How Do People Take into Account Weight, Strength and Quality of Segregated vs. Aggregated Data? Experimental Evidence.*

Carlo Kraemer[a]

Martin Weber[b*]

[a] Lehrstuhl für Allgemeine Betriebswirtschaftslehre, Finanzwirtschaft, insb. Bankbetriebslehre; Universität Mannheim; 68131 Mannheim; Germany.

[b] Lehrstuhl für Allgemeine Betriebswirtschaftslehre, Finanzwirtschaft, insb. Bankbetriebslehre; Universität Mannheim; 68131 Mannheim; Germany; CEPR; London; UK; weber@bank.BWL.uni-mannheim.de. Tel: +49-621-1811532; Fax: +49-621-1811534.

* The authors gratefully acknowledge the financial support for this research, which was provided by the Deutsche Forschungsgemeinschaft (grant We993/7-3). Helpful comments were received from Thomas Langer and an anonymous referee.

How Do People Take into Account Weight, Strength and Quality of Segregated vs. Aggregated Data? Experimental Evidence.

Abstract

In this experimental study we investigated how people aggregate two sets of signals about the state of the world to reach a single probability judgment. The signal sets may differ in the way signals are presented, in their number as well as their quality. By varying the presentation mode of the signals we investigated how people deal with segregated and aggregated evidence. We investigated whether subjects sufficiently take into account weight (number of signals), strength (composition) and quality of the information provided. The results indicate that consideration of the weight and strength of signals strongly depends on the type of their presentation. Particular patterns can be identified which determine if weight and/or strength are either under- or overweighted.

JEL C25, C91, D8

Keywords: weight and strength of information, Bayes' rule, heuristics

Imagine you are in a situation where you want to update your initial beliefs based upon a variety of information sources. For instance, one such situation would be if you wanted to assess the probability of the stock market going up the next trading day given such information as research reports, market prices and colleagues' opinions. The question is how to evaluate the existing evidence and how to combine these multiple pieces of evidence coming from different sources in order to reach a single probability judgment.

In many situations, it is possible to distinguish a set of available evidence by the dimensions of extremeness (strength), credibility (weight) and quality of the pieces it consists of. For example, strength can refer to the proportion of colleagues who think that the market will go up the next day; weight can refer to the total number of opinions and quality can refer to the knowledge of a specific colleague. Hence, strength expresses how representative the evidence is of a specific hypothesis, whereas weight expresses its statistical reliability and quality expresses the reliability of a single observation.

Another important distinction is whether evidence represents already aggregated information or whether evidence just consists of multiple pieces of information which still have to be aggregated. Market prices, for example, can be viewed as the aggregated opinions of all market participants; a set of opinions from one's colleagues, however, still has to be aggregated to reach a judgment.

All these characteristics, i.e. strength, weight and quality as well as whether evidence is already aggregated or not, affect posterior beliefs which are determined based on all available sources of information. For instance, a posteriori beliefs should not only be based on the fact that all colleagues think that the market will go up but it is also important to incorporate the aspect of how many colleagues express this opinion, as larger samples allow for more reliable inference. Thus, it certainly makes a difference if just one or two colleagues express this opinion or the

entire research department does. Furthermore, it is also important to consider how knowledgeable the specific colleagues are, i.e. are they veterans or rookies. Finally, weight of evidence has to be treated differently when evidence represents an aggregate of opinions rather than a set of opinions.

Previous studies

In normative theory Bayes' rule dictates how the different characteristics of a set of evidence have to be used to update prior beliefs. Experimental literature however indicates that subjects often do not follow Bayes' rule when updating their beliefs. Griffin and Tversky (1992) have investigated subjects' consideration of strength and weight of evidence when subjects observed a sample of tosses from a biased coin and then had to quote their confidence that the coin is biased in favor of heads. In their experiments strength refers to the proportion of signals in the sample, which support a specific hypothesis, and weight is defined as the total number of signals contained. They found that subjects focus too much on strength or extremeness of the information and insufficiently take into account its predictive validity (weight). As a result, subjects are underconfident when evidence is of low strength and high weight but are overconfident when evidence is of high strength and low weight. Thus, they do not focus enough on the statistical reliability of the evidence but too much on its extremeness.

Other than Griffin and Tversky (1992), Kraemer and Weber (2002) as well as Kraemer (2002) found situations in which subjects did account for the weight of information, even though it is irrelevant from a rational point of view. In their study subjects sequentially assessed the probability that one of two states of the world had occurred. The probability assessments were passed from one subject to the next and subjects had an opportunity to buy signals about the unknown state of nature. They found that subjects expressed greater confidence in probabilities

observed later in the sequence even though all predecessors were perfect Bayesians simulated by the computer. In this sense, participants focused too much on weight, because given a fixed probability, the number of signals on which the a posteriori probability is based, is irrelevant. Hence, the results indicate that subjects who faced aggregated evidence focused too much on weight of information.

Non-rational consideration of strength and weight can also affect aggregate market outcomes as shown in a series of experimental asset markets conducted by Nelson, Bloomfield, Hales and Libby (2001). The value of the traded assets in their experimental markets depended on a rational posterior probability, which was manipulated analogously to the experiments of Griffin and Tversky (1992). Bloomfield, Hales and Libby (2001) found that excessive focus on strength and insufficient consideration of weight influenced individual estimates of securities' values as predicted by previous experiments. Furthermore, this miscalibration persisted at the aggregate market level, leading to biased market prices. Bloomfield, Libby and Nelson (2000) explain this overestimation of unreliable evidence and underestimation of reliable evidence by a model of "moderated confidence". In their experiments confidence which is "moderated" towards investor's prior expectation was observed in a market setting and seemed to be robust to experience and additional information making information reliability more salient.

Barberis, Shleifer and Vishny (1998) constructed a model which explains underreaction to earnings changes and overreaction to long-term earnings trends as found in capital markets (see, for example, Cutler, Poterba and Summers 1991) by assuming that agents excessively focus on strength of information and insufficiently account for its weight. In their model subjects insufficiently take into account last period's earnings as an indicator of future earnings and focus too much on the history of earnings which is uninformative because earnings follow a random walk. Subjects rather believe that the earnings process shifts between a regime of trending, i.e.

earnings changes are followed by similar changes, and a regime of mean-reversion, i.e. earnings changes are followed by opposite changes. Bloomfield and Hales (2002) demonstrated in an experiment that subjects indeed believe in mean reversion or trending even though they observe a random walk.

Motivation and overview

Previous studies provided two main findings. First, the initial research by Griffin and Tversky (1992) showed that subjects insufficiently take into account weight and quality of evidence and focus too much on strength when they receive multiple observations, i.e. segregated information. Second, Kraemer and Weber (2002) as well as Kraemer (2002) found that subjects do take into account weight of evidence when they face aggregated information even though this is not rational from a Bayesian point of view. Hence, subjects seem to excessively focus on weight given aggregated information.

Based on these findings our paper has four main goals. First, we will try to replicate the findings of previous studies when subjects are given segregated evidence. Also, we will investigate how behavior is influenced when subjects receive explicit information about strength and weight and when evidence is presented sequentially. Second, we will investigate more thoroughly if subjects indeed focus excessively on weight when given aggregated information in an individual decision making task.[1] Third, if we find support for this hypothesis, we will then try to explain exactly **when** either under- or overweighting of relevant parameters such as weight and

[1] In the experiments of Kraemer and Weber (2002) subjects observed probabilities which were the result of an aggregation process of multiple computer simulated agents. One might argue that the excessive focus on weight simply arises because of a distrust in the computer simulated agents. Therefore, we will investigate the consideration of weight in an individual decision making task.

strength occurs and **why**. Finally, we will investigate how subjects consider quality of evidence and if the findings are robust to the subjects' experience.

Our experiment differs from previous ones in two important ways. First, we will investigate how updating behavior is adjusted given different presentation modes, thereby demonstrating that the presentation mode determines when and to what extent subjects overweight strength and insufficiently consider weight of evidence and vice versa. Second, we will provide subjects not only with one set of evidence but with two. By letting subjects aggregate two sets of information, we can analyze how they aggregate information presented in different ways or of different quality. Hence, we will have a multitude of additional experimental manipulations compared to previous studies which enables us to answer the question **why** subjects use false weightings of relevant parameters. We will, for example, be able to observe the subjects' behavior when they have to combine a set of aggregated evidence with a set of segregated evidence in order to see if the subjects treat those presentation modes differently.

In line with previous findings, we found too much emphasis on strength and too little emphasis on weight in situations in which evidence consists of multiple pieces of information, i.e. evidence is presented in segregated form. When subjects face aggregated data this relation is reversed, i.e. subjects excessively focus on weight. These findings appear to be caused by the fact that subjects do not sufficiently differentiate between the posterior probability following from a set of information and the proportion of signals which support this specific hypothesis.[2] In

---

[2] Even though this finding appears to be related to the literature investigating Bayes' updating when probabilities are presented in terms of frequencies (see, for example, Gigerenzer and Hoffrage 1995, Cosmides and Tooby 1996), our experiment addresses a different issue. In these previous experiments subjects update an a priori probability given a **single** observation. The authors show that when input information (likelihoods and a priori probability) as well as output information (a posterior probability) are expressed in frequency terms subjects are much more likely to calculate the correct Bayesian posterior than subjects having to update their beliefs based on information given in probability terms. In our experiments subjects receive **multiple** observations. Relative frequency or proportion refers to the proportion of a

addition to a biased consideration of strength and weight, subjects also insufficiently take into account the information quality leading in our experiment to underconfidence. Interestingly, expertise in Bayes' updating does not seem to affect our findings.

We will proceed as follows. In section 1 we will show how a rational Bayesian individual should control for weight, strength and quality of evidence when updating her probability distribution. We will present the hypothesis for our experiment in section 2. Section 3 outlines the experimental design used to test these hypothesis. This is followed by a presentation of the results in section 4 and the conclusion in section 5.

