Beyond what you see: How expectations and experience shape information search, preferences, and judgments.

Inauguraldissertation
zur
Erlangung der Würde
einer Doktorin der Philosophie
vorgelegt der
Fakultät für Psychologie
der Universität Basel

von

Janine Christin Hoffart

aus Heiligenhaus, Deutschland

Basel, 2018

UNI
BASEL

Genehmigt von der Fakultät für Psychologie

auf Antrag von

Prof. Dr. Jörg Rieskamp
Prof. Dr. Benjamin Scheibehenne

Basel, den 21.03.2018

_____
Prof. Dr. Roselind Lieb

UNI
BASEL

This dissertation framework is based on four manuscripts:

Hoffart, J. C., Rieskamp, J., & Dutilh, G. (2017). *The influence of sample size on preferences from experience*. Manuscript submitted for publication.

Hoffart, J. C., Olschewski, S., & Rieskamp, J. (2018). *Reaching for the star ratings: A partly Bayesian account of how people integrate consumer ratings*. Working paper.

Hoffart, J. C., Rieskamp, J., & Dutilh, G. (in press). How environmental regularities affect people's information search in probability judgments from experience. *Journal of Experimental Psychology: Learning, Memory, and Cognition*.

Hoffart, J. C., & Scheibehenne, B. (2018). *Perceived riskiness of clinical trials: How incentives influence probability expectations of side effects*. Manuscript submitted for publication.

Declaration

I, Janine Christin Hoffart (born November 18 1987 in Velbert, Germany) hereby declare the following:

i My cumulative dissertation is based on four manuscripts: one accepted for publication (Manuscript 3 of this dissertation), one under revision (Manuscript 1 of this dissertation), one submitted for publication (Manuscript 4 of this dissertation), and one in preparation (Manuscript 2 of this dissertation). I contributed substantially and independently to all manuscripts in this dissertation and have been primarily responsible for the ideas, data collection, analyses, and writing of the manuscripts.

ii I only used the resources indicated.

iii I marked all the citations.

Basel, 20th January, 2018

Janine Hoffart

Acknowledgments

Many people have supported me on the journey of my PhD. I am grateful to all of them. I thank Jörg Rieskamp and Gilles Dutilh for their supervision, time, and funding. I thank my collaborators Benjamin Scheibehenne –who also is the second reviewer of my dissertation– and Sebastian Olschewski –who also has been the best office colleague during the past years– for their great ideas and their enthusiasm. I thank all my (PhD) colleagues – Sebastian, Marin, Laura, Michael, Oli, Rebecca, Bettina, Janina, Jana, Ash, and Nathaniel – for the stimulating discussions we had and for making my time at the Center for Economic Psychology and at several conferences fun. I thank my *NRW* friends, especially Kristina, Lari, Janina, and Bernie. Although, many kilometers separate us, I can always count on you. More than once, you brought me back on track and reminded me that there is a life outside the PhD craziness. I thank my *Basel* friends, especially Anissa, Alisa, Eva, and Claudia who helped me to feel at home in Basel. I thank my family, Mama, Papa, Christian (+ Christiane, Luisa and Niklas) and Tobias (+ Anna) for their endless support and trust in me. Last but not least, I thank Johannes for his patience, and encouragement. You have been my rock during the past years.

Abstract

When people make decisions under uncertainty, they can relate to their *expectations* about and *experience* with the choice options. In decisions from experience, where people actively search for information, they typically rely on a few observations. I approached the question of *why* people feel confident to rely on only a few observations by investigating *what* influences information search, preferences, and judgments. In Manuscript 1, we investigated how people's valuations of gambles change with growing experience. In two experiments, people observed different numbers of outcomes (sample sizes) of unknown gambles and then indicated for how much money they would sell each gamble. We contrasted a Bayesian model that predicts that with growing sample size selling prices change with a model that predicts that selling prices are stable across different sample sizes. People differed in how they treated sample size: Roughly half of the people integrated sample size as predicted by a Bayesian model and half did not. This finding was replicated in the study reported in Manuscript 2, in which people chose between two hotels based on summaries of customer ratings. We found that people's ability to deal with statistical information correlated with how well a Bayesian model described their data. In Manuscripts 3 and 4, we studied how people's expectations influence judgments and information search. In real life, larger rewards typically occur with lower probabilities than smaller rewards. In two experiments reported in Manuscript 3, people judged how likely they thought they were to win different monetary amounts in psychological studies. Following the environmental regularity, people made higher estimates for smaller than for larger rewards. In a third experiment where people made probability judgments from experience, they searched less when this expectation about a negative probability–reward correlation was met. In Manuscript 4, we demonstrated that also in situations involving real-life decisions, people expect structural regularities. People judged how many participants they expect to experience side effects in clinical trials with different incentives. When incentives were larger, people expected more people to experience side effects than when incentives were smaller. In sum, in this dissertation, I showed that people differ in how they integrate sample–size based uncertainty and that both experiences made in the laboratory and expectations gained outside the laboratory influence information search and judgments.

Introduction

Inescapably, risk is part of life: Should one cross the street when the traffic light is red or wait until it turns green and risk missing the tram? Should one invest one's retirement money in risky stocks or play it safe and invest it in government bonds? Minute by minute, humans make similar decisions where they do not know *with certainty* what outcomes they will receive. How people make such decisions has been studied extensively in the past century (e.g., Kahneman & Tversky, 1979; Von Neumann & Morgenstern, 1947). Most empirical findings that describe how people choose between risky prospects are based on experimental paradigms assessing *decisions from description*, the *drosophila* of decision-making research (Hertwig, 2012). How people, here, "Mrs. Thomas", make decisions from description, is often investigated with risky gambles for which all possible outcomes and the associated outcome probabilities are described. For instance, the choice between Gamble A and Gamble B can be described as follows:

Gamble A: $82 for sure

or

Gamble B: $100 with probability .8 or $10 with probability .2

This formulation of decision problems describes several choices Mrs. Thomas faces on a daily basis: For instance, every morning, she decides whether to take her umbrella when leaving the house. To make this decision, she looks up the weather forecast that precisely informs her about the probability and the amount of precipitation during the day. However, in many situations—for instance, when she decides where to eat lunch—Mrs. Thomas cannot easily look up the outcome distributions of the set of choice options she can choose from. Luckily, in familiar situations, Mrs. Thomas can relate to her past *experience* when making decisions.

The Role of Experience in Judgments and Decision Making

In his influential book *Risk, Uncertainty, and Profit*, Knight (1921) made a theoretical distinction between three different classes of situations involving risk and uncertainty:

Situations in the first class involve *a priori probabilities*. These are situations in which probabilities of outcomes are known or can be precisely estimated, as is the case in decisions from description. Situations in the second class involve *estimates*. These are situations in which probabilities of outcomes are unknown and cannot be estimated meaningfully. Situations in the third class involve *statistical probabilities*. These lie on the continuum between the first two categories and involve situations in which probabilities are not known but can be approximated empirically. Statistical probabilities can, for instance, be approximated from experience by relying on observed past outcomes (Camilleri & Newell, 2013; Hau, Pleskac, & Hertwig, 2010; Lopes, 1983). Also when Mrs. Thomas decides where to have lunch by remembering past visits at different restaurants, this decision is based on statistical probabilities.

In the 1950s and 1960s, researchers devoted much attention to topics involving statistical probabilities, such as the question of how people learn probability and outcome structures over time (e.g., Edwards, 1961; Estes, 1950; Katz, 1962; Myers & Katz, 1962; Suydam, 1965). By the end of the century, experimental psychology had relegated this topic to second place and instead was focusing mostly on research involving *risk* (i.e., a priori probabilities) and complete *uncertainty* (i.e., estimates; cf. Hertwig, 2012). But as in many situations people make judgments and decisions on the basis of statistical probabilities, recently researchers have again started to study in depth how people make decisions between risky gambles from experience (e.g., Hertwig, Barron, Weber, & Erev, 2004). An often-used paradigm to investigate how people make decisions from experience is the sampling paradigm (Hertwig et al., 2004). In this framework, people learn how the choice options are structured by actively exploring the gambles' outcome distributions by sampling outcomes. At the beginning of a trial, gambles are presented as empty boxes that decision makers can sample from. Every time decision makers click on a box, they receive a nonconsequential outcome from the box they clicked on. Decision makers typically draw outcomes from the boxes until they feel confident making a judgment or decision on the basis of the observed outcomes. Such experience-based paradigms have been receiving much attention, as people's choices have been found to differ systematically from choices made from description: Whereas when

making decisions from description people behave as if they overweight rare events (Kahneman & Tversky, 1979), when making decisions from experience people behave as if they underweight rare events (e.g., Hau et al., 2010; Hau, Pleskac, Kiefer, & Hertwig, 2008; Hertwig et al., 2004; Ungemach, Chater, & Stewart, 2009; Wulff, Mergenthaler-Canseco, & Hertwig, 2017). This finding has raised the question of how people form preferences based on the sampled outcomes in experience-based paradigms.

Information Integration From Experience

In recent years, several theories have been proposed that describe how people integrate the sampled outcomes to form beliefs about gamble values. Generally, two classes of models can be distinguished: associative learning models—for example, the *instance-based learning model* (e.g., Dutt & Gonzalez, 2012; Gonzalez & Dutt, 2011; Lejarraga, Dutt, & Gonzalez, 2012) and the *value updating model* (Hertwig et al., 2004)—that describe how over the course of sampling, the propensity for choosing a gamble changes based on the outcomes they observe; and heuristics—for example, the *natural mean heuristic* (Hertwig & Pleskac, 2010)—that describe how people use cognitive shortcuts to simplify the sampled information.

Mostly, such learning models and heuristics have been used to describe how people make decisions in *free-sampling* paradigms where decision makers sample until they feel confident making a decision. Interestingly, people typically rely on relatively few outcomes (e.g., Hau et al., 2008; Hertwig et al., 2004; Rakow, Demes, & Newell, 2008). Consequently, the total sample often does not represent a gamble's outcome distribution correctly (Hadar & Fox, 2009). Whereas several attempts have been made to explain *why* people feel confident making decisions based on small samples (e.g., Hau et al., 2008; Hertwig et al., 2004; Rakow et al., 2008), little effort has been made to understand *whether* and *how* larger, as compared to smaller, samples alter people's preferences (but see, e.g., Hau et al., 2008). In the first part of my dissertation, I focused on this question by investigating how *sample size* (i.e., number of sampled outcomes) influences preferences.

The Role of Sample Size in Experience-Based Paradigms

Typically, Mrs. Thomas eats Thai food for lunch. But one rainy Monday, the Thai take-out restaurant was closed and Mrs. Thomas decided to try out a new Indian restaurant. She heard people speaking well of this new restaurant. Despite her positive expectations, Mrs. Thomas was very disappointed with the food. Will she ever eat at the Indian restaurant again?

Bayesian information integration: A way of dealing with uncertainty. In the past, Mrs. Thomas learned that single recommendations and single experiences do not always reflect underlying qualities accurately. And as many colleagues had recommended the Indian restaurant, Mrs. Thomas considers her bad experience as being unrepresentative of the quality of the restaurant. She will give it another try and eat lunch there again in the future.

Such behavior can be explained by the principles of Bayesian information integration, which suggest that people incorporate prior beliefs (i.e., expectations) and new information to form posterior beliefs. Figure 1 shows this principle by illustrating how Mrs. Thomas updates her prior beliefs about the reward probability of a risky gamble with two outcomes (zero and reward) based on new observations. Initially, she may assume that the reward probability most likely is 50%. However, there is uncertainty about this estimate as displayed in the first column of Figure 1. A prior belief may be relatively weak as illustrated in Figure 1 (a) and hence widely distributed, indicating that Mrs. Thomas can very well imagine that the empirical reward probability might strongly differ from 50%. In a different situation however, her prior belief may be stronger and she is more certain that the empirical reward probability is close to 50%. This is indicated by a belief distribution that is more peaked around 50% as illustrated in Figure 1 (d). Mrs. Thomas then samples new information in the form of outcomes (second column of Figure 1). She integrates this information with her prior belief to form a posterior belief (Figure 1 (b), (c), (e), and (f)). Bayesian principles postulate that first, the stronger the prior beliefs are, the more information is needed to overcome these prior beliefs (compare Figure 1 (b) and (c) where the prior belief was weak as illustrated in (a) with Figure 1 (e) and (f) where the prior belief was stronger as illustrated in (d)). And second, the more information is available, the more closely people's posterior beliefs will converge to the underlying empirical reward probability of a choice option (compare Figure 1 (c) and (f) where many new

observations were made with Figure 1 (b) and (e) where only few observations were made).

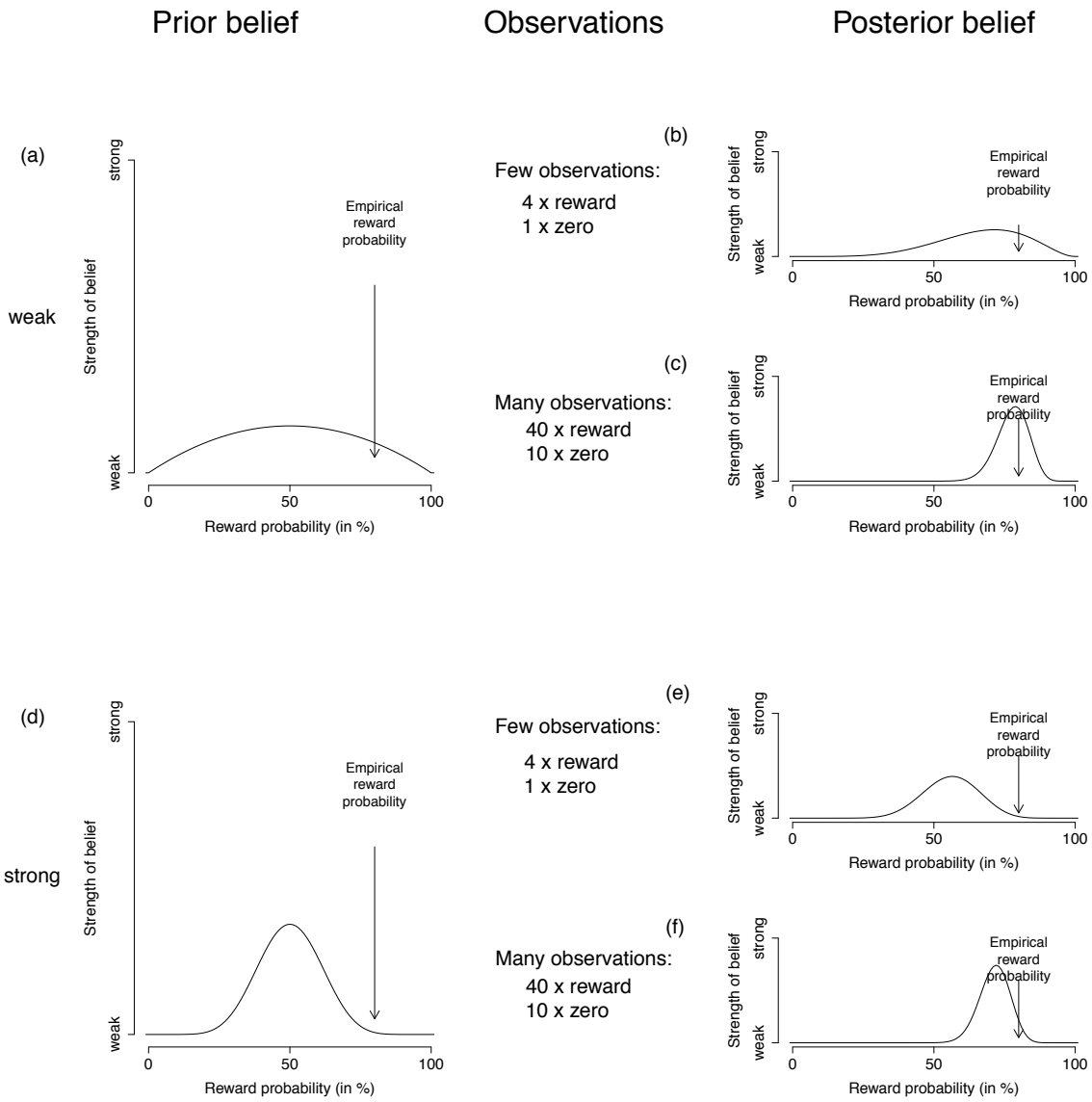

Figure 1. Illustration of the Bayesian updating process. The first column displays Mrs. Thomas's (the Bayesian updater) expectation about a two-outcome (zero; reward) gamble's reward probability. (a) displays a weak and (d) a strong prior belief, both peaked at 50%. Beliefs get updated with new information (second column) to form posterior belief distributions (third column). The arrows display the true reward probability of 80% that is assumed to be unknown to Mrs. Thomas and has to be learned through experience.

In many situations, people's behavior can be described well with Bayesian principles: when learning concepts and words (e.g., Tenenbaum, Griffiths, & Kemp, 2006; Xu & Tenenbaum, 2007), in conditioning (e.g., Courville, Daw, & Touretzky, 2006) and object perception tasks (e.g., Kersten, Mamassian, & Yuille, 2004), when judging the performance of students (Fiedler, 2000), or when making consumer choices (e.g., De Martino,

Bobadilla-Suarez, Nouguchi, Sharot, & Love, 2017). However, people seem not always to integrate information in line with Bayesian principles.

Belief in the law of small numbers: Ignoring sample size. Mrs. Thomas did not eat lunch alone. At the Indian restaurant, she met her close friend Mr. Amos. Mr. Amos also did not enjoy the food, but he decided never to eat at this restaurant again.

Mr. Amos's decision reflects the idea of *the belief in the law of small numbers*. This principle describes empirical findings that people often ignore sample size when making judgments: They rely on only a few experiences and treat these as if they comprehensively described the outcome-generating distribution (Tversky & Kahneman, 1971). For instance, Griffin and Tversky (1992) showed that when people observed a number of coin tosses, they ignored sample size (i.e., number of coin tosses) when judging whether coins were biased. Also, in judgment tasks, people were often found not to weight sample sizes sufficiently (e.g., Kutzner, Read, Stewart, & Brown, 2016; Obrecht, Chapman, & Gelman, 2007).

In the first manuscript of my dissertation, we investigated whether and how people integrate sample sizes when making gamble valuations from experience. In particular, we studied whether people integrate sample-size-based uncertainty in their judgments as proposed by Bayesian principles or whether they ignore sample-size-based uncertainty as proposed by the principle of the belief in the law of small numbers.

Manuscript 1: The Influence of Sample Size on Preferences From Experience

Hoffart, J. C., Rieskamp, J., & Dutilh, G. (2017). *The influence of sample size on preferences from experience*. Manuscript submitted for publication.

People often make judgments and decisions based on their own experiences. How people solve such tasks has been investigated experimentally with experience–based tasks. As described above, in such tasks people typically learn about the outcome distributions of gambles by repeatedly drawing outcomes from the unknown gambles (Hertwig et al., 2004). This way of learning about the gambles' outcomes and associated probabilities differs from the decisions-from-description approach, where people see complete summaries of the gambles' outcomes and probabilities.

One of the most prominent findings when comparing decisions from description with decisions from experience is that people's choices systematically differ between the two paradigms. While people behave as if they overweight rare events when making decisions from description (Kahneman & Tversky, 1979), they behave as if they underweight rare events when making decisions from experience (Hertwig et al., 2004; Wulff et al., 2017). The explanation suggested for this difference in people's choice patterns is often that, when making decisions from experience, people typically rely on relatively few observations (e.g., Hau et al., 2008; Hertwig et al., 2004; Rakow et al., 2008).

Although small sample sizes are frequently mentioned as an explanation for the description–experience gap, it is unclear how sample size *itself* influences preferences and judgments in experience-based tasks. To address this question, we conducted two experiments in which people sampled from unknown gambles. Every gamble had two outcomes, a gain of varying magnitude and a zero outcome. In each trial, people got to sample a *predefined* number of times from *one gamble at a time*. People then indicated for how much they would sell the gamble and how confident they were in their response. In Experiment 1, people saw the gambles' outcomes before they started sampling; in Experiment 2, they did not. Hence, in this second experiment, we further increased the uncertainty compared to Experiment 1 and applied a design that more closely resembles classic experience-based tasks where before they start sampling, people most of the times do not know a gamble's outcomes. Crucially, we presented individual gambles repeatedly with different sample sizes and ensured that the observed outcomes in every trial represented a gamble's empirical outcome distribution. The main goal of this study was to clearly compare how the amount of information influences preferences. As a side goal, we also wanted to compare decisions from experience with decisions from description in the domain of gamble valuations because so far, most studies have focused on comparing decisions from description with decisions from experience (but see N. J. S. Ashby & Rakow, 2014; Golan & Ert, 2014; Pachur & Scheibehenne, 2012). To do so, in a second block people indicated their selling prices for the gambles from description, in other words, in a situation in which all outcomes and associated probabilities were described.

