True Overconfidence: The Inability of Rational Information Processing to Account for Apparent Overconfidence

Christoph Merkle, Martin Weber*

This version: April 2011 - First version: June 2009

Abstract

The better-than-average effect describes the tendency of people to perceive their skills and virtues as being above average. We derive a new experimental paradigm to distinguish between two possible explanations for the effect, namely rational information processing and overconfidence. Experiment participants evaluate their relative position within the population by stating their complete belief distribution. This approach sidesteps recent methodology concerns associated with previous research. We find that people hold beliefs about their abilities in different domains and tasks which are inconsistent with rational information processing. Both on an aggregated and an individual level, they show considerable overplacement. We conclude that overconfidence is not only apparent overconfidence but rather the consequence of a psychological bias.

JEL-Classification Codes: C90, G10

Keywords: Overconfidence, Better-than-average effect, Overplacement, Bayesian updating, Belief distribution

*Merkle: Department of Banking and Finance, University of Mannheim, L5, 2, 68131 Mannheim, Germany, chmerkle@mail.uni-mannheim.de. Weber: Department of Banking and Finance, University of Mannheim, L 5, 2, 68131 Mannheim, Germany and CEPR, London, UK. We thank Juan Dubra, Markus Glaser, Robin Hogarth, Don Moore, Daniel Smith, participants of the IAREP/SABE 2009 joint conference and the SPUDM22 conference, and three anonymous referees for their comments. All errors are our own.

Introduction

Overconfidence is not just an artifact of psychological experiments but seems present in many real life situations where considerable stakes are involved. Overconfident decision making has been observed in financial markets (Odean, 1998), corporations (Malmendier and Tate, 2005), with business entries (Cooper, Woo, and Dunkelberg, 1988) or even marriages (Mahar, 2003). Indeed, overconfidence is perhaps the behavioral bias most readily embraced by academic researchers in economics and finance. In particular the better-than-average effect, which is the tendency of people to rate their skills and virtues favorably relative to a comparison group, yields direct predictions for economic decision making.

In a recent paper, Benoît and Dubra (2009) challenge the notion of overconfidence as it was previously analyzed in psychology and economics. The subject of their criticism is the conventionally used research methodology to demonstrate the better-than-average effect. In a signaling framework, Benoît and Dubra show that rational information processing can lead to the very results formerly interpreted as evidence for overconfidence. This does not rule out true overconfidence as an explanation for these findings, but instead also allows for straightforward Bayesian updating as an alternative explanation.[1]

Despite this setback for the overconfidence literature, it is not sufficient to take a methodological viewpoint on the matter; we have to ask ourselves about the psychological reality of this bias and its relation to other self-serving biases. The assertion that people are overconfident is an appealing explanation for behavior, both on the financial markets and elsewhere. In contrast, rational updating is demanding in terms of people's information processing capacity and the underlying signal structure necessary to produce apparent overconfidence. It therefore seems worthwhile to design a research strategy that would be able to demonstrate

[1] In accordance with Benoît and Dubra, we use the term "true overconfidence" for truly biased self-evaluations. In contrast "apparent overconfidence" stands for data that seems to reflect overconfidence, but where it is not possible to prove the presence of a better-than-average effect. The term "apparent overconfidence" thus includes cases of true overconfidence and other possible causes such as rational information processing.

the presence of true overconfidence by improving previous research methodology in such a way that it becomes capable of withstanding the critisism of Benoît and Dubra (2009).

We identify the aggregation of beliefs as the feature most damaging to the interpretational value of the traditional experimental setting. The simplest setup asks people to judge whether they believe themselves to be above average in a certain domain, as for example in the famous account on driving ability by Svenson (1981). More advanced designs ask participants to specify the percentile of a distribution they believe themselves to belong to (e.g. Dunning, Meyerowitz, and Holzberg, 1989). Both approaches have the common feature that as they only retrieve a single estimate, a lot of information gets lost, thus leaving room for alternative explanations. Many sets of beliefs can produce the same result when aggregated in this manner; Bayesian posteriors and true overconfidence are just two of them.

We design two experiments to elicit more detailed beliefs of participants concerning a number of domains that have previously been associated with overconfidence, overoptimism, or underconfidence. Self-evaluations are given along a quantile scale that describes the ability distribution relative to a peer group. Along this scale, participants provide estimates representing their subjective probability of themselves falling into each skill quantile. The extended assessment allows us to directly test whether the findings are in line with rational information processing.

Our central result is that considerable overconfidence is present in the belief distributions of experiment participants. We test various conditions for population averages of these probability distributions and find them incompatible with rational information processing. Bayesian updating can be rejected as an explanation for apparent overconfidence at conventional significance levels. Most people find it highly probable that they rank among the higher quantiles of the ability distribution and not at all likely that they are below average. On an individual level, they often fall short of their expectations, and especially the unskilled exhibit pronounced overconfidence. We conclude that true overconfidence is the main driver of our results.

Types of overconfidence

While this is not the place to review an abundant overconfidence research in psychology and economics (consider e.g. Glaser and Weber, 2010), it is nevertheless useful to divide the field into three subareas which can be summarized following Moore and Healy (2008):

1. Judgments of one's absolute performance or ability (overestimation)
2. Confidence in the precision of one's estimates (miscalibration or overprecision)
3. Appraisal of one's relative skills and virtues (better-than-average effect or overplacement)

Overestimation is diagnosed if people's absolute evaluation of their own performance (e.g. correct answers in a knowledge test) exceeds their actual performance (Lichtenstein, Fischhoff, and Phillips, 1982; Moore and Healy, 2008). Miscalibration or overprecision denotes the observation that people choose overly narrow confidence intervals when asked for a range that is supposed to contain a true value with a certain probability (Alpert and Raiffa, 1982; Russo and Schoemaker, 1992). Overplacement often occurs when people try to evaluate their competence in a certain domain relative to others. Typically, most people rate themselves above average, which is why this effect is also called better-than-average effect (Alicke and Govorun, 2005). The relationship between these different forms of overconfidence is discussed for instance, in Glaser, Langer, and Weber (2009), Healy and Moore (2007), and Larrick, Burson, and Soll (2007).

Apart from the aforementioned, overoptimism (Weinstein, 1980) and illusion of control (Langer, 1975) are associated with overconfidence in a broad interpretation of the term. We will concentrate on the better-than-average effect (overplacement) and occasionally on overoptimism, as the elicitation techniques for these biases are similar.

Criticism which has been raised against all types of overconfidence is usually directed either at research methodology and experimental design or the underlying concept itself; the list of authors in psychology who have questioned the reality of over-

confidence or the research design includes Gigerenzer (1991), Gigerenzer, Hoffrage, and Kleinbölting (1991), Juslin (1994), Erev, Wallsten, and Budescu (1994), Dawes and Mulford (1996), and Klayman, Soll, González-Vallejo, and Barlas (1999). In economics—where the rationality assumption was long prevalent—the emphasis was a different one: in recent years, various approaches were pursued to reconcile overconfidence with rational behavior (Bénabou and Tirole, 2002; Brocas and Carrillo, 2002; Compte and Postlewaite, 2004; Healy and Moore, 2007; Köszegi, 2006; Santos-Pinto and Sobel, 2005; Van Den Steen, 2004; Zábojník, 2004). These models differ mainly in their assumptions, their relevance for different forms of overconfidence and the degree of rationality they are based on. In many ways, this literature has contributed to improving and clarifying methodology, but the debate whether overconfidence exists at all is far from being settled.