1 Rational Bayesian Behavior

We restrict the following presentation to a simplified world. This or similar simplified frameworks are often used to study Bayesian updating.[3] Furthermore, the same information structure is widely used in the literature on information cascades, which investigates information aggregation given a simple information structure and exogenous sequence of aggregation.[4]

The simplified world can take on two states of nature, labeled **A** and **B**. Both states of the world are a priori equally likely ($p(A) = p(B) = 0.5$). Subjects receive independent signals giving them a hint about the state of the world and have the task to update the probability that state A resp. state B has occurred. Signals can be either ***a*** or ***b*** signals. Signals are such that

---

specific signal type in this *set* of observations. We then investigate how subjects treat the characteristics (weight, strength and quality) of the signal *set* when updating their beliefs. Hence, we do not investigate how different representations of likelihoods and a priori information (given a **single** observation) influence updating performance but instead we investigate how different representations of the characteristics of a *set* of information (given **multiple** observations) influence updating behavior. These characteristics, such as number of observations and proportion of a specific observation type can only be defined when updating is based on multiple observations! Nevertheless, updating performance in our experiments might change if likelihood information and a apriori probabilities were expressed in frequency terms, but we do not investigate this issue here.

[3] See e.g. Griffin and Tversky (1992), Kraemer and Weber (2002), Rabin (2003).

[4] See e.g. Bikhchandani, Hirshleifer and Welch (1992), Anderson and Holt (1997), Hung and Plott (2001).

$p(a|A) = p(b|B) > 0.5 > p(b|A) = p(a|B)$. This means that an *a*-signal provides a hint that state A has occurred and is correct with probability $p(a|A)$ and false with probability $p(a|B)$. Analogously, a *b*-signal indicates that state B has occurred and is correct with probability $p(b|B)$, but is false with probability $p(b|A)$.

First, we want to define weight, strength and quality of information in our simplified world. Let *S* be a set of signals containing n^a *a*-signals and n^b *b*-signals, i.e. $S = \left\{(a)^{n^a}, (b)^{n^b}\right\}$, then:

Definition 1:

Weight is defined as the total number of signals $N = n^a + n^b$ in the signal set *S*.

Strength is defined as the proportion of *a*-signals $pp = \frac{n^a}{N}$ in the signal set *S*.

Quality is defined as the probability $q = p(a|A) = p(b|B)$, that a signal contained in the signal set *S* correctly indicates the occurred state.

Now, we want to illustrate how weight, strength and quality of the provided information affect the Bayesian a posteriori probability of state A. Using Bayes' rule the posterior log odds given a signal set *S* are:

$$log\left(\frac{p(A|S)}{p(B|S)}\right) = log\left(\frac{p(a|A)^{n^a - n^b}\, p(A)}{p(a|B)^{n^a - n^b}\, p(B)}\right) \tag{1}$$

$$= (n^a - n^b)\, log\left(\frac{p(a|A)}{p(a|B)}\right) + log\left(\frac{p(A)}{p(B)}\right) \tag{2}$$

One can see that apart from the quality of the signals $p(a|A)$ and the base rate $p(A)$,[5] the only factor which affects the rational a posteriori probability is the difference between the number of

[5] Note that $p(a|B) = 1 - p(a|A)$ and $p(B) = 1 - p(A)$.

a- and *b*-signals. Holding the difference between *a*- and *b*-signals constant, the total number of signals ($N = n^a + n^b$) is irrelevant. This means that the same inferences can be drawn by observing 102 *a*-signals and 100 *b*-signals as from observing just 2 *a*-signals and no *b*-signal.

Rewriting equation (2) shows how the posterior log odds are affected by the different information parameters:

$$\log\left(\frac{p(A|S)}{p(B|S)}\right) = N\left(2\frac{n^a}{N} - 1\right)\log\left(\frac{p(a|A)}{p(a|B)}\right) + \log\left(\frac{p(A)}{p(B)}\right) \tag{3}$$

Weight Strength Quality Base rate

Ceteris paribus, as the number of signals or the proportion of *a*-signals increases, the a posteriori probability that state A has occurred also increases.

After having presented how weight, strength and quality affect the a posteriori probability, we can now define the general presentation modes for the evidence given to the subjects in our experiment.

Definition 2:

Segregated evidence is defined as signal set *S*, which is presented such that subjects can directly observe or infer which signals are contained.

Aggregated evidence is defined as signal set *S*, which is presented such that subjects are given the rational a posteriori probability distribution over states which can be inferred from the signal set.

The presentation mode has a crucial influence on how subjects should take into account the different parameters weight, strength and quality. When subjects are given segregated evidence, then weight, strength and quality of the information should be incorporated into posterior

probabilities according to the right hand side of equation (3). A different situation is given when subjects receive aggregated information. In this case, subjects observe the outcome of the Bayes' updating represented by the left hand side of equation (3), i.e. weight, strength and quality of the information are already incorporated into this probability. Hence, weight is irrelevant when subjects observe aggregated information.

After having presented how a rational Bayesian individual would take into account the different parameters weight, strength and quality, we will now focus on our experimental hypothesis which we derive from the findings of previous studies.

2 Hypothesis

The experiments by Griffin and Tversky (1992) directly lead to our first hypothesis of how subjects treat segregated evidence:

> *Hypothesis 1:* Insufficient consideration of weight and excessive focus on strength given segregated evidence
>
> *If the evidence provided to the subjects is segregated, i.e. subjects know the signals contained in the signal set, subjects insufficiently take into account the total number of signals (weight) and focus too much on the signal proportions (strength) in the sample.*

In our experiment we will try to provide support for this hypothesis by first attempting to replicate the findings of Griffin and Tversky (1992). In addition, we will investigate a treatment in which subjects receive explicit information concerning strength and weight of evidence in order to see if this influences the consideration of these parameters. Furthermore, we will investigate how subjects combine two sets of segregated evidence in order to learn more about

the updating mode which leads to the erroneous consideration of the relevant parameters observed by Griffin and Tversky (1992).

Even though from a rational point of view, the sequential presentation of evidence should not influence updating behavior, various studies suggest that subjects' updating behavior is influenced when evidence is presented sequentially (see, for example, Anderson 1981, Davis 1984, Hogarth and Einhorn 1992, Rabin 2003). Empirical evidence as well as theoretical models provide no specification of how a consideration of weight and strength of evidence is affected in general, nevertheless we will investigate a treatment in which subjects sequentially observe the signals contained in a set of signals in order to verify the following hypothesis:

> *Hypothesis 1a:* Influence of sequential presentation of evidence on consideration of weight and strength
>
> *If the evidence provided to the subjects is presented sequentially, i.e. subjects observe a sequence of signals contained in the signal set, their subjective probability distribution over states differs from their subjective probability distribution when evidence is presented at once.*

Kraemer and Weber (2002) investigated information acquisition in a simple sequential aggregation process. The authors observed that, even though a fixed probability represented the same information regardless of the number of signals the updating was based on,[6] subjects' demand for additional information decreased with the number of predecessors. Kraemer (2002) showed that this effect persisted even when participants could observe the signals of their predecessors. Transferring this observation to our experiment suggests that subjects view a

---

[6] Predecessors acted like perfect Bayesian individuals, and therefore the information content of a fixed probability does not depend on the number of predecessors or the number of signals, as shown in section 1.

probability as more informative when it is based on more information. This leads to the following hypothesis:

Hypothesis 2: Relevance of weight given aggregated evidence

Even if the evidence provided to the subjects is already aggregated, i.e. subjects are given the a posteriori probability resulting from the evidence, their subjective probability distribution over states depends on the total number of signals.

In our experiment we will first try to replicate the finding that subjects do consider weight of evidence when given aggregated information. We will present subjects with a single set of evidence in aggregated form and will check if they treat this probability differently when it is based on different amounts of information. In order to provide more reliable proof for the hypothesis we will also investigate how subjects combine two sets of already aggregated evidence into one single probability judgment. Finally, we will investigate how subjects aggregate a set of segregated evidence with a set of aggregated evidence in order to answer the question why subjects weight the relevant parameters erroneously.

Griffin and Tversky (1992) also explored how subjects take into account the discriminability of the hypothesis, which can be interpreted as the quality of the provided information as well. Their experiments revealed that subjects insufficiently took into account the quality of the evidence (resp. the dicriminability of the hypothesis). Transferring this observation to our experiment would suggest that subjects will insufficiently take into account the quality of the provided signals and leads to our final working hypothesis:

Hypothesis 3: Insufficient consideration of quality given segregated evidence

If the evidence provided to the subjects is segregated, i.e. subjects know the signals contained in the signal set, subjects insufficiently take into account the quality of the provided signals.

Other than Griffin and Tversky (1992), we will not investigate the influence of the quality of the evidence by varying the discriminability of the hypothesis. Instead, we will provide subjects with two sets of information, each containing signals of different quality and will explicitly tell them about the quality of the signals. Doing so, we will make the difference in quality between the signals of the two signal sets more salient and can therefore investigate whether the insufficient consideration of quality persists under these conditions.

3 Experimental Design and Procedures

3.1 Design

We tested the hypothesis by observing subjects' behavior in several updating tasks. The experimental world was chosen in analogy to the simplified world described in section 1. As a reminder, the experimental world can take on one of two states of nature, labeled **A** and **B**. Both states are a priori equally likely. Subjects are provided with independent signals indicating which state has likely occurred. Signals can be either ***a***- or ***b***-signals.