To derive predictions for how sample size might influence people's preferences, we

developed two models: one that captures the idea of Bayesian information integration and another that captures the idea of the belief in the law of small numbers. In a nutshell, the Bayesian updating model assumes that before sampling, people expect every reward probability between 0% and 100% to be equally likely. As they sample, they update their belief distribution about the reward probability with the new incoming information. People's selling prices depend on the mean of the posterior probability of the reward distribution[1]. Only attending the reward probability is possible as the second outcome is always zero. Crucial predictions of the model are (a) that people's selling price changes with growing sample size and will approach their *true selling price* the more outcomes they observe, and (b) that people will be more confident about their valuations with growing sample size. Both prediction are derived from the fact that in the model, it is assumed that before sampling, people's beliefs about the reward probabilities are uniformly distributed and become increasingly peaked over the empirical reward probability as they sample more.

In contrast to the Bayesian updating model, the model capturing the belief in the law of small numbers assumes that people always treat the sample of outcomes they observed as being representative of the gamble's outcome distribution. In our experiments, we ensured that in every trial, the frequency of rewards and zero outcomes corresponded to a gamble's empirical outcome distribution. Therefore, neither selling prices nor confidence of believers in the law of small numbers should change across sample sizes. Alternatively, although not directly following from the model specifications, believers in the law of small numbers may perceive fewer outcomes as easier to process and therefore are more confident in their judgments when sample sizes are small. This hypothesis was based on the findings of Griffin and Tversky (1992), who asked people to judge whether coins were biased based on previous observations. They found that people treated a sample of outcomes as *always* representative of the outcome distribution irrespective of its size but still were more confident with their judgments when the number of coin tosses was small.

When we modeled the selling prices on an aggregate level, we did not find an effect of sample size on selling prices. However, when we modeled the data on an individual level, we

[1]Precisely, the selling price is simplified by converting the product of the mean of the posterior probability of the reward distribution and the utility of the gain back to the monetary scale.

found that slightly less than half of the participants were best described by the Bayesian model. The other half followed the principles of the belief in the law of small numbers. Also, people's confidence judgments differed between the groups: Bayesian updaters (i.e., people better described by the Bayesian model) were more confident in their judgments when they had observed more samples. In contrast, confidence of believers in the law of small numbers did not change across sample sizes. Further, we found a difference between valuations from description and valuations from experience. In both situations, people behaved as if they overweighted rare events. However, people did so more in valuations from experience than in valuations from description. The finding that people behaved as if they overweighted rare events more strongly from experience than from description contradicts the description–experience gap found in the choice literature: There, people typically chose as if they overweighted rare events from description and as if they underweighted rare events from experience (but see Glöckner, Hilbig, Henninger, & Fiedler, 2016, for a previous observation of a reversed description–experience gap).

In sum, our main findings show that individuals used different strategies when making valuations from experience: While some people behaved according to Bayesian principles, others did not. This finding confirms previous research showing that different people apply different strategies in a range of tasks (e.g., F. G. Ashby, Maddox, & Lee, 1994; Fischbacher, Hertwig, & Bruhin, 2013; Lewandowsky & Farrell, 2011). Arguably, it is more complex to integrate sample size and valence of outcomes in a Bayesian way than to ignore sample size altogether. Therefore, we hypothesize that potentially statistical numeracy, that is, how well people understand statistical information, is related to which strategy people use. Supporting this idea, previous research showed that statistical numeracy correlates with performance in Bayesian reasoning tasks (Brase & Hill, 2017). We tested this prediction in the study reported in Manuscript 2, where we explored how people make consumer decisions based on other customers' experiences, described in the form of user ratings.

Other People's Experience

Often Mrs. Thomas and Mr. Amos have not had any experiences with the set of alternatives they can choose from. For instance, when booking a hotel for their next vacation in a country they have never visited, they cannot relate to previous experiences with the hotels. However, both excellently navigate the Internet and know how to consult Web pages to look up summaries of other people's satisfaction with stays at the hotels. Such rating summaries, at the minimum, contain information about how people on average have rated a product (quality) and how many people have rated the product (sample size). Similarly to a situation in which people make choices from their own experiences, people can use these summaries in a Bayesian way to update their beliefs about competing products and form preferences. From a statistical perspective not only average ratings but also the number of ratings provide helpful information as average customer ratings that are based on many people's opinions provide more reliable estimates of product qualities. Following Bayesian principles, a few positive ratings signal greater *downside risk* than many positive ratings. This means that the few ratings may not represent the quality accurately and the product may in reality be much worse than suggested. However, if a product has been rated many times, the decision maker can be more certain that the true quality is close to the average ratings. This implies that when choosing between two well-rated products, a decision maker should be more likely to choose the more-often rated product. However, when choosing between two poorly rated products, a decision maker should choose the product that has been rated less often. In this situation, few ratings signal greater *upside potential* than many ratings. This means that the less often rated product may be in reality much better than suggested by the few ratings.

Manuscript 2: Reaching for the Star Ratings: A Partly Bayesian Account of How People Integrate Consumer Ratings

Hoffart, J. C., Olschewski, S., & Rieskamp, J. (2017). *Reaching for the star ratings: A partly Bayesian account of how people integrate consumer ratings*. Working paper.

Online platforms allow users to easily retrieve information about how previous customers have liked products (Fang, Wen, George, & Prybutok, 2016). The growing

importance of such online platforms is confirmed by findings that show that people consult reviews of other people when planning trips (Gretzel & Yoo, 2008). Positive reviews increase hotel bookings (Ye, Law, & Gu, 2009) and purchases of books (Chen, 2008). On online-rating platforms, customer experiences are described with summaries about how much they have liked the products on a scale of 1 (worst possible rating) to 5 (best possible rating) points. On most platforms, one can see how individual customers have rated a product, how many people have rated the product, and what the average rating is.

As described above, Bayesian principles suggest an interaction between average ratings and the number of ratings (i.e., sample size): On the lower end of the scale, people should prefer options with fewer ratings, even if these were on average rated slightly worse than the more often rated options. This pattern changes on the higher end of the scale, where Bayesian principles suggest that people should prefer more often rated options even if on average they have been rated slightly worse than less often rated options. Alternative hypotheses suggest that people ignore the number of ratings (belief in the law of small numbers) or even treat the number of ratings as a cue for quality. This latter hypothesis follows from the assumption that people heuristically treat the popularity of a product as a signal of its quality (Powell, Yu, DeWolf, & Holyoak, 2017). Generally, people like to make choices that conform with other people's choices (Schöbel, Rieskamp, & Huber, 2016; Zhang, Ye, Law, & Li, 2010). Normative social influence as an explanation suggests people deem it desirable to be approved by others and to conform with other people's opinions and behavior. Informal social influence, on the other hand, describes how people infer useful information from the choices of other people (Deutsch & Gerard, 1955).

Previous literature has provided mixed results with respect to how people integrate the quality ratings of products (i.e., average ratings) and the number of times the products have been rated: Some studies have suggested that people ignore sample size altogether (Obrecht et al., 2007), others have found evidence that people integrate sample size and quality in line with Bayesian principles (De Martino et al., 2017), and still others have suggested that people treat sample size as a cue indicating the quality of products (Powell et al., 2017).

We propose that a reason for these diverging findings with regard to how people

integrate the number of ratings in their choices may be that individual people use different strategies. This has been shown in Manuscript 1 of this dissertation and in other tasks (e.g., F. G. Ashby et al., 1994; Fischbacher et al., 2013; Lewandowsky & Farrell, 2011). To investigate how people integrate sample size and average ratings when making choices based on online ratings, we conducted a study in which we asked people to choose between two hotels based on summaries containing information about how previous visitors had rated the hotels. We varied the hotels' average ratings and how often the hotels had been rated. Crucially, we manipulated whether the less often rated hotel was rated as being equally good as, better than (+ .5 points), or worse than (- .5 points) the more often rated hotel. In addition, we investigated people's choices in trials where one hotel was rated very well (bad) and the other hotel was not rated yet. We analyzed people's data by testing three different cognitive models for each individual participant: a Bayesian updating model predicting an interaction between sample size and average ratings, a model ignoring sample size, and a model treating sample size *and* average rating as quality cues where higher values signal higher quality. An interaction between number of ratings and the average ratings, as predicted by Bayesian principles, arguably involves more cognitive effort than ignoring sample size or treating it as a constant cue for quality. Therefore, we hypothesized that people's scores in a statistical numeracy test (the Berlin numeracy test, Cokely, Galesic, Schulz, Ghazal, & Garcia-Retamero, 2012) correlates with how well their choices can be described by the Bayesian model. This hypothesis is in line with previous findings showing that people higher in numeracy solve tasks more in line with Bayesian principles (e.g., Brase & Hill, 2015; Chapman & Liu, 2009; Rheinberger & Hammitt, 2015).

On the aggregate level, people chose according to Bayesian principles in most choice situations—however, not in all. In particular, in situations where a slightly better hotel had also been rated more often, people's choices deviated from Bayesian principles. In these situations, Bayesian principles suggest that on the lower end of the scale, people prefer the less often rated hotel and on the higher end of the scale, the more often rated hotel. However, people preferred the better hotel across all average ratings. When we modeled the data on the individual participant level, in line with our predictions, the strategies that people used

differed substantially between individuals. Interestingly, only very few people ignored sample size altogether. Instead roughly 40% of the participants were qualitatively best described by the Bayesian model and 50% by the model treating sample size as a quality cue. As predicted, people's score in the numeracy test positively correlated with how well the Bayesian model described their choices.

Although quantitatively most people were best described by a model that treats the number of ratings as indicating a hotel's quality, qualitatively in many choice situations, the choices suggested otherwise: When the plotted the model predictions against people's choice proportion separately for the best fitting model, the data of people best fit by the model suggesting that people treat sample size as cue for quality, deviated from the model predictions in some choice situations substantially. We identified that qualitatively the data were better described by a strategy assuming that (a) when the cues (i.e., average rating and number of ratings) were not–conflicting (the on average better rated hotel was also rated more often, or both hotels were rated equally but one hotel more often), people chose the hotel with higher cue values. This choice pattern conforms with the model treating sample size and average ratings as independent cues for quality. And (b) when the cues were conflicting (the on average better rated hotel was rated less often) or one cue was not available (one hotel had not been rated yet), people chose the less often or never rated hotel on the lower end of the scale and the more often rated hotel on the higher end of the scale. This choice pattern conforms with Bayesian principles. On the basis of these observations, we developed a post hoc Bayesian decision-tree model that assumes that in some situations, people choose in line with Bayesian principles and in others they do not. In sum, the model suggests that people take the cue values *average rating* and *number if ratings* as an indicator of quality. If the cues do not conflict, people choose the option with the higher cue value. Only if information is missing or the cues point in opposing directions do people engage in more detailed elaboration of the choice options that we described with the Bayesian model.

We tested this Bayesian decision tree against the full Bayesian model, the model that incorporated the belief in small numbers, and the model treating sample size as a quality cue. The Bayesian decision tree described the choice behavior of roughly half of the people best.

Interestingly, another 30% were—quantitatively and qualitatively—best described by the full Bayesian model, meaning that only 20% ignored sample size or treated it as a quality cue consistently. As a critical test of our decision-tree model, we applied it to recently published data of two experiments reported in Powell et al. (2017). In their original study, the authors suggested that people *always* treat the number of ratings as a cue signaling quality. However, when we modeled the data on the individual participant levels, again we found considerable differences in the strategies people use. In particular, we identified that many people (30% Experiment 1 and 36% Experiment 2) were best described by the Bayesian decision tree and that their preferences followed Bayesian information integration in conflict situations.

In sum, in the first part of my dissertation, I investigated how people integrate sample-size-based uncertainty in their judgments. I established that interindividual differences as well as situational factors influence whether people integrate information according to Bayesian principles. Crucially, in the second manuscript, we identified cue congruence and statistical numeracy as factors that influence whether people engage in more detailed Bayesian reasoning. This finding can help explain conflicting results reported in the literature with regards to how people integrate the number of customer ratings when making decisions between consumer products (compare, for example, De Martino et al., 2017; Obrecht et al., 2007; Powell et al., 2017). In the second part of my dissertation, I focused on investigating how people's expectations further influence judgments under uncertainty and search effort in experience-based tasks.

Experience and the Structure of the Environment

As early as in the first half of the 20th century, Brunswik (1947, 1955) claimed that psychological processes are adapted to the environment. Today there is acknowledgment of this proposition in the growing consensus that human cognition can be considered boundedly rational (e.g., Gigerenzer & Selten, 2001). To economize on scarce cognitive resources, in several situations people have learned to rely on simple yet productive heuristics (e.g., Gigerenzer, Todd, & the ABC Research Group, 1999; Todd & Gigerenzer, 2012).

Coming back to our example, Mrs. Thomas and Mr. Amos have learned that their

everyday environment follows structural regularities. For instance, when making investment decisions, they expect that options with higher risk (i.e., options with more variable payouts) in the long run on average give higher returns than options with lower risk. This expectation is based on the risk–return trade-off that has been observed in financial markets and forms the basis of Markowitz's (1952) mean–variance model. Sunden and Surette (1998) have reported that people who are more willing to trade risk for return also report more often investing mostly in stocks than people that are less willing to trade risk for return. To date, many researchers have applied risk–return models to predict behavioral and neural data (e.g., Mohr, Biele, Krugel, Li, & Heekeren, 2010).

Also, rewards and probabilities are tied in several domains. Early on, Edwards (1962) claimed that "our world is so constructed that the more desirable objects are harder to get" (p. 49). Pleskac and Hertwig (2014) followed up on this notion and empirically demonstrated that reward probabilities and reward magnitudes are indeed systematically correlated in many real-life domains, such as roulette games or horse racing: Larger rewards are typically less likely than smaller rewards. Furthermore, they showed that people have adapted to this regularity and exploit their knowledge, making use of a *risk–reward heuristic*. In a behavioral experiment, people gave estimates of the chances of winning in monetary gambles. To solve this task, people knew the reward magnitudes, in other words, how much they could win in the gamble, and that costs were associated with gambling. Importantly, the reward magnitude varied between conditions but the costs were fixed and the same in all conditions. People's estimates of the reward probabilities followed the ecological regularity and they estimated the chances of winning larger monetary rewards as being lower than the chances of winning smaller monetary rewards (Pleskac & Hertwig, 2014). Generally, this expectation is in line with the concept of fair bets that assumes that across risky situations, the expected values (costs – outcomes $\times$ outcome probabilities) are constant (Osherson, 1995).

In the work reported in Manuscript 3, we extended previous findings on the risk–reward heuristic and tied it to work in the experience-based literature. First, we showed that also in gambling tasks in which no costs are associated with gambling, people expect rewards and probabilities to be negatively correlated. Further, we investigated whether

people's expectations about regularities in the environment influence their search effort in experience-based probability judgment tasks.

Manuscript 3: How Environmental Regularities Affect People's Information Search in Probability Judgments From Experience

Hoffart, J. C., Rieskamp, J., & Dutilh, G. (in press). How environmental regularities affect people's information search in probability judgments from experience. *Journal of Experimental Psychology: Learning, Memory, and Cognition*.

Brunswik (1947, 1955) argued that experimental research needs to take people's adaptation to natural environments into account when designing experiments. People enter experiments with beliefs and expectations about the microcosm of the experimental situation and these expectations likely influence their behavior. Indeed, in several tasks, for instance, function learning, people's behavior is influenced by whether their expectations are met by statistical regularities of experimental stimuli (Busemeyer, Byun, Delosh, & McDaniel, 1997). Crucially, following Bayesian principles, when the environment contradicts people's expectations, people need very much or strong evidence to overcome their beliefs (Alloy & Tabachnik, 1984).

Surprisingly, in the decisions-from-experience literature, so far it has mostly been neglected that expectations about statistical regularities of stimuli in experiments may influence people's behavior (but see Lejarraga, Hertwig, & Gonzalez, 2012; Mehlhorn, Ben-Asher, Dutt, & Gonzalez, 2014). As described above, in several real-life domains, probabilities and rewards are negatively correlated and people expect this regularity to hold in psychological experiments (Pleskac & Hertwig, 2014). In line with people's expectations, in decision-making tasks, competing gambles in choice trials typically show a negative correlation between probabilities and rewards. If rewards and probabilities were independent, the expected values of gamble pairs might differ strongly and choices between gambles would not be insightful for researchers (Rieskamp, 2008). Therefore, it is reasonable to investigate how people make choices between pairs of gambles whose rewards and probabilities are correlated. However, such a correlation may influence people's search behavior in

experience-based tasks. For instance, consider the situation in which a decision maker can choose between two gambles both offering two outcomes, one zero and one reward. Yet, the decision maker knows neither the reward magnitudes nor the reward probabilities. To learn about them, she can sample outcomes from the options' outcome distributions. Assume the decision maker samples five times from Gamble A and draws five zeros. She also samples five times from Gamble B and draws five twos. Following the principle of fair bets and assuming that expected values are similar across competing gambles (Osherson, 1995), the decision maker may assume that Gamble A likely offers a large gain with low probability, although she has never actually observed a reward. This implies that with few samples, the decision maker may have created a more detailed representation of the gambles than would be possible by only attending to the objectively available information. Consequently, people may stop their information search earlier when they learn that rewards and probabilities are correlated than in situations where rewards and probabilities are uncorrelated and still feel confident to make informed choices.

In this study, we first investigated whether people believed that rewards and probabilities are correlated in experimental tasks when no direct costs are associated with gambling. We asked people how likely it would be that they would receive different reward amounts if they were participating in psychological gambling studies. Across two experiments, we showed that in psychological studies people expect larger rewards to occur with lower probabilities than smaller rewards also when *no direct costs* were associated with gambling. This finding generalizes the risk–reward heuristic reported by Pleskac and Hertwig (2014) who have demonstrated that people rely on this heuristic in situations in which gambling was associated with *explicit costs*. After having established that people enter psychological experiments with the expectation that larger rewards occur less often than smaller rewards, we studied whether the existence and/or the direction of a correlation between probabilities and rewards influenced people's search effort and their information integration in an experience-based probability judgment task. We asked people to sample outcomes from two-outcome gambles (reward vs. zero) with the goal to estimate a gamble's reward probability. Between three groups, we manipulated whether probabilities and rewards

were negatively correlated (representative), positively correlated (nonrepresentative), or uncorrelated (nonrepresentative). By having two conditions in which rewards and probabilities were correlated, we could control whether *any* correlation or selectively a *representative* correlation influences behavior.

We observed that people in the representative condition (negative correlation between rewards and probabilities) sampled less than people in the other conditions. Importantly, differences in learning did not explain the results: At the end of the study we asked people to estimate reward probabilities for new rewards *without* sampling. There, people's responses followed the probability–reward relationships they had observed before. Additionally, across all conditions, the accuracy of people's probability estimates was comparable. We furthermore modeled people's judgments with two models that, similar to the models described in the manuscripts above, captured the belief in the law of small numbers and Bayesian updating. In this task, most people's responses were best described by a model capturing the belief in the law of small numbers and only a few people integrated information in a Bayesian way.

In sum, our results show that people's search effort is directly influenced by their expectations. This finding sheds light on why in experience-based tasks people may often engage in less search than would be required by normative principles: Arguably, they infer information about one gamble from knowledge of the other gamble. Our results demonstrate how important it is to acknowledge that human cognition is adapted to natural environments (e.g., Brunswik, 1947, 1955; Gigerenzer, Hoffrage, & Kleinbolting, 1991; Gigerenzer & Hug, 1992) and that people's expectations influence their behavior in the laboratory. In the work reported in Manuscript 4, we expanded the focus of the present research and showed that also outside the domain of monetary gambles people treat payoffs as indicators of probabilities.