Criticism by Benoît and Dubra

The reasoning of Benoît and Dubra (2009) to some extent combines the two mentioned strands of criticism. They identify a problematic feature in the conventional procedure to demonstrate the better-than-average effect, namely relative imprecise inquiries for an appraisal of relative skills and virtues. Based on a parsimonious signaling model, they then employ rational Bayesian argumentation to illustrate that this kind of research cannot show overconfidence in the form of the better-than-average effect. We will now examine their reasoning in detail.

Probably the most prominent account of the better-than-average effect is given in Svenson (1981), who finds that a great majority of subjects rated themselves to be safer drivers than the median driver (77% of his Swedish and 87% of his US sample). He explains his findings by a general tendency of people to view themselves more favorably than they view others, possibly accompanied by cognitive effects such as low availability of negative memories. Similar results could be reproduced for other domains, for example for people

evaluating their personal virtues relative to others (Alicke, Klotz, Breitenbecher, Yurak, and Vredenburg, 1995).

These overplacement studies have a common research methodology, which often simply consists of asking participants whether they view themselves as better as or worse than the median or average of a comparison group with respect to some skill or virtue. Researchers occasionally require more precise estimates, i.e. other quantiles (often percentiles or deciles) are used instead of the median. Overconfidence is usually diagnosed if significantly more than half of the participants place themselves above the median, or more generally if more than x% place themselves above the (100-x)-percentile.

Some concerns regarding this design were raised earlier; for instance, people may interpret the skill in question differently or they may lack information about its distribution within the population. Additionally, the sample of participants might not be representative of the population, and the meaning of "average" can be understood in various ways. These problems can nevertheless be addressed by a more careful experimental design including precise and unambiguous formulation of questions and a fairly large and representative choice of subjects. Combined with the assumption that participants use best estimates of their own and others' abilities and skills, the general result remains valid—it seems intuitive that no more than a certain fraction of the population can rate themselves above a respective percentile.

However, Benoît and Dubra (2009) show that exactly this is possible even when people update their beliefs in a perfectly rational manner. In order to illustrate this, we shall briefly reproduce their example capitalizing on Svenson's study of driving ability here. In a uniformly distributed population of low, medium and highly skilled drivers, people are assumed to evaluate their driving skills depending on whether they have previously had an accident or not. Probabilities for causing an accident are given as $p_L = 0.8$, $p_M = 0.4$ and $p_H = 0$ for the different groups. If drivers do not know their initial skill level but interpret the occurrence of an accident as a signal, they will update their beliefs according to Bayes'

law. Given the prior probabilities, all people who did not experience an accident will arrive at a posterior probability of $\frac{5}{9}$ that they are of high skill; it seems reasonable for this group to rate themselves above average. As 60% of all drivers have had no accident, this implies that these 60% are expected to regard themselves as highly skilled. Beyond this concrete example, Benoît and Dubra (2009) show that—within the traditional experimental design—almost any distribution of respondents on the percentile scale can be explained by rational information processing.

There are some immediate concerns with the Benoît-Druba-criticism. One concern refers to the way people deal with signals in classic overconfidence domains. A very early study of Preston and Harris (1965) suggests that even drivers hospitalized after an accident exhibit the same overplacement patterns when asked for their driving performance. The authors find no evidence that participants adjust their evaluation according to the received signal. Benoît and Dubra (2009) discuss this study and some more recent evidence on how people interpret adverse signals. Although the results are quite mixed, a general self-serving bias in the perception of signals is well documented (Bradley, 1978; Zuckerman, 1979). People tend to ascribe bad outcomes to external forces rather than to their own performance or ability. For this reason, it is at the least unclear whether good and bad signals are perceived symmetrically within a Bayesian model.

The framework of Benoît and Dubra also imposes some requirements on signal structure. It is obvious that if the number of signals becomes large (or alternatively the quality of signals very good), overconfidence can no longer be explained rationally. With perfect signals and people allocating themselves reasonably to the percentiles, the distribution will correspond to underlying probabilities. If one extends the aforementioned driving setup by an additional period and maintains the same probabilities as before, this becomes clear: after observing two signals, only 47% would reasonably consider themselves as highly skilled and 27% each as medium or low skill drivers. After ten periods, the Bayesian result is practically indistinguishable from the real probabilities. If, for instance, ten signal observations are in-

terpreted as ten years of driving experience, there is no room for overconfidence afterwards. While it is possible to construct different examples that still show rational overconfidence after many periods, this comes at the cost of a highly asymmetric signal structure with the rare occurrence of (very) negative signals. Indeed, the asymmetric signal structure is one key ingredient to the emergence of rationally explicable overconfidence in Benoît and Dubra (2009), but this is a good portrayal of reality only for some domains (e.g. driving).

However, with respect to signal frequency and quality, feedback is far from perfect in many situations—even in financial markets where new information arrives almost continuously. It would therefore be premature to dismiss the Benoît-Dubra-criticism solely on these grounds. We instead design an experiment to distinguish between two possible sources of apparent overconfidence, namely rational information processing and true overconfidence.

Derivation of an alternative experimental design

The distinction of how people arrive at overconfident judgments is crucial, as Bayesian updaters are not biased in a special direction; whether they appear over- or underconfident is simply a result of prior probabilities and signal distribution. Any claim that human beings are persistently overconfident must be based on a non-rational formation of beliefs. Overconfidence is consequently mostly modeled as over- or underreaction relative (and thus distinct) to rational Bayesian updating (cp. e.g. Odean, 1998).

Classic experiments, however, are unable to distinguish between apparent and true overconfidence. To overcome this problem, Benoît and Dubra (2009) propose using a stronger requirement to test for overplacement. Based on their proof that maximally 2*x% can rate themselves rationally among the top x% of the population, they suggest using this hurdle for future experiments. This rule is unsuitable for the often used median condition and represents a very strong requirement to find overconfidence for other percentiles. For example (following this logic), more than 60% of the subjects must place themselves among the top

30% of a population before one can deduce a better-than-average effect. Although many studies observe overconfidence among their participants, it is rarely this pronounced. Even for the large levels of overconfidence observed by Svenson (1981) this rule allows to identify overconfidence only for some intervals of his US subject sample.

Furthermore, this rule applies only in the case that people indeed use the median of their own beliefs to arrive at their rating. Another perfectly prudent way of answering such a question is to take the average of one's beliefs; in that case almost any possible distribution of self-evaluations can be rationalized (for a proof, see again Benoît and Dubra, 2009).

The difficulty in showing true overconfidence within the traditional framework lies in the aggregation of information resulting from subjects placing themselves in one specific half, decile, quartile or other category. In the example mentioned, drivers that experienced no accident had a posterior probability of $\frac{5}{9}$ of being of high skill, $\frac{1}{3}$ of being of medium skill and $\frac{1}{9}$ of being of low skill. This distributional information gets lost if one observes only a point estimate. Figure 1 shows how a rich belief distribution containing information about probabilities for all deciles is represented by a single rating. First, subjects enjoy some discretion concerning how they determine this rating given their beliefs, i.e. how to summarize their beliefs in a single parameter. Second, and more importantly, the resultant ratings yield much less information to distinguish between true overconfidence and alternative explanations. This is why one can often only speak of "apparent overconfidence" in such situations.

— Please insert FIGURE 1 approximately here —

The experimental setting proposed here is to ask subjects directly for the probabilities with which they would place themselves in the different quantiles (e.g. deciles). This avoids the complication of different aggregation methods and preserves the additional information coming from people's distributional beliefs. The setup imposes clear restrictions on what is possible under rational Bayesian updating. Posterior probabilities calculated by Bayes' law,

weighted by their occurrence, must add up to the relative frequencies within the population. In a quantile framework, these real probabilities are simply defined by the chosen partition of the scale: for instance, for any decile, there are 10% of the population who belong to that decile. To determine whether the conditional beliefs for a state A are consistent with Bayesian updating, one has to check whether

$$\sum_i P(A|S_i) \times P(S_i) = P(A), \tag{1}$$

where the signals S_i form a disjoint partition of the universe.