In order to investigate how subjects process weight, strength and quality of evidence when information is either segregated or aggregated we provided subjects with sets of signals and asked them to state their subjective posterior probability distribution over states. In a first sequence of updating tasks we provided subjects with only a single set of signals and asked them for their updated probability. Then, we gave subjects several updating tasks in which they had to aggregate two sets of signals about the state of the world. By giving subjects two instead of just one signal set we are able to study the subjects' updating behavior when the information

contained in a signal set is already aggregated.[7] For example, we can provide subjects with the rational posterior probability distribution following from one signal set and in addition provide them with a second set of not yet aggregated signals. Analyzing their behavior in this situation reveals how subjects treat the signal set of which they know the correct posterior probability distribution and thus how subjects treat this set of aggregated signals. Furthermore, by giving subjects two sets of signals we can investigate additional effects. For example, we are able to study how subjects combine two sets of segregated signals. We can investigate whether subjects calculate their subjective posterior distribution based on all signals in one step or whether they first aggregate each signal set and then combine both sets.

The experiment consisted of 51 rounds. In each of these rounds, the state of the world was determined and subjects were provided with information about this state. Their task was to aggregate this information and quote their subjective probability that state A resp. state B had occurred. The signals, which subjects received in each round, were divided into two signal sets. The signal sets could contain different numbers of signals, compositions of signals and qualities of signals. Furthermore, both sets could differ in the way the contained signals were presented.

Formally, each signal set S_i, $i = 1,2$, contained N_i $(N_i \geq 0)$ signals s_{ij}, $j=1,\ldots,N_i$. A signal s_{ij} could be either an ***a***- or ***b***-signal, i.e. $s_{ij} \in \{a,b\}$. All signals s_{ij}, $j=1,\ldots,N_i$, in signal set i were of the same quality $q_i \geq 0.5$. The quality q_i was indicated to the subjects in percentage terms and was communicated as follows. Signal s_{ij} of quality q_i can be interpreted as a random draw from an urn, with replacement, containing $100*q_i$ signals which indicate the state that has actually occurred (i.e. $p_i(a|A) = p_i(b|B) = q_i$), and $100-100*q_i$ signals of the opposing state (i.e.

---

[7] When subjects receive just a single set of signals in aggregate form, then there is nothing to be done left. Combining two sets of signals, of which at least one is in aggregated form, represents a new aggregation task and enables us to study subjects updating behavior when given aggregated information.

$p_i(b|A) = p_i(a|B) = 1 - q_i$). Alternatively, signal s_{ij} of quality q_i correctly indicates the state of the world with probability q_i and indicates the "wrong" state with probability $1 - q_i$.

The presentation mode M_i of signal set S_i determined how the signals in the signal set were presented. There were four different presentation modes:

- $M_i = Sig$: All signals s_{ij}, $j = 1,\ldots,N_i$, contained in signal set S_i were presented at once. All *a*-signals were shown in one row and all *b*-signals were shown in a second row.
- $M_i = Seq$: The signals s_{ij}, $j = 1,\ldots,N_i$, contained in signal set S_i were presented sequentially.
- $M_i = PP$: Subjects could not observe the signals s_{ij}, $j = 1,\ldots,N_i$, contained in signal set S_i. Instead, subjects were told that the signal set S_i contains N_i signals of which pp_i percent are *a*-signals.
- $M_i = Prob$: Subjects could not observe the signals s_{ij}, $j = 1,\ldots,N_i$ contained in signal set S_i. Instead, subjects were told that the signal set S_i contains N_i signals, leading to a probability p_i that state A has occurred.

The different presentation modes served to test our different hypothesis formulated in section 2. A signal set in presentation modes ***Sig***, ***Seq*** and ***PP*** represents **segregated information** as defined in the "Hypothesis" section. In these presentation modes subjects can either directly observe or infer which signals are contained in the signal set.[8] By studying how people process signal sets in these three presentation modes we can test **hypothesis 1**. The Sig presentation mode

---

[8] In the *PP* presentation mode, subjects cannot observe the signals but they are told that the signal set S_i contains N_i signals with a proportion of *a*-signals of pp_i. Using the provided information, subjects can directly infer that the signal set contains $n_i^a = pp_i \cdot N_i$ *a*-signals and $n_i^b = (1 - pp_i) \cdot N_i$ *b*-signals, so that this presentation mode also qualifies as segregated evidence.

was used to replicate the findings by Griffin and Tversky (1992). The *PP* presentation mode provided subjects with explicit information concerning the strength of evidence contained in the signal set and thereby tested whether this explicit information intensified the focus on strength of evidence. The *Seq* presentation mode served to test the influence of a sequential presentation of the signals (**hypothesis 1a**).

A signal set in ***Prob*** presentation mode represents **aggregated information**. Subjects are given the rational a posteriori probability resulting from this signal set, so that the parameters weight, strength and quality of information are already incorporated. By investigating how subjects process a signal set in *Prob* presentation mode we were able to test **hypothesis 2**.

Finally, we tested **hypothesis 3** by providing subjects with two signal sets in *Sig* presentation mode, where the quality of the signals is different between signal sets. In this case we did not investigate effects of sequential presentation or explicit information concerning strength. Table 1 illustrates how the presentation modes are related to the hypothesis we wanted to test.

To ensure that subjects had an incentive to submit meaningful probability judgments, they were paid according to their stated probabilities in each round. Because a correct Bayesian posterior could be calculated in every situation given in the experiment, it was not necessary to use a scoring rule; instead one could pay subjects according to the proximity of their probability estimate to the corresponding Bayesian posterior. We paid subjects relative to the absolute difference of their stated probability to the rational probability, which was calculated applying Bayes' rule to the given information. The smaller the difference, the higher the payments. If subjects submitted probability p_{sub} and the rational Bayesian probability equaled p_{Bayes}, then the following formula determined the payment P_{cu} in currency units in this round:

$$P_{cu} = 1000cu - 2000cu \cdot | p_{sub} - p_{Bayes} | \qquad (4)$$

By paying subjects according to the distance between their stated probability to the rational probability instead of using a scoring rule, we eliminated effects from gambling as well as from risk attitude. For example, subjects had no incentive to gamble if they had only weak information, because the payment formula penalized them for deviating from the rational probability. Furthermore, decisions were not influenced by gambler's fallacy, because payments were deterministic rather than probabilistic. Finally, it was much easier for the subjects to understand this payment mechanism than a scoring rule. In order to avoid learning effects and to keep subjects from inferring the correct Bayesian posterior in situations which were later repeated by changing only the presentation mode for the signals, subjects received no feedback at the end of each round on the respective state of the world or their payment.

Since we were interested in how people act in specific situations, the chosen signals were not determined randomly. Instead, we constructed 51 sets of signals, each one representing one round. As stated earlier, these sets of signals were split into two subsets representing the two information sets subjects received in a round. In order to avoid ordering effects in the data, the sequence of rounds was determined randomly.

3.2 Procedures

The experiment was conducted using a computer program. At the beginning subjects had to submit some personal information, e.g. if they had attended the class "Rational Decision Making" or if they are undergraduate or graduate students. The instructions were then displayed on the computer screen. Subjects faced no time restriction to read the instructions and could ask questions. After the subjects had completed reading the instruction the experimental rounds started. Figure 1 illustrates a sample screen.

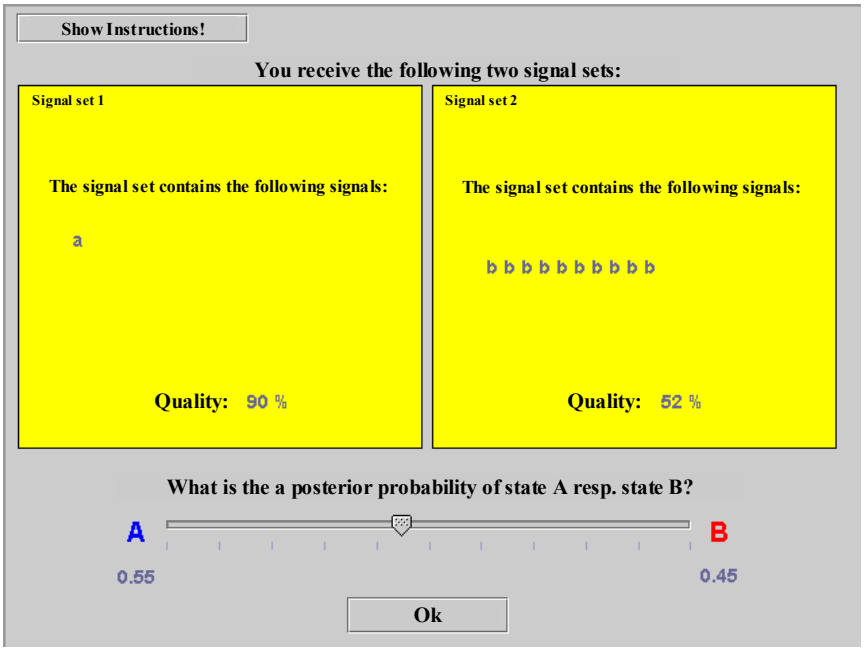


Figure 1: Sample program window.

The two boxes represent the two signal sets which the subjects have to aggregate. In this example both signal sets are presented in *Sig* mode. This means that the subjects simultaneously observe all signals contained in the two signal sets represented by the left and right box. To avoid ordering effects the computer randomly determined the signal set to be displayed on either side of the program window.

After observing the signals, subjects had to state their subjective probability on state A resp. state B using the sliding bar underneath the boxes. The probabilities indicated below the "A" and "B"-label were updated with every alternation of the slider and indicated which probability is submitted when subjects press the "Ok"-button.

51 subjects participated in the experiment at the University of Mannheim in January, 2002. 46 of whom were graduate students. All except for one participant were studying either Business

Administration or Economics. 24 of the 51 participants had attended a class in "Rational Decision Making" in which Bayes' theorem was extensively covered.

At the end of the experiment the total currency units a subject had earned in the 51 experimental rounds were converted to Euro at a fixed rate of 0.00025 Euro/cu (1000 cu = 0.25 Euro). Subjects on average earned 9.97 Euro, ranging from 7 Euro to 12 Euro.