Manuscript 4: Perceived Riskiness of Clinical Trials: How Incentives Influence Probability Expectations of Side Effects

Hoffart, J. C., & Scheibehenne, B. (2017). *Perceived riskiness of clinical trials: How incentives influence probability expectations of side effects*. Manuscript submitted for publication.

In clinical trials, it is common practice to reimburse participants for their participation. However, several clinicians and researchers argue that such incentives – in particular when they are very large – may be "undue" and coercive as they entrap people into participating in clinical trials against their better judgments (e.g., Macklin, 1981; McNeill, 1997). In line with this notion, also ethics guidelines of for example the guidelines published by the Council for International Organizations of Medical Sciences (CIOMS, 2016) explicitly state that "the level of compensation should not be related to the level of risk that participants agree to undertake" (pp. 53–54).

However, as we showed in Manuscript 3, in gambles people often assume that probabilities and rewards are negatively correlated and expect larger rewards to occur with lower probabilities than smaller rewards. Also, clinical trials can be described as gambles: They offer *sure* rewards (i.e., incentive for participation) that may come with costs (i.e., side effects) with *unknown* probabilities (i.e., probabilities of side effects). If, following the principles of fair bets and in line with the risk–reward heuristic, people in clinical trial believe that the expected value is constant across different trials, they may assume that participation in trials is riskier (i.e., probability of suffering from side effects is higher) when the incentives are larger. We contrasted this hypothesis with two alternative hypotheses.

In opposition to the risk–reward heuristic, the first alternative hypothesis was based on the finding that people sometimes estimate probabilities of highly desirable outcomes as higher than probabilities of less desirable outcomes (Irwin, 1953). Such a desirability bias has also been observed when people judge the riskiness of investment options: More profitable options were also judged as less risky than less profitable options (Kempf, Merkle, & Niessen-Ruenzi, 2014). On the basis of these findings, people may also judge the riskiness of clinical trials as *lower* when the incentives are higher. The second alternative hypothesis predicted that—in line with ethics guidelines—people would treat incentives and probabilities of side effects as uncorrelated and hence that people's risk assessments are not influenced by the magnitude of incentives for participation in clinical trials.

We addressed these opposing hypotheses with an online experiment in which participants read a vignette about a clinical trial investigating whether a new Ebola vaccine

causes side effects. Critically, we described that mild or very severe side effects could occur but pointed out that no one knew for sure *if* these side effects would occur. This implies that no precise probabilities could be associated with the side effects. Between subjects, we manipulated whether participation in the hypothetical trial would be reimbursed with $500 or $10,000. Participants of our study estimated how many of 1,000 women who would be participating in the clinical trial would experience *mild* and *very severe* side effects. These judgments indicated as how risky people perceived the trial.

As expected from the risk–reward–heuristic, participants estimated the number of women who would experience both mild and very severe side effects as being larger in the high-incentive ($10,000) than in the low-incentive ($500) condition. This result demonstrates that the risk–reward heuristic also seems to governs probability estimates in the domain of losses (i.e., side–effects). Furthermore, our results contribute to the on-going debate (e.g., Grady, 2005) on how research participants should be compensated.

Discussion

Often in life, people make decisions under (relative) uncertainty where precise outcomes and/or the associated outcome–probabilities are unknown. To make such decisions, people can relate to their own or other people's experiences. For instance, Mrs. Thomas makes a decision from experience when she decides whether to have dinner at a restaurant where she has eaten before or for which she can look up reviews on the Internet. In the past decade, there has been growing interest in understanding how people make judgments and decisions from experience (e.g., N. J. S. Ashby & Rakow, 2014; Barron & Yechiam, 2009; Camilleri, 2011; Camilleri & Newell, 2009, 2011; Ert & Trautmann, 2014; Glöckner, Fiedler, Hochman, Ayal, & Hilbig, 2012; Golan & Ert, 2014; Hadar & Fox, 2009; Hau et al., 2010, 2008; Hertwig et al., 2004; Hertwig, Barron, Weber, & Erev, 2006; Hertwig & Erev, 2009; Hertwig & Pleskac, 2008, 2010; Mehlhorn et al., 2014; Wulff et al., 2017). In other situations, people cannot directly relate to their own or other people's experiences. However, in environments where structural regularities exist, expectations—gained from previous experience with similar choice options—can simplify choices (e.g., Pleskac & Hertwig, 2014).

In this dissertation, in four manuscripts, I aimed to shed light on how people's experiences and expectations in the form of past outcomes shape information search, preferences, and judgments in situations of (relative) uncertainty.

How Sample Size Influences Judgments and Decisions from Experience

In three manuscripts (1–3), we investigated whether the amount of information (i.e., the number of past outcomes) influences judgments and decisions. In particular, we contrasted a Bayesian perspective on dealing with sample-size-based uncertainty with the ideas that people ignore sample size (Manuscripts 1–3) and treat sample size as indicating quality (Manuscript 2). The short answer to the question of whether people integrate sample size in their judgments is *yes, many people do, but not everyone, and not always*.

Individual differences. In Manuscript 1, we investigated whether sample size, that is, the number of outcomes people observe from unknown gambles, influences how people value the gambles. In short, by modeling people's selling prizes on the individual level, we compared a Bayesian account that ingrates sample-based uncertainty with a model ignoring sample size and treating the relative observed outcome frequency as representative of the true outcome distribution. Crucially, the strategies that people used varied considerably between people and slightly less than half of the people followed Bayesian principles while the other half did not. Had we modeled the data only on the aggregate level, we would have missed the fact that many people—but not all—integrated sample size in line with Bayesian principles. We are not the first to suggest that modeling data on an individual level can provide important insights that otherwise would be overlooked if only aggregate data are modeled (F. G. Ashby et al., 1994; Lewandowsky & Farrell, 2011).

In Manuscript 2, we investigated how people integrate other people's experience in the form of summaries about customer ratings. Confirming the results presented in Manuscript 1, a considerable proportion of people integrated average ratings and the number of ratings according to Bayesian principles. We found that people who were better described by the Bayesian model also scored higher in a statistical numeracy task. Future research could follow up on this finding and investigate whether it is indeed numeracy that moderates people's

strategies or an (unrelated) different factor. For instance, people who are very engaged in the task may be more motivated to both solve the numeracy task *and* think about the choices between the hotels. Furthermore, future research should explore whether numeracy is also related to how well a Bayesian model describes individuals' behavior in a task where people sequentially learn about options, as in Manuscript 1.

In Study 3 of Manuscript 3, we manipulated the relationship between probabilities and rewards in three conditions and people made probability judgments from experience by repeatedly sampling outcomes from unknown gambles. In contrast to the findings reported in Manuscripts 1 and 2, almost no one integrated sample-size-based uncertainty in their judgments. Instead, most people relied on the relative observed outcome frequency when making judgments. However, in each condition some people sampled more than others. N. J. Ashby (2017) recently showed that in decisions-from-experience tasks, people higher in numeracy search more than people lower in numeracy. This finding may help to explain why we failed to observe differences in how people integrated sample size in their judgments. With growing sample sizes, the predictions made by the Bayesian model converge with the predictions made by the model ignoring sample size and treating the relative frequency of the observed outcomes as representative for the outcome probabilities. When the predictions of two models are very similar, a simpler model (as the model capturing the belief in the law of small numbers) will be preferred as it has fewer free parameters. Consider the hypothesis that people who are high in numeracy, integrate information more in line with Bayesian principles (as suggested in Manuscript 2 of this dissertation) *also* search more (as suggested by N. J. Ashby (2017)). If this hypothesis holds true, people who in reality integrate information in line with Bayesian principles, might be classified by the non–Bayesian model. This is because the model predictions will be very similar and due to its simplicity the non–Bayesian model would be preferred. Future research should test this hypothesis, for instance by assessing people's numeracy and by investigating how numeracy scores interact with search effort and model classifications. In an experiment, people could, for instance, in some trials be interrupted in their search at a stage where the models make sufficiently different predictions and prompted for their probability estimates.

An alternative explanation for the opposing finding may be rooted in differences in the task structures between the experiments reported in Manuscript 3 and those reported in Manuscripts 1 and 2.

Task-related and situational differences. Not only individual differences but also task-related differences and situational differences may have modulated the extent to which sample size influences judgments and decisions. In Study 3 of Manuscript 3, people decided freely how many samples they wanted to draw in order to make their judgments. In contrast, in Manuscript 1, people always sampled a fixed number of outcomes and in Manuscript 2 people saw a fixed number of previous-user ratings. Arguably, when people freely decide how much they want to sample, they may feel more *committed* to the outcomes they observed. More precisely, people may sample until they assume that the number of outcomes they have drawn is a correct representation of the empirical outcome distribution. Alternatively, response strategies may differ between preferential tasks where no correct answer can be defined (e.g., tasks in Manuscripts 1 and 2) and nonpreferential tasks, where a distinct correct answer can be defined (e.g., the probability–judgment task in Experiment 3 of Manuscript 3). Future research could test these two above mentioned hypotheses experimentally, for instance, by manipulating whether sampling is fixed or free and whether people solve a nonpreferential or preferential task. If it is the form of sampling (free vs. fixed) that influences whether people integrate information in a Bayesian way, evidence for Bayesian information integration in both non–preferential judgment tasks *and* preferential tasks should be observed. However, if it is task format (i.e., preferential vs. non–preferential task) that influences whether people integrate information in a Bayesian way, evidence for Bayesian information integration for both fixed- and free sampling tasks should be observed.

In Manuscript 2, we observed that in the same task, in some choice situations almost everyone chooses according to Bayesian principles and in other situations far fewer people do so. We developed a Bayesian decision tree that describes the behavior of people following a partly Bayesian strategy: When the cues *average ratings* and *number of ratings* are inconsistent, people choose in line with Bayesian principles. On the other hand, when the cues are consistent, people choose the option with higher values on both cues in line with a model

treating average ratings and sample size as indicating quality. This model captures the empirical observation that people —in line with Bayesian principles— prefer a slightly-worse-but-more-often-rated option when both options are rated well and that people —opposing Bayesian principles— prefer a slightly-better-and-more-often-rated option when both options are rated poorly. Similarly to the latter, Kutzner et al. (2016) reported that in two-alternative choice tasks, people's choices deviated from Bayesian principles. Using an overview of previous outcomes (i.e., "red" or "blue" draws) of two options, people judged which option would provide a higher chance of reaching a predefined goal (e.g., "Reach N blue outcomes in the next X draws"). In line with Bayesian principles, the authors identified that under certain circumstances it is beneficial to choose the option about which less information is available. Consider a situation where the goal is to reach 80 "blue" outcomes in 100 trials. For option A, one "red" observation has been made. For option B, 200 "red" and 10 "blue" observations have been made. The relative frequency of "blue" to "red" outcomes is lower for option A ($\frac{0}{1}$) than for option B ($\frac{10}{210}$). However, due to the larger uncertainty related to the outcome distribution of option A, this option gives a higher chance of reaching the goal. Yet, people almost never choose a slightly worse option (i.e., relatively more "red" than "blue" outcomes) that provides better chances of reaching the goal. The authors concluded that "forgoing a better option for a worse but more uncertain one might be hard to justify. Having chosen an employee because we did not know much about her is a difficult argument to make, especially if the employee underperforms" (Kutzner et al., 2016, p. 8).

A similar argument can be used to explain our findings of Manuscript 2 that on the lower end of the scale, people's choices violate predictions made by Bayesian principles *although* they may understand that few outcomes entail more uncertainty. Evidence that people indeed understand the impact of sample-size-based uncertainty can be deduced from a post-choice questionnaire that we applied in the study of Manuscript 2. There, we asked people again to choose between two hotels: one an often-rated hotel with an average of 1.5 points and the other a rarely rated hotel with an average of 1 point. Before making this choice, people wrote down reasons that supported choosing either the first or the second hotel. In this situation, the proportion of people who chose—in line with Bayesian principles—the slightly

worse and less often rated hotel was, at 50%, much higher than the proportion in comparable trials in the choice phase (24%) —where people did not provide reasons to support the choice options. Arguably, when thinking about reasons, many people engaged in more detailed elaboration about the choice options and hence overcame the initial urge to choose the better and more often rated option. This finding should be further elaborated in future research. For instance, larger incentives may motivate people to engage in more detailed elaboration of choice options. High incentives are often involved in consumer choices: When people consult online ratings to book a hotel or make purchasing decisions, they naturally face higher costs and higher potential incentives than in the laboratory as they experience the consequences of their actions once they visit the booked hotel or receive the chosen product. Future research could exploit this fact by replicating our experiment with real consequences, for instance, with a task where people can choose between products based on real customer ratings and then receive one product after the experiment.

In sum, interindividual differences as well as task-related and situational differences contribute to whether and how people integrate the amount of information they have in their judgments and decisions. This observation may explain why previous research has provided conflicting answers to the question of whether people's information integration can be described by Bayesian principles. More research is needed to better and more precisely understand *who* uses *which* strategy under *what* circumstances.

How Expectations Influence Information Search and Judgments

As a second goal of my dissertation, I investigated how people's expectations influence how they search for information and make judgments. As the world is complex, people have learned to detect regularities in the environment. These regularities allow humans to easily navigate in familiar environments and also in new environments where the expected regularity holds true.

Because in many real-world environments, rewards and probabilities are negatively correlated, individuals expect this regularity to hold in behavioral studies as well. In Manuscript 3, we showed that when judging unknown reward probabilities of known reward

amounts, people expect larger rewards to occur with lower probabilities than smaller rewards. Furthermore, in situations where the experimental environment met this expectation, people searched less for information than in situations where this expectation was not met. This finding illustrates how expectations may influence not only judgments under uncertainty but also the effort people make when learning about unknown options. Future research could link these findings more closely to traditional decisions-from-experience paradigms. There, people sample outcomes from more than one choice option and then choose between these options. For instance, future studies could explore people's beliefs about the choice options' outcome structures after they have sampled. Given the results reported in Manuscript 3, it can be hypothesized that people extract information from one gamble to infer (unknown) information about the other. This in turn can help to explain why people often rely on only a few observations: They might have a more detailed understanding of the choice options that expected only based on the samples outcomes.

In Manuscript 4, we extended the finding that people expect rewards and probabilities to be correlated to a judgment situation in a more applied context. When judging how likely side effects in clinical trials are to occur, people expect incentives for participating in trials to be tied to the likelihood of experiencing side effects. When incentives are larger, people expect more side effects than when incentives are smaller. Similar to the findings in Manuscript 3, these result suggest that people believe in fair bets (Osherson, 1995). Further, in contrast to the work by Pleskac and Hertwig (2014) and our work in Manuscript 3, here people estimated the probability of losses (i.e., side effects) and not gains. In sum, our findings demonstrate that the risk–reward heuristics can be applied beyond monetary gambles in the gain domain.

Congruence between beliefs and the structure of the environment. Outside the laboratory (Edwards, 1962; Pleskac & Hertwig, 2014), rewards and probabilities are often negatively correlated. In Manuscript 3 we showed that a similar regularity exists between competing two-outcome (zero and reward) gamble by analyzing past gambles that were used to study decisions from experience. However, in clinical trials it is not clear if a relationship between risks and reward exists: Only a few research organizations have policies in place defining how subjects of clinical trials should be reimbursed and even fewer organizations

were, when asked explicitly, confident in estimating how many of their studies' research subjects were paid at all (Dickert, Emanuel, & Grady, 2002). Given ethical guidelines, rewards *must strictly not* be related to the riskiness of the trials (e.g., CIOMS, 2016). However, third factors may still contribute to a relationship between risks and rewards. For instance, ethical guidelines also state that participants of clinical trials shall be compensated for their time (e.g., CIOMS, 2016). If, for instance, riskier trials are on average also more time consuming and hence incentivized with more money, a positive correlation between risk and incentives results. In sum, because we have not conducted an analysis of the environment, we cannot conclude whether the positive relationship between probabilities of side effects and incentives in clinical trials that people expect is adapted to the environment. If probabilities of side effects are positively correlated with incentives, people's expectations would be well adapted to the environment and help them make informed decisions under uncertainty. However, under the assumption that in clinical trials incentives and risk are uncorrelated or even negatively correlated, people's expectations conflict with the structure of the environment. Such a mismatch between expectations and environment is not unusual: For instance, when judging how risky it is to invest in different investment options, people often believe that riskier options offer lower returns than less risky options (e.g., Kempf et al., 2014; Shefrin, 2001). In the case of clinical trials, this mismatch between beliefs and the risk–reward structure may have dramatic consequences. If people expect incentives to foreshadow risks and base decisions on this expectation but in reality this connection does not exist, people may participate in (low incentivised) trials although they would not have agreed to do so if they had been aware of the true risks.

The above-stated argument needs to be considered with caution: Our data do not allow to elaborate to what extent incentives affect people's willingness to participate in clinical trials when they face this decision for real. More research is needed to answer this question. Future studies could for instance monitor participation rates and risk perception of participants in real trials to study the extent to which the magnitude of incentives influences both factors.

Conclusion

In four manuscripts, I showed that information search, judgments, and decisions can be influenced by both people's expectations about regularities in the task structure and the information that is available to make judgments and decisions. Further, I identified, that people differ substantially in the strategies they use. When only modeling data on the aggregate level, important data structures were overlooked. Hence, to fully understand how people make judgments and decisions, researchers need to take people's expectations as well as the reasons why certain individuals make use of different strategies into account.

References

Alloy, L. B., & Tabachnik, N. (1984). Assessment of covariation by humans and animals: The joint influence of prior expectations and current situational information. *Psychological Review*, *91*, 112–149. doi: http://dx.doi.org/10.1037/0033-295X.91.1.112

Ashby, F. G., Maddox, W. T., & Lee, W. W. (1994). On the dangers of averaging across subjects when using multidimensional scaling or the similarity-choice model. *Psychological Science*, *5*, 144–151. Retrieved from `http://www.jstor.org/stable/40063087`

Ashby, N. J. (2017). Numeracy predicts preference consistency: Deliberative search heuristics increase choice consistency for choices from description and experience. *Judgment and Decision Making*, *12*, 128–139.

Ashby, N. J. S., & Rakow, T. (2014). Forgetting the past: Individual differences in recency in subjective valuations from experience. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *40*, 1153–1162. doi: http://dx.doi.org/10.1037/a0036352

Barron, G., & Yechiam, E. (2009). The coexistence of overestimation and underweighting of rare events and the contingent recency effect. *Judgment and Decision Making*, *4*, 447–460. Retrieved from `http://journal.sjdm.org/9729b/jdm9729b.pdf`

Brase, G. L., & Hill, W. T. (2015). Good fences make for good neighbors but bad science: A review of what improves Bayesian reasoning and why. *Frontiers in Psychology*, 340. doi: http://dx.doi.org/10.3389/fpsyg.2015.00340

Brase, G. L., & Hill, W. T. (2017). Adding up to good Bayesian reasoning: Problem format manipulations and individual skill differences. *Journal of Experimental Psychology: General*, *146*, 577–591. doi: http://dx.doi.org/10.1037/xge0000280

Brunswik, E. (1947). *Systematic and representative design of psychological experiments. With results in physical and social perception.* Berkeley, CA: University of California Press.

Brunswik, E. (1955). Representative design and probabilistic theory in a functional psychology. *Psychological Review*, *62*, 193–217. doi: https://dx.doi.org/10.1037/h0047470

Busemeyer, J. R., Byun, E., Delosh, E. L., & McDaniel, M. A. (1997). Learning functional relations based on experience with input-output pairs by humans and artificial neural networks. In K. Lamberts & D. Shanks (Eds.), *Knowledge, concepts, and categories* (pp. 405–437). New York: Psychology Press.

Camilleri, A. R. (2011). *The psychological mechanisms underpinning experience-based choice* (Unpublished doctoral dissertation). The University of New South Wales, Sydney, Australia.

Camilleri, A. R., & Newell, B. R. (2009). The role of representation in experience-based choice. *Judgment and Decision Making*, *4*, 518–529. Retrieved from `http://journal.sjdm.org/9713/jdm9713.pdf`

Camilleri, A. R., & Newell, B. R. (2011). Description- and experience-based choice: Does equivalent information equal equivalent choice? *Acta Psychologica*, *136*, 276–284. doi: http://dx.doi.org/10.1016/j.actpsy.2010.11.007

Camilleri, A. R., & Newell, B. R. (2013). Mind the gap? Description, experience, and the continuum of uncertainty in risky choice. In N. Srinivasan & C. Pammi (Eds.), *Progress in brain research: Vol. 202. Decision making: Neural and behavioural approaches* (pp. 55–71). Oxford, UK: Elsevier.