To make this restriction clearer, we will again refer to the driving skill example: in the example, 60% of the population had no accident. These people share the beliefs mentioned above: $P(H|no\,accident) = \frac{5}{9}$, $P(M|no\,accident) = \frac{1}{3}$, and $P(L|no\,accident) = \frac{1}{9}$. Among the 40% who experienced an accident, posterior probabilities are 0 for being highly skilled, $P(M|accident) = \frac{1}{3}$ and $P(L|accident) = \frac{2}{3}$ for being of medium and low skill. To translate the example into equation (1), signal S_1 corresponds to "no accident" and signal S_2 to "accident". Together, these two signals describe all possible scenarios. If we plug in the different skill levels for A, we arrive at the following equations:

$$\begin{aligned}
P(H|no\,accident) \times P(no\,accident) + P(H|accident) \times P(accident) &\stackrel{!}{=} P(H) \\
P(M|no\,accident) \times P(no\,accident) + P(M|accident) \times P(accident) &\stackrel{!}{=} P(M) \\
P(L|no\,accident) \times P(no\,accident) + P(L|accident) \times P(accident) &\stackrel{!}{=} P(L)
\end{aligned}$$

We know from the given distribution of driving skill within the population that $P(H) = P(M) = P(L) = \frac{1}{3}$. This provides us with three conditions that have to hold when beliefs are updated rationally. As the posterior probabilities stated above were calculated by Bayes' law, the conditions are of course met.

Note that in the example, the signal and the probability of the signal for each group were known; this is not necessarily the case. In an experimental setting, $P(S_i)$ are unobservable

and signals may be much more complicated than the binary "accident" versus "no accident". We treat the S_i as elements of a set of possible signals S. One may think of these signals as idiosyncratic life-time experiences in a certain domain. We do not impose any further restrictions on these signals except for the standard assumption that signal realizations for experiment participants are randomly drawn from S.

We now define K ability quantiles Q_k to provide a common understanding of skill levels. The probability $P(Q_k)$ of falling into each quantile is evaluated conditional on the observed signal S_i and subjects in an experiment will thus report $P(Q_k|S_i)$. Inserting this into equation 1 we obtain:

$$\sum_i P(Q_k|S_i) * P(S_i) \stackrel{!}{=} P(Q_k) = \frac{1}{K} \tag{2}$$

The right-hand side of equation 2 is defined by the choice of the scale's partition. Conditional probabilities $P(Q_k|S_i)$ may differ from $1/K$ but—weighted by the probability of the signals in the population—they must equal $P(Q_k)$. As $P(S_i)$ corresponds to the fraction of subjects observing signal S_i, the average reported probability for a quantile must again equal 1/K. We arrive at K equations of this form as the condition needs to be satisfied for each quantile.

In an experimental setting, concrete signals and signal probabilities are usually unknown. However, under random sampling for the number of participants $n(S_i)$ observing each signal S_i it holds that $E[n(S_i)] = P(S_i) * n$; we can thus replace $P(S_i)$ by $E[n(S_i)/n]$. Moving the expectation operator outside the sum, we arrive at equation 3:

$$E\left[\sum_i P(Q_k|S_i) * \frac{n(S_i)}{n}\right] \stackrel{!}{=} \frac{1}{K} \tag{3}$$

Our final simplification is to assume that $n(S_i)$ equals one, which corresponds to the notion of idiosyncratic signals—we nevertheless allow for several subjects to observe the same signal or for elements of S not to be observed.

In the experiments we will mostly use a decile setup. Equation 3 then generates ten conditions of the form:

$$E\left[\frac{1}{n}\sum_{j=1}^{n} P(Q_k|S_j)\right] = 0.1 \qquad k = 1, ..., 10 \tag{4}$$

The left-hand side represents the average reported probability for each ability decile, which in expectation equals 0.1 for a population of perfect Bayesian updaters. It enables us to compare the realized average belief distribution in the experiments to the uniform benchmark distribution.

To test for true overconfidence, we will additionally rely on two conditions: first, the average reported probability mass for the upper half of the ability quantiles should not exceed 50%. This follows directly from equation 1 if A represents the state of "being above average". Likewise, in the decile setup of equation 4, the probability of the union of the top five deciles must in expectation equal 0.5 across participants. In contrast, true overconfidence would predict that

$$\frac{1}{n}\sum_{k=6}^{10}\sum_{j=1}^{n} P(Q_k|S_j) > 0.5 \tag{5}$$

Of course, similar relations also exist for the top 30%, top 20% and other fractions of the scale. Since the better-than-average effect takes its name from the notion of being above average, we will mainly concentrate on 5 but report other results occasionally.

As a second indicator of overconfidence, we consider the mean of the individual belief distribution means. This mean of means should correspond to the middle point of the ability scale in a population of rational updaters, which follows from the definition of the mean

for individual belief distributions $\sum_{k=1}^{K} P(Q_k|S_i) * k$ and the aggregation using the signal weights:

$$\sum_i P(S_i) * \sum_{k=1}^{K} P(Q_k|S_i) * k = \sum_{k=1}^{K} \sum_i P(Q_k|S_i) * P(S_i) * k = \sum_{k=1}^{K} P(Q_k) * k = \frac{K+1}{2} \quad (6)$$

The second equality uses equation 2 and shows that the mean of the belief distribution means must equal $(K+1)/2$. For a decile scale, the mean of individual belief distribution means should thus be at 5.5. True overconfidence predicts a mean of means > 5.5.

Experiment one

Method

Participants

Experiment one was conducted in 2008 at the University of Mannheim. 68 business students completed the paper-based questionnaire; 69% of the participants were male, the median age was 24. We excluded four participants who left blank substantial portions of the questionnaire.

Procedure

Subjects answered questions about their skills and abilities in several domains. We selected various domains of skills and abilities to reflect different levels of overconfidence. Subjects were asked for their performance as a student, their abilities in choosing investments, their ability to get along with other students, their programming skills, their sense of humor and their risk of suffering a heart attack before the age of 40.

The ability to get along with other people is a domain in which people are prone to high overconfidence (Moore and Cain, 2007). The same applies to sense of humor (Kruger and Dunning, 1999). In contrast, the performance as a student is regularly objectively reported

by a relatively efficient feedback mechanism (grades), thus less overconfidence is expected here. For investment abilities, we anticipate considerable overconfidence in line with the behavioral finance literature (e.g. Odean, 1998). Computer skills were previously related to underconfidence (Kruger, 1999), the risk of a heart attack is one of many incidents where overoptimism has been observed (Weinstein, 1980).

Appendix A contains part of the questionnaire used in the experiment, which explains to subjects the meaning of the quantile scale and how to fill out the input fields. They were asked to state their probabilities for the quantiles of the scale according to their belief distribution. We alternately used a quartile and a decile scale; the decile scale has the advantage of being more precise at the expense of being more demanding to complete.

It is crucial for our analysis of belief distributions that people understood the scale and answered the question for their probabilities of falling into each quantile reasonably. They were informed by the instructions that the probabilities had to add up to one (see Appendix A). For 96% of the entries, people obeyed this rule. In the remaining cases, the sum of probabilities was almost always close to one, suggesting mistakes in calculation and not in comprehension; we nevertheless exclude these cases.

Additionally, we asked subjects for a point estimate along the quantile scale following the traditional approach of demonstrating overconfidence; they thus made both judgments as illustrated in figure 1. This enables us to compare the two evaluations and—most likely—infer how subjects tend to aggregate their beliefs. We varied the order of point and probability judgments and for two domains we used a between-group design in which one group was asked only for probabilities while the other stated only a point estimate; this allows us to test for order effects and interdependencies between the two types of evaluations.

