derived from different trend prediction tasks are significantly positively correlated whereas performance measures are not.

3. In trend prediction tasks professional traders are generally more overconfident than students.

Our basic results that hold across both subject pools are remarkably similar to those presented by (Budescu and Du 2007). In their two studies they find slight underconfidence and slight overconfidence in direct probability estimates and strong overconfidence when subjects state 90% confidence intervals. Also related to our findings is the result that subjects are internally consistent across the two response modes.

These findings demonstrate that a theorist has to be careful when deriving assumptions that are based on psychological evidence. This is in line with Hirshleifer's argument that "it is often not obvious how to translate preexisting evidence from psychological experiments into assumptions about investors in real financial settings. Routine experimental testing of the assumptions and conclusions of asset-pricing theories is needed to guide modeling."[32] Our finding is particularly interesting as we consider two types of trend prediction that have obvious practical relevance and are both near at hand to facets of the general phenomenon that people are biased when detecting trends. In addition, they are both realistic forms of trend prediction tasks that are separately modeled in theoretical analyses of financial markets. In a similar vein, (Budescu and Du 2007) argue that recent experimental results challenged the generality of the overconfidence claim and questioned whether financial models should assume overconfidence when predicting investors' behaviors, or whether they should be context-specific (i.e., incorporate either over - or underconfidence depending on the response mode) to improve their predictive validity. If a model is built in which investors observe asset prices and draw conclusions about future price developments, how should possible biases in trend prediction be incorporated? Should a theorist rely on the result of our first experimental task, which indicates that subjects underestimate the mathematically correct probability indicating underconfidence?

[32] (Hirshleifer 2001), p. 1577.

Or should a theorist model overconfidence in this setting by assuming confidence intervals that are too narrow or, in other words, underestimation of the variance? It is important to stress that the answer to this question does not depend on a specific informational setting (in our experiment both trend prediction tasks were embedded in the *same* informational setting) but solely on the specific task considered.

The second result is of interest as the documented positive correlation of the degrees of overconfidence suggest robust individual differences in reasoning (see also (Parker and Fischhoff 2005), (Schunk and Betsch 2006), (Stanovich and West 1998), and (Stanovich and West 2000)). In this context, it should be emphasized that trend prediction by probability estimates and trend prediction by confidence intervals are completely different tasks and not simply a different way of measuring the same aspects.[33] Moreover, this result is of interest as it shows that individuals are not generally better or worse in different judgment tasks, but just more or less overconfident. Overconfidence rather than performance seems to be the robust individual characteristic.

The third result is important because it shows that professional experience does not necessarily eliminate biases in job-related judgment tasks. This contradicts the argument that overconfidence might not play a significant role in financial markets since professionals that dominate the markets are less susceptible to such biases than lay people.

[33] This is best illustrated by pointing out that in the trend probability estimation task the mathematically correct answer depends neither on the path length nor directly on the variance of the price processes, but only on the odds k, measuring the relative strengths of both trends. For the well calibrated confidence intervals, however, the variance of the processes and the prediction time horizons are most relevant.

References

Agnew, J. R., and L. R. Szykman, 2005, "Asset Allocation and Information Overload: The Influence of Information Display, Asset Choice, and Investor Experience," *Journal of Behavioral Finance*, 6(2), 57–70.

Alevy, J. E., M. S. Haigh, and J. List, 2007, "Information Cascades: Evidence from a Field Experiment with Financial Market Professionals," *Journal of Finance*, 62(1), 151–180.

Anderson, L. R., and C. A. Holt, 1997, "Information Cascades in the Laboratory," *American Economic Review*, 87(5), 847–862.

Andersson, P., J. Edman, and M. Ekman, 2005, "Predicting the World Cup 2002 in soccer: Performance and confidence of experts and non-experts," *International Journal of Forecasting*, 21(3), 565–576.

Barberis, N., A. Shleifer, and R. Vishny, 1998, "A model of investor sentiment," *Journal of Financial Economics*, 49(3), 307–343.

Benos, A. V., 1998, "Aggressiveness and survival of overconfident traders," *Journal of Financial Markets*, 1(3-4), 353–383.

Bernard, V. L., and J. K. Thomas, 1989, "Post-Earnings-Announcement Drift: Delayed Price Response or Risk Premium," *Journal of Accounting Research*, 27, 1–36, Supplement 1989.