Chapman, G. B., & Liu, J. (2009). Numeracy, frequency, and Bayesian reasoning. *Judgment and Decision Making*, *4*, 34–40. Retrieved from `http://www.sjdm.org/journal/8708/jdm8708.pdf`

Chen, Y.-F. (2008). Herd behavior in purchasing books online. *Computers in Human Behavior*, *24*, 1977–1992. doi: http://dx.doi.org/10.1016/j.chb.2007.08.004

Cokely, E. T., Galesic, M., Schulz, E., Ghazal, S., & Garcia-Retamero, R. (2012). Measuring risk literacy: The Berlin numeracy test. *Judgment and Decision Making*, *7*, 25–47.

Council for International Organizations of Medical Sciences (CIOMS). (2016). *International ethical guidelines for health-related research involving humans.* Geneva, Switzerland: CIOMS.

Courville, A. C., Daw, N. D., & Touretzky, D. S. (2006). Bayesian theories of conditioning in a changing world. *Trends in Cognitive Sciences*, *10*, 294–300. doi:

http://dx.doi.org/10.1016/j.tics.2006.05.004

De Martino, B., Bobadilla-Suarez, S., Nouguchi, T., Sharot, T., & Love, B. C. (2017). Social information is integrated into value and confidence judgments according to its reliability. *Journal of Neuroscience*, *37*, 6066–6074. doi: http://dx.doi.org/10.1523/jneurosci.3880-16.2017

Deutsch, M., & Gerard, H. B. (1955). A study of normative and informational social influences upon individual judgment. *The journal of abnormal and social psychology*, *51*, 629–636. doi: https://doi.org/10.1037/h0046408

Dickert, N., Emanuel, E., & Grady, C. (2002). Paying research subjects: An analysis of current policies. *Annals of Internal Medicine*, *136*, 368–373. doi: http://dx.doi.org/10.7326/0003-4819-136-5-200203050-00009

Dutt, V., & Gonzalez, C. (2012). The role of inertia in modeling decisions from experience with instance-based learning. *Frontiers in Psychology*, *3*(177), 1–10. doi: http://dx.doi.org/10.3389/fpsyg.2012.00177

Edwards, W. (1961). Probability learning in 1000 trials. *Journal of Experimental Psychology*, *62*, 385–394. doi: http://dx.doi.org/10.1037/h0041970

Edwards, W. (1962). Utility, subjective probability, their interaction, and variance preferences. *Journal of Conflict Resolution*, *6*, 42–51. doi: http://dx.doi.org/10.1177/002200276200600106

Ert, E., & Trautmann, S. T. (2014). Sampling experience reverses preferences for ambiguity. *Journal of Risk and Uncertainty*, *49*, 31–42. doi: http://dx.doi.org/10.1007/s11166-014-9197-9

Estes, W. K. (1950). Toward a statistical theory of learning. *Psychological Review*, *57*, 94–107. doi: http://dx.doi.org/10.1037/0033-295x.101.2.282

Fang, J., Wen, C., George, B., & Prybutok, V. R. (2016). Consumer heterogeneity, perceived value, and repurchase decision-making in online shopping: The role of gender, age, and shopping motives. *Journal of Electronic Commerce Research*, *17*, 116–131. Retrieved from `https://search.proquest.com/docview/1792212900`

Fiedler, K. (2000). Beware of samples! A cognitive-ecological sampling approach to

judgment biases. *Psychological Review*, *107*, 659-676. doi: http://dx.doi.org/10.1037//0033-295x.107.4.659

Fischbacher, U., Hertwig, R., & Bruhin, A. (2013). How to model heterogeneity in costly punishment: Insights from responders' response times. *Journal of Behavioral Decision Making*, *26*, 462–476. doi: http://dx.doi.org/10.1002/bdm.1779

Gigerenzer, G., Hoffrage, U., & Kleinbolting, H. (1991). Probabilistic mental models—A Brunswikian theory of confidence. *Psychological Review*, *98*, 506–528. doi: http://dx.doi.org/10.1037/0033-295x.98.4.506

Gigerenzer, G., & Hug, K. (1992). Domain-specific reasoning: Social contracts, cheating, and perspective change. *Cognition*, *43*, 127–171. doi: http://dx.doi.org/10.1016/0010-0277(92)90060-u

Gigerenzer, G., & Selten, R. (2001). *Bounded rationality: The adaptive toolbox* (G. Gigerenzer & R. Selten, Eds.). Cambridge, MA: MIT Press.

Gigerenzer, G., Todd, P. M., & the ABC Research Group. (1999). *Simple heuristics that make us smart*. New York, NY: Oxford University Press.

Glöckner, A., Fiedler, S., Hochman, G., Ayal, S., & Hilbig, B. (2012). Processing differences between descriptions and experience: A comparative analysis using eye-tracking and physiological measures. *Frontiers in Psychology*, 3. doi: http://dx.doi.org/10.3389/fpsyg.2012.00173

Glöckner, A., Hilbig, B. E., Henninger, F., & Fiedler, S. (2016). The reversed description-experience gap: Disentangling sources of presentation format effects in risky choice. *Journal of Experimental Psychology: General*, *145*, 486–508. doi: http://dx.doi.org/10.1037/a0040103

Golan, H., & Ert, E. (2014). Pricing decisions from experience: The roles of information-acquisition and response modes. *Cognition*, *136*, 9–13. doi: http://dx.doi.org/10.1016/j.cognition.2014.11.008

Gonzalez, C., & Dutt, V. (2011). Instance-based learning: Integrating sampling and repeated decisions from experience. *Psychological Review*, *118*, 523–552. doi: http://dx.doi.org/10.1037/a0024558

Grady, C. (2005). Payment of clinical research subjects. *Journal of Clinical Investigation*, *115*, 1681–1687. doi: http://dx.doi.org/10.1172/jci25694

Gretzel, U., & Yoo, K. H. (2008). Use and impact of online travel reviews. *Information and Communication Technologies in Tourism 2008*, 35–46. doi: http://dx.doi.org/10.1007/978-3-211-77280-5_4

Griffin, D., & Tversky, A. (1992). The weighing of evidence and the determinants of confidence. *Cognitive Psychology*, *24*, 411–435. doi: http://dx.doi.org/10.1016/0010-0285(92)90013-r

Hadar, L., & Fox, C. R. (2009). Information asymmetry in decision from description versus decision from experience. *Judgment and Decision Making*, *4*, 317–325. Retrieved from `http://journal.sjdm.org/9331/jdm9331.html`

Hau, R., Pleskac, T., & Hertwig, R. (2010). Decisions from experience and statistical probabilities: Why they trigger different choices than a priori probabilities. *Journal of Behavioral Decision Making*, *68*, 48–68. doi: http://dx.doi.org/10.1002/bdm.665

Hau, R., Pleskac, T. J., Kiefer, J., & Hertwig, R. (2008). The description–experience gap in risky choice: The role of sample size and experienced probabilities. *Journal of Behavioral Decision Making*, *21*, 493–518. doi: http://dx.doi.org/10.1002/bdm.598

Hertwig, R. (2012). The psychology and rationality of decisions from experience. *Synthese*, *187*, 269–292. doi: http://dx.doi.org/10.1007/s11229-011-0024-4

Hertwig, R., Barron, G., Weber, E. U., & Erev, I. (2004). Decisions from experience and the effect of rare events in risky choice. *Psychological Science*, *15*, 534–539. doi: http://dx.doi.org/10.2139/ssrn.1301100

Hertwig, R., Barron, G., Weber, E. U., & Erev, I. (2006). The role of information sampling in risky choice. In K. Fiedler & P. Juslin (Eds.), *Information sampling and adaptive cognition* (pp. 72–91). New York, NY: Cambridge University Press.

Hertwig, R., & Erev, I. (2009). The description-experience gap in risky choice. *Trends in Cognitive Sciences*, *13*, 517–523. doi: http://dx.doi.org/10.1016/j.tics.2009.09.004

Hertwig, R., & Pleskac, T. J. (2008). The game of life: How small samples render choice simpler. In N. Chater & M. Oaksford (Eds.), *The probabilistic mind: Prospects for*

rational models of cognition (pp. 209–236). Oxford, UK: Oxford University Press.

Hertwig, R., & Pleskac, T. J. (2010). Decisions from experience: Why small samples? *Cognition*, *115*, 225–237. doi: http://dx.doi.org/10.1016/j.cognition.2009.12.009

Irwin, F. W. (1953). Stated expectations as functions of probability and desirability of outcomes. *Journal of Personality*, *21*, 329–335. doi: http://dx.doi.org/10.1111/j.1467-6494.1953.tb01775.x

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, *47*, 263–291. doi: http://dx.doi.org/10.2307/1914185

Katz, L. (1962). Monetary incentive and range of payoffs as determiners of risk taking. *Journal of Experimental Psychology*, *64*, 541–544. doi: http://dx.doi.org/10.1037/h0047013

Kempf, A., Merkle, C., & Niessen-Ruenzi, A. (2014). Low risk and high return—Affective attitudes and stock market expectations. *European Financial Management*, *20*, 995–1030. doi: http://dx.doi.org/10.1111/eufm.12001

Kersten, D., Mamassian, P., & Yuille, A. (2004). Object perception as Bayesian inference. *Annual Reviews*, *55*, 271–304. doi: http://dx.doi.org/10.1146/annurev.psych.55.090902.142005

Knight, F. H. (1921). *Risk, uncertainty, and profit.* New York, NY: Sentry Press.

Kutzner, F. L., Read, D., Stewart, N., & Brown, G. (2016). Choosing the devil you don't know: Evidence for limited sensitivity to sample size–based uncertainty when it offers an advantage. *Management Science*, *63*, 1519–1528. doi: http://dx.doi.org/10.1287/mnsc.2015.2394

Lejarraga, T., Dutt, V., & Gonzalez, C. (2012). Instance-based learning: A general model of repeated binary choice. *Journal of Behavioral Decision Making*, *25*, 143–153. doi: http://dx.doi.org/10.1002/bdm.722

Lejarraga, T., Hertwig, R., & Gonzalez, C. (2012). How choice ecology influences search in decisions from experience. *Cognition*, *124*, 334–342. doi: http://dx.doi.org/10.1016/j.cognition.2012.06.002

Lewandowsky, S., & Farrell, S. (2011). *Computational modeling in cognition: Principles and*

practice. Thousand Oaks, CA: Sage.

Lopes, L. L. (1983). Some thoughts on the psychological concept of risk. *Journal of Experimental Psychology: Human Perception and Performance*, *9*, 137-144. doi: http://dx.doi.org/10.1037/0096-1523.9.1.137

Macklin, R. (1981). 'Due' and 'undue' inducements: On paying money to research subjects. *IRB: Ethics & Human Research*, *3*, 1–6. doi: http://dx.doi.org/10.2307/3564136

Markowitz, H. (1952). Portfolio selection. *Journal of Finance*, *7*, 77–91.

McNeill, P. (1997). Paying people to participate in research: Why not? *Bioethics*, *11*, 390–396. doi: http://dx.doi.org/10.1111/1467-8519.00079

Mehlhorn, K., Ben-Asher, N., Dutt, V., & Gonzalez, C. (2014). Observed variability and values matter: Toward a better understanding of information search and decisions from experience. *Journal of Behavioral Decision Making*, *27*, 328–339. doi: http://dx.doi.org/10.1002/bdm.1809

Mohr, P. N. C., Biele, G., Krugel, L. K., Li, S. C., & Heekeren, H. R. (2010). Neural foundations of risk-return trade-off in investment decisions. *NeuroImage*, *49*, 2556–2563. doi: http://dx.doi.org/10.1016/j.neuroimage.2009.10.060

Myers, J. L., & Katz, L. (1962). Range of payoffs and feedback in risk taking. *Psychological Reports*, *10*, 483–486. doi: http://dx.doi.org/10.2466/pr0.1962.10.2.483

Obrecht, N. A., Chapman, G. B., & Gelman, R. (2007). Intuitive *t* tests: Lay use of statistical information. *Psychonomic Bulletin & Review*, *14*, 1147–1152. doi: http://dx.doi.org/10.3758/bf03193104

Osherson, N., D. (1995). Probability judgment. In E. E. Smith & D. N. Osherson (Eds.), *Thinking* (pp. 35–75). Cambridge, MA: MIT Press.

Pachur, T., & Scheibehenne, B. (2012). Constructing preference from experience: The endowment effect reflected in external information search. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*, 1108–1116. doi: http://dx.doi.org/10.1037/a0027637

Pleskac, T. J., & Hertwig, R. (2014). Ecologically rational choice and the structure of the environment. *Journal of Experimental Psychology: General*, *143*, 2000–2019. doi:

http://dx.doi.org/10.1037/xge0000013

Powell, D., Yu, J., DeWolf, M., & Holyoak, K. J. (2017). The love of large numbers: A popularity bias in consumer choice. *Psychological Science*, *10*, 1432-1442. doi: http://dx.doi.org/10.1177/0956797617711291

Rakow, T., Demes, K. A., & Newell, B. R. (2008). Biased samples not mode of presentation: Re-examining the apparent underweighting of rare events in experience-based choice. *Organizational Behavior and Human Decision Processes*, *106*, 168–179. doi: http://dx.doi.org/10.1016/j.obhdp.2008.02.001

Rheinberger, C. M., & Hammitt, J. K. (2015). *Dinner with Bayes: On the revision of risk reliefs*. (Report No. TSE 15-574). Retrieved from: http://tse-fr.eu/pub/29293.

Rieskamp, J. (2008). The probabilistic nature of preferential choice. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*, 1446-1465. doi: http://dx.doi.org/10.1037/a0013646

Schöbel, M., Rieskamp, J., & Huber, R. (2016). Social influences in sequential decision making. *PloS ONE*, *11*, 1–23. doi: https://doi.org/10.1371/journal.pone.0146536

Shefrin, H. (2001). Do investors expect higher returns from safer stocks than from riskier stocks? *Journal of Psychology and Financial Markets*, *2*, 176–181. doi: http://dx.doi.org/10.1207/s15327760jpfm0204_1

Sunden, A. E., & Surette, B. J. (1998). Gender differences in the allocation of assets in retirement savings plans. *American Economic Review*, *88*, 207–211. Retrieved from `http://www.jstor.org/stable/116920`

Suydam, M. M. (1965). Effects of cost and gain ratios, and probability of outcome on ratings of alternative choices. *Journal of Mathematical Psychology*, *2*, 171–179. doi: http://dx.doi.org/10.1016/0022-2496(65)90022-2

Tenenbaum, J. B., Griffiths, T. L., & Kemp, C. (2006). Theory-based Bayesian models of inductive learning and reasoning. *Trends in Cognitive Sciences*, *10*, 309–318. doi: http://dx.doi.org/10.1016/j.tics.2006.05.009

Todd, P. M., & Gigerenzer, G. (2012). *Ecological rationality: Intelligence in the world*. New York, NY: Oxford University Press.

Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin*, *76*, 105–110. doi: http://dx.doi.org/10.1037/h0031322

Ungemach, C., Chater, N., & Stewart, N. (2009). Are probabilities overweighted or underweighted when rare outcomes are experienced (rarely)? *Psychological Science*, *20*, 473–479. doi: http://dx.doi.org/10.1111/j.1467-9280.2009.02319.x

Von Neumann, J., & Morgenstern, O. (1947). *Theory of games and economic behavior* (2nd ed.). Princeton, NJ: Princeton University Press.

Wulff, D. U., Mergenthaler-Canseco, M., & Hertwig, R. (2017). A meta-analytic review of two modes of learning and the description-experience gap. *Psychological Bulletin*, Advance online publication. doi: http://dx.doi.org/10.1037/bul0000115

Xu, F., & Tenenbaum, J. B. (2007). Word learning as Bayesian inference. *Psychological Review*, *114*, 245–272. doi: http://dx.doi.org/10.1037/0033-295X.114.2.245

Ye, Q., Law, R., & Gu, B. (2009). The impact of online user reviews on hotel room sales. *International Journal of Hospitality Management*, *28*, 180–182. doi: http://dx.doi.org/10.1016/j.ijhm.2008.06.011

Zhang, Z., Ye, Q., Law, R., & Li, Y. (2010). The impact of e-word-of-mouth on the online popularity of restaurants: A comparison of consumer reviews and editor reviews. *International Journal of Hospitality Management*, *29*, 694–700.

The influence of sample size on preferences from experience

Janine Christin Hoffart

Jörg Rieskamp

Gilles Dutilh

University of Basel, Department of Psychology, Center for Economic Psychology

Author Note

This research is supported by a grant (SNF # 143854) from the Swiss National Science Foundation to the second and third author.

Correspondence concerning this article should be addressed to Janine Christin Hoffart. University of Basel, Department of Psychology, Missionsstrasse 62a, 4055 Basel, Switzerland. E-mail: janine.hoffart@unibas.ch

Abstract

People often evaluate risky options in situations in which they need to learn from experience how likely certain outcomes are. When people can actively sample information about outcomes and outcome probabilities to make decisions, they typically draw small samples. To understand why people rely on small samples, we studied how the amount of information influences preferences for gambles. In two studies, people drew different numbers of samples from risky gambles and then indicated their selling prices for these gambles and their confidence in their judgments. Remarkably, the results show that averaged across people, neither valuations nor confidence in the valuations changed substantially with sample size. However, on an individual level approximately half of the participants could be classified as Bayesian learners and the other half as frequentist learners. Both valuations and confidence of Bayesian learners changed with sample size as predicted by Bayesian principles. In contrast, sample size did not influence valuations or confidence of frequentist learners. These results illustrate the variability in how people learn from sampled information and provide an explanation for why sample size often does not affect judgments. We also found a difference in valuations based on descriptive information versus valuations based on experienced outcomes. In both conditions, people behaved as if they overweighted rare events.

Keywords: D–E gap, valuations from experience

The influence of sample size on preferences from experience

In everyday life, people often form preferences on the basis of past experiences. How strong these preferences are arguably depends on the amount of previous experience. Someone who has visited a restaurant for many years and has always had positive experiences likely has a stronger positive preference for this restaurant than someone who visited the restaurant only once and had a positive experience. Thus, even if the relative frequency of positive experiences is similar, the strength of preferences can differ. We examined how people's preferences in an experience-based judgment situation evolve as a function of growing experience.

The D–E Gap and Search Effort in Experience-Based Tasks

How people develop preferences based on information that they accumulate over time has been studied with experience-based judgment and decision-making tasks (e.g., Hau, Pleskac, Kiefer, & Hertwig, 2008; Hertwig, Barron, Weber, & Erev, 2004). Experience-based tasks differ from descriptive tasks in the way information is presented. In descriptive tasks, people make judgments about gambles based on numerical, graphical, or text-based summaries of the gambles' outcome distributions. In experience-based tasks, on the other hand, people learn about the outcome distributions by repeatedly drawing samples from them until they feel confident to make a judgment or decision. Interestingly, past research suggests that people's decisions systematically differ between description and experience (the so-called D–E gap): When people make decisions from description, they choose as if they overweight rare events (Kahneman & Tversky, 1979). At the same time, when they make similar decisions based on experience, they choose as if they underweight rare events (Hertwig et al., 2004).

The D–E gap has often been attributed to limited search effort in experience-based tasks and resulting *sampling error*. Such sampling error occurs when the observed outcome sequence deviates from the objective outcome probabilities (Hadar & Fox, 2009). For instance, Hertwig et al. (2004) reported that participants drew a median of only 15 observations per decision. With such a low number, the skew of a gamble's binomial sampling distribution implies that more people undersample than oversample a rare event. Indeed,

averaged across trials, in Hertwig et al.'s 2004 study, the rare event was undersampled by 78% of the participants. Other studies have reported similar sampling errors (e.g., Hau et al., 2008; Rakow, Demes, & Newell, 2008).