4 Results

This section presents the results of our experiment and is organized as follows. First, we examine how subjects dealt with the **weight and strength** of **segregated** evidence (hypothesis 1). We then focus on subjects' consideration of **weight and strength** when they were given **aggregated** information (hypothesis 2). This is followed by an analysis of the consideration of **quality** of the provided signals (hypothesis 3). Finally, we study whether **expertise** in Bayes' updating influenced the results.

*4.1 Consideration of **weight and strength** given a **single** set of **segregated** evidence*

Initially, we look at how subjects controlled for weight and strength of evidence when they were provided with segregated data. In order to do so we first investigate subjects' updating behavior when they received a single signal set in *Sig* and *PP* presentation modes (hypothesis 1). In addition, we examine possible influences of a sequential signal presentation by giving the signals to the subjects in sequential order (hypothesis 1a).

We varied the weight of evidence by giving subjects a single signal set S_1 consisting of a changing number of signals. Subjects were provided with signal sets containing $N_1=5$, $N_1=15$ or $N_1=25$ signals for each presentation mode *Sig*, *Seq* and *PP*. In addition, we observed their updating behavior when they received just one *a*-signal in *Sig* presentation mode to get a baseline

case. The signals contained in the signal sets were all of quality $q_1 = 60\%$. All signal sets were constructed in such a way that the rational a posteriori probability that state A had occurred, which can be inferred from the signal set, was either 0.4 or 0.6. Table 1 shows the signal sets, which were given to the subjects. In order to be able to directly compare all given situations we transformed the reported probabilities in the case of $N_1=15$ to the probabilities that state B had occurred.[9] For example, if a subject reported a probability of 0.42 that state A had occurred when she observed 7 *a*- and 8 *b*-signals, we transformed her subjective probability to 0.58 to be able to compare her judgment in this case to the subjective probabilities reported in the other situations.

Table 1: Signal sets, segregated evidence.

S_1	M_1=*Sig*	M_1=*PP*	M_1=*Seq*	$p(A\|S_1)$
A	X			0.6
aaa bb	X	x	x	0.6
aaaaaaa bbbbbbbb	X	x	x	0.4
aaaaaaaaaaaa bbbbbbbbbbbb	X	x	x	0.6

Figure 2 illustrates the average subjective probabilities in all given situations. In addition, the light colored bars illustrate the proportion of *a*-signals in the signal set representing the strength of evidence (see definition 1). If subjects focus exclusively on strength and entirely neglect the other dimensions of the evidence such as weight and quality subjects' quoted probabilities should coincide with the light colored bars.

What can be clearly noted is the fact that even though from a rational point of view where all situations are equal, people tend to reach a probability judgment closer to 0.5 as the number of

[9] This means that after transformation the situation in which subjects receive 7 *a*-signals and 8 *b*-signals is equal to a situation in which subjects receive 8 *a*-signals and 7 *b*-signals.

signals increases. The median data provides a similar picture. Subjects' behavior in the *Sig* presentation mode, which is comparable to the experimental design in Griffin and Tversky (1992), is in line with their findings.

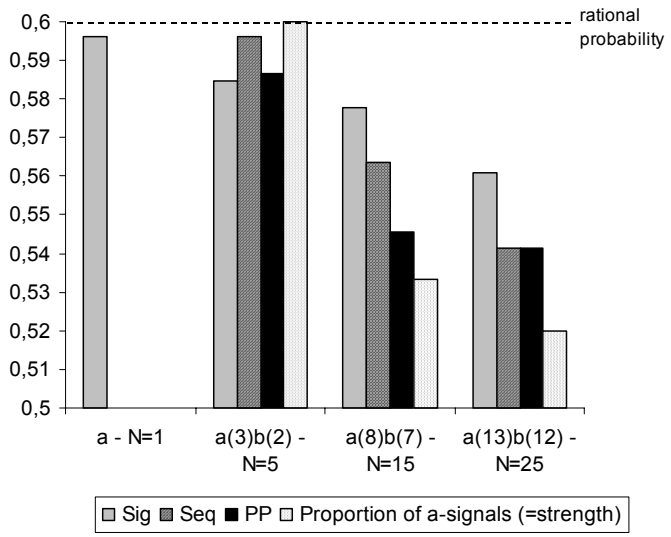


Figure 2: Average quoted probabilities facing segregated evidence, one signal set.

In general, subjects' behavior exhibits a strong focus on the composition of signals in the signal set. Since the strength of evidence, i.e. the proportion of *a*-signals, decreases as the number of signals increases, this excessive focus on strength leads to the observed pattern of behavior. This is especially evident in situations in which subjects were explicitly provided with information concerning the strength of evidence (*PP* presentation mode). In this case stated probabilities were closest to the proportion of *a*-signals in the signal set as indicated in figure 2. Table 2 illustrates the results from a Wilcoxon test when the quoted probabilities are compared across different presentation modes.

Table 2: Significance of the difference between average quoted probabilities in different presentation modes (Wilcoxon Zs).

Number of signals	*Sig* ↔ *PP*	*Sig* ↔ *Seq*	*PP* ↔ *Seq*
$N_1=5$	-0.37	-1.271	-0.787
$N_1=15$	-2.15	-0.884	-0.895
$N_1=25$	-2.49	-1.981	-0.155

The difference between the *Sig* and the *PP* presentation mode is insignificant when the signal set contains $N_1=5$ signals but significant on a 5% level when it contains $N_1=15$ and $N_1=25$ signals (see table 2). This means that the subjects' focus on strength is pronounced when they receive explicit information regarding the strength of evidence. Furthermore, it can be observed that sequential presentation also accentuates the focus on strength.[10] The difference in stated probability between the *Sig* and the *Seq* presentation modes is insignificant when the signal set contains $N_1=5$ and $N_1=15$ signals but significant when the judgment is based on $N_1=25$ signals (see table 2). Differences between the *PP* and the *Seq* presentation modes are all insignificant.

Table 3: Significance of the difference between average quoted probabilities and proportion of *a*-signals (=strength) (t values).

Number of signals	*Sig*	Seq	*PP*
$N_1=5$	-1.44	-0.479	-1.297
$N_1=15$	3.546	3.27	1.008
$N_1=25$	5.072	2.477	2.597

[10] This might be due to the fact that the constructed sequence of signals which subjects observe (see table 2) had a high alternation frequency between *a*- and *b*-signals with only short streaks.

Nevertheless, subjects do take into account weight of information as indicated in table 3. When the signal set contained $N_1=5$ signals the stated probability is close to the strength of information (see also figure 2). Then, as the number of signals increases ($N_1=15$ and $N_1=25$) the stated probabilities are all greater than the proportion of *a*-signals representing the strength of information. In *Sig* and *Seq* presentation modes the differences are significant according to a t-test (see table 3). In *PP* presentation mode, in which participants receive explicit information concerning the strength of evidence, the difference is only significant when the signal set contained 25 signals (see table 3).

In conclusion, we can say that our findings indicate that subjects faced with segregated data anchored on the proportion of signals and insufficiently adjusted for the weight of information with an increased number of signals. The data furthermore indicates that when presenting the signals sequentially or giving explicit information regarding strength to the participants this focus on strength increased.

Analogous to Griffin and Tversky we determined subjects' weighting of the different parameters of the provided information, using the following functional form for the rational posterior log log odds:[11]

$$log\left(log\left(\frac{p(A\mid S_1)}{p(B\mid S_1)}\right)\right) = \underbrace{log(N_1)}_{\text{Weight}} + \underbrace{log\left(\frac{n_1^a - n_1^b}{N_1}\right)}_{\text{Strength}} + \underbrace{log\left(log\left(\frac{p_1(a\mid A)}{p_1(a\mid B)}\right)\right)}_{\text{Quality}} \qquad (5)$$

The first term on the right hand side represents weight, the second term strength and the last term the quality of information. Let w_k be the subjective probability for state A that participant *k*

---

[11] For calculations refer to the "Hypothesis" section. We dropped the term representing the base rates, because here both states of the world are a priori equally likely. Hence, $log\left(\frac{p(A)}{p(B)}\right) = 0$.

submitted after observing signal set S_1. Then we ran the following regression in order to determine the weighting of the different parameters:

$$log\left(log\left(\frac{w_k}{1-w_k}\right)\right) = a + b \cdot log(N_1) + c \cdot log\left(\frac{n_1^a - n_1^b}{N_1}\right) + d \cdot log\left(log\left(\frac{p_1(a \mid A)}{p_1(a \mid B)}\right)\right) + u_k \quad (6)$$

In all situations in which subjects were provided with only one signal set S_1, the quality of the signals was constant. Therefore, we drop the term $d \cdot log\left(log\left(\frac{p_1(a \mid A)}{p_1(a \mid B)}\right)\right)$. Furthermore, in all situations except for one the signal set contained one more *a*-signal than *b*-signals. In one situation the signal set contained one more *b*-signal than *a*-signals. As stated above, we transformed the quoted probabilities w_k^* in this situation to $w_k = 1 - w_k^*$, so that this situation can be directly compared to the other situations. The regression then simplifies to:

$$log\left(log\left(\frac{w_k}{1-w_k}\right)\right) = a + b \cdot log(N_1) + c \cdot log\left(\frac{1}{N_1}\right) + u_k \quad (7)$$

$$= a + e \cdot log(N_1) + u_k \qquad \text{with } e = b - c \quad (8)$$

When subjects update their prior beliefs according to Bayes' rule, then $b = 1$ and $c = 1$, leading to $e = 0$. If $e > 0$ ($e < 0$) subjects excessively (insufficiently) take into account weight of information and insufficiently (excessively) take into account its strength. Since we wanted to verify if presenting the signals sequentially or providing subjects with the proportion of *a*-signals influences their relative weighting between strength and weight, we included dummies for the *PP* as well as for the *Seq* presentation modes. The following table illustrates the results of the OLS regression:

Table 4: Regression results, segregated evidence

Parameter	Value	σ	*p*-value
A	-0.720	0.039	<0.001
E	-0.25	0.038	<0.001
PP	-0.0308	0.039	0.43
SEQ	0.028	0.039	0.473
Adjusted $R^2 = 0.097$			

In line with the findings of Griffin and Tversky (1992), we found that subjects focused excessively on strength of information and insufficiently took into account its weight. The presentation mode did not seem to influence the relationship between the consideration of weight and strength. The constant *a* captures the effects which are not explained by the difference in regression coefficients *b* and *c*, e.g. the absolute level of the coefficients. Since this also influenced the subjects' updating behavior, the coefficient was significantly different from zero.[12]

In conclusion we can say that in line with previous results we found insufficient consideration of weight of evidence and excessive focus on strength. Even though the regression provides no additional support, the reported probabilities suggest that this effect is pronounced when subjects receive explicit information on strength of evidence and when evidence is presented sequentially.