Results and discussion

Classic overconfidence

Subjects classifying themselves by point estimates into ability quantiles for the different domains corresponds to the traditional way of testing for overplacement. We can conclude whether there is apparent overconfidence or not and—more importantly—later compare whether true overconfidence shows up in the same domains. Table 1 shows the results.

— Please insert TABLE 1 approximately here —

In line with earlier research, participants appear to be exceedingly overconfident when judging their sense of humor or their ability to get along with others. Most people see themselves above average (96% and 84%, respectively) and many place themselves in the 7th or 8th decile of the ability distribution. However, one needs to keep in mind that even this extreme case is not sufficient for proving the existence of true overconfidence; for instance, with respect to sense of humor, exactly 60% place themselves among the top 30% of the distribution, which does not violate the condition set up by Benoît and Dubra (2009). As discussed before, their requirement is too strong for the ratios typically found in overconfidence experiments.

For study performance (a domain where better feedback is available), overconfidence is less pronounced, but still mean, median and percentage of participants viewing themselves to be above average are significantly greater than the corresponding neutral values. We find only slight underconfidence for programming skills and a neutral result for investing abilities. A young student population accustomed to computers may feel more competent in programming while at the same time having little financial market experience. Kruger (1999) shows that the self-assessed ability in a domain is a strong driver of overplacement. In particular Glaser et al. (2009) find higher levels of overconfidence for finance professionals than for students. The result for risk of a heart attack is again as anticipated: most

participants are overoptimistic and assess their personal risk as lower compared to their peers; in fact, 75% of participants believe that their risk is below average.

We emphasize that—for the purpose of this paper—it is less important whether the results for each domain match precisely the expectations derived from the literature as we are primarily interested in the relationship between apparent and true overconfidence.

Probability assessments

The key data of our study consist of the probability assessments supplied by experiment participants. While participants exhibit many shapes of belief distributions—some skewed and others symmetric, some flat and others very steep—the distributions have in common that they are unimodal, i.e. exhibiting a single probability peak in one quantile or several adjacent quantiles sharing the same probability. We almost never observe a first peak followed by a drop in probability followed by another peak; this makes sense intuitively as one mostly feels either skillful or not for any given domain. The tightness of the distribution hints at how sure subjects are about their self-assessment: very often, subjects indicate a zero probability for several deciles, i.e. they are sure that they could conceivably fall only into a certain range of the scale.

When we ask for the whole distribution of beliefs, it is no longer possible for a rational population to be predominantly above average in the sense that neither the mean of the individual distribution means, nor a major part of the probability mass of the distributions can be significantly above average. In a decile framework, the mean of individual distribution means must be at 5.5, in a quartile setup it must lie at 2.5; the aggregated probability mass must be split equally between the lower and upper half of the quantiles. In fact, in a population of true Baysians, equation 2 has to hold in expectation for every quantile. This is far more restrictive than the direct assessment analyzed before, where people only indicated to which quantile of the distribution they believed themselves to belong to.

— Please insert TABLE 2 approximately here —

The results for the probability assessment in table 2 appear similar to those of table 1. We again find strong overconfidence for sense of humor and the ability to get along with others, and somewhat weaker overconfidence for study performance, all significant at 1%-level (t-test). The rightmost column of table 2 refers directly to equation 5. Values significantly above 50% indicate true overconfidence (overplacement). Overplacement is found in all expected domains with the exception of "financial market investment". Analogously, underplacement can be diagnosed for programming skills and risk of a heart attack (overoptimism). We did not find any order effects, neither for the order of domains nor for the order of probability estimate and point estimate. The between-subject domains ("sense of humor" and "programming skills") reveal that the degree of overplacement is similar between point estimates and probability estimates even if compared across groups.

To test whether the elicited probabilities coincide approximately with the rational benchmark (and thus might have been derived by Bayesian updating), we use a chi-square goodness-of-fit test and a Kolmogorov-Smirnov test for equality of distributions. Table 3 shows the p-values of these tests for the skills and abilities used in experiment one.

— Please insert TABLE 3 approximately here —

For four domains, both tests reject rational information processing at 1%-level, for the remaining two domains at least the chi-square test is significant.[2] The in general pronounced asymmetric shape of the average belief distribution cannot be reconciled with the normative result derived in section 4. While in the theoretical driving skill example optimistic assessments of those who received a positive signal were counterbalanced by the beliefs of those with a negative signal, this does not seem to happen in the experiment. We will analyze individual belief distributions in more detail in experiment two.

[2]The differences between the two tests arise from the fact that the chi-square test penalizes any deviation from the distribution, while the Kolmogorov-Smirnow test is sensitive to deviations in the cumulative distribution function.

Limitations of experiment one

It has been argued that ambiguity is a problem in questions concerning skills and virtues (Dunning et al., 1989; Van Den Steen, 2004). People might interpret the skill in question differently, consequently allowing everyone to rightfully reach the conclusion that they are above average with respect to their subjective definition of the skill in question. There also were no monetary incentives in experiment one, primarily because a convincing incentive scheme was not available. However, it has been demonstrated that behavioral biases may disappear with proper incentivization, although evidence in psychological and economic experiments is mixed (Camerer and Hogarth, 1999; Hertwig and Ortman, 2001). To account for these possibilities, we design a second experiment in which subjects have to evaluate their performance in incentivized laboratory tasks.

Experiment two

Method

Participants

Experiment two took place at the University of Mannheim in 2010. 50 students of various faculties (50% business and economics) were recruited via an online recruitment system for economic experiments (ORSEE; Greiner, 2004). 48% of the participants were male, the median age was 24. Experiment two was computer-based and programmed in z-tree (Fischbacher, 2007).

Procedure

In this experiment we elicited probability assessments for four tasks conducted in the laboratory. We chose tests for intelligence, memory, creativity, and general knowledge as tasks for the experiment (see Appendix B); these domains should represent meaningful and desirable qualities for our subjects.

Trait desirability often goes along with self-serving ability assessments (Alicke, 1985); in particular, Burks, Carpenter, Goette, and Rustichini (2009) find overplacement in an IQ test and Moore and Healy (2008) demonstrate a similar effect in knowledge tests. It has been further shown that in social comparisons, easy tasks typically produce more pronounced overplacement than do difficult tasks (Larrick et al., 2007; Moore and Healy, 2008), as people seem to focus on their own result and do not fully account for the task being easy or difficult for most of the other participants as well.[3] We thus expect true overconfidence of subjects for the test domains, probably moderated by increasing task difficulty.

As subjects had given appropriate responses in the more precise decile framework in experiment one, we used solely this design in experiment two. The wording of the experimental instructions remained the same (see Appendix A). It was automatically checked whether probabilities summed to one and subjects were prompted to correct their entries if not. Two participants repeatedly failed to correct their answers and were excluded from the analysis.

Participants completed tasks prior to their evaluation of probabilities, implying that at the time they had to make their judgments, there was little ambiguity about which performance they should evaluate. We used a quadratic scoring rule to incentivize subjects (Selten, 1998), as a quadratic scoring rule makes it optimal for (risk-neutral) subjects to submit their true belief distributions. If anything, risk-averse subjects would bias their response to a more uniform distribution which would counteract our results. In overconfidence research, Moore and Healy (2008) apply the quadratic scoring rule in a similar experimental design. Participants were told that tied scores would be resolved by chance.

[3]This finding is a reversed form of the classic hard-easy effect (Lichtenstein et al., 1982; Juslin, Winman, and Olsson, 2000).