Since the initial finding that people often sample insufficiently, researchers have tried to reduce or eliminate sampling error. Attempts have involved encouraging participants to draw larger samples by increasing the incentives (Hau et al., 2008); showing fixed samples that represent the gambles' objective probability distributions (Ungemach, Chater, & Stewart, 2009); showing fixed—relatively large—numbers of random samples, as larger total samples naturally reduce sampling error (Hau, Pleskac, & Hertwig, 2010; Hau et al., 2008); presenting gamble descriptions that match the outcome distribution that a "partner" participant had sampled (Rakow et al., 2008); and constraining statistical analysis to trials in which the total sample closely represented the gambles' underlying outcome distributions (e.g., Camilleri & Newell, 2009, 2011). Typically, the goal of the attempt to eliminate sampling error was to allow for an unbiased comparison between decisions from description and decisions from experience. The results are mixed: In some studies, the authors found a diminished D–E gap (Hau et al., 2010, 2008; Ungemach et al., 2009) or failed to find a D–E gap (Camilleri & Newell, 2009, 2011; Glöckner, Fiedler, Hochman, Ayal, & Hilbig, 2012; Rakow et al., 2008).

While the question of whether people process information from description differently from information from experience is of great importance, in the present work, we approached it only marginally. Instead, we examined how valuations of risky prospects from experience are influenced by different sample sizes. To do so, we manipulated how many samples people could draw from the prospects' outcome distributions and measured people's preferences (i.e., selling prices).

How Does Sample Size Influence Preferences?

The question of how sample size influences judgments has interested psychologists for decades (e.g., Griffin & Tversky, 1992; Tversky & Kahneman, 1971). We investigated how sample size influences preferences from experience, by contrasting two existing theories: *the belief in the law of small numbers* (Tversky & Kahneman, 1971) and *Bayesian updating*.

Belief in the Law of Small Numbers

Tversky and Kahneman (1971) formulated the belief in the law of small numbers: an exaggerated belief about the likelihood of small sequences of random draws representing outcome distributions. For instance, when people were asked to mentally produce a short sequence of random coin tosses, the produced sequences resembled a relative frequency of .5 more often than expected by chance (Tversky & Kahneman, 1971). Griffin and Tversky (1992) showed that people tend to neglect the role of sample size in their judgments. They showed participants sequences of coin tosses using coins with different biases. Then, participants estimated the probability that a coin was biased and expressed their confidence in their estimate. Interestingly, Griffin and Tversky (1992) varied the total number of coin tosses (i.e., sample size). They found that participants neglected sample size and based their estimates mostly on the observed relative frequency of head and tail outcomes. Surprisingly, participants expressed more confidence in their judgments when the number of coin tosses was small. Since these initial findings, several studies have confirmed that people are often less sensitive to sample size than normative models suggest (e.g., Kutzner, Read, Stewart, & Brown, 2016; Obrecht, Chapman, & Gelman, 2007).

Bayesian Updating

In contrast to the psychological principle of the belief in the law of small numbers, Bayesian principles suggest that sample size influences judgments. The Bayesian view offers a normative solution to how people deal with sample sizes. People can treat sample size as a measure of uncertainty, knowing that small samples can lead to a biased representation of the true outcome distribution and larger samples to a more veridical description of the true outcome distribution. Recent research has provided evidence that people indeed treat sample size as a measure of uncertainty, for instance, when they report preferences for consumer goods (De Martino, Bobadilla-Suarez, Nouguchi, Sharot, & Love, 2017) or make judgments about group differences (Obrecht, Chapman, & Suárez, 2010) or the performance of students (Fiedler, Walther, Freytag, & Plessner, 2002).

Cognitive Models and Hypothesis

To investigate how people deal with accumulating evidence, we tested whether they integrate samples by following Bayesian principles or by following the belief in the law of small numbers. For this purpose, we formalized two models that mathematically implement the two theories described above.

To understand the models, it is crucial to recognize the following facts about our experimental methods. In our experiments, we asked participants to indicate their valuations (i.e., selling prices) of two-outcome gambles where one of the outcomes was always zero. People were able to sample a *predefined* number of outcomes from the gamble's outcome distribution and then indicated their selling price for this gamble. Because there were only two outcomes, of which one was always zero, the following models were defined by estimating the probability of the nonzero gain outcome. Note that extensions of the models for more general cases with several outcomes are possible. Importantly, the sample sequences that people saw were always representative of the gamble's outcome distribution. Individual gambles were presented with different sample sizes.

Both models that we propose are expected utility models. According to expected utility theory, people's preferences about risky options can be represented by subjective utility functions. The subjective value of a risky option is defined as the sum of the subjective values of the outcomes weighted by the outcome probabilities. If outcome probabilities are not provided, as in experience-based tasks, they can be replaced by the decision makers' beliefs. Crucially, the two models that we propose differ with regard to how people develop beliefs about the outcome probabilities.

Relative frequency model. The relative frequency (RF) model formalizes how people who believe in the law of small numbers deal with sample sizes. It suggests that people simply calculate the relative frequency of all observed samples when forming a belief about the gain probability to make a valuation of the gamble.

Formally, the belief B_{RF} about the gain probability after t observations is given by $B_{RF,t} = \frac{1}{t} \times \sum_{t=1}^{t} f(t)$, where $f(t)$ is a sign function that equals 1 when the observed outcome is a gain and zero otherwise. In our experiments, we presented gambles with two outcomes,

one a gain and the other zero. Thus, we can formulate the valuation after sample t ($V_{RF,t}$) as $V_{RF,t} = \sqrt[\alpha]{B_{RF,t} \times u(g)}$, where $u(g)$ describes the subjective utility of the gain defined as $u(g) = g^{\alpha}$. The RF model predicts that sample size in itself does not influence valuations.

Following Griffin and Tversky (1992), we also asked participants for their confidence in their valuations. We tested whether people would be more confident about their responses when the sample size was small, as found by Griffin and Tversky (1992).

Bayesian value updating model. The Bayesian value updating (BVU) model implements a Bayesian view of how people deal with uncertainty as they gain new information about risky options. The model assumes that people update probability beliefs by following Bayesian principles. For the two-outcome gambles, we formulated the BVU model as follows: A person's belief about the gain probability after sampling t outcomes is represented by a beta distribution, $B_{BVU,t} \sim Beta(a, b)$. Before sampling ($t = 0$), the model assumes that participants judge the two outcomes as equally probable, $B_{BVU,t} \sim Beta(1, 1)$. As people sample, they update their initial belief about the gain probability by adding a value of 1 to the *a* parameter when observing a gain and by adding a value of 1 to the *b* parameter when observing a zero. The belief at time t about the probability of a gain is simplified as the mean of this distribution at that moment: $B_{BVU,t} = \frac{a_t}{a_t + b_t}$. Similar to the formulation of the RF model, the valuation is then formulated as $V_{BVU,t} = \sqrt[\alpha]{B_{BVU,t} \times u(g)}$.

The BVU model predicts that sample size influences valuations. This is because as sample size increases, the beta-distributed belief shifts from the average 50% in the direction of the true underlying probability of the gain. The model predicts that valuations decrease as a function of sample size when the true gain probability is smaller than 50% and increase as a function of sample size when the true gain probability is larger than 50%. Furthermore, the probability belief gets increasingly peaked around the true probability of the gain as sample size grows. Therefore, the model also predicts that confidence in valuations increases with increasing sample size.

Study 1

We defined two main goals and one subordinate goal for the experimental method of Studies 1 and 2. First, we aimed to measure people's strength of preference. Therefore, we prompted participants to value single gambles. This approach is in line with recent studies (e.g., Ashby & Rakow, 2014; Golan & Ert, 2014; Pachur & Scheibehenne, 2012) and offers the advantage of measuring preference more precisely than with choices. This is important, as detecting subtle differences in strength of preference requires fine-grained measurement scales. Second, we aimed to lay bare the precise effects of sample size. Therefore, we presented each gamble with various sample sizes. To allow for an unbiased comparison of different sample sizes, we eliminated sampling error by presenting representative samples. However, the order of individual samples within a sequence was random.

As as third, subordinate goal we sought to clearly compare valuations from experience with valuations from description. Therefore, we added a block where people made valuations from description. However, as our main goals involved studying how preferences change as a function of sample size, we refrained from manipulating the order of the experience and description blocks. Instead, all people made valuations first from experience and afterward in a separate block from description.

In Study 1, we examined how people form valuations from experience when they know the possible outcomes before they start sampling. In a situation where the outcomes are known, learning is simplified because participants only need to learn the outcome probabilities from experience.

Method

Participants. Forty people (31 women, 9 men) from the participant pool of the University of Basel between 18 and 40 years old ($M = 23.41$ years, $SD = 4.86$) participated in the study. Participants could choose between a show-up fee of 10 CHF (approximately \$11.10 at the time of the experiment) or course credit. Additionally, each participant received a bonus payment that depended on the outcome of a randomly selected trial ($M = \$6.38$, $SD = \$7.62$).

Materials. Participants repeatedly valued six different gambles that all consisted of two outcomes, a gain and a zero. Some gambles provided a small gain with high probability (p-bet gamble type, notation following Lichtenstein and Slovic (1971)); other gambles provided a high gain with low probability ($-bet gamble type, notation following Lichtenstein and Slovic (1971)). Table 1 gives an overview of the gambles. There were three pairs of gambles that were matched with regard to their expected value. Each pair consisted of one p-bet and one $-bet gamble. This implies that the gambles showed a relationship between probabilities and returns: High gains were less likely than small gains. There was, however, no perfect linear relationship between gain amount and probability; for example, a gain of 2.40 CHF (Gamble 5) was less likely than a higher gain of 4.70 CHF (Gamble 6). In the experience condition, the gambles were represented with four categories of sample size: extra small (xs = 5, 6, or 7), small (s = xs $\times$ 2), medium (m = xs $\times$ 3), and large (l = xs $\times$ 6). These sample sizes allowed us to represent rare events of 20% (1/5), roughly 17% (1/6), and roughly 14% (1/7).

Table 1

Gambles Used in Studies 1 and 2

Gamble	Gamble type	Sample-size category	Sample size	p(gain)	EV
1	$-bet	xs	5	1/5(16.00)	3.2
		s	10		
		m	15		
		l	30		
2	$-bet	xs	6	1/6(12.00)	2
		s	12		
		m	18		
		l	36		
3	$-bet	xs	7	1/7(28)	4.03
		s	14		
		m	21		
		l	42		
4	p-bet	xs	5	4/5(4.00)	3.2
		s	10		
		m	15		
		l	30		
5	p-bet	xs	6	5/6(2.40)	2
		s	12		
		m	18		
		l	36		
6	p-bet	xs	7	6/7(4.70)	4.03
		s	14		
		m	21		
		l	42		

Note. Sample size denotes the total number of observations with which the gambles were described in the experience condition. The heading p(gain) describes the probability (p) with which a gain occurred and the gain amount in Swiss francs (gain). The probability is expressed as the ratio of the relative frequency of the number of gain observations to the number of observations in the smallest sample size category (xs) of this gamble. EV = expected value; xs = extra small; s = small; m = medium; l = large.

For illustration of the different sample sizes, consider Gamble 1 in Table 1. It is a $-bet that offers a gain of 16 with a probability of 1/5 (20%). The different sample-size categories for this gamble are xs (5 observations, one gain), s (10 observations, two gains), m (15 observations, three gains), and l (30 observations, six gains).

Participants valued each gamble three times in each sample-size category. To avoid memory effects between trials, we included 18 distraction trials consisting of six distraction gambles. These gambles had the same outcomes as the gambles in Table 1 except that the gains of $-bets occurred with a probability of 75% and the gains of p-bets with a probability of 25%. These distraction gambles were presented with sample sizes of 4, 8, and 16. Since their only purpose was to avoid memory effects, we did not analyze these trials. The presentation order of all trials was fully randomized.

Furthermore, participants valued each gamble three times in the descriptive condition. In this condition, participants got a description of the outcomes and the probability with which the outcomes occurred. The probabilities were rounded to one decimal place, for example, 1/5 = 20%, 1/6 = 16.7%, 1/7 = 14.3%. Again, presentation order of all trials was fully randomized.

Procedure. Participants read printed instructions and completed a questionnaire that ensured their understanding of the instructions. They valued gambles first from experience and afterward from description. In the experience phase, participants learned about a gamble's outcome distribution by repeatedly sampling outcomes. Prior to and throughout each trial, participants saw the possible outcomes displayed above a virtual urn. Gains were always displayed as a black "marble," a solid black circle in which the gain amount was written in numerals. The zero outcome was presented as a white marble, a solid white circle in which the

numeral 0 was written. Participants knew that in each trial, they had to sample a predefined number of marbles. They did not know the precise number, which changed from trial to trial. The order of trials and the sequence of black and white marbles within a trial was randomized for individual participants. Participants pressed the space bar to see a marble, which was then revealed for 250 ms. After participants had drawn all outcomes in a trial, they were prompted for (1) their valuation of the gamble and (2) their confidence in their valuation on a 7-point Likert scale, ranging from 0 (*very unconfident*) to 6 (*very confident*).

To measure valuations, we elicited selling prices by following the Becker–DeGroot–Marschak (BDM) method (Becker, DeGroot, & Marschak, 1964). For each gamble, participants stated their selling prices. The selling price was defined as the lowest amount of money for which they would give up the right to gamble. Prices were allowed to range between 0.00 CHF and the current gain in Swiss francs, with possible increments of 0.10 CHF. A BDM auction was incentivized as follows: The selling price that a participant entered was compared to a random value drawn from all possible selling prices in the trial. If this random number was larger than the stated selling price, the participant "sold" the gamble and received a monetary amount that equaled the random number. If the random number was equal to or smaller than the selling price, the participant got to gamble. For the BDM, the optimal strategy was always to report the true selling price.[1] Detailed instructions as well as an example of the BDM auction were presented to explain the auction's mechanism. We used a short questionnaire to ensure that all participants had understood how the auction worked.

Results

For all analyses we used the software R (R Core Team, 2014). For the linear model comparisons, we used the BayesFactor package (Morey & Rouder, 2014). We based our inferences on the Bayes factor (BF_{ij}), which quantifies how much more or less likely the data

[1] To understand this logic, consider the following example: A person has a true selling price for a given gamble of $3. If this person sets her selling price too low (e.g., $2) and the randomly generated number is $2.10, she sells her gamble for a lower price than her true selling price. If, however, she sets her selling price too high (e.g., $5) and the randomly generated number is $4, she keeps and plays her gamble, although she actually would have preferred to receive the $4.

are under Model i (M_{i}) than Model j (M_{j}). A Bayes factor of 1 ($BF_{\text{ij}} = 1$) means that the data do not discriminate between the two models M_{i} and M_{j}. A Bayes factor of 5 (i.e., $BF_{\text{ij}} = 5$) means that the data are 5 times more likely under M_{i} than under M_{j}. A Bayes factor of $1/5$ ($BF_{\text{ij}} = 1/5$) means that the data are 5 times more likely under M_{j} (note that $BF_{\text{ji}} = 1/BF_{\text{ij}}$). In all analyses below, we included participant as a random factor if not mentioned otherwise.

Sample size. Table 2 displays the mean and median valuations from experience, that is, the monetary amount elicited with the BDM method, separately for each gamble and each sample-size category. The data suggest that sample size does not affect valuations from experience consistently. A model comparison supports this indication. Model M_0, which predicts valuations from experience as a function of the random factor expected value and the fixed factor gamble type (p-bet vs. $-bet), is preferred over a model that takes into account sample size as an additional fixed factor ($BF_{01} = 409$) and a model that additionally takes into account the interaction between sample size and gamble type ($BF_{01} > 1,000$).

Table 2

Valuations, Study 1

Gamble	Condition	Sample size	*p*(gain)	EV	Median	Mean	D–E	BF_{10}
1	E	5 (xs)	1/5(16.00)	3.20	5.00	5.16	0.56	4.43
		10 (s)			4.55	5.30	0.70	64.94
		15 (m)			5.00	5.34	0.74	24.12
		30 (l)			5.00	5.29	0.69	147.10
	D				4.00	4.60		
2	E	6 (xs)	1/6(12.00)	2.00	4.00	4.33	0.72	132.21
		12 (s)			4.00	4.31	0.70	625.86
		18 (m)			4.00	4.04	0.43	4.46
		36 (l)			4.00	3.99	0.38	1.85
	D				3.00	3.61		
3	E	7 (xs)	1/7(28.00)	4.03	6.00	7.56	0.99	4.32
		14 (s)			6.70	8.40	1.83	905.21
		21 (m)			6.20	7.92	1.35	141.95
		42 (l)			6.00	7.68	1.11	15.99
	D				5.00	6.57		

4	E	5 (xs)	4/5(4.00)	3.20	3.00	2.80	-0.28	5,765.48
		10 (s)			3.20	2.92	-0.16	3.24
		15 (m)			3.00	2.80	-0.28	57.41
		30 (l)			3.00	2.95	-0.13	1.41
	D				3.20	3.08		
5	E	6 (xs)	5/6(2.40)	2.00	2.00	1.77	-0.10	8.03
		12 (s)			2.00	1.75	-0.12	14.56
		18 (m)			2.00	1.73	-0.14	28.34
		36 (l)			2.00	1.81	0.06	0.71
	D				2.00	1.87		
6	E	7 (xs)	6/7(4.70)	4.03	4.00	3.46	-0.18	2.66
		14 (s)			4.00	3.55	-0.09	0.34
		21 (m)			4.00	3.58	-0.06	0.19
		42 (l)			4.00	3.70	0.06	0.22
	D				4.00	3.64		

Note. Median and mean valuations across participants. The heading p(gain) describes the probability (p) with which a gain occurred and the gain amount in Swiss francs (gain). The probability is expressed as the ratio of the relative frequency of the number of gain observations to the number of observations in the smallest sample size category (xs) of this gamble. D–E describes the difference between mean valuations from description (D) and those from experience (E). Bayes factor BF_{10} quantifies the evidence for a linear model (M_1) that predicts that valuations differ between description and experience over a linear model (M_0) that predicts no such difference. Both models contain participant as a random factor. EV = expected value; xs = extra small; s = small; m = medium; l = large. Gambles 1, 2, and 3 represent $-bets and Gambles 4, 5, and 6 represent p-bets.

Modeling procedure. To study the role of sample size more closely, we examined how well the RF model and the BVU model modeled participants' valuations in the experience-based condition. To test the psychological plausibility of both models, we also compared them to a baseline model. This baseline model predicts random responses, where each response between zero and the gain amount is equally likely.

Before estimation, we rescaled both the observed and the model-predicted valuations by

dividing them by the possible gain of the gamble in a trial. Consequently, all data points lay in the interval between 0 and 1. All models were estimated by applying maximum likelihood methods to participants' individual data. To compute the likelihood, we assumed that observed valuations followed a truncated normal distribution around the model's predicted valuation with a standard deviation that was estimated as a free parameter for each participant. Thus, the RF and BVU models, but not the baseline model, were equipped with one additional free standard deviation parameter. We searched for the set of parameter values that minimized the deviance, defined as the negative log likelihood of the data given the model and its parameter values. To identify the best model parameters, we used a brute-force grid-search approach. We searched the parameter space for the utility parameter α between 0 and 3 in steps of 0.025 and for the standard deviation of the truncated normal distribution between 0.00001 and 0.3 in steps of 0.01. We compared the models based on Bayesian model weights that we computed from the Bayesian information criterion (Kass & Raftery, 1995; Lewandowsky & Farrell, 2011).

Modeling results. Figure 1 displays the relative evidence for each of the models. It shows for how many participants (x axis) each model is favored and how strong this evidence is (grayscale). The data of 19 participants (47.5%) were best described by the BVU model. Evidence for individual participants is predominantly very strong or strong. For 18 participants (45%), the RF model performed best. The data of three participants (7.5%) were fit best by the baseline model.

Next, we explored the data using the results of the quantitative model comparison. For this purpose, we plotted individual valuations against the predictions made by the best fitting model separately for every participant given the optimal parameter estimates for that participant. This plot gives an indication of whether the models qualitatively capture data patterns. Figure 2 shows that the models generally capture the data well. However, inspecting the figure reveals that in some cases, even the better fitting model does not capture the data patterns well. More precisely, this holds for four participants (marked with an x in Figure 2). For these participants the model is rejected because of the lack of qualitative fit.