*4.2 Consideration of **weight and strength** given **two** sets of **segregated** evidence*

We looked at how subjects aggregate two sets of signals in *Sig* presentation mode to shed more light on the updating procedure of the subjects. When subjects face two instead of just one signal set they have to conduct an additional aggregation step to reach a probability judgment. By

[12] The adjusted R^2 is rather small since the regression only captures the difference between regression weights that subjects put on weight and strength of information and the influence of the presentation mode on this difference.

analyzing this additional step we can gain additional insight on which parameters of the information subjects focus. Without having any specific model in mind, we hypothesize that when subjects excessively focus on the strength of information they might aggregate the signals contained in a signal set with a strong focus on strength and then aggregate the two signal sets by simply averaging their strength.

To investigate this hypothesis we divided the signal sets containing $N_I=5$ and $N_I=25$ which were used for the analysis when subjects receive a single signal set in two different ways. First, the signal set containing 3 *a*- and 2 *b*-signals (resp. 13 *a*- and 12 *b*-signals) was split up into two homogenous signal sets, one containing 3 *a*-signals and the other containing 2 *b*-signals (resp. one containing 13 *a*-signals and the other containing 12 *b*-signals). Second, we split up the signal set into two heterogeneous signal sets, one containing 1 *a*-signal and 1 *b*-signal and the other containing 2 *a*-signals and 1 *b*-signal (resp. one containing 6 *a*-signals and 6 *b*-signals and the other containing 7 *a*-signals and 6 *b*-signals).

Even though, from a rational point of view, the split has no influence on the a posteriori probability distribution, our above stated heuristic leads to different probabilities. If we assume that subjects only consider strength in their updating procedure, the heuristic leads to a 0.5 probability that state A has occurred when subjects are given two homogenous signal sets (A(n+1)/B(0) and A(0)/B(n)). On the other hand the heuristic leads to a probability of 0.58 (0.52) if the signal set containing $N_I=5$ ($N_I=25$) signals is split up into two heterogeneous sets.

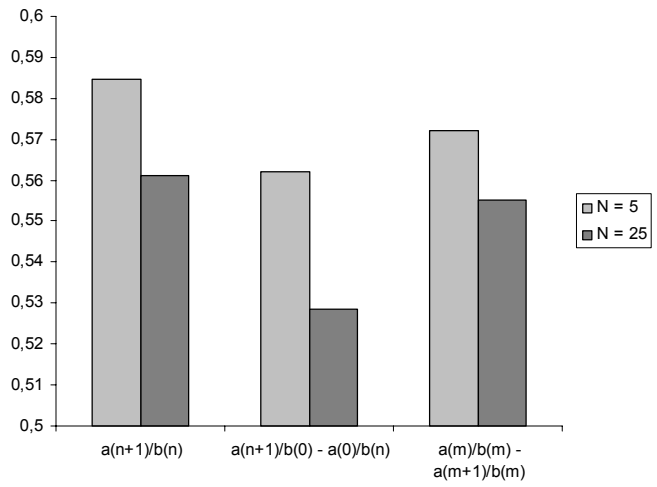


Figure 3: Splitting effects, segregated evidence.

Figure 3 illustrates the average probabilities we observed in the experiment. One can see that the split of one heterogeneous signal set into two heterogeneous signal sets did not influence the subjects' updating behavior. But splitting the signal sets into two homogenous signal sets led to a probability closer to 0.5 (difference for $N_1 = 5$ ($N_1 = 25$) is significant on a 10% (1%) level based on a Wilcoxon test). The data supports our heuristic as the split into two homogenous sets resulted in a subjective probability which is significantly closer to 0.5. Hence, the data provides further evidence that subjects excessively focus on strength.

Altogether, updating behavior based on segregated evidence clearly supports hypothesis 1 which claims that subjects excessively focus on the strength of evidence. Even though they take into account the weight of information, the adjustment is insufficient.

*4.3 Consideration of **weight and strength** given a **single** set of **aggregated** evidence*

Now we take a look at the situation when we provide subjects with already aggregated information but in addition inform them about the weight of information. We do this by

investigating subjects' updating behavior when they are given signal sets in *Prob* presentation mode. From a rational point of view, once the signals are aggregated into a probability judgment using Bayes' law, weight is irrelevant. A fixed probability judgment conveys the same information regardless of the number of signals it is based on. The probability judgment only depends on the difference between *a*- and *b*-signals as shown in section 1.

We gave subjects the same signal sets S_1 containing $N_1=5$, $N_1=15$ and $N_1=25$ signals used in the analysis of updating based on segregated data. But instead of showing them the signals or giving them the proportion of *a*-signals, we gave them the probability of state A which results from the signals in the signal set along with the number of signals in the signal set (which is, of course, irrelevant). Opposed to the situations in which subjects faced segregated data, the average stated probability does not decrease as the number of signals increases. Stated probabilities are close to the rational probability of 0.6. Nevertheless, in the case in which the signal set contained only $N_1=5$ signals, the probability is significantly smaller than 0.6 on a 10% level based on a t-test. In the other cases the difference is not significant. ($t_{N=5} = 1.733$; $t_{N=15} = 1.659$; $t_{N=25} = 0.239$).

So far this result is not very surprising, since subjects can first observe the probability they are later asked for. So it seems obvious that the stated probability is close to the rational probability. Nevertheless, redoing the regression illustrated above to determine the difference in regression coefficients which subjects address to strength and weight of evidence yields an interesting result.[13] Table 5 shows the results:

---

[13] When subjects receive a signal set in *Prob* presentation mode, they do not have explicit information regarding the strength of evidence. Nevertheless since subjects have information regarding the weight of evidence, the regression can determine which relative weighting of strength and weight best explains the quoted probabilities.

Table 5: Regression results, aggregated evidence.

Parameter	Value	σ	*p*-value
a	-0.950	0.093	<0.001
e	0.147	0.083	0.08

Adjusted $R^2 = 0.016$

It can be seen that $e > 0$, which means that subjects put more emphasis on the weight of information than on its strength. This contrasts the findings in the other presentation modes, in which subjects put more emphasis on the strength of information than on its weight.

*4.4 Consideration of **weight and strength** given **two** sets of **aggregated** evidence*

In order to gain a more reliable insight into the way people treat weight of information when they are given aggregated information, we provided subjects with two signal sets both in *Prob* presentation mode and increased the signal count in one of the signals sets holding the probability which results from this signal set constant. A rational Bayesian individual would aggregate both signal sets without focusing on the total number of signals. Taking into account the above results, however, we presume that subjects do incorporate the weight of the signal sets into their judgment leading to a subjective probability which is biased towards the set that contains more signals. This means in the aggregation procedure subjects put more weight on the signal set which contains more signals. Figure 4 illustrates the average probabilities subjects submitted in the experiment along with the corresponding rational probabilities.[14] The label on the x-axis indicates which signal sets subjects observed. For example, p=0.6/N=25 represents a signal set containing 25 signals which results in a 0.6 probability that state A has occurred.

---

[14] The rational probabilities are calculated using Bayes' rule. E.g. aggregating a signal set with a posterior probability $p_1=0.77$ and a signal set with a posterior probability $p_2=0.4$ leads to an a posterior probability based on both sets of $p_{Bayes}=0.77*0.4/(0.77*0.4+0.23*0.6)=0.69$.

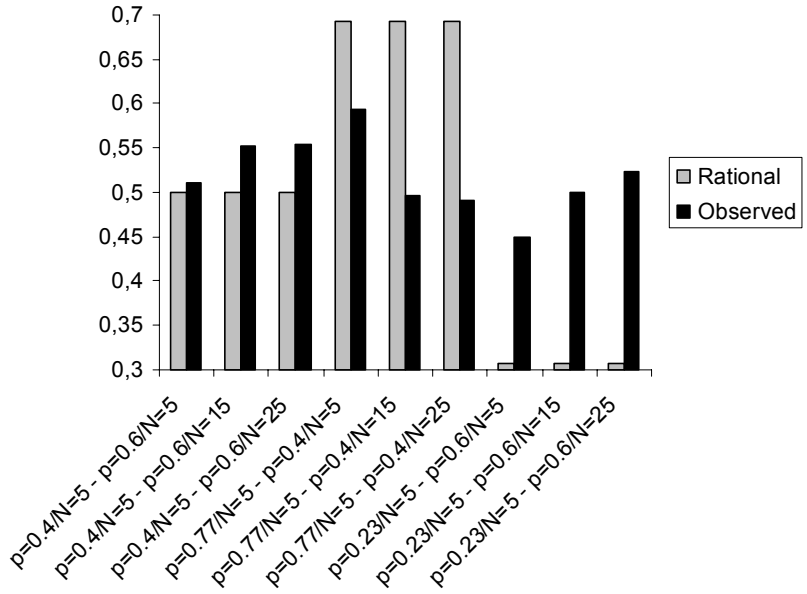


Figure 4: Average quoted probabilities facing aggregated evidence, two signal sets.

p=x/N=y refers to a signal set containing *y* signals which leads to a rational a posteriori probability of *p=x*.