Results and discussion

Probability assessments

Participants find it most likely that they rank between the sixth and ninth decile for the tasks in experiment two. Table 4 shows the average probabilities that participants stated for the ability scale used. Relatively few subjects believe their performance to be in the very top decile compared to their peers. While the seventh and eighth decile are the most popular choice (with average probabilities assigned to these deciles mostly exceeding 0.15), participants submit very low probabilities for the bottom deciles, sometimes as low as between 0.01 and 0.03 for the domains of intelligence and memory. Kruger and Dunning (1999) show that for many tasks, very few people believe that they perform very badly compared to their peers. We add that people do not even find it *probable* that they could be bad.

— Please insert TABLE 4 approximately here —

Table 5 generalizes these results to the statistics of the distribution that we are especially interested in. In three out of four domains, we find significant overplacement of participants measured by both the mean of the distribution means and the average probability mass above the middle point of the scale. The extent of overplacement is comparable to the untested abilities of experiment one. Ambiguity may have contributed to overconfidence in domains such as "sense of humor" (cp. Dunning et al., 1989), but true overconfidence is present also in controlled tasks with little interpretational flexibility.

— Please insert TABLE 5 approximately here —

Extending the analysis to other thresholds than the average or middle point of the quantile scale reveals patterns which were already suggested by the descriptive statistics. For the top 40%, we still find overplacement similar to the presented results for being above

average. For high quantiles, however, overconfidence becomes markedly weaker or even disappears altogether. Only in one domain of experiment two (memory test), overplacement is still significant for the top 20%; the better-than-average effect thus appears to be a "slightly-better-than-average effect".

The pattern of overplacement found is in line with the reversed hard-easy effect for relative judgments (Larrick et al., 2007; Moore and Healy, 2008). The right column of table 5 shows the proportion of correct answers given in the tasks. With only 32.8% correct responses, the creativity test clearly qualifies as hard, and we find no significant overplacement; on the other hand, overplacement is most pronounced in the easiest task (memory). We explain this finding by the egocentric nature of relative judgments (Kruger, 1999; Moore and Cain, 2007): subjects react more strongly to variations in their own performance than to possible variations in the performance of other participants.

We again test whether the submitted probabilities coincide approximately with the rational benchmark, and hence might have been derived by Bayesian updating. Table 6 shows the p-values of the chi-square test and of the Kolmogorov-Smirnov test for the tasks used in experiment two.

— Please insert TABLE 6 approximately here —

For three out of four domains, rational information processing can clearly be rejected. The asymmetry of belief distributions already visible in table 4 is again not compatible with the uniform distribution postulated by Bayesian updating. The statistical tests suggest that deviations are too strong to be a result of imperfect sampling of experiment participants alone, as this type of randomness should be small in magnitude and not systematic in a manner observed in the presented results.

Overconfidence on an individual level

Besides population averages, the controlled tasks allow us to test for overconfidence on an individual level. In the probability framework, participants provide a range of deciles they

believe to be possible for themselves with different probabilities. If the actual result is below or above this range, this is far stronger evidence for over- or underplacement than if it were to fall short of—or exceed—a point estimate. Table 7 shows the fraction of participants who ranked below their worst expectation or above their best expectation. For the domains of intelligence, memory, and general knowledge, about a third of the subjects end up in a decile below all of the deciles they had assigned a probability greater than zero, i.e. they perform worse than they had even considered possible. The other extreme—reaching a decile above one's best expectation—happens far less often (between 2% and 10%). Except for the domain of creativity, the difference between the two proportions is strongly significant (z-test). This impression of asymmetry is backed by the fraction of subjects reaching a decile below their mean expectation. It is rather common for subjects to fall short of their mean expectation, especially for the intelligence and memory test.

— Please insert TABLE 7 approximately here —

We have previously in parts explained the failure of average belief distributions to match the rational benchmark by the pronounced overplacement of unskilled participants (cp. Kruger and Dunning, 1999; Ehrlinger, Johnson, Banner, Dunning and Kruger, 2008): those who receive a negative signal should adjust their probabilities accordingly, and (as in the driving skill example) should hence submit high probabilities for low quantiles. We thus now examine the belief distributions of two specific groups, namely the skilled and the unskilled, where we define the groups as those who finish in the top three and bottom three deciles in each task, respectively. We assume that participants hold neutral priors before they enter the tasks.[4] A good or bad performance in the task should then inflate their subjective probabilities of falling into low or high quantiles, at least if subjects interpret their task performance correctly and update their beliefs in a rational manner.

— Please insert TABLE 8 approximately here —

[4]This is a conservative assumption; if anything, experience should already have shifted the skilled and unskilled towards more realistic priors.

However, table 8 shows that unskilled participants recognize their negative signal only partially: with the exception of the memory task, they understand that they are less likely to reach the top 30% but only slightly and occasionally increase their probability for the bottom 30%. For two tasks, the intelligence and memory test, they state even smaller probabilities than the neutral prior probability of 0.3 for the bottom 30%. Skilled participants seem to react more strongly to their positive signal: while they assign high probabilities to the top 30%, they regard the bottom 30% as almost impossible. This asymmetry in signal processing is one major cause for the disparity between the belief distributions and the rational benchmark. The right column of table 8 displays the average probability subjects assign to the correct decile, i.e. the decile they actually fall into according to their task performance. With the exception of creativity, skilled participants here submit higher probabilities and are thus better able to recognize their true skill level; they consequently earn more under the quadratic scoring rule regime. This finding is consistent with the idea that poor performers also lack metacognitive skills (Ehrlinger et al., 2008). While part of the result may be due to a regression effect (cp. Burson, Larrick, and Klayman, 2006), it cannot explain the asymmetry we observe between the judgments of unskilled and skilled participants.

General discussion

We propose a new methodology to measure overconfidence. Experiment participants evaluate their relative position within the population for different skills and tasks by stating their complete corresponding belief distribution; they provide probability estimates for each decile or quartile instead of a single point estimate. This approach avoids many problems that were shown to be detrimental to previous research in overconfidence. Belief distributions yield clear restrictions as to what is possible for a population of rational Bayesian updaters.

There is considerable overplacement in the belief distributions of experiment participants. Probability estimates closely resemble results based on traditionally used point estimates. Population averages for different characteristics of belief distributions are inconsistent with Bayesian updating. Participants on average state high probabilities for quantiles above average while they regard it as unlikely that they should fall into the bottom quantiles. Because of this pattern, the aggregated belief distribution fails to match the rational benchmark. Individual level results confirm these observations, with people often underperforming even their worst expectations. Overplacement is particularly pronounced for unskilled participants who apparently do not fully account for the negative signals they receive.

Causes of true overconfidence

We believe that motivational and non-motivational factors account for the existence of overconfidence (and in particular overplacement). It has been argued that positive illusions contribute to mental health and well-being (Taylor and Brown, 1988). They foster self-esteem and enhance the motivation to act (Bénabou and Tirole, 2002). Among the non-motivational factors, selective recruitment of information, focalism, and egocentrism have been put forward (cp. Alicke and Govorun, 2005). As discussed before, ambiguity, desirability, and controllability of the judgment item moderate the degree of overplacement. We favor these explanations as they examine the psychological roots of the phenomenon and seem more plausible than a logically rigorous but less realistic model.

Healy and Moore (2007) provide such a rational explanation for the occurrence of the reversed hard-easy effect, which we observe in experiment two. In their model, people hold incorrect prior beliefs but then update these beliefs rationally. If subjects perform better or worse than their expectation, they will attribute this partly to chance and partly to their ability; we cannot fully exclude this possibility. However, we use tests that relate to abilities and virtues such as memory or creativity for which subjects should have more

accurate prior beliefs (compared to the trivia quizzes used in Healy and Moore, 2007). To additionally reduce surprise potential in these tasks, we mostly use questions of a type that people may have seen before, e.g. typical IQ-test questions. In the post-experiment questionnaire, subjects rate the tasks according to their perceived reliability to test for the ability in question. Predominantly high ratings support the impression that the test content was in line with the expectation of participants.