The fact that similar numbers of participants were best described by the BVU model

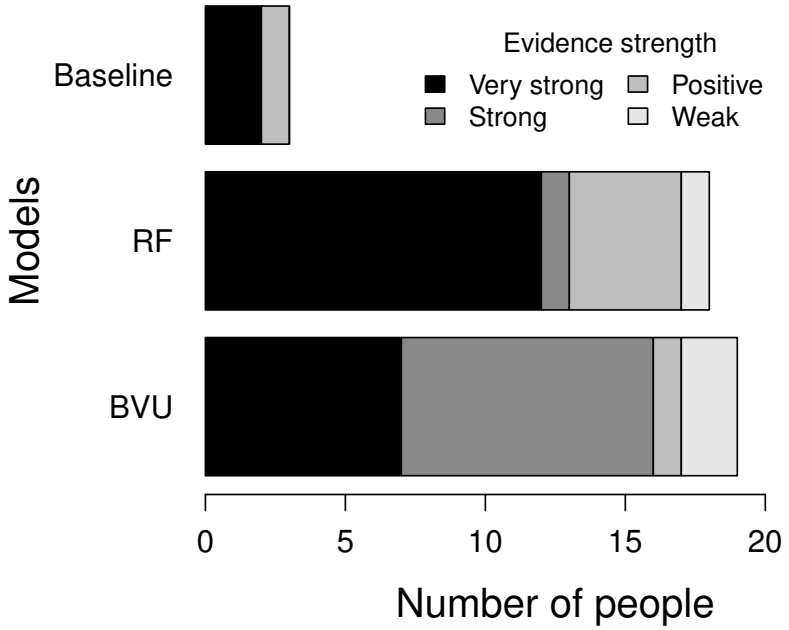

Figure 1. Number of people who were best fit by each model. Evidence strength is indicated by shades of gray. Baseline = Baseline model; BVU = Bayesian value updating model; RF = relative frequency model.

(Bayesian learners) and the RF model (frequentist learners) can explain the lack of an average effect of sample size on the valuations that we observed in the previous analysis. This is because only Bayesian learners are expected to attend to sample size.

Therefore, we next qualitatively investigated how the valuations of frequentist and Bayesian learners differ. We expected an effect of sample size only for Bayesian and not for frequentist learners. As explained previously, the BVU model predicts that valuations decrease as a function of sample size for $-bets and increase for p-bets. Figure 3 shows the mean ratings for $-bets and p-bets separately for people quantitatively and qualitatively best described by either the BVU or the RF model. The figure shows that the mean valuations of Bayesian learners indeed differed from the mean valuations of frequentist learners: The mean valuations of Bayesian learners slightly decreased as a function of sample size for $-bets and increased as a function of sample size for p-bets. The effect appears to be more pronounced for p-bets than

for $-bets. The mean valuations of frequentist learners did not show consistent variation.

Effect of gamble type and sampling order. The mean and median ratings in Table 2 show that valuations for $-bets (Gambles 1–3) were higher than for p-bets (Gambles 4–6), even if the gambles had the same expected value. This finding is supported by comparing model M_0, which assumes that gamble valuation can be predicted as a function of the random factor expected value, and model M_1, which makes the additional assumption that valuations differ between p-bets and $-bets ($BF_{10} > 1,000$).

To test also for recency or primacy effects, we compared how well the first half and the second half of the sample sequence in each trial predicts valuations. A linear model M_0 that predicts valuations as a function of the random factor gamble (1–6) outperforms model M_1, which also includes the mean of the first half of the observed samples as a fixed factor ($BF_{01} = 16.6$) and model M_2, which includes the mean of the second half of observed samples ($BF_{02} = 6.7$). In summary, our analysis did not provide evidence for recency or primacy effects.

Confidence ratings. Participants' mean confidence ratings of their valuations from experience in the extra small, small, medium, and large sample-size categories were xs = 4.11 ($SD = 1.10$); s = 4.15 ($SD = 1.04$); m = 4.14 ($SD = 1.03$); and l = 4.16 ($SD = 1.06$). Sample size did not influence participants' confidence systematically: M_0, which predicts confidence rating as a function of a random participant effect, was strongly preferred over M_1, which in addition includes the sample-size category as a predictor ($BF_{01} = 737$).

Because only the BVU model predicts increasing confidence ratings as a function of sample size, we repeated the previous analysis separately for Bayesian and frequentist learners. Interestingly, the analysis of those participants who were quantitatively and qualitatively best described by the BVU model revealed a small effect of sample size (mean confidence ratings: xs = 3.87, $SD = 1.05$; s = 4.05, $SD = 0.94$; m = 4.03, $SD = 0.94$; l = 4.09, $SD = 0.96$; $BF_{10} = 2.94$). The analysis of those participants who were quantitatively and qualitatively best described by the RF model did not show an effect of sample size on confidence ratings (mean confidence ratings: xs = 4.23, $SD = 1.23$; s = 4.19, $SD = 1.18$; m = 4.17, $SD = 1.17$; l = 4.13, $SD = 1.22$; $BF_{01} = 164$).

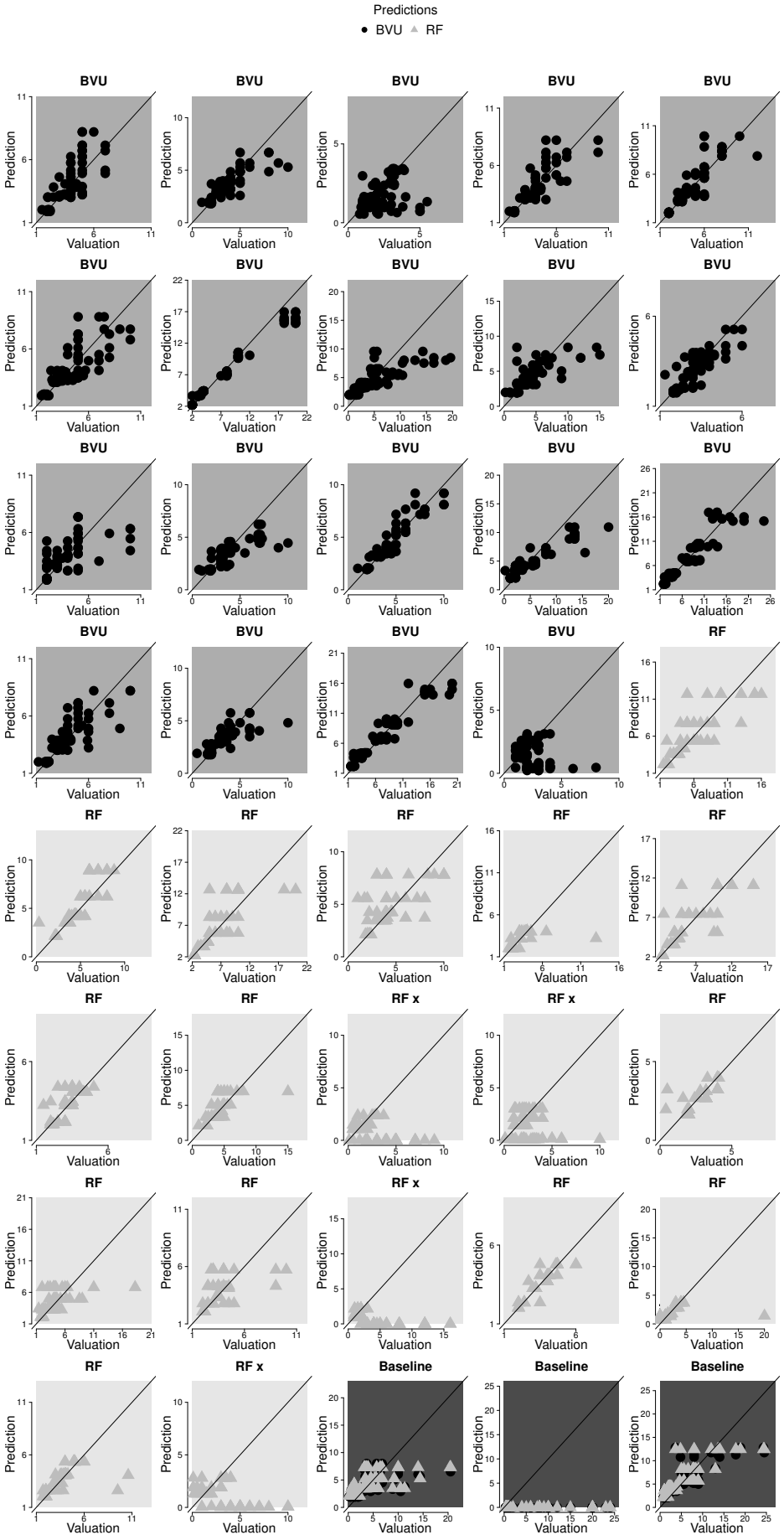

Figure 2. Individual valuations plotted against model predictions given the optimal parameter estimates for each individual. The titles reflect the model that best fitted the data and produced the predictions. BVU = Bayesian value updating; RF = relative frequency.

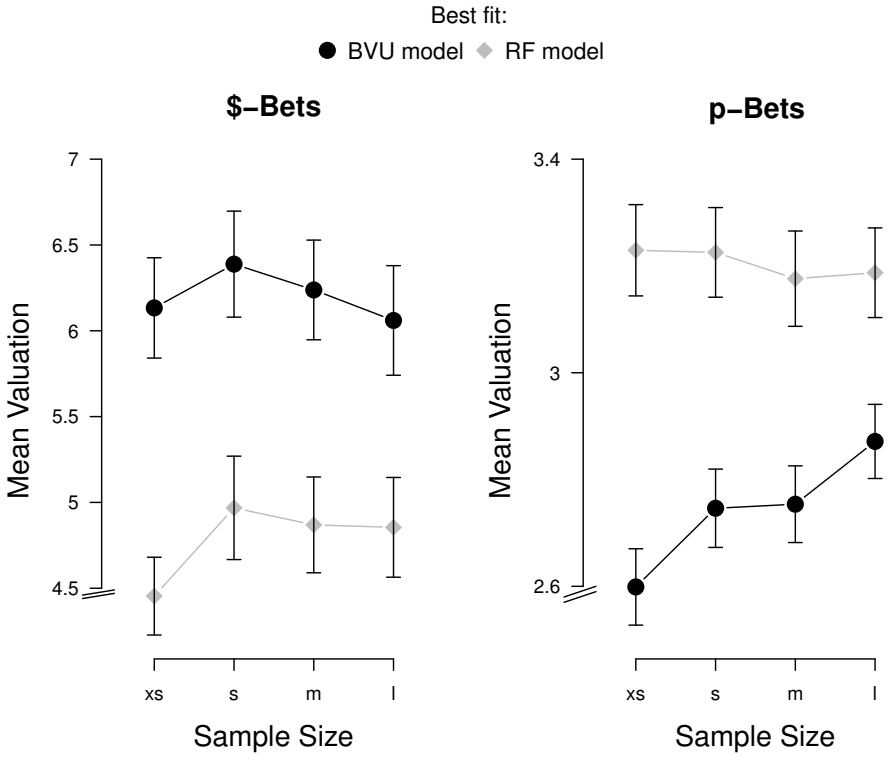

Figure 3. Mean valuations of participants who were best described by the Bayesian value updating (BVU) model (black) and the relative frequency (RF) model (gray). Error bars indicate the standard error of the mean. Sample sizes: xs = extra small; s = small; m = medium; l = large.

Description versus experience. Table 2 displays the mean and median valuations from description. It also shows the difference between the mean valuations in the experience and descriptive conditions (column D–E, separately for different sample sizes). Further, it provides the Bayes factors, quantifying the evidence in favor of a difference between valuations from experience and those from description.

Valuations made from description and experience differed for most of the gambles and sample sizes (see Table 2, rightmost column). In particular, participants attached a higher value to experienced than to described $-bets (Gambles 1–3) but attached a higher value to described than to experienced p-bets (Gambles 4–6). Thus, we found a D–E gap that is the opposite of the classic D–E gap observed in choice paradigms. In our study, participants valued gambles as if they overweighted rare events from description *and* experience. This

effect was even stronger when people made valuations from experience.

We also compared participants' confidence ratings of valuations from experience in each sample-size category to those from description ($M = 4.04$, $SD = 1.08$). Separately for each sample-size category, we compared M_0, which predicts confidence as a function of random participant effects, with M_1, which takes condition as an additional fixed factor into account. The analyses suggest that participants were slightly more confident about their ratings from experience than from description for small ($BF_{10} = 2.5$), medium ($BF_{10} = 1.3$), and large ($BF_{10} = 4.8$) sample sizes. For the extra small sample sizes ($BF_{10} = 0.3$), confidence judgments did not differ.

Discussion of Study 1

Study 1 shows that the size of the sample overall did not have an effect on valuations from experience. However, our quantitative model comparison shows that roughly half of the participants were best described by a BVU model that assumes that people treat sample size as a measure of uncertainty. The data of the other half of participants were best described by the RF model that disregards sample size. These results explain why we failed to find an effect of sample size across the whole data set.

Qualitatively, the valuations of people best described by the RF model differed from those of people best described by the BVU model. Figure 3 illustrates the trend that valuations of people best described by the BVU model changed as a function of sample size. Especially for p-bets, valuations increased as sample size grew. For participants best described by the RF model, valuations appear not to have varied systematically with sample size.

Valuations also differed when made from description versus experience. Interestingly, this difference is the opposite of the classic D–E gap that has been reported in choice paradigms: We found evidence that people behaved as if they overweighted rare events when making valuations from description *and* experience. Notably, this overweighting of rare events was even stronger for the experience condition. Importantly, people assigned higher valuations to $-bets (i.e., gambles that provide a high reward with low probability) than p-bets (i.e., gambles that provide a small reward with low probability) even when both gambles had

similar expected values. This finding is in line with recent research that suggested people overweight extreme outcomes in experience-based tasks (Ludvig, Madan, McMillan, Xu, & Spetch, 2017).

We found that participants' confidence in their valuations was generally high ($M = 4.12$, $SD = 1.06$ on a scale of 0 to 6). Importantly, only when we restricted the analysis to participants who were best described as Bayesian learners did we find that confidence increased with sample size as predicted by Bayesian models. We did not find an effect of sample size for participants who were best described by the RF model. If there was any trend, it seemed that confidence decreased with growing sample size, in line with Griffin and Tversky (1992). Further, we found evidence that people were more confident in their valuations from experience than from description.

Study 2

In Study 1, participants knew the outcomes before they started sampling. This is different from typical studies on decisions from experience, in which people have to learn the value of the outcomes during the sampling process. Therefore, in Study 2, we changed our experimental method such that the outcomes were learned from sampling.

Method and Results

Before participants started sampling, they knew only that each gamble consisted of two outcomes, of which one was zero and the other a gain. In contrast to Study 1, the precise gain amounts were not displayed above the virtual urn before or during the sampling stage. In all other aspects, Study 2 was identical to Study 1.

Forty people (38 women, 2 men) between 18 and 27 years old ($M = 21.1$ years, $SD = 2.05$) from the subject pool of the University of Basel participated. We followed the incentive scheme of Study 1. On average, participants earned a bonus of 7.16 CHF ($SD = 6.91$). For the data analysis, we compared the same statistical models as described in Study 1.

Table 3

Valuations, Study 2

Gamble	Condition	Sample size	p(gain)	EV	Median	Mean	D–E	BF_{10}
1	Experience	5 (xs)	1/5(16)	3.20	5.15	6.29	0.86	11.13
		10 (s)			5.00	6.39	0.96	354.66
		15 (m)			5.95	6.38	0.95	78.07
		30 (l)			6.00	6.53	1.10	171.43
	Description				4.00	5.43		
2	Experience	6 (xs)	1/6(12)	2.00	4.00	4.55	0.24	0.26
		12 (s)			4.00	4.72	0.41	1.25
		18 (m)			5.00	4.88	0.57	3.99
		36 (l)			4.95	4.78	0.47	3.58
	Description				3.35	4.31		
3	Experience	7 (xs)	1/7(28)	4.03	8.95	10.66	1.40	5.80
		14 (s)			10.00	10.26	1.01	1.04
		21 (m)			9.70	10.37	1.12	1.67
		42 (l)			10.00	11.18	1.93	220.94
	Description				6.00	9.25		
4	Experience	5 (xs)	4/5(4)	3.20	3.30	2.91	-0.29	13.53
		10 (s)			3.20	2.94	-0.26	25.50
		15 (m)			3.20	3.02	-0.18	1.74
		30 (l)			3.20	3.09	-0.11	0.42
	Description				3.50	3.20		
5	Experience	6 (xs)	5/6(2.4)	2.00	2.00	1.82	-0.10	1.32
		12 (s)			2.00	1.85	-0.07	0.37
		18 (m)			2.00	1.85	-0.07	0.34
		36 (l)			2.10	1.95	0.03	0.20
	Description				2.00	1.92		
6	Experience	7 (xs)	6/7(4.7)	4.03	4.00	3.68	-0.16	0.60
		14 (s)			4.10	3.78	-0.06	0.17
		21 (m)			4.15	3.85	0.01	0.14
		42 (l)			4.20	3.90	0.06	0.18
	Description				4.10	3.84		

Note. Median and mean valuations across participants. The heading p(gain) describes the probability (p) with which a gain occurred and the gain amount in Swiss francs (gain). The probability is expressed as the ratio of the relative frequency of the number of gain observations to the number of observations in the smallest sample size category (xs) of this gamble. D–E describes the difference between mean valuations from and experience (E). Bayes factor BF_{10} quantifies the evidence for a linear model (M_1) that predicts that valuations differ between description and experience over a linear model (M_0) that predicts no such difference. Both models contain participant as a random factor. EV = expected value; xs = extra small; s = small; m = medium; l = large. Gambles 1, 2, and 3 represent $-bets and Gambles 4, 5, and 6 represent p-bets.

Sample size. Table 3 displays mean and median valuations per gamble and sample-size category. Similar to in Study 1, mean valuations did not differ much between sample sizes: Again, valuations were best predicted by the random factor expected value and the fixed factor gamble type (additional fixed factor sample size: $BF_{01} = 285$; additional interaction between sample size and gamble type: $BF_{01} > 1,000$).

Modeling results. The behavior of 13 participants (32.5%) was best described by the BVU model. However, for the majority of participants (24 people, 60% of all participants), the RF model performed best. The baseline model was preferred for three participants (7.5%). Generally, the results illustrate the psychological plausibility of both models, but the dominance of the RF model shows that most participants were not very sensitive to the sample sizes.

As in Study 1, we examined the data qualitatively. In Figure 5, participants' valuations are plotted against the predicted valuations of the best fitting model given the optimal parameter estimates for each individual. Again the predictions capture the data well.

Figure 6 shows the mean valuations of the frequentist and Bayesian learners separately for $-bets and p-bets. Again, the two groups differed. As predicted by the BVU model, the valuations of Bayesian learners changed as a function of sample size. Especially for p-bets, valuations increased with sample size.

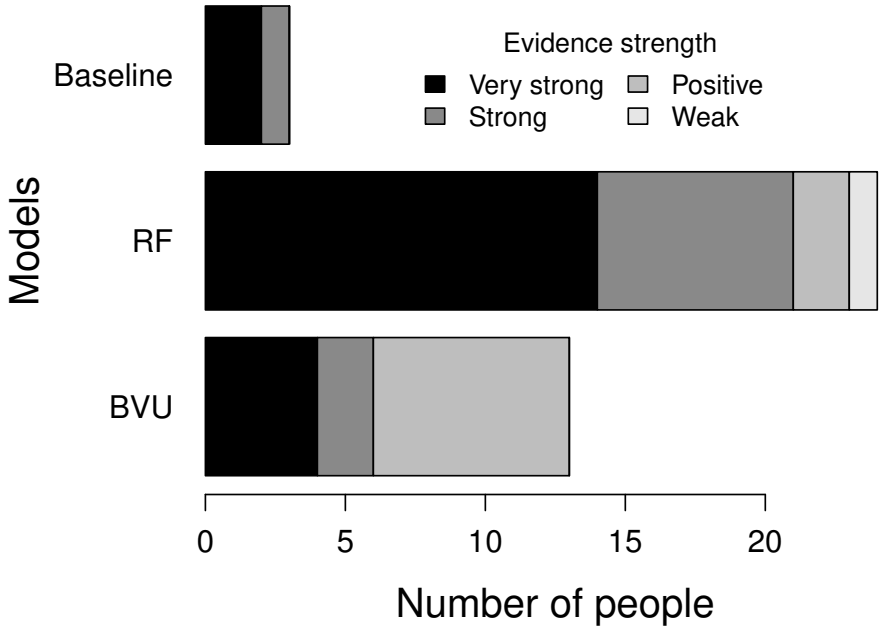

Figure 4. Number of people who were best fit by each model. Evidence strength is indicated by shades of gray. Baseline = Baseline model; BVU = Bayesian value updating model; RF = relative frequency model.