The figure indicates that there is a clear effect from increasing the number of signals. Stated probabilities move in the direction of the signal set containing more signals, even though the number of signals is irrelevant from a rational point of view. Furthermore, we observe that when the smaller signal set allowed for strong inference ($p_1=0.23$ or $p_1=0.77$), this signal set was generally underweighted. This means that apart from the irrational consideration of weight the probabilities were biased towards 0.5 in this case.

The results concerning updating based on aggregated data are in line with the position effect found by Kraemer and Weber (2002) and Kraemer (2002). As already stated, in these experiments subjects purchased less information when a fixed probability was communicated by more predecessors. Here, subjects put more weight on larger sets of signals than on smaller ones when both sets lead to the same probability. This overweighting of inference from larger signal

sets also reduces the value of additional information if subjects are rewarded for correctly predicting the occurred state.

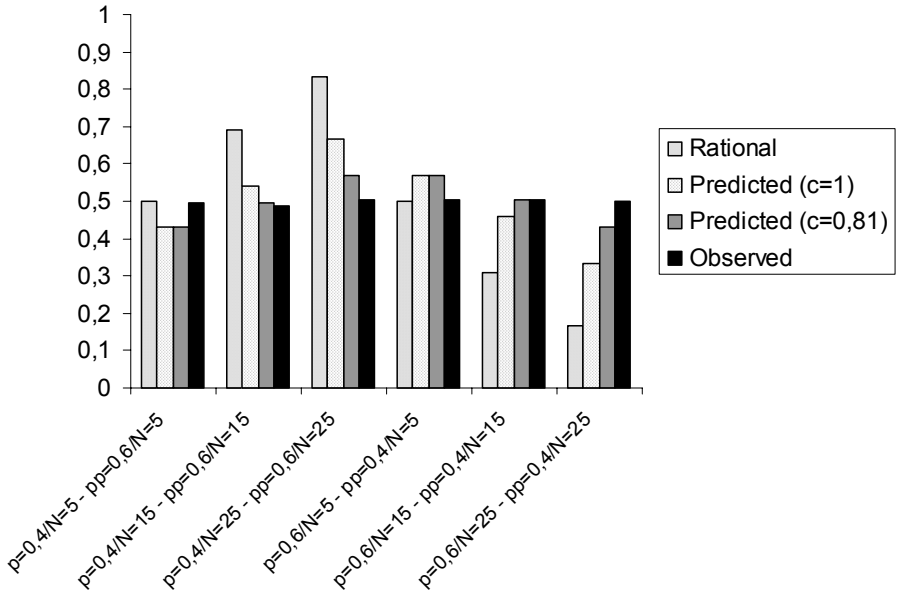
In conclusion we can say that our results on the consideration of weight given aggregated information clearly support hypothesis 2. Subjects do take into account weight even though a rational Bayesian individual would neglect this information. Together with the findings on the treatment of segregated information this raises the question how subjects combine aggregated and segregated evidence and if their consideration of weight differs in those cases. Investigating this can furthermore add to the understanding of how subjects interpret probabilities. For example, consideration of weight given aggregated data could be explained if subjects have some kind of frequency understanding of probabilities.

*4.5 Consideration of **weight and strength** given **one** set of **segregated** and **one** set of **aggregated** evidence*

In order to investigate this question we provided subjects with two signal sets, one in *PP* presentation mode and the other in *Prob* presentation mode. Both sets of signals contained the same number of signals. Then we equally increased the number of signals in both sets but kept the proportion resp. the resulting a posteriori probability constant. By increasing the number of signals in both signal sets we can examine whether subjects differently take into account weight when given segregated data compared to aggregated data. A Bayesian individual would take into account the weight of information only when inferring from the signal set in *PP* presentation mode and would discard the weight of the signal set in *Prob* presentation mode.

Based on our previous findings we are able to predict subjects' behavior in these situations. We would expect that subjects do take into account weight of evidence when they infer from the set presented in *PP* presentation mode as well as when they infer from the signal set in *Prob*

presentation mode. We have estimated the difference between *b* (regression coefficient for weight of evidence) and *c* (regression coefficient for strength of evidence) in equation (14) when participants face segregated information (*b - c = -0.25*) and when subjects face aggregated information (*b - c = 0.147*). In order to determine predicted probabilities we have to make an assumption regarding the absolute value of either *b* or *c*. We assume two different absolute values for *c* to perform a sensitivity analysis. First, we assume that *c = 1*, which constitutes to the rational weighting of strength. Then we can predict the probabilities we would expect in situations in which subjects are given one signal set in *PP* presentation mode and another one in *Prob* presentation mode. In addition, we predict probabilities assuming *c = 0.81*, which is the regression coefficient that Griffin and Tversky (1992) estimated in their study. Figure 5 illustrates the predicted probabilities along with the observed subjects' probabilities and the rational probabilities.[15]



[15] Rational probabilities are calculated using the probability resulting from a signal set in *Prob* presentation mode as a base rate and updating this probability according to the signals contained in the other signal set.

Figure 5: Observed and predicted probabilities facing two signal sets (M1=Prob, M2=PP). p=x1/N=y1 refers to a signal set containing y1 signals, which leads to a rational a posteriori probability of p=x1. pp=x2/N=y2 refers to a signal set containing y2 signals with a proportion of a-signals of pp=x2.

The figure reveals that in all situations participants on average seemed to believe that both sets of signals exactly set each other off, leading to a subjective probability of 0.5. Observed probabilities are biased in the direction we would expect from our previous findings. The predicted biased probabilities fit the data much better than the rational prediction. This is especially true for the prediction based on the parameter *c* taken from the Griffin and Tversky (1992) experiment. The fact that average reported probabilities were extremely close to 0.5 might be explained by the composition of the signal sets. The parameters of the signal sets were such that if someone is prone to the biases described above then a probability of 0.5 (which is of course false in some of the situations) was an apparent solution. We conclude that subjects believe that a signal set S_1 consisting of N_1 signals, which leads to an a posteriori probability of p_1 represents the same information as a signal set S_2 containing N_2 signals with a proportion of *a*-signals $pp_2 = p_1$. For example, subjects believe that a signal set containing 9 *a*-signals and 6 *b*-signals (*pp=60%*) is equal to a signal set which contains 15 signals and results in an a posteriori probability of 0.6 for state A.

This observation supports the hypothesis that subjects have a frequency understanding of probabilities, which is in line with findings that subjects use averaging heuristics in Bayes' updating tasks when sequentially processing data (Beach, Wise and Barclay 1970; Lopes 1985, 1987, Hogarth and Einhorn 1992). For example, Lopes (1985) shows that when signals are presented sequentially then subjects tend to average across observations to calculate their

subjective posterior probability instead of updating the already known information according to Bayes rule. This leads to posterior probabilities which resemble the frequency of those signals in the sample which support the hypothesis.[16]

This frequency understanding also explains why subjects take into account weight if they are presented with aggregated information. Considering weight is reasonable when subjects focus on signal frequencies and since they do not differentiate between frequencies and probabilities in aggregation tasks, this leads to the observed updating patterns.

*4.6 Consideration of **quality** of the data*

We now look at the implications of quality of information on subjects' updating procedures. First we look at how subjects aggregated two signals of different quality. On the one hand, we gave them two conforming signals of varying quality, and on the other hand they received two contradicting signals of varying quality. Subjects received two signal sets, each set containing one of the signals. Figure 6 illustrates the average probabilities which subjects quoted.

16 The frequency understanding of probabilities found in our experiments refers to the fact that subjects simply set the posterior probability equal to the relative frequency of those signals in the sample which support the tested hypothesis. This updating mode is only feasible when evidence consists of multiple observation. Hence, this is different from experiments which investigate how Bayesian updating is effected by presenting likelihoods and base rates in terms of frequencies (see, for example, Gigerenzer and Hoffrage 1995, Cosmides and Tooby 1996). So in our experiments frequency understanding rather refers to an updating mode given multiple observations than to a way of presenting the given information and leads to worse rather than better updating performance. See also footnote 2.

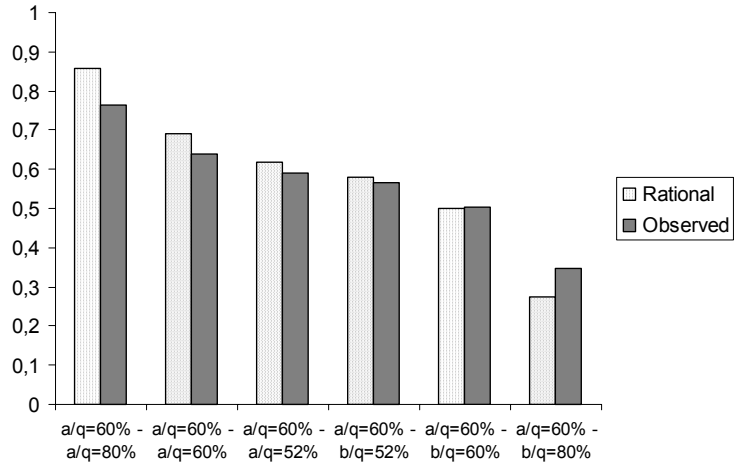


Figure 6: Consideration of quality given two signals.

a/q=x refers to a signal set containing one a-signal with a quality of x percent. b/q=x refers to a signal set containing one b-signal with a quality of x percent.

The figure points out that subjects insufficiently take into account quality. All observed probabilities are biased towards 0.5 even when the quality of the signals was low (*q=52%*). This means that subjects were generally underconfident. This contradicts the findings by Griffin and Tversky (1992), who found slight overconfidence when the information was of weak quality. Furthermore, the adjustments subjects made when the quality of the signals was changed were generally too small as well. The difference in observed probabilities was generally lower than the difference in rational probabilities.