Bloomfield, R., and J. Hales, 2002, "Predicting the next step of a random walk: experimental evidence of regime-shifting beliefs," *Journal of Financial Economics*, 65(3), 397–414.

Budescu, D. V., and N. Du, 2007, "The Coherence and Consistency of Investors Probability Judgments," *Management Science*, pp. –, forthcoming.

Caballé, J., and J. Sákovics, 2003, "Speculating against an overconfident market," *Journal of Financial Markets*, 6(2), 199–225.

Cesarini, D., Ö. Sandewall, and M. Johannesson, 2006, "Confidence interval estimation tasks and the economics of overconfidence," *Journal of Economic Behavior and Organization*, 61(3), 453–470.

Daniel, K., D. Hirshleifer, and A. Subrahmanyam, 1998, "Investor Psychology and Security Market Under- and Overreactions," *Journal of Finance*, 53(6), 1839–1885.

——— , 2001, "Overconfidence, Arbitrage, and Equilibrium Asset Pricing," *Journal of Finance*, 56(3), 921–965.

Daniel, K., D. Hirshleifer, and S. H. Teoh, 2002, "Investor psychology in capital markets: evidence and policy implications," *Journal of Monetary Economics*, 49(1), 139–209.

Deaves, R., E. Lüders, and M. Schröder, 2007, "The Dynamics of Overconfidence: Evidence from Stock Market Forecasters," Working paper, Center for European Economic Research (ZEW), Mannheim.

Dhar, R., and N. Zhu, 2006, "Up Close and Personal: Investor Sophistication and the Disposition Effect," *Management Science*, 52(5), 726–740.

Diamond, D. W., and R. E. Verrecchia, 1981, "Information aggregation in a noisy rational expectations economy," *Journal of Financial Economics*, 9(3), 221–235.

Edwards, W., 1968, "Conservatism in Human Information Processing," in *Formal Representation of Human Judgment*, ed. by B. Kleinmuntz. Wiley, New York, pp. 17–52.

Elliott, W. B., F. D. Hodge, and K. E. Jackson, 2007, "The Association Between Nonprofessional Investors' Information Choices and Their Portfolio Returns: The Importance of Investing Experience," *Contemporary Accounting Research*, pp. –, forthcoming.

Erev, I., T. S. Wallsten, and D. V. Budescu, 1994, "Simultaneous Over- and Underconfidence: The Role of Error in Judgment Processes," *Psychological Review*, 101(3), 519–527.

Fama, E. F., 1965, "The Behavior of Stock-Market Prices," *Journal of Business*, 38(1), 34–105.

Feng, L., and M. S. Seasholes, 2005, "Do Investor Sophistication and Trading Experience Eliminate Behavioral Biases in Financial Markets?," *Review of Finance*, 9(3), 305–351.

Gervais, S., and T. Odean, 2001, "Learning to Be Overconfident," *Review of Financial Studies*, 14(1), 1–27.

Gigerenzer, G., U. Hoffrage, and H. Kleinbölting, 1991, "Probabilistic Mental Models: A Brunswikian theory of confidence," *Psychological Review*, 98(4), 506–528.

Glaser, M., T. Langer, J. Reynders, and M. Weber, 2007a, "Framing Effects in Stock Market Forecasts: The Difference Between Asking for Prices and Asking for Returns," *Review of Finance*, 11(2), 325–357.

——— , 2007b, "Scale Dependence of Overconfidence in Stock Market Volatility Forecasts," Working paper, University of Mannheim.

Glaser, M., T. Langer, and M. Weber, 2005, "Overconfidence of Professionals and Lay Men: Individual Differences Within and Between Tasks?," Working paper, University of Mannheim.

Glaser, M., M. Nöth, and M. Weber, 2004, "Behavioral Finance," in *Blackwell Handbook of Judgment and Decision Making*, ed. by D. J. Koehler, and N. Harvey. Blackwell, Malden, Mass., pp. 527–546.

Glaser, M., and M. Weber, 2003, "Momentum and Turnover: Evidence from the German Stock Market," *Schmalenbach Business Review*, 55(4), 108–135.