Remaining caveats

A caveat to the proposed methodology is that participants may have difficulties with meaningfully completing the probability evaluation. People possess underlying beliefs about their skills but may not be able to express them in a probability distribution. We tried to address this concern by a careful analysis of what subjects actually do in the experiment: individual responses seem reasonable (as submitted belief distributions are unimodal without jumps or breaks), but this is of course only indicative evidence. Additionally, the incentive scheme in experiment two should motivate subjects to represent their true beliefs as closely as possible.

We further do not measure priors directly in experiment two and also cannot observe the signals participants receive from having completed the tasks. It is thus hard to determine precisely at what point during the information processing procedure the biases occur. However, in combination with experiment one (which elicits unconditional beliefs in several domains) our impression is that both priors and interpretation of signals are biased.

Implications

Overconfidence is among the behavioral biases most readily adopted by academic researchers in economics and finance. In the literature, it is related to excessive trading volume (Barber and Odean, 2000; Glaser and Weber, 2007; Odean, 1998), to the emergence of stock market

bubbles (Scheinkman and Xiong, 2003; Shiller, 2002), to corporate investment decisions (Gervais, Heaton, and Odean, 2003; Malmendier and Tate, 2005), and to the predictability of market returns (Daniel, Hirshleifer, and Subrahmanyam, 1998). Most of these articles take overconfidence as a given result from psychology and not as a subject of further scrutiny. For instance, Odean (1998) states that "a substantial literature in cognitive psychology establishes that people are usually overconfident and, specifically, that they are overconfident about the precision of their knowledge (p. 1888)." Some caution seems to be appropriate here: whereas excessive trading for instance is an observed reality, its link to overconfidence is established only on argumentative grounds; it relies on the existence of overconfidence as a robust feature of human behavior.

Consequently, if the existence of overconfidence is challenged in psychology, this will directly affect the mentioned research in economics and finance. Alternative explanations appear less compelling in many situations, thus without overconfidence these results lose much of their appeal. On that account, our findings contribute to behavioral explanations built on overconfidence remaining intact. They still might inspire some research to directly relate behavioral phenomena to economic reality.

Conclusion

The evidence collected suggests that the theoretically valid criticism of Benoît and Dubra (2009) has only little practical consequences for overconfidence research. In general, apparent overconfidence represents underlying true overconfidence which is reflected in belief distributions. It is not necessary to discard the literature on the better-than-average effect or to redo the entire research with a methodology that is robust against this objection. For future research, scientists may want to adopt a design like ours to avoid potential concerns.

Appendix A: Questionnaire

This is an example for the questionnaire in experiment one. Phrasing differed slightly when a point estimate was requested first. Experiment two used similar wording, except that "decile" was replaced by the more intuitive "rank" as participants were invited in groups of ten (in this case deciles and ranks are of course equivalent).

Question 3: Get along with other students

We would like to know how you think about your ability to get along with other students. We ask you to compare your ability with that of the other participants in this experiment. You should by means of a scale evaluate your estimated position within this group. The partition of the scale represents the ten deciles (i.e. 10%) of all students in ascending order. That is, the first decile contains the 10% of students who get along worst with other students, the second decile the following 10% up to the tenth decile, where you find the 10% of students getting along best with other students. You should state for each decile state the probability with which you think you are among this rank group. A value of 0.3 in one of the ten boxes thus means that you assume to belong to this decile of students with a probability of 30%. A value of 0 states that you definitely not belong to this ability group, while a value of 1 indicates that you are absolutely sure to be among this 10%-category.

Please enter the probabilities in the third row of the table and notice that the values must add up to 1.

worst 10%									best 10%
1	2	3	4	5	6	7	8	9	10

Appendix B: Sample questions of the experimental tasks

The intelligence test and the memory test in experiment two were taken from the Italian psychology platform www.nienteansia.it, translated to German and adjusted where necessary. Questions in the general knowledge domain come from Studenten Pisa, a knowledge test administered by German news magazine "Der Spiegel". The test for creativity is a self-designed remote associates test (Mednick, 1968). We reproduce here three sample questions for each task.

Intelligence test (21 questions)

How does this series of numbers continue? 1 - 4 - 10 - 22 - 46 - 94 - ...

A 188, B 190, C 200, D 47

Which of the following words does not fit the rest?

A Mouse, B Whale, C Snake, D Cat, E Seal

Please complete the sentence: "Car is to chassis as body is to..."

A Skin, B Blood, C Brain, D Skeleton

Solutions: B, C, D

Memory test (18 questions)

(Participants first read an excerpt from Oscar Wilde's short story "The Remarkable Rocket".)

How long had the King's son waited for his bride?

A One month, B One year, C Two years

What nationality was the bride?

A Russian, B Finnish, C None of these

What means of transportation did the bride use?

A Coach, B Sledge, C Ship

Solutions: B, A, B

Creativity test (12 questions)

(Participants are asked to think of a word that relates to the other three words. We do not present original examples as the task is very language specific.)

Bass—Complex—Sleep

Chamber—Staff—Box

Desert—Ice—Spell

Solutions: deep, music, dry

General knowledge test (24 questions)

In which century did the Thirty Years' War take place?

A 16th century, B 17th century, C 18th century, D 19th century

In which city is the novel "Buddenbrooks" situated?

A Lübeck, B Danzig, C Husum, D Kiel

Which sensory cells in the human eye are responsible for color vision?

A Cones, B Rods, C Plugs, D Studs

Solutions: B, A, A

References

Alicke, M. D. (1985). Global Self-Evaluation as Determined by the Desirability and Controllability of Trait Adjectives *Journal of Personality and Social Psychology*, 49, 1621–1630.

Alicke, M. D., & Govorun, O. (2005). The better-than-average effect. In M. D. Alicke, D. A. Dunning, & J. I. Krueger (Eds.), *The self in social judgment* (pp. 83–106). New York, NY: Psychology Press.

Alicke, M. D., Klotz, M. L., Breitenbecher, D. L., Yurak, T. J., & Vredenburg, D. S. (1995). Personal contact, individuation, and the better-than-average effect *Journal of Personality and Social Psychology*, 68, 804–825.

Alpert, M., & Raiffa, H. (1982). A progress report on the training of probability assessors. In A. Tversky & D. Kahneman (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 294–305). New York, NY: Cambridge University Press.

Barber, B. M., & Odean, T. (2000). Trading is hazardous to your wealth: The common stock investment performance of individual investors *The Journal of Finance*, 55, 773–806.

Bénabou, R., & Tirole, J. (2002). Self-confidence and personal motivation *The Quaterly Journal of Economics*, 117, 871–915.

Benoît, J., & Dubra, J. (2009). Overconfidence? *Working paper*, available at SSRN: http://ssrn.com/abstract=1088746.

Bradley, G. W. (1978). Self-serving biases in the attribution process: A reexamination of the fact or fiction question *Journal of Personality and Social Psychology*, 36, 56–71.

Brocas, I., & Carrillo, J. D. (2002). Are we all better drivers than average? Self-perception and biased behavior *CEPR discussion paper No. 3603*.

Burks, S. V., Carpenter, J. P., Goette, L., & Rustichini, A. (2009). Is Overconfidence a Judgment Bias? Theory and Evidence *Working paper*.

Burson, K. A., Larrick, R. P., & Klayman, J. (2006). Skilled or unskilled, but still unaware of it: How perceptions of difficulty drive miscalibration in relative comparisons *Journal of Personality and Social Psychology*, 90, 60–77.