Rational posterior log odds can be written as follows when subjects are given two signal sets, S_1 and S_2, each containing one signal of quality q_1 resp. q_2:

$$log\left(\frac{p(A \mid S_1 S_2)}{p(B \mid S_1 S_2)}\right) = log\left(\frac{p(S_1 \mid A)}{p(S_1 \mid B)}\right) + log\left(\frac{p(S_2 \mid A)}{p(S_2 \mid B)}\right) \tag{9}$$

$$= (n_1^a - n_1^b) log\left(\frac{p_1(a \mid A)}{p_1(a \mid B)}\right) + (n_2^a - n_2^b) log\left(\frac{p_2(a \mid A)}{p_2(a \mid B)}\right) \quad (10)$$

Since signal set S_1 always contained one *a*-signal of quality $q_1 = 60\%$ the equation simplifies to:

$$= 0{,}176 + (n_2^a - n_2^b) log\left(\frac{p_2(a \mid A)}{p_2(a \mid B)}\right) \quad (11)$$

Signal set S_2 always contained a single *a*- or *b*-signal. This means $(n_2^a - n_2^b)$ is either 1 or –1. Therefore, we can estimate the weight that subjects place on the quality of the evidence in signal set S_2 by estimating the following linear regression:[17]

$$log\left(\frac{w_k}{1 - w_k}\right) = a + d\left[(n_2^a - n_2^b) log\left(\frac{p_2(a \mid A)}{p_2(a \mid B)}\right)\right] + u_k \quad (12)$$

Table 6 illustrates the results of the OLS regression:

Table 6: Regression results, quality of the evidence

Parameter	Value	σ	p-value
A	0.127	0.009	<0.001
D	0.656	0.025	<0.001

Adjusted $R^2 = 0.691$

17 Note that weight here cannot be directly compared to the weights which we have estimated for strength and weight of evidence. Here, weight is estimated on the level of log odds using a linear model whereas above we used an exponential model. The use of a linear model is possible here, because we varied only a single factor, namely the quality of the signals. Above multiple factors were varied at the same time, namely strength and weight of evidence.

The results clearly indicate that subjects underweight the quality of the information. The parameter *d* is significantly smaller than 1 (*t = (1-0.656)/0.025 = 13.76*) which represents rational Bayesian weighting of signal quality.

Finally, we investigated how quality influences judgment when the signal sets contain more than one signal. Figure 7 shows the results.

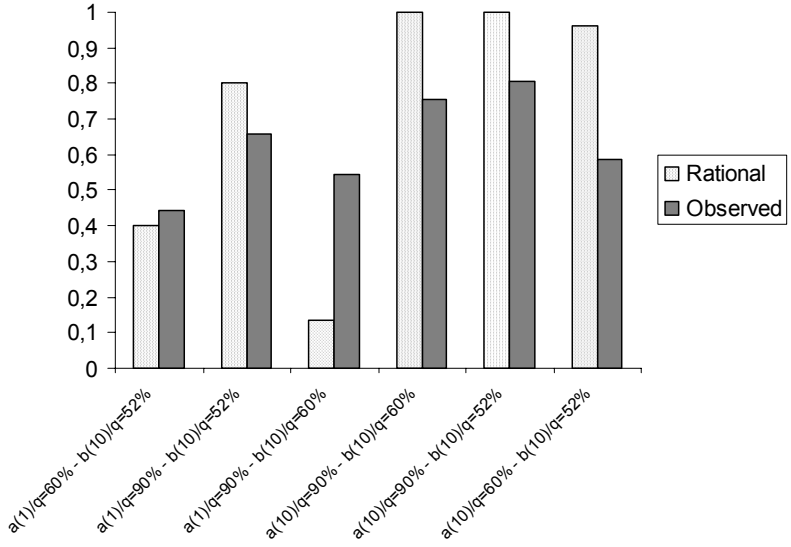


Figure 7: Consideration of quality given two signal sets.

In situations in which one signal set contained just a single signal, adjustment for quality is again insufficient, conforming to the findings when both signal sets contain just one signal. In addition it can be observed that participants are again underconfident as all probabilities are biased towards 0.5.

The three situations on the right hand side of figure 7 reveal a different picture. In these three situations the signal set containing ten *a*-signals leads to a rational a posteriori probability close to 1 which is more or less unaffected by the ten *b*-signals of weak quality. One can see that the participants' subjective probabilities are still conservative relative to the rational probabilities, but

adjustment for quality in terms of change in stated probabilities is too large. This raises the presumption that subjects seem to linearly adjust for quality even though from a rational point of view adjustment should be smaller if one of the pieces of evidence leads to strong inferences, i.e. rational a posteriori probabilities are close to 0 or 1.

In conclusion we can say that the data concerning consideration of quality supports hypothesis 3 in favor of rational Bayesian behavior. Nevertheless, if consideration of quality is viewed as an adjustment of beliefs due to a change in quality of the information, then consideration of quality can sometimes be too extreme.

*4.7 Influence of **expertise** on behavior*

Finally, we want to take a look if expertise in Bayes' updating influences our results. As already noted in section 3.2, 24 of the 51 participants had attended a course in decision analysis. Bayes' rule was one major topic in this course and was extensively covered. Subjects' expertise in Bayes' updating was also supported by the results in the final exam, in which the Bayes' task yielded more than average points. Apart from that both groups were similar. All 24 participants who had attended the decision analysis class were graduate students in business administration. Of the remaining 27 subjects, 5 were undergraduates, 1 was a graduate student in economics and 1 was a non-business student. Since all graduate students were recruited from only 2 classes which are attended at more or less the same stage of graduate studies in business administration, age and prior knowledge should be comparable between both groups.

Redoing the whole analysis separately for the attendants' as well as the non-attendants' group, reveals that attendants on average made the same judgment errors as non-attendants. On some occasions their probabilities are even more biased than those of the non-attendants.

Participants in the class on "Rational Decision Making" performed slightly better in the experiment, on average earning 2.5% more than the rest of the subjects. Nevertheless, the difference in earnings is not statistically significant according to a Mann-Whitney test. Since earnings were directly connected to deviations from rational behavior and were not influenced by any random effects, this provides evidence that attendants were not better calibrated.

In conclusion we can say that the data provides no evidence that expertise in Bayes' updating influences our findings as all identified biases seem to be resistant to learning. Even though this finding seems to be surprising at the first glance, it is consistent with the finding by Griffin and Tversky (1992), who showed that the difficulty of a task has only limited influence on the degree of overconfidence. Furthermore, various previous studies revealed that training can only reduce biases to a limited degree (see, for example, Kagel and Levin 1986; Nisbett, Fong, Lehman and Cheng 1987 and Larrick, Morgan and Nisbett 1990). So it seems reasonable that when subjects are only trained to apply Bayes theorem to simple updating tasks given a single observation, they still commit the same errors as untrained subjects in the more complex situations given in our experiment, even though compared to real life situations the situations are still quite simple.

5 Conclusions

In this experiment we investigated updating behavior when subjects are provided with two sets of signals. Each set contains signals, which provide a hint about the state of world. All signals in a signal set are of the same quality. By altering the way signals of a signal set are presented we tried to find out how people adjust for weight, strength and quality when they are given segregated and aggregated evidence.

Conforming to the findings by Griffin and Tversky (1992) we found that when subjects' updating is based on segregated information they excessively focus on the strength of information

and insufficiently take into account its weight. Sequential presentation as well as explicit information on strength seems to accentuate this discrepancy. Nevertheless, in situations in which weight has only a weak influence on the rational a posteriori probability this relation can be reversed.

When individuals are given aggregated information, i.e. subjects are provided with a posteriori probabilities which can be rationally inferred from the evidence, then subjects also take into account weight which is not rational. The reason for this behavior seems to be a frequency understanding of probabilities.

Consideration of quality of the information is also insufficient, regardless if only one or multiple signals are incorporated. Expertise does not have a significant impact on the biases observed in our setting.

In conclusion we can say that in aggregation tasks people seem to anchor on a summary statistic of the provided evidence such as proportion of specific signals or probability and then adjust for the weight of information. Depending on how information is presented this can lead to both under- and overrepresentation of information weight.

Appendix: Instructions

Thank you for your participation in this experiment on decision making. This experiment will probably last about one hour. Please follow the instructions very carefully in order to earn as much money as possible.

Course of a round

This experiment consists of several rounds. In each round you will be provided with information and subsequently have to make a judgment based on this information.

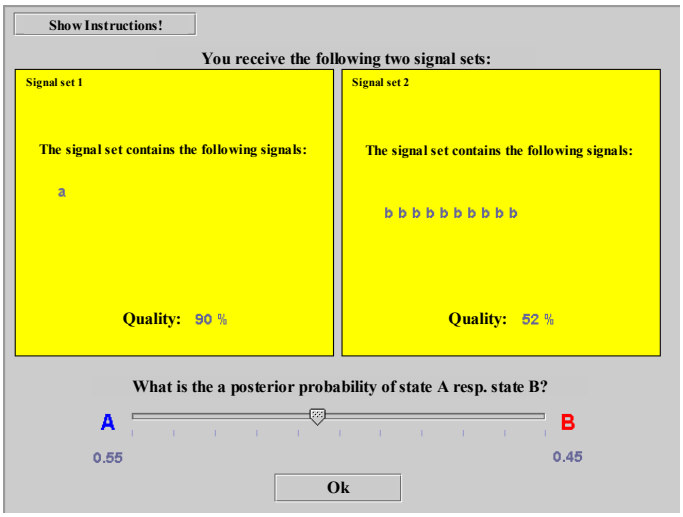
The experimental world can take on two states of nature, labeled **A** and **B**. Both states are a priori equally likely, i.e. p(A) = 0.5 and p(B) = 0.5. At the beginning of each round the computer randomly selects one of these states by using a random number generator. But you do not know which state has been selected. Note that the computer determines the state of nature at the beginning of EACH round. This means, that the current round does not depend on the preceding and/or subsequent rounds.