——— , 2005, "September 11 and Stock Return Expectations of Individual Investors," *Review of Finance*, 9(2), 243–279.

——— , 2007, "Overconfidence and Trading Volume," *Geneva Risk and Insurance Review*, 32(1), 1–36.

Gneezy, U., and J. Potters, 1997, "An Experiment on Risk Taking and Evaluation Periods," *Quarterly Journal of Economics*, 112(2), 631–645.

Gompers, P. A., and A. Metrick, 2001, "Institutional Investors and Equity Prices," *Quarterly Journal of Economics*, 116(1), 229–259.

Graham, J. R., and C. R. Harvey, 2003, "Expectations of equity risk premia, volatility and asymmetry," Working paper, Fuqua School of Business, Duke University.

Griffin, D., and A. Tversky, 1992, "The Weighing of Evidence and the Determinants of Confidence," *Cognitive Psychology*, 24(3), 411–435.

Grossman, S. J., and J. E. Stiglitz, 1980, "On the impossibility of informationally efficient markets," *American Economic Review*, 70(3), 393–408.

Haigh, M. S., and J. A. List, 2005, "Do Professional Traders Exhibit Myopic Loss Aversion? An Experimental Analysis," *Journal of Finance*, 60(1), 523–534.

Harrison, G. W., and J. A. List, 2004, "Field Experiments," *Journal of Economic Literature*, 42(4), 1009–1055.

Hellwig, M. F., 1980, "On the aggregation of information in competitive markets," *Journal of Economic Theory*, 22(3), 477–498.

Hilton, D. J., 2001, "The Psychology of Financial Decision-Making: Applications to Trading, Dealing, and Investment Analysis," *Journal of Psychology and Financial Markets*, 2(1), 37–53.

Hirshleifer, D., 2001, "Investor Psychology and Asset Pricing," *Journal of Finance*, 56(4), 1533–1597.

Hirshleifer, D., and G. Y. Luo, 2001, "On the survival of overconfident traders in a competitive securities market," *Journal of Financial Markets*, 4(1), 73–84.

Hong, H., J. C. Stein, and J. Yu, 2007, "Simple Forecasts and Paradigm Shifts," *Journal of Finance*, 62(3), 1207–1242.

Jegadeesh, N., and S. Titman, 1993, "Returns to Buying Winners and Selling Losers: Implications for Stock Market Efficiency," *Journal of Finance*, 1(48), 65–91.

——— , 2001, "Profitability of Momentum Strategies: An Evaluation of Alternative Explanations," *Journal of Finance*, 2(56), 699–720.

Jones, C. P., and R. H. Litzenberger, 1970, "Quarterly Earnings Reports and Intermediate Stock Price Trends," *Journal of Finance*, 25(1), 143–148.

Jonsson, A.-C., and C. M. Allwood, 2003, "Stability and variability in the realism of confidence judgments over time, content domain, and gender," *Personality and Individual Differences*, 34(4), 559–574.

Jorgensen, M., K. H. Teigen, and K. Molokken, 2004, “Better sure than safe? Overconfidence in judgement based software development effort prediction intervals,” *Journal of Systems and Software*, 70(1-2), 79–93.

Juslin, P., 1994, “The Overconfidence Phenomenon as a Consequence of Informal Experimenter-Guided Selection of Almanac Items,” *Organizational Behavior and Human Decision Processes*, 57(2), 226–246.

Juslin, P., P. Wennerholm, and H. Olsson, 1999, “Format Dependence in Subjective Probability Calibration,” *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 25(4), 1038–1052.

Juslin, P., A. Winman, and P. Hansson, 2007, “The Naive Intuitive Statistician: A Nave Sampling Model of Intuitive Confidence Intervals,” *Psychological Review*, 114(3), 678–703.

Juslin, P., A. Winman, and H. Olsson, 2000, “Naive Empiricism and Dogmatism in Confidence Research: A Critical Examination of the Hard-Easy Effect,” *Psychological Review*, 107(2), 384–396.

Keren, G., 1991, “Calibration and probability judgements: Conceptual and methodological issues,” *Acta Psychologica*, 77(3), 217–273.

Klayman, J., J. B. Soll, C. Gonzáles-Vallejo, and S. Barlas, 1999, “Overconfidence: It Depends on How, What, and Whom You Ask,” *Organizational Behavior and Human Decision Processes*, 79(3), 216–247.