Camerer, C. F., & Hogarth, R. M. (1999). The effects of financial incentives in experiments: A review and capital-labor-production framework *Journal of Risk and Uncertainty*, 19, 7–42.

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

Cooper, A. C., Woo, C. Y., & Dunkelberg, W. C. (1988). Enterpreneurs' perceived chances for success *Journal of Business Venturing*, 3, 97–108.

Daniel, K., Hirshleifer, D., & Subrahmanyam, A. (1998). Investor psychology and security market under- and overreactions *The 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 and Human Decision Processes*, 65, 201–211.

Dunning, D., Meyerowitz, J. A., & Holzberg, A. D. (1989). Ambiguity and self-evaluation: The role of idiosyncratic trait definitions in self-serving assessments of ability *Journal of Personality and Social Psychology*, 57, 1082–1090.

Ehrlinger, J., Johnson, K., Banner, M., Dunning, D., and Kruger, J. (2008). Why the unskilled are unaware: Further explorations of (absent) self-insight among the incompetent *Organizational Behavior and Human Decision Processes*, 105, 98–121.

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.

Fischbacher, U. (2007). Z-tree - Zurich toolbox for readymade economic experiments *Experimental Economics*, 10, 171–178.

Gervais, S., Heaton, J. B., & Odean, T. (2003). Overconfidence, investment policy, and executive stock options *Rodney L. White Center for Financial Research Working Paper No. 15-02*, available at SSRN: http://ssrn.com/abstract=361200.

Gigerenzer, G. (1991). How to make cognitive illusions disappear: Beyond "Heuristics and Biases" *European Review of Social Psychology*, 2, 83–115.

Gigerenzer, G., Hoffrage, U., & Kleinbölting, H. (1991). Probabilistic mental models: A Brunswikian theory of confidence *Psychological Review*, 98, 506–528.

Glaser, M., Langer, T., & Weber, M. (2009). Overconfidence of Professionals and Laymen: Individual Differences Within and Between Tasks?, *Working paper*.

Glaser, M., & Weber, M., (2007). Overconfidence and Trading Volume *The GENEVA Risk and Insurance Review*, 32, 1–36.

Glaser, M., % Weber, M. (2010). Overconfidence. In H. K. Baker & J. Nofsinger (Eds.), *Behavioral Finance: Investors, Corporations, and Markets*, (pp. 241–258). Hoboken, NJ: John Wiley & Sons.

Greiner, B. (2004). An Online Recruitment System for Economic Experiments. In K. Kremer & V. Macho (Eds.), *Forschung und wissenschaftliches Rechnen 2003* (pp. 79–93), Göttingen: Ges. fr Wiss. Datenverarbeitung.

Healy, P. J., & Moore, D. A. (2007). Bayesian overconfidence *Working paper*, available at SSRN: http://ssrn.com/abstract=1001820.

Hertwig, R., & Ortman, A. (2001). Experimental practices in economics: A methodological challenge for psychologists? *Behavioral and Brain Sciences*, 24, 383–403.

Juslin, P. (1994). The overconfidence phenomenon as a consequence of informal experimenter-guided selection of almanac items *Organizational Behavior and Human Decision Processes*, 57, 226–246.

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.

Klayman, J., Soll, J. B., González-Vallejo, C., & Barlas, S.(1999). Overconfidence: It depends on how, what, and whom you ask *Organizational Behavior and Human Decision Processes*, 79, 216–247.

Köszegi, B. (2006). Ego Utility, Overconfidence, and Task Choice *Journal of the European Economic Association*, 4, 673–707.

Kruger, J. (1999). Lake Wobegon be gone! The "Below average effect" and the egocentric nature of comparative ability judgements," *Journal of Personality and Social Psychology*, 77, 221–232.

Kruger, J., & Dunning, D. (1999). Unskilled and Unaware of It: How Difficulties in Recognizing One's Own Incompetence Lead to Inflated Self-Assessments *Journal of Personality and Social Psychology*, 77(6), 1121–1134.

Langer, E. J. (1975). The illusion of control *Journal of Personality and Social Psychology*, 32, 311–328.

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

Lichtenstein, S., Fischhoff, B., & Phillips, L. D. (1982). Calibration of probabilities: The state of the art to 1980. In A. Tversky & D. Kahneman (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 306–351). New York, NY: Cambridge University Press.

Mahar, H. (2003). Why are there so few prenuptial agreements? *Harvard Law School John M. Olin Center for Law, Economics and Business, Discussion Paper No. 436.*

Malmendier, U., & Tate, G. (2005). CEO overconfidence and corporate investment *The Journal of Finance*, 60, 2661–2700.

Moore, D. A., & Cain, D. M. (2007). Overconfidence and underconfidence: When and why people underestimate (and overestimate) the competition *Organizational Behavior and Human Decision Processes*, 103, 197–213.

Moore, D. A., & Healy, P. J. (2008). The trouble with overconfidence *Psychological Review*, 115, 502–517.

Odean, T. (1998). Volume, volatility, price, and profit When all traders are above average *The Journal of Finance*, 53, 1887–1934.

Preston, C. E., & Harris, S. (1965). Psychology of drivers in traffic accidents *Journal of Applied Psychology*, 49, 284–288.

Russo, J. E., & Schoemaker, P. J. H. (1992). Managing overconfidence *Sloan Management Review*, 33, 7–17.

Santos-Pinto, L., & Sobel, J. (2005). A model of positive self-image in subjective assessments *The American Economic Review*, 95, 1386–1402.

Scheinkman, J. A., & Xiong, W. (2003). Overconfidence and Speculative Bubbles *Journal of Political Economy*, 111, 1183–1219.

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

Shiller, R. J. (2002). Bubbles, Human Judgment and Expert Opinion *The Financial Analysts Journal*, 58, 18–26.

Svenson, O. (1981). Are we all less risky and more skillful than our fellow drivers? *Acta Psychologica*, 94, 143–148.

Taylor, S. E., & Brown, J. D. (1988). Illusion and Well-Being: A Social Psychological Perspective on Mental Health *Psychological Bulletin*, 103, 193–210.

Van den Steen, E. (2004). Rational overoptimism (and other biases) *The American Economic Review*, 94, 1141–1151.

Weinstein, N. D. (1980). Unrealistic Optimism about Future Life Events *Journal of Personality and Social Psychology*, 39, 806–820.

Zábojník, J. (2004). A model of rational bias in self-assessments *Economic Theory*, 23, 259–282.

Zuckerman, M. (1979). Attribution of success and failure revisited, or: The motivational bias is alive and well in attribution theory *Journal of Personality*, 47, 245–287.