Signals

The computer gives you a hint which state has actually been selected by providing you with signals. There are **a** and **b** signals. The signals provide a hint which state has occurred, but are not perfect. This means signals can be false. An a-signal gives a hint that state A has occurred, whereas a b-signal provides a hint that state B has occurred. The signals' reliability is indicated by their quality. For example, if the computer displays that the signals have a quality of 60% then signals indicate the state that has actually occurred in 60% of the cases and they indicate the opposing state in 40% of the cases. In other words, a signal that has a quality of 60% can be

interpreted as a random draw from an urn containing 60% “correct” signals and 40% “wrong” signals (draw WITH replacement).

In order for you to be able to make a judgment on the state which has been drawn, you receive multiple signals as described above. The signals you receive are divided into two subsets. Nevertheless, keep in mind that all signals give you a hint concerning the SINGLE state of the world that has been drawn at the beginning of the round. The following figure illustrates the display of the two signal sets:



In the left box you can observe signal set 1 and in the right box signal set 2. One of the signal sets could as well be empty. In this case you have to make a judgment based on only a single signal set.

All signals contained in a signal set are of the same quality. But the quality of the signals can vary between the two signal sets. Therefore please pay close attention to the display of the quality!

The signals contained in each signal set can be presented in 4 different ways. Which presentation modes are feasible along with a short explanation can be found at the end of the instructions.

Your task is to make a judgment concerning the probability that state **A** resp. state **B** has occurred based on BOTH signal sets. In order to quote your subjective probability, please use the slider underneath the two signal sets. Click on the slider and hold down the mouse button to adjust the slider. The closer you move the slider towards the left end of the scale (towards A), the greater your confidence that state A has occurred (your quoted probability for state A rises). Analogously, you express a greater confidence in state B by moving the slider towards the right end of the scale.

Once you have entered your subjective probability you can approve your input by pressing the "Ok"-button and a new round with a NEW DRAW of the state of nature begins.

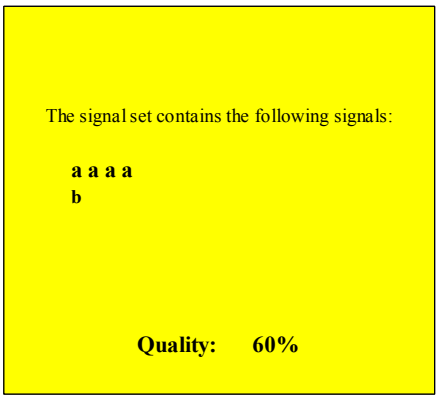
In order for you to have an incentive to spend effort in the updating task and to quote meaningful probabilities, your probability statement is paid for. In each round, given the signals you receive, the rational ("correct") probability p_{rat} that state A resp. state B has occurred can be calculated using laws of probability. Your payment in currency units (cu) is higher the smaller the distance between your quoted probability p_{quote} and this rational probability p_{rat}. The following formula determines your payment in a round:

$$Payment\ in\ cu = 1000 - 2000 \cdot | p_{quote} - p_{rat} |$$

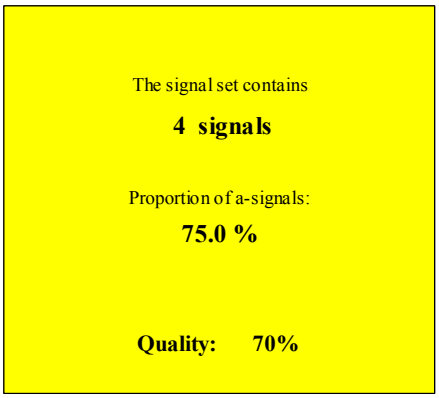
The absolute difference between your quoted probability p_{quote} and the rational probability p_{rat} is multiplied by 2000 and the resulting product is subtracted from 1000.

NOTE: You receive no information concerning your payment in each round. Instead, the computer sums all payments and at the end of the experiment your total earnings are disclosed. At the end of the experiment the total of all your payments is converted to Euro (€) at a rate of 0.00025 €/cu. This means, 1000 cu equal 25 cents.

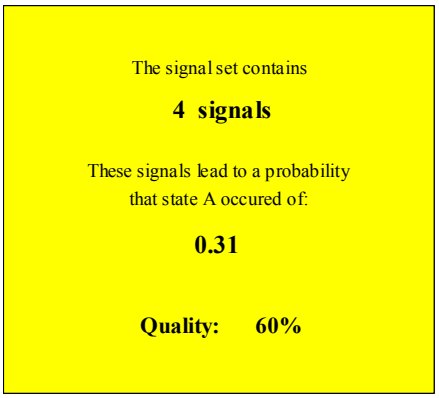
Presentation modes



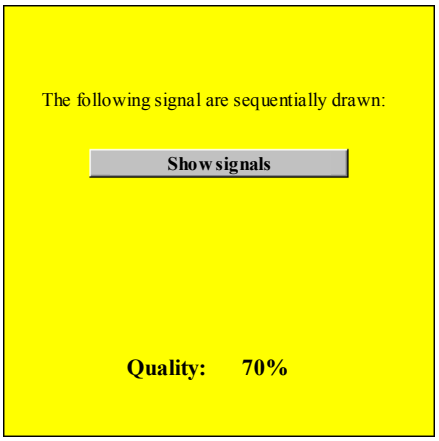
In this mode the signals are all presented at once.
A quality of 60% means that each signal can be interpreted as a draw from an urn which contains 60% "correct" and 40% "wrong" signals. (Draw WITH replacement)



In this mode you can observe the number of signals in the signal set along with the proportion of *a*-signals.
A quality of 70% means that each signal can be interpreted as a draw from an urn which contains 70% "correct" and 30% "wrong" signals. (Draw WITH replacement)



In this mode you can observe the number of signals in the signal set along with the probability that state A has occurred which can be inferred from these signals. This probability is calculated by aggregating the signals in the signal set using laws of probability.
A quality of 60% means that each signal can be interpreted as a draw from an urn which contains 60% "correct" and 40% "wrong" signals. (Draw WITH replacement)



In this mode you can observe the signals contained in the signal set. The signals are drawn sequentially. To start the signal draws please click on the "Show signals"-button.
A quality of 70% means that each signal can be interpreted as a draw from an urn which contains 70% "correct" and 30% "wrong" signals. (Draw WITH replacement)

References

Anderson, N. H. (1981). "*Foundation of information integration theory"*. New York: Academic Press.

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

Barberis, N. C., A. Shleifer, and R. W. Vishny. (1998). "A Model of Investor Sentiment." *Journal of Financial Economics*, 49, pp. 307 - 343.

Beach, L. R., J.A. Wise, and S. Barclay. (1970). "Sample proportions and subjective probability revision." *Organizational Behavior and Human Performance*, 5, pp. 183 - 190.

Bikhchandani, S., D. Hirshleifer, and I. Welch. (1992). "A Theory of Fads, Fashion, Custom, and Cultural Change as Informational Cascades." *Journal of Political Economy*, 100, pp. 92 - 1026.

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, pp. 397 - 414.

Cosmides, L. and J. Tooby. (1996). "Are humans good intuitive statisticians after all? Rethinking some conclusions from the literature on judgment under uncertainty". *Cognition*, 58, pp. 1 - 73.

Cutler, D. M., J.M. Poterba, and L. H. Summers. (1991). "Speculative dynamics". *Review of Economic Studies*, 58, pp. 529 - 546.

Davis, J. H. (1984). "Order in the courtroom". In: Miller, D. J., Blackman, D. G. Blackman and A. J. Chapman, eds., *Perspective in psychology and law*, New York: John Wiley.

Gigerenzer, G. and U. Hoffrage. (1995). „How to Improve Bayesian Reasoning Without Instruction: Frequency Formats". *Psychological Review*, 102, pp. 684 - 704.

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

Hogarth, R. M. and H. J. Einhorn. (1992). "Order Effects in Belief Updating: The Belief-Adjustment Model". *Cognitive Psychology*, 24, pp. 1 - 55.

Hung, A. A. and C. R. Plott. (2001). "Information Cascades: Replication and an Extension to Majority Rule and Conformity-Rewarding Institutions". *American Economic Review*, 91, pp. 508 – 1520.

Kagel, J. and D. Levin. (1986). "The winner's curse and public information in common value auctions". *American Economic Review*, 76, pp. 894 - 920.

Kraemer, C. (2002). "What causes different confidence in equal probabilities in sequential aggregation tasks?" Working Paper, University of Mannheim.

Kraemer, C. and M. Weber. (2002). "To Buy or not to Buy: Why do People Buy too Much Information?" Working Paper, University of Mannheim.

Larrick, R. P., J.N. Morgan, and R. E. Nisbett. (1990). "Teaching the use of cost-benefit reasoning in everyday life". *Psychological Science*, 1, pp. 362 - 370.

Lopes, L. L. (1985). "Averaging rules and adjustment processes in Bayesian inference". *Bulletin of the Psychonomic Society*, 23, pp. 509 - 512.

Lopes, L. L. (1987). "Procedural Debiasing". *Acta Psychologica*, 64, pp. 167 - 185.

Nelson, M. W., R. Bloomfield, J.W. Hales, and R. Libby. (2001). "The Effect of Information Strength and Weight on Behavior in Financial Markets". *Organizational Behavior and Human Decision Processes*, 86, pp. 168 - 196.

Nisbett, R. E., G.T. Fong, D.R. Lehman, and P. W. Cheng. (1987). "Teaching reasoning". *Science*, 238, pp. 625 - 631.

Rabin, M. (2003). "Inference by Believers in the Law of Small Numbers". *Quarterly Journal of Economics*, forthcoming.

Tversky, A. and D. Kahneman. (1971). "The belief in the law of small numbers". *Psychological Bu*lle*tin*, 76, pp. 105 - 110.