Koehler, D. J., L. Brenner, and D. Griffin, 2002, “The Calibration of Expert Judgment: Heuristics and Biases Beyond the Laboratory,” in *Heuristics and Biases: The Psychol-*

ogy of Intuitive Judgment, ed. by T. Gilovich, D. Griffin, and D. Kahneman. Cambridge University Press, Cambridge, pp. 686–715.

Kyle, A. S., 1985, “Continuous Auctions and Insider Trading,” *Econometrica*, 53(6), 1315–1336.

——— , 1989, “Informed Speculation with Imperfect Competition,” *Review of Economic Studies*, 56(3), 317–356.

Kyle, A. S., and F. A. Wang, 1997, “Speculation Duopoly with Agreement to Disagree: Can Overconfidence Survive the Market Test?,” *Journal of Finance*, 52(5), 2073–2090.

Lakonishok, J., A. Shleifer, and R. W. Vishny, 1992, “The impact of institutional trading on stock prices,” *Journal of Financial Economics*, 32(1), 23–43.

Lawrence, M., P. Goodwin, M. O’Connor, and D. Önkal, 2006, “Judgmental forecasting: A review of progress over the last 25 years,” *International Journal of Forecasting*, 22(3), 493–518.

Lichtenstein, S., B. Fischhoff, and L. D. Phillips, 1982, “Calibration of probabilities: The state of the art to 1980,” in *Judgment under uncertainty: Heuristics and Biases*, ed. by D. Kahneman, P. Slovic, and A. Tversky. Cambridge University Press, Cambridge, pp. 306–334.

Locke, P. R., and S. C. Mann, 2005, “Professional trader discipline and trade disposition,” *Journal of Financial Economics*, 76(2), 401–444.

Massey, C., and G. Wu, 2005, “Detecting Regime Shifts: The causes of under- and overreaction,” *Management Science*, 51(6), 932–947.

Murphy, A. H., and B. G. Brown, 1984, "A Comparative Evaluation of Objective and Subjective Weather Forecasts in the United States," *Journal of Forecasting*, 3(4), 369–393.

Murphy, A. H., and R. L. Winkler, 1984, "Probability Forecasting in Meterology," *Journal of the American Statistical Association*, 79(387), 489–500.

Nofsinger, J. R., and R. W. Sias, 1999, "Herding and Feedback Trading by Institutional and Individual Investors," *Journal of Finance*, 54(6), 2263–2295.

Odean, T., 1998, "Volume, Volatility, Price, and Profit When All Traders Are Above Average," *Journal of Finance*, 53(6), 1887–1934.

Parker, A. M., and B. Fischhoff, 2005, "Decision-making Competence: External Validation through an individual-difference Approach," *Journal of Behavioral Decision Making*, 18(1), 1–27.

Poon, S.-H., and C. Granger, 2004, "Practical Issues in Forecasting Volatility," *Financial Analysts Journal*, 61(1), 45–56.

Rapoport, A., and T. S. Wallsten, 1972, "Individual decision behavior," *Annual Review of Psychology*, 23, 131–176.

Rouwenhorst, K. G., 1998, "International Momentum Strategies," *Journal of Finance*, 53(1), 267–284.

——— , 1999, "Local Return Factors and Turnover in Emerging Markets," *Journal of Finance*, 54(4), 1439–1464.

Russo, J. E., and P. J. H. Schoemaker, 1992, "Managing Overconfidence," *Sloan Management Review*, 33(2), 7–17.

Sarin, R. K., and M. Weber, 1993, "Effects of Ambiguity in Market Experiments," *Management Science*, 39(5), 602–615.

Schunk, D., and C. Betsch, 2006, "Explaining heterogeneity in utility functions by individual differences in decision modes," *Journal of Economic Psychology*, 27(3), 386–401.

Shapira, Z., and I. Venezia, 2001, "Patterns of behavior of professionally managed and independent investors," *Journal of Banking and Finance*, 25(8), 1573–1587.

Shiller, R. J., 2001, "Bubbles, Human Judgment, and Expert Opinion," Cowles Foundation Discussion Paper, Yale University.