Effect of gamble type and sampling order. In line with Study 1, the mean and median ratings in Table 3 show higher valuations for $-bets (Gambles 1–3) than for p-bets (Gambles 4–6) even if the gambles had the same expected value (BF_{10} >1,000). We did not find evidence for any recency or primacy effects. M_0, which predicts valuations as a function of the factor gamble (1–6), was slightly preferred over models that also include the mean of the first half of the observed samples (M_1) or the mean of the second half of the observed samples (M_2) as predictors ($BF_{01} = 3.4$, $BF_{02} = 1.9$).

Confidence ratings. The mean confidence ratings in the extra small, small, medium, and large sample-size categories were xs = 4.01 ($SD = 1.23$); s = 4.09, ($SD = 1.14$); m = 4.04 ($SD = 1.19$); and l = 4.16 ($SD = 1.19$). M_0, which predicts confidence rating as a function of just a random participant effect, was again preferred over a model that also includes sample-size category as predictor ($BF_{01} = 11.2$).

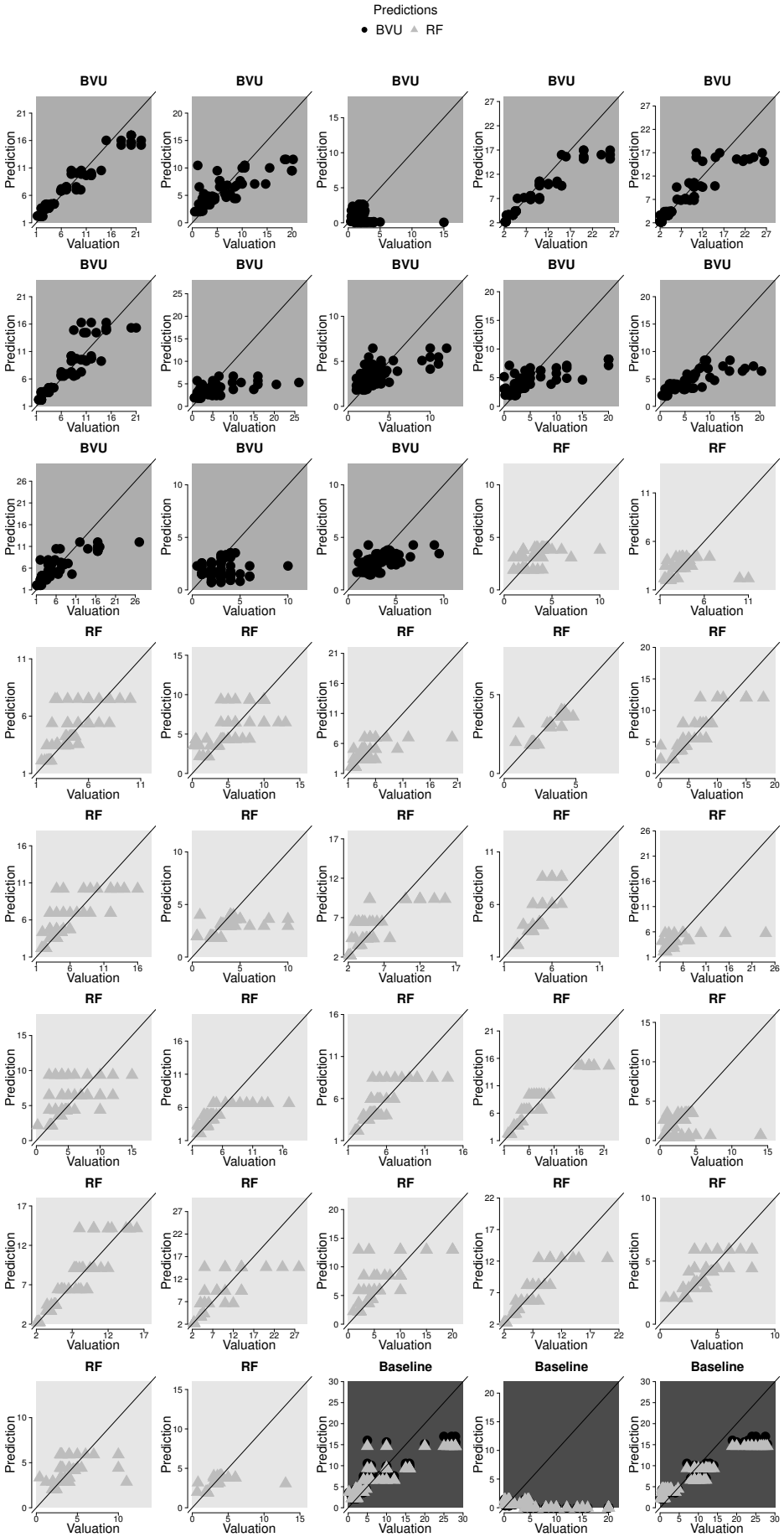

Figure 5. Individual valuations plotted against model predictions given the optimal parameter estimates for each individual. The titles reflect the model that best fitted the data and produced the predictions. BVU = Bayesian value updating; RF = relative frequency.

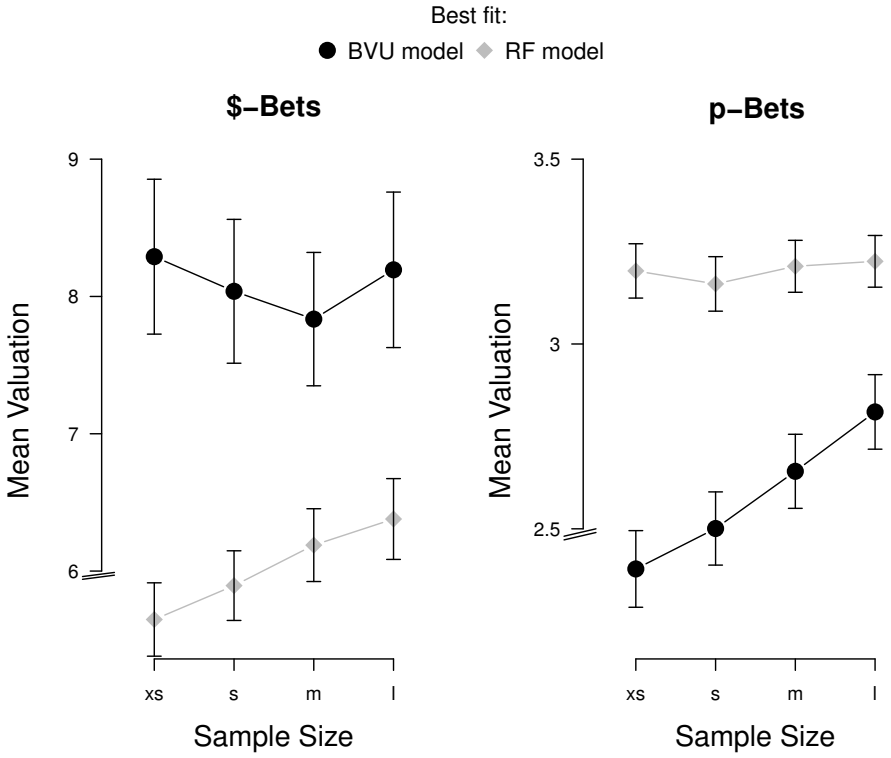

Figure 6. Mean valuations of people who were best fit by the Bayseian value updating (BVU) model (black) and the relative frequency (RF) model (gray). Sample sizes: xs = extra small; s = small; m = medium; l = large. Error bars indicate the standard error of the mean.

Considering only those data sets that were best described by the BVU model, confidence ratings increased as a function of sample size (mean confidence ratings: xs = 3.84, $SD = 1.31$; s = 3.95, $SD = 1.16$; m = 4.02, $SD = 1.13$; l = 4.1, $SD = 1.28$). However, the simpler model that predicts no influence of sample size was still slightly preferred over the model that predicts an effect of sample size ($BF_{01} = 1.3$). Similar to in Study 1, the data sets that were quantitatively and qualitatively best described by the RF model did not show an effect of sample size (mean confidence ratings: xs = 4.04, $SD = 1.19$; s = 4.14, $SD = 1.09$; m = 4.02, $SD = 1.21$; l = 4.13, $SD = 1.13$; $BF_{01} = 72$).

Description versus experience. Table 3 displays the mean and median valuations from experience and from description. It also shows the difference between the mean valuations in the experience and descriptive conditions (column D–E, separately for different sample sizes).

In line with Study 1, valuations made from experience and description differed. Participants valued experienced $-bets higher than described $-bets. In contrast, participants valued described p-bets higher than experienced p-bets. These observations are again in line with the assumption that the probabilities of rare events were overweighted in both conditions, and that this effect was more pronounced for valuations from experience.

Similar to in Study 1, the Bayes factors (rightmost column of Table 3) generally suggest a difference between description and experience. However, the finding is, especially for p-bets, less pronounced than in Study 1. Again, the gap is in contrast to the classic D–E gap reported in the decision-making literature.

Participants' mean confidence from description was 3.99 ($SD = 1.12$). Only for large sample sizes were participants more confident about their valuations from experience than they were about valuations from description ($BF_{10} = 53$). There is no evidence for this effect for extra small ($BF_{10} = 0.1$), small ($BF_{10} = 0.7$), or medium ($BF_{10} = 0.1$) sample sizes.

Discussion of Study 2

Study 2 tested whether the results of Study 1 are generalizable to typical sampling paradigms where people have to learn the outcome values from experience. The results of Studies 1 and 2 are remarkably similar: Participants behaved as if they overweighted rare events, when making valuations from both experience and description. In line with Study 1, we found that people could be classified as Bayesian and frequentist learners.

General Discussion

We studied how sample sizes influence people's valuations of risky gambles in two studies. Whereas in Study 1 the possible gain amount was known before people started sampling, in Study 2 the gain amount had to be learned through sampling. The main goal of these studies was to investigate how people form preferences from experience by comparing gamble valuations after having observed different numbers of outcomes, that is, after encountering different sample sizes. Further, we studied how valuations differ when based on description versus experience.

Belief in the Law of Small Numbers Versus Bayesian Updating

Across participants, Studies 1 and 2 showed that the number of samples people drew did not influence their valuations of gambles from experience. People's confidence in valuations was also constant across sample sizes.

It has been previously shown that people believe in the representativeness of short sequences of outcomes (Griffin & Tversky, 1992; Tversky & Kahneman, 1971). The RF model that uses relative frequencies of observed outcomes as probability estimates captures this logic. Accordingly, people attend only to the relative frequency of outcomes when they form gamble valuations and neglect the size of the samples. The model best described the behavior of 40% of the participants in Study 1 and 60% of the participants in Study 2, which suggests that a substantial proportion of people believe in the law of small numbers. This finding helps explain why people often do not sample much in experience-based tasks: If people believe that a short sequence of outcomes represents a prospect's outcome distribution comprehensively, they do not need to sample much, as they believe that sampling more will not yield new information.

However, not all participants were best described by the RF model: 42.5% of the participants in Study 1 and 30% of the participants in Study 2 were classified as Bayesian learners. Such people treat sample size as information about the uncertainty that the sample sequence entails. As sample size grows, people's prior beliefs about the outcome probabilities have a decreasing impact and valuations of gambles change accordingly.

Generally, the classification of people into two different subgroups suggests that people use different strategies when they update information from experience. This classification is supported by people's confidence judgments: When we analyzed confidence judgments separately for people who were classified as believers in the law of small numbers and Bayesian learners, we found that only Bayesian learners were more confident about their judgments for larger than for smaller samples.

The question arises of why some people behave according to Bayesian principles and others do not. Future research should focus on examining what factors influence whether people behave according to Bayesian principles. One possible candidate that mediates what

strategy people use is numerical literacy. Previous research has, for instance, shown that numerical literacy correlates with performance in Bayesian reasoning tasks (Brase & Hill, 2017).

Description Versus Experience in Choice and Valuation

In risky choice tasks, people have been found to behave as if they overweight rare events from description and underweight rare events from experience. In our valuation tasks participants behaved as if they overweighted rare events from both description and experience. Thus, our findings are not in line with the classic D–E gap.

Here, it is noteworthy that in contrast to experience-based choice tasks, in our valuation task, participants only had to focus on one gamble in each trial. This task structure differs from two-gamble choice tasks. This difference in structure might elicit different cognitive processes. Indeed, Lichtenstein and Slovic (1971) showed that people's preferences based on description can be reversed when elicited by value judgments rather than choices: Someone who chooses Gamble A over Gamble B might still assign a higher value to Gamble B than to Gamble A. One of the first explorations of preference reversals in valuations from experience was done by Golan and Ert (2014). They reported a difference between valuations from description and valuations from experience. Yet, this difference pointed in a direction opposite to our findings: Participants overweighted rare events more strongly when doing so from description than from experience. There is, however, a crucial difference between our experimental design and the experimental design of Golan and Ert (2014). We presented representative sample sequences, whereas they allowed participants to draw as many samples as they wished. As we argued in the Introduction, in free-sampling paradigms, the experienced relative frequencies of outcomes do not necessarily match the underlying probability. Such sampling error might have contributed to the difference in the results. Alternatively, the difference could reflect that information is processed differently in free- and forced-sampling paradigms. For instance, people may pay more attention to a sampled outcome that resulted from a voluntary sampling decision than to a sampled outcome that resulted from a forced sampling decision.

Buying versus selling prices. In our experiments, we assessed preferences by asking people for selling prices for gambles. The endowment effect describes that people attach higher value to goods when they sell them than when they buy them (Thaler, 1980). Pachur and Scheibehenne (2012) demonstrated this endowment effect in an experience-based task. It would be interesting to replicate our experiments with buying instead of selling prices and more rigorously compare the results with buying and selling prices from description. However, this was beyond the scope of the present study. Here, our main focus was to investigate whether preferences are influenced by sample sizes. This question can be answered independently of response format. Irrespective of whether people are asked for selling prices or buying prices or make repeated choices, the Bayesian model predicts an effect of sample size on valuations.

Confidence from description and experience. People were more confident with their gamble valuations from experience than from description. This finding is remarkable given that from a normative perspective one could argue that people feel more confident when making valuations from description than when making valuations from experience. This is because valuations from description do not entail any uncertainty about the outcome distributions, whereas valuations from experience may entail such uncertainty.

Our findings do support research by Bradbury, Hens, and Zeisberger (2014). They reported that people who made investment decisions in the laboratory felt better informed and were more confident about their decisions from experience than those from description. Potentially, people perceive and process experienced information differently than described information (Kahneman, 2009): While information that is described is more abstract, experienced information has direct salience or impact. When people draw a sample, they not only observe the outcome but also experience emotional reactions resulting from that observation. Thus, people do not only learn plain facts about the outcome distribution; they also gain a more vivid understanding of this distribution. This more concrete understanding of the gamble can help people access their preferences and thus creates higher confidence in valuations.

Conclusion

In this study, we investigated how people form preferences from experience. Our results offer a new perspective on the finding that people sample very little when forming preferences from experience: A substantial proportion of people appear to believe in the law of small numbers; that is, their valuations do not change as a function of sample size.

References

Ashby, N. J. S., & Rakow, T. (2014). Forgetting the past: Individual differences in recency in subjective valuations from experience. *Journal of experimental psychology. Learning, memory, and cognition*, *40*, 1153–1162.

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

Bradbury, M. A., Hens, T., & Zeisberger, S. (2014). Improving investment decisions with simulated experience. *Review of Finance*, *19*, 1019–1052.

Brase, G. L., & Hill, W. T. (2017). Adding up to good Bayesian reasoning: Problem format manipulations and individual skill differences. *Journal of Experimental Psychology: General*, *146*, 577–591.

Camilleri, A. R., & Newell, B. R. (2009). The role of representation in experience-based choice. *Judgment and Decision Making*, *4*, 518–529.

Camilleri, A. R., & Newell, B. R. (2011). Description- and experience-based choice: does equivalent information equal equivalent choice? *Acta psychologica*, *136*, 276–284.

De Martino, B., Bobadilla-Suarez, S., Nouguchi, T., Sharot, T., & Love, B. C. (2017). Social Information is Integrated into Value and Confidence Judgments According to its Reliability. *Journal of Neuroscience*, *37*, 6066–6074.

Fiedler, K., Walther, E., Freytag, P., & Plessner, H. (2002). Judgment Biases in a Simulated Classroom – A Cognitive-Environmental Approach. *Organizational Behavior and Human Decision Processes*, *88*, 527–561.

Glöckner, A., Fiedler, S., Hochman, G., Ayal, S., & Hilbig, B. (2012). Processing differences between descriptions and experience: A comparative analysis using eye-tracking and physiological measures. *Frontiers in psychology*, *3*, 1–15.

Golan, H., & Ert, E. (2014). Pricing decisions from experience: The roles of information-acquisition and response modes. *Cognition*, *136*, 9–13.

Griffin, D., & Tversky, A. (1992). The weighing of evidence and the determinants of confidence. *Cognitive psychology*, *24*, 411–435.

Hadar, L., & Fox, C. R. (2009). Information asymmetry in decision from description versus

decision from experience. *Judgment and Decision Making*, *4*, 317–325.

Hau, R., Pleskac, T., & Hertwig, R. (2010). Decisions from experience and statistical probabilities: Why they trigger different choices than a priori probabilities. *Journal of Behavioral Decision Making*, *68*, 48–68.

Hau, R., Pleskac, T. J., Kiefer, J., & Hertwig, R. (2008). The description–experience gap in risky choice: the role of sample size and experienced probabilities. *Journal of Behavioral Decision Making*, *21*, 493–518.

Hertwig, R., Barron, G., Weber, E. U., & Erev, I. (2004). Decisions from experience and the effect of rare events in risky choice. *Psychological Science*, *15*, 534–539.

Kahneman, D. (2009). The myth of risk attitudes. *The Journal of Portfolio Management*, *36*, 1.

Kahneman, D., & Tversky, A. (1979). Prospect Theory: An Analysis of Decision under Risk. *Econometrica*, *47*, 263–291.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*, 773–795.

Kutzner, F. L., Read, D., Stewart, N., & Brown, G. (2016). Choosing the Devil You Don't Know: Evidence for Limited Sensitivity to Sample Size–Based Uncertainty When It Offers an Advantage. *Management Science*, *63*, 1519–1528.

Lewandowsky, S., & Farrell, S. (2011). *Computational modeling in cognition: Principles and practice.* Thousand Oaks, CA: Sage.

Lichtenstein, S., & Slovic, P. (1971). Reversals of preference between bids and choices in gambling decisions. *Journal of Experimental Psychology*, *89*, 46–55.

Ludvig, E. A., Madan, C. R., McMillan, N., Xu, Y., & Spetch, M. L. (2017). Edge proximity, not distinctiveness, produces overweighting of extreme outcomes in risky decisions from experience. *Retrieved from psyarxiv.com/my4hd*.

Morey, R. D., & Rouder, J. N. (2014). BayesFactor: Computation of Bayes factors for common designs [Computer software manual]. Retrieved from `http://CRAN.R-project.org/package=BayesFactor` (R package version 0.9.9)

Obrecht, N. A., Chapman, G. B., & Gelman, R. (2007). Intuitive t tests: Lay use of statistical information. *Psychonomic Bulletin & Review*, *14*, 1147–52.

Obrecht, N. A., Chapman, G. B., & Suárez, M. T. (2010). Laypeople do use sample variance: The effect of embedding data in a variance-implying story. *Thinking & Reasoning*, *16*, 26–44.

Pachur, T., & Scheibehenne, B. (2012). Constructing preference from experience: the endowment effect reflected in external information search. *Journal of experimental psychology. Learning, memory, and cognition*, *38*, 1108–1116.

R Core Team. (2014). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `http://www.R-project.org/`

Rakow, T., Demes, K. A., & Newell, B. R. (2008). Biased samples not mode of presentation: Re-examining the apparent underweighting of rare events in experience-based choice. *Organizational Behavior and Human Decision Processes*, *106*, 168–179.