Table 1: Point estimates of own skills and virtues on quantile scales (Experiment 1)

Domain	Expectation	Scale	n	Mean	Median	% above avg.
Study performance	overplacement (> 5.5)	deciles	64	6.39^{***}	6^{***}	78.1^{***}
Financial markets	overplacement (> 2.5)	quartiles	61	2.56	3	54.1
Sense of humor	overplacement (> 5.5)	deciles	25	7.80^{***}	8^{***}	96.0^{***}
Programming skills	underplacement (< 2.5)	quartiles	25	2.32	2	32.0^{*}
Getting along with others	overplacement (> 5.5)	deciles	64	7.08^{***}	7^{***}	84.4^{***}
Risk of heart attack	overoptimism (< 2.5)	quartiles	63	1.83^{***}	2^{***}	25.4^{***}

Notes: The table shows the tested skill domains, the experimental expectation (including its numerical meaning), and the partition of the scale used for each domain in experiment one. It contains number of observations, mean, median and percentage of subjects that placed themselves above average. For the decile scale the midpoint is 5.5, for the quartile scale 2.5. We use a two-tailed t-test (mean), Wilcoxon signed-rank test (median), and binominal probability test (percentage above average). $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

Table 2: Population averages of probability assessments for skills and virtues (Experiment 1)

Domain	Scale	n	Mean of distr. means	Total probability mass above average
Study performance	deciles	64	6.20***	69.0%***
Financial markets	quartiles	63	2.51	50.0%
Sense of humor	deciles	39	7.08***	80.6%***
Programming skills	quartiles	38	2.15***	36.2%***
Getting along with others	deciles	64	7.08***	80.6%***
Risk of heart attack	quartiles	64	2.02***	31.0%***

Notes: The table shows the tested skill domains and scale used for each domain in experiment one. It contains number of observations, mean of individual distribution means, and the total probability mass above average in %. For the decile scale the midpoint is 5.5, for the quartile scale 2.5. Two-sided t-test: $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

Table 3: Tests for compatibility of average belief distributions with the prediction of rational information processing (Experiment 1)

	p-values	
Domain	χ^2-test	KS-test
Study performance	0.000	0.002
Financial markets	0.005	0.351
Sense of humor	0.000	0.001
Programming skills	0.041	0.211
Getting along with others	0.000	0.000
Risk of heart attack	0.004	0.009

Notes: The table reports two test whether the average probabilities submitted by experiment participants correspond to the theoretical prediction of Bayesian updating. It shows the p-values of a chi-square test with nine degrees of freedom (deciles) and 3 degrees of freedom (quartiles) and the p-values of a Kolmogorow-Smirnov test for all domains of experiment one.

Table 4: Average probabilities assigned to deciles (Experimtent 2)

	worst 10%				decile scale					best 10%
Domain	1	2	3	4	5	6	7	8	9	10
Intelligence	0.014	0.020	0.045	0.073	0.120	0.140	0.189	0.185	0.129	0.085
	(0.061)	(0.061)	(0.105)	(0.094)	(0.127)	(0.126)	(0.152)	(0.147)	(0.146)	(0.165)
Memory	0.016	0.023	0.014	0.025	0.076	0.115	0.167	0.200	0.189	0.176
	(0.065)	(0.074)	(0.049)	(0.055)	(0.131)	(0.135)	(0.149)	(0.132)	(0.169)	(0.245)
Creativity	0.056	0.084	0.088	0.115	0.125	0.147	0.155	0.132	0.058	0.040
	(0.130)	(0.135)	(0.119)	(0.125)	(0.113)	(0.128)	(0.153)	(0.135)	(0.101)	(0.082)
Knowledge	0.033	0.071	0.078	0.062	0.073	0.107	0.152	0.188	0.153	0.085
	(0.083)	(0.131)	(0.135)	(0.093)	(0.096)	(0.129)	(0.159)	(0.167)	(0.153)	(0.162)

Notes: The table shows the average probabilities assigned to the deciles of the ability scale for the four tasks of experiment two. Standard deviatons are in parentheses.

Table 5: Population averages of probability assessments in experimental tasks (Experiment 2)

Domain	Scale	n	Mean of distr. means	Total probability mass above average	Proportion of correct responses
Intelligence	deciles	48	6.74***	72.8%***	69.5%
Memory	deciles	48	7.50***	84.7%***	85.9%
Creativity	deciles	48	5.52	53.2%	32.8%
Knowledge	deciles	48	6.45***	68.4%***	67.8%

Notes: The table shows the experimental tasks and the scale used for each domain of experiment two. It contains number of observations, the mean of individual distribution means, the total probability mass above average in %, and the proportion of correct responses for each task. For the decile scale the midpoint is 5.5. Two-sided t-test: $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

Table 6: Test for compatibility of average belief distributions with the prediction of rational information processing (Experiment 2)

	p-values	
Domain	χ^2-test	KS-test
Intelligence	0.000	0.002
Memory	0.000	0.000
Creativity	0.013	0.343
Knowledge	0.007	0.033

Notes: The table reports two test whether the average probabilities submitted by experiment participants correspond to the theoretical prediction of Bayesian updating. It shows the p-values of a chi-square test with nine degrees of freedom (deciles) and 3 degrees of freedom (quartiles) and the p-values of a Kolmogorow-Smirnov test for all domains of experiment two.

Table 7: Individual overplacement in experimental tasks (Experiment 2)

	Percentage of subjects ranked...		
Domain	...below worst expectation	...below mean expectation	...above best expectation
Intelligence	27.1%	70.8%	10.4%**
Memory	37.5%	70.8%	2.1%***
Creativity	10.4%	45.8%	8.3%
Knowledge	31.2%	56.3%	4.2%***

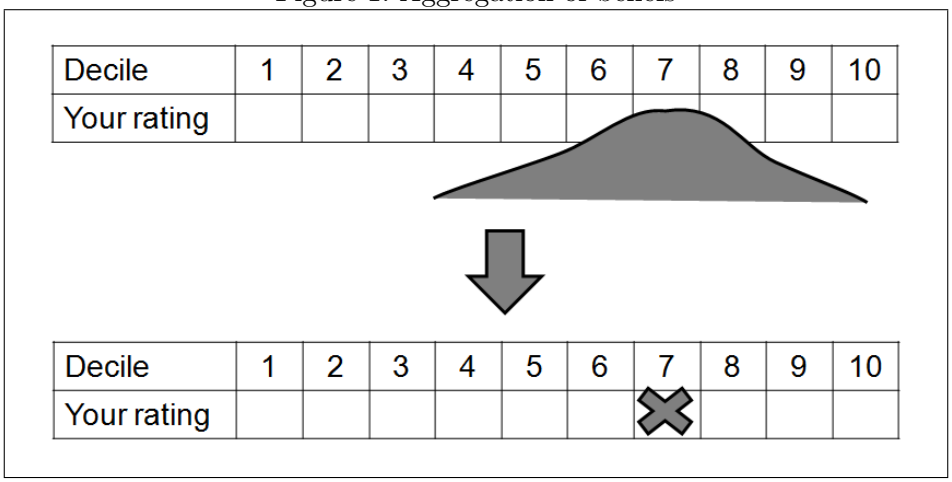
Notes: The table shows the proportion of subjects for which their actual decile rank is below their worst expectations, below their mean expectations, and above their best expectations in experiment two. Worst (best) expectations are defined as the lowest (highest) decile for which subjects submit a probability > 0. Asterisks stand for significant differences between the proportion below worst expectation and above best expectation using a two-sample z-test of proportion. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

Table 8: Probability assessment of skilled and unskilled participants (Experiment 2)

Domain	Skill level	Estimated probability to be in top 30%	Estimated probability to be in bottom 30%	Estimated probability for actual decile
Intelligence	skilled	0.54^{**}	0.04^{***}	0.20^{**}
	unskilled	0.20	0.19	0.06^{*}
Memory	skilled	0.83^{***}	0.00^{***}	0.33^{***}
	unskilled	0.39	0.07^{***}	0.04^{**}
Creativity	skilled	0.42	0.05^{***}	0.15
	unskilled	0.12^{**}	0.46^{*}	0.18^{*}
Knowledge	skilled	0.65^{***}	0.03^{***}	0.22^{**}
	unskilled	0.17	0.42^{*}	0.08

Notes: The table shows the average estimated probabilities of skilled and unskilled participants for different ranges of the decile scale in experiment two. We test for deviations from neutral prior probability using a Wilcoxon signed-rank test. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

Figure 1: Aggregation of beliefs



Notes: The top panel of the figure shows a belief distribution over ten deciles a person may possess about a skill or ability. The typical assessment of the better-than-average effect asks people to aggregate this belief distribution in a one point-estimate.