Shleifer, A., and L. H. Summers, 1990, "The Noise Trader Approach to Finance," *Journal of Economic Perspectives*, 4(2), 19–33.

Slovic, P., and S. Lichtenstein, 1971, "Comparison of Bayesian and regression approaches to the study of information processing in judgment," *Organizational Behavior and Human Performance*, 6(6), 649–743.

Soll, J. B., and J. Klayman, 2004, "Overconfidence in Interval Estimates," *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 30(2), 299–314.

Stanovich, K. E., and R. F. West, 1998, "Individual differences in Rational Thought," *Journal of Experimental Psychology: General*, 127(2), 161–188.

——— , 2000, "Individual differences in reasoning: Implications for the rationality debate," *Behavioral and Brain Sciences*, 23(5), 645–726.

Wang, F. A., 1998, "Strategic trading, asymmetric information and heterogeneous prior beliefs," *Journal of Financial Markets*, 1(3-4), 321–352.

——— , 2001, "Overconfidence, Investor Sentiment, and Evolution," *Journal of Financial Intermediation*, 10(2), 138–170.

Winkler, R. L., and A. H. Murphy, 1973, "Experiments in the laboratory and the real world," *Organizational Behavior and Human Performance*, 10(2), 252–270.

Table 1: **Trend Prediction by Probability Estimates**

This table presents the number of probability estimates above (overconfident), equal to (well calibrated), and below the correct probability for the three time periods $t = 5$, $t = 10$, and $t = 20$. doc_1 is the difference between the subjective and the mathematically correct probability. doc_2 is the difference between the implied price change and the observed price change. Median values of these measures are also presented.

	t=5	t=10	t=20
Number of observations (all observations)	528	515	524
Overconfident ($doc_1 > 0$)	175	140	106
	(33.14%)	(27.18%)	(20.23%)
Well calibrated ($doc_1 = 0$)	60	25	17
	(11.36%)	(4.85%)	(3.24%)
Underconfident ($doc_1 < 0$)	293	350	401
	(55.50%)	(67.96%)	(76.53%)
doc_1 (medians; all distributions)	-4.12	-5.26	-8.75
doc_1 (medians; $k = 1.5$)	-3.00	-2.14	-5.23
doc_1 (medians; $k = 1.75$)	-3.64	-5.51	-9.35
doc_1 (medians; $k = 2$)	-5.67	-7.64	-9.81
doc_2 (medians; all distributions)	-0.42	-1.00	-2.43
doc_2 (medians; $k = 1.5$)	-0.30	-0.47	-1.43
doc_2 (medians; $k = 1.75$)	-0.28	-1.07	-3.00
doc_2 (medians; $k = 2$)	-0.60	-1.00	-3.00
Number of observations (medians per subject)	93	93	93
Overconfident ($doc_1 > 0$)	28	19	13
	(30.11%)	(20.43%)	(13.98%)
Well calibrated ($doc_1 = 0$)	7	3	0
	(7.53%)	(3.23%)	(0.00%)
Underconfident ($doc_1 < 0$)	58	71	80
	(62.37%)	(76.34%)	(86.02%)
doc_1 (medians of medians per subject; all distributions)	-5.45	-5.03	-7.46
Wilcoxon signed-rank test ($H_0 : \quad doc_1 = 0$)	$p = 0.0011$	$p = 0.0000$	$p = 0.0000$
doc_1 (medians of medians per subject; $k = 1.5$)	-2.57	-2.54	-5.11
doc_1 (medians of medians per subject; $k = 1.75$)	-3.64	-6.60	-9.93
doc_1 (medians of medians per subject; $k = 2$)	-8.06	-9.17	-9.90
doc_2 (medians of medians per subject; all distributions)	-0.55	-0.81	-2.39
Wilcoxon signed-rank test ($H_0 : \quad doc_2 = 0$)	$p = 0.0183$	$p = 0.0000$	$p = 0.0000$
doc_2 (medians of medians per subject; $k = 1.5$)	-0.40	-0.49	-1.40
doc_2 (medians of medians per subject; $k = 1.75$)	-0.46	-1.07	-2.95
doc_2 (medians of medians per subject; $k = 2$)	-0.60	-1.25	-3.48

Table 2: **Trend Prediction by Confidence Intervals**

This table presents the number of confidence intervals classified as overconfident or underconfident by the measures doc_3 and doc_4. doc_3 is defined as: $doc_3 = 90\% - (p_{high} - p_{low})$ with positive doc_3-values corresponding to overconfidence, negative doc_3-values to underconfidence. In the computation of doc_4 we replace the correct trend probability derived from the price level at $t = 20$ by the trend probability stated by the individual in $t = 20$. Median values of these measures are also presented.