Thaler, R. (1980). Toward a positive theory of consumer choice. *Journal of Economic Behavior & Organization*, *1*, 39–60.

Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological bulletin*, *76*, 105–110.

Ungemach, C., Chater, N., & Stewart, N. (2009). Are probabilities overweighted or underweighted when rare outcomes are experienced (rarely)? *Psychological science*, *20*, 473–479.

Reaching for the star ratings: A partly Bayesian account of how people use consumer ratings

Janine Christin Hoffart

Sebastian Olschewski

Jörg Rieskamp

University of Basel

Author Note

The authors report no conflict of interest. We thank Konstantin Escher, who contributed valuable ideas and feedback. Correspondence concerning this article should be addressed to Janine Christin Hoffart. University of Basel, Department of Psychology, Missionsstrasse 62a, 4055 Basel, Switzerland. E-mail: janine.hoffart@unibas.ch

Abstract

Online consumer ratings reflect how many people and how well they have rated a product. Not only the average ratings but also the number of ratings (i.e. the sample size) provide helpful information as many ratings signal higher reliability than few ratings. Yet, previous research has provided mixed results with regards to whether people pay attention to the sample size of statistical information: Some studies support the notion that people integrate sample size as suggested by Bayesian principles, others that people ignore sample size and again others that people treat sample size as indicating the quality of products. We propose that individual differences in how people deal with sample size can account for conflicting results of previous research. Therefore, in the present work, on the individual level, we examined to what extend people pay attention to the number of consumer ratings when making preferential choices. In an experiment, people choose between two competing hotels based on previous visitors' ratings. For each individual participant, we tested three different cognitive models describing how sample size and average ratings are processed: First, a Bayesian updating model that predicts an interaction between average ratings and sample size; second, a naive statistician model ignoring sample size, and third a weighted additive model using the average and the sample size as two independent cues. The results indicate substantial individual differences of how consumer ratings were used. In line with the hypothesis that the interaction predicted by the Bayesian model is cognitively more complex, people higher in statistical numeracy were better described by the Bayesian model. Interestingly, none of the models were able to capture important qualitative findings, namely that people only attend sample size in some choice situations but not in all. To account for these findings, we further developed a heuristic Bayesian decision tree that describes in which situations behaviour accords to Bayesian principles. The model captures the data pattern well and also performed well a reanalysis of data published in a previous study.

Keywords: Consumer ratings, decision making, Bayesian updating, belief in the law of small numbers

Reaching for the star ratings: A partly Bayesian account of how people use consumer ratings

In times of digitization, the internet becomes increasingly important for shopping, ordering food or renting hotel rooms (Fang, Wen, George, & Prybutok, 2016). Several online platforms allow to easily retrieve information about products. For one, people can identify objective features of products (e.g., does a hotel have a sauna or not) but they can also consult reviews to learn how satisfied previous consumers were. Indeed, people who use the internet for planning trips likely read other travelers' reviews about similar trips (Gretzel & Yoo, 2008). Such reviews influence people's preferences and purchasing decisions: Positive reviews raise the number of hotel bookings (Ye, Law, & Gu, 2009) and book purchases (Chen, 2008).

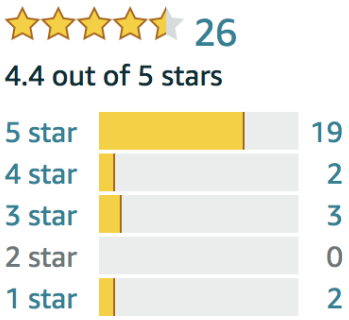

Figure 1. Example of how consumer reviews are often displayed in online shops.

Typically, rating information is displayed with summaries that indicate how much the reviewers liked a product on a scale from one star (worst possible rating) to five stars (best possible rating); how many people have rated the product, and the average rating of all reviews. Figure 1 illustrates a typical presentation format of customer ratings.

As online ratings only recently became influential through digitization, it has not been fully understood how people exploit rating information when making decisions. In particular, the question raises whether and how people integrate average ratings and the number of ratings. The concept of statistical power suggests that, when inferring product qualities, people should not only attend the average rating of the products. Instead, also the number of people who rated them as well as the dispersion of the raters' judgments provide important information (Obrecht, Chapman, & Gelman, 2007). For example, when a single person

recommends (not recommends) a restaurant, this recommendation is less reliable than when many people have recommended (not recommended) the restaurant. The single rater may have had an extraordinary good (or bad) experience at the restaurant, or is friends (enemies) with the owner. Such outliers are less influential, when many ratings are available. Therefore, average ratings that are based on many opinions are a more reliable information source.

A Bayesian View on Sample Size

Bayesian principles offer a solution for how people can deal with the uncertainty related to online ratings. Abstractly, when following Bayesian information integration people integrate *prior beliefs* with *new information* to *posterior beliefs*. When much new information about a product in form of ratings can be retrieved, posterior beliefs are stronger than when only little new information is observed. This implies that a decision maker who has observed many ratings has greater trust that the average rating about a product represents the true quality of this product well.

Concretely, Bayesian principles suggest that when confronted with two well rated choice options, the more often rated option is the better choice[1]. Consider the choice between Hotel A that has on average been rated with five stars by two people and Hotel B that has on average been rated with 4.5 stars by 200 people. In this situation, the average rating of Hotel B likely is - due to the larger number of ratings - a more reliable measure of the hotel's true quality than the average rating of Hotel A. Hence, although on average rated better, the few ratings of Hotel A indicate more downside potential than the many ratings of Hotel B. This means that the true quality of Hotel A may we much worse than five stars.

In contrast, when confronted with two poorly rated options, Bayesian principles suggest that the less often rated option is the better choice: People should prefer a Hotel C that has been rated with one star by two people over Hotel D that has been rated with 1.5 stars by one-hundred people. This is because Hotel C with its higher uncertainty offers upside-potential, i.e., in reality it may be of better quality. Hotel D on the other hand has been rated poorly by many people indicating that the true quality of the hotel can be expected to be

[1]This prediction arguably changes when the decision maker has a very strong prior belief. However, here we assume relatively weak and uniformly distributed prior beliefs.

low.

In sum, Bayesian principles predict an interaction between average ratings and the number of ratings: On the lower end of the scale, Bayesian updaters prefer rarely rated products and on the higher end of the scale often rated products.

There is evidence that people learn according to Bayesian principles in many domains as for instance in conditioning (e.g. Courville, Daw, & Touretzky, 2006), concept- and word-learning (e.g. Tenenbaum, Griffiths, & Kemp, 2006; Xu & Tenenbaum, 2007), and object perception (Kersten, Mamassian, & Yuille, 2004).

Closer related to the question whether and how people integrate the size of samples, in simulated classrooms, people consider how much they know about students when judging their abilities: If a student has performed poorly (well) multiple times, people judged the ability of the student as lower (higher) than when not much was known about the student's past performance (Fiedler, Walther, Freytag, & Plessner, 2002). Similarly, people were more confident that two group means differ when they based their judgments on large samples than when they based their judgments on small samples (Obrecht, Chapman, & Suárez, 2010). In a consumer context, people's preferences for goods changed as predicted by Bayesian principles when people received information about how other people have rated the goods (De Martino, Bobadilla-Suarez, Nouguchi, Sharot, & Love, 2017). In line with these findings, Khare, Labrecque, and Asare (2011) showed that people find poorly rated movies (2 stars) more attractive when they have been rated less often. However, this pattern changes when the movies have been rated well (4 stars) and people found movies that have been rated very often more attractive than movies than have only been rated a few times.

On Limited Integration of Sample Size

While Bayesian information integration provides a solution for how people can deal with uncertainty, many empirical studies have also shown that people often do not act in line with Bayesian principles. In the following, we focus on two non-Bayesian ways of how sample size can be treated: The first postulates that people ignore sample size as described by the *belief in the law of small numbers* (Tversky & Kahneman, 1971). The second postulates

that people take sample size as a indicator for quality (Powell, Yu, DeWolf, & Holyoak, 2017).

Ignorance of Sample Size: Belief in the Law of Small Numbers

The belief in the law of small numbers describes the finding that people often ignore sample size when making judgments. They treat short sequences of outcomes as if these represented the theoretical outcome distributions comprehensively (Tversky & Kahneman, 1971). Griffin and Tversky (1992) found that people neglected sample size when they judged whether a coin was biased after having observed only few coin tosses. In another study, people gave confidence judgments about which of two hotels is better based on how previous customers have rated the hotels on average and how many people have rated the hotels (Obrecht et al., 2007). People gave little weight to the number of ratings and mainly considered average ratings when making their judgments. Also, in a task where participants choose between risky gambles based on different numbers of previous outcomes with the goal to reach a predefined number of points, people ignored sample size (Kutzner, Read, Stewart, & Brown, 2016).

In sum, believers in the law of small numbers will not pay attention to the number of ratings and therefore always choose a better rated option. If two options are rated equal, preference will be random. Contrasting this argumentation, recently it has been argued that people treat sample size as a indicator for quality.

Social Inference: Sample Size as Quality Indicator

Relying on social information helps humans to solve problems efficiently (Call & Tomasello, 1995). Online reviews allow a relatively new form of social learning: People can retrieve preferences of thousands of people within seconds. Empirically, the number of customer reviews predict sales for new products. This holds especially for experience goods (Zhu & Zhang, 2010). In general, people are often influenced by other people, either by normative or informational social influence (Schöbel, Rieskamp, & Huber, 2016). According to normative influence, people are motivated to convey a specific impression to others, to obtain other people's approval and avoid rejections. According to informational social influence on the other hand, people infer useful and valid information from the actions and

opinions of others (Deutsch & Gerard, 1955). Experimentally Powell et al. (2017) investigated which of two products people preferred based on average consumer ratings and the number of ratings. In their choices, people treated the average ratings and the number of ratings as two independent indicators for quality and chose as if they assumed that the more popular a product and the better it was rated, the better its quality was.

In sum, people who treat sample size as indicating quality, will—given that the differences in average ratings are relatively small—choose a more popular over a less popular product if the average rating is poor.

Information Processing and Interindividual Differences

As outlined above, when it comes to the question of how people integrate other consumers' experiences, the results of previous research were inconclusive: Some researchers found that people integrate information according to Bayesian principles (e.g., De Martino et al., 2017); others reported that people ignore the number of ratings (e.g., Obrecht et al., 2007); and others that people treat the number of ratings as indicator for quality (e.g., Powell et al., 2017). A potential reason for these diverging findings is that different people use different strategies when forming preferences as has been reported elsewhere (e.g., Payne, Bettman, & Johnson, 1988; Rieskamp & Hoffrage, 2008).

Payne et al. (1988) argued that strategies differ with respect to the effort that is required to execute them. When integrating number of ratings and average ratings, Bayesian information integration arguably is more complex than the other two strategies: Believers in the law of small numbers ignore the number of ratings completely. People who treat the number of ratings as quality *always* treat it as indicating quality. Finally, Bayesian updaters treat it as positive indicator in some situations (on the high end of the scale) and as negative indicator in others (ratings on the lower end of the scale). This interaction arguably presumes that people need to be able to deal with statistical information. Therefore, the selection of different inference strategies should depend on people's statistical numeracy. We expect that people who make more choices in line with Bayesian predictions have higher statistical numeracy than people who ignore the number of ratings or treat it as indicator for quality.

This prediction is in line with previous research where it has been found that people higher in numeracy better perform in Bayesian reasoning tasks than people lower in numeracy(e.g., Brase & Hill, 2015; Chapman & Liu, 2009).

Models

We formalized three models that describe different decision strategies people may use: These are first, *Bayesian information integration*; second *the belief in the law of small numbers*, and third the notion that people treat both the *number of ratings* and the *average ratings* as indicators for quality.

When someone is facing a decision between option A and option B, the models propose that people make their choices, based on subjective values (V) that they assign to the options. How people determine the subjective values differs between models. To account for the stochasticity in decision making, each model (m) defines the probability of choosing option A over option B as:

$$p(A,B)_m = \frac{1}{1+e^{-\varphi \times [V(A)_m - V(B)_m]}}$$

, where $\varphi >= 0$ is a sensitivity parameter that determines how sensitively people react to differences in the subjective values of the options. Very large values of φ imply that people consistently choose the better option according to the model, even when there are only small differences.

Bayesian Information Integration

To implement the idea of Bayesian information integration, we defined the Bayesian updating (BU) model which assumes that at each point in time a person's belief (B) about how ratings of a option (e.g. hotel) are distributed can be described by a Dirichlet distribution with one parameter (h_i) for each possible rating: $B \sim Dir(h_1, h_2, ..., h_N)$. Before having observed ratings (t = 0), people may assume that all possible ratings are equally likely, i.e. $B_{t=0} \sim Dir(h_1 = s, h_2 = s, ..., h_N = s)$. The parameter s described how strong the prior belief is, if s is very large, much information is needed to overcome the prior belief. In our

study, we fixed s to 1. The model describes how people update their beliefs when receiving rating information (t = 1), by adding the number of times that the option (e.g. hotel) of interest has been rated with 1, 2, ... , and N stars to the according parameters of the belief distribution. Importantly, we assumed that when making choices people's subjective beliefs about the true quality of the option can be summarized by the mean of the posterior distribution. This implies that the subjective value of option j (V_j) equals $V_j = \frac{\sum_{i=1}^{N} h_i \times i}{\sum_{i=1}^{N} i}$.

Belief in the Law of Small Numbers: A Naive Sampling Heuristic

We created a model capturing the belief in the law of small numbers (BLSN) by assuming that people ignore the number of ratings completely and that their subjective value of an option can be simply represented by the average rating of the option. If one of the options has not been rated yet, the model predicts that people choose randomly by assigning the average rating of the reviewed option also to the never–rated option[2].

Number of Ratings as Quality Indicator: A Weighted Additive Model

Similarly to Powell et al. (2017), we implemented a weighted additive model that treats the number of ratings as indicating quality. The model assumes that people people's subjective value of a product is the weighted sum of the average rating and the dummy coded number of ratings (1 for the hotel with more ratings, 0 for the other Hotel). Similar to the BLSN model, if one hotel has not been rated yet, its average rating is the same as the average rating of the hotel that has received ratings. Two free parameters, w_{avg} and w_N indicate how important the average rating (w_{avg}) and the number of ratings (w_N) are. Both parameters are restricted to values larger than zero.

Choice Classes and Predictions

Rating information can differ with regards to the number of ratings and the average rating. We will refer to these as *cues* and to different combinations of cues as choice classes.

[2] An alternative prediction is that in uncertainty trials, where one hotel has not been rated yet, this hotel's rating is assumed to be equal to the average rating (Garcia-Retamero & Rieskamp, 2009). We have implemented this assumption and found that the results of our model comparison almost do not change.

To test the different models against each other, we created four choice classes for which the models make diverse predictions: The more often rated product was on average rated better (+ .5 stars), worse (- .5 stars) or equal than the less often rated product. Further, we investigated uncertainty trials in which one product has not been rated yet and the other product has been rated often very poorly (i.e., one star) or very well (i.e., five stars) [3].

General predictions. As will be explained in detail below, in particular for Bayesian principles, it is important to distinguish between predictions *generally* made by Bayesian principles, and predictions *specifically* made by the Bayesian model we implemented.

Without further model specifications and assuming weak prior beliefs, Bayesian principles predict an interaction between sample size and average ratings: If the difference in average ratings is relatively small, people prefer options with fewer ratings on the lower end of the scale and options with more ratings on the higher end of the scale. However, at which point of the rating scale (i.e., which average rating), people switch from preferring the less often to preferring the more often rated option, depends on how the theory is implemented and which prior beliefs are assumed. The principle of the belief in the law of small numbers *always* predicts that people choose the on average better rated hotel and make random choices when both hotels were rated equally or one hotel has not been rated yet. If people *always* treat sample size as *strong* indicator for quality, they will choose the hotel with more ratings. We will analyze these theoretical predictions *without* making specific model assumptions by analyzing on the aggregate level whether people's behaviour is *generally* in line with the three principles.

Specific model predictions. Further, on the individual level, we test the specific models we proposed. The predictions outlined below and in Figure 2 are specific for the models, we implemented and the decision problems we used. In the WA model, the weights assigned to average ratings and number of ratings are free parameters. Here we assumed that the weight for the number of ratings is positive, meaning that more ratings indicate better quality.

[3] In principle, options can be also rated equally often but one hotel can be rated better than the other. Here, we do not investigate such trials as all theories that be compare predict that people would choose the better rated hotel.

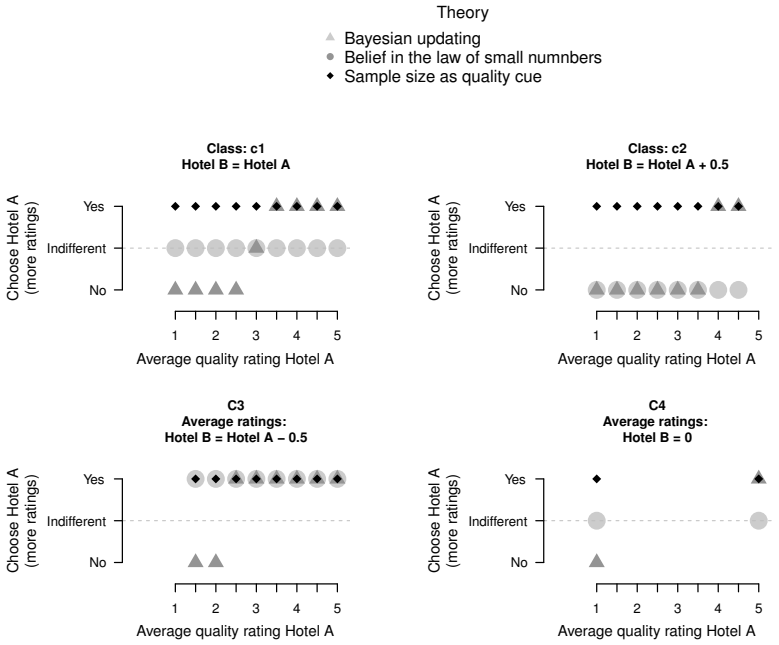

Figure 2. Qualitative predictions made by the three theories split by choice classes.

Choice class 1: Same average rating but different number of ratings. When both options have the same number of ratings, the BU model predicts that if the average ratings are below three stars, people prefer the less often rated hotel. However, if the average ratings are above three stars, people prefer the more often rated hotel. When both hotels have an average rating of three stars, the model predicts random choices. The WA model predicts that people always prefer the more often rated hotel and the BLSN model predicts random choices.

Choice class 2: Worse hotel has been rated more often. When the slightly worse hotel (-.5 stars) has been rated more often, the BU model predicts that people choose the better – but less often rated hotel – when its average rating is four stars or lower. However, for trials that include 4.5 stars, people choose the more often rated *but* slightly worse hotel. The WA model predicts that people always prefer the slightly worse but more often rated hotel. The BLSN model predicts that people always choose the slightly better but less often rated hotel.

Choice class 3: Better hotel has been rated more often. When the slightly better hotel (+.5 stars) has been rated more often, the BU model predicts that people choose the worse – and less often rated hotel – in all trials including 1.5 stars. However, in trials where both hotels have been rated on average better than 1.5 stars, people's preferences switch and people

choose the better – and more often rated – hotel. The BLSN and the WA model predict that people always choose the slightly better and more often rated hotel.

Choice class 4: One hotel has not been rated yet. When one hotel has not been rated yet and the second hotel has been rated often, the BU model predicts that people choose the non–rated hotel when the other hotel has been rated poorly and the often–rated hotel when it has been rated well. The WA model predicts that people choose the hotel that has received more ratings. The BLSN model predicts random choices. Figure 2 illustrates the model predictions separately for each choice class.

Method

Our study received ethical approval of the institutional review board of the University of Basel. The study and the hypotheses have been preregistered and can be found on https://osf.io/8cpf4/.

Participants

118 people, recruited via *Clickworker*, an online platform based in Germany, participated in the study. [4] We excluded the data of the first eight participants as due to an programming error, wrong stimuli were presented and the responses were not saved. Furthermore, we excluded four data-sets from further analyses: One person participated twice; two people did not answer the open ended questions in German, hence, we did not know whether they spoke German and understood the instructions. The remaining 106 participants (51 women, 55 men) were on average 37.35 years old ($SD = 11.13$, range = 18–62). All persons (one person preferred not to answer) had