All observations	doc_3	doc_4
Number of observations	552	524
$doc_i > 0, \quad i \in \{3,4\}$ (overconfident)	335	341
	(60.69%)	(65.08%)
$doc_i < 0, \quad i \in \{3,4\}$ (underconfident)	217	183
	(39.31%)	(34.92%)
$doc_i, \quad i \in \{3,4\}$ (median of all observations)	10.03	9.11

Medians per subjects	doc_3	doc_4
Number of observations	93	93
$doc_i > 0, \quad i \in \{3,4\}$ (overconfident)	61	67
	(65.59%)	(72.04%)
$doc_i < 0, \quad i \in \{3,4\}$ (underconfident)	32	26
	(34.41%)	(27.96%)
$doc_i, \quad i \in \{3,4\}$ (median of median per subject)	8.01	13.22
Wilcoxon signed-rank test ($H_0: \quad doc_i = 0, \quad i \in \{3,4\}$)	$p = 0.0000$	$p = 0.0000$

Table 3: **Trader versus Lay People**

This table presents the various overconfidence measures defined in Subsection 4.1 and Subsection 4.2 separately for the two subject groups: professional traders and lay people (students). The first part of Table 3 shows the medians of all observations. Accordingly, the second column shows figures that are also part of Table 1 and Table 2. The next two columns present the median doc_1-value in $t = 5$ as well as all the other doc_i-values, $i \in \{1, 2, 3, 4\}$, for the two groups of subjects separately. The last column of Table 3 presents the p-value of a Kruskal-Wallis test. Null hypothesis is equality of populations. The second part of Table 3 shows the median doc_i-value of the medians per subject. In contrast to Table 1 we calculate the median of doc_1 and doc_2 over the three dates $t = 5$, $t = 10$, and $t = 20$ for ease of exposition.

	All subjects	Professionals	Students	p-value of Kruskal-Wallis test H_0 : Equality of populations
doc_1 (medians; all observations; $t = 5$)	-4.12	0.00	-5.00	0.0094
doc_1 (medians; all observations; $t = 10$)	-5.26	-1.93	-6.97	0.0001
doc_1 (medians; all observations; $t = 20$)	-8.75	-8.00	-9.22	0.0314
doc_2 (medians; all observations; $t = 5$)	-0.42	0.00	-0.58	0.0023
doc_2 (medians; all observations; $t = 10$)	-1.00	-.042	-1.05	0.0001
doc_2 (medians; all observations; $t = 20$)	-2.43	-2.04	-2.49	0.0772
doc_3 (medians; all observations)	10.03	16.68	5.88	0.0313
doc_4 (medians; all observations)	9.11	13.35	8.07	0.1616
doc_1 (medians of medians)	-5.00	-1.46	-6.86	0.0089
doc_2 (medians of medians)	-0.92	-0.36	-1.09	0.0062
doc_3 (medians of medians)	8.01	17.42	4.79	0.2277
doc_4 (medians of medians)	13.22	15.31	8.51	0.4388

Table 4: **Correlation of Overconfidence Measures**

This table presents Spearman rank correlations as well as significance levels of correlations of doc_i-measures, $i \in \{1, 2, 3, 4\}$. doc_1 and doc_2 are the medians over the three dates $t = 5$, $t = 10$, and $t = 20$ per subject. doc_3 and doc_4 are the medians per subject.

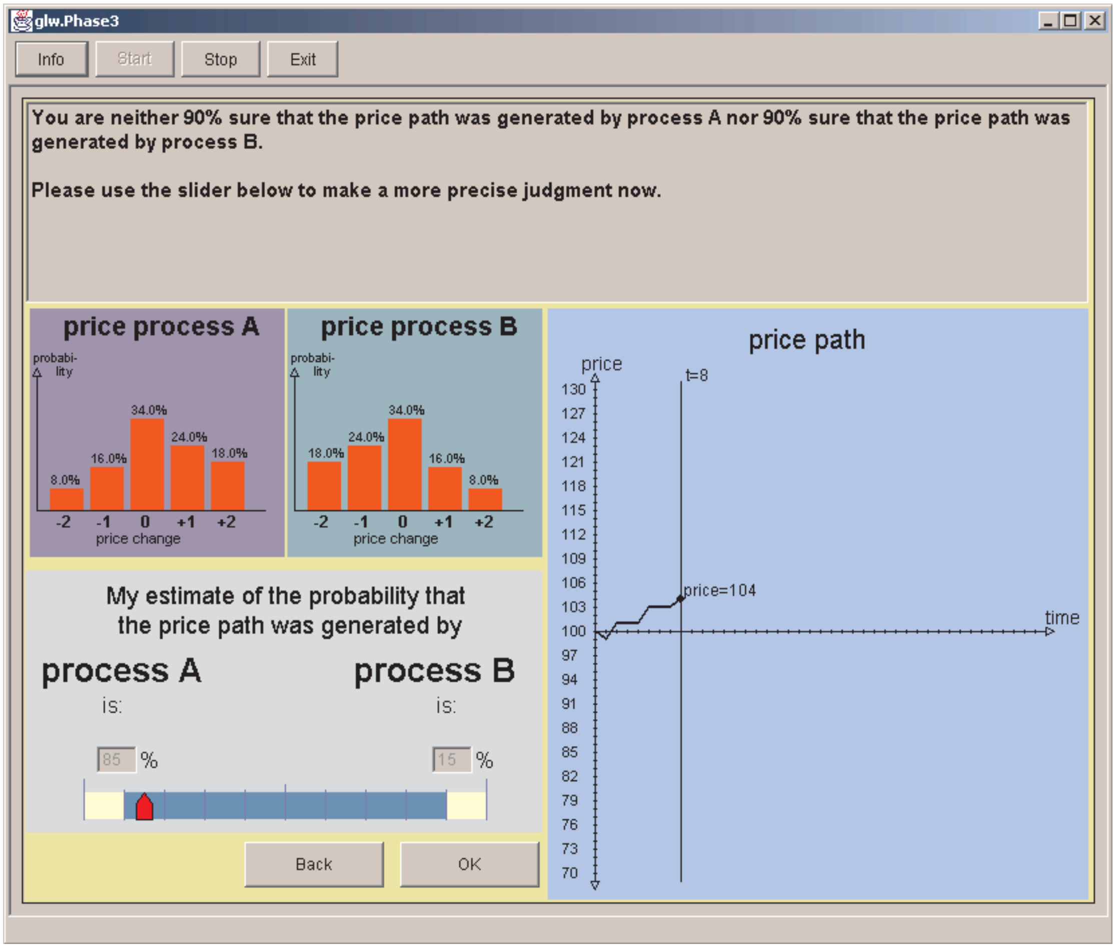
	doc_1	doc_2	doc_3	doc_4
doc_1	1			
doc_2	0.90 (< 0.0001)	1		
doc_3	0.32 (0.0021)	0.40 (0.0001)	1	
doc_4	0.24 (0.0217)	0.32 (0.0016)	0.98 (< 0.0001)	1

Table 5: **Correlation of Performance Measures**

This table presents Spearman rank correlations as well as significance levels of correlations of $perf_i$-measures, $i \in \{1, 2, 3, 4\}$. $perf_1$ and $perf_2$ are the medians over the three dates $t = 5$, $t = 10$, and $t = 20$ per subject. $perf_3$ and $perf_4$ are the medians per subject.

	$perf_1$	$perf_2$	$perf_3$	$perf_4$
$perf_1$	1			
$perf_2$	0.69 (< 0.0001)	1		
$perf_3$	0.15 (0.1518)	0.21 (0.0442)	1	
$perf_4$	0.16 (0.1167)	0.11 (0.2741)	0.93 (< 0.0001)	1

Figure 1: **Screen Shot**



You are neither 90% sure that the price path was generated by process A nor 90% sure that the price path was generated by process B.

Please use the slider below to make a more precise judgment now.

My estimate of the probability that the price path was generated by

process A is: 85 %

process B is: 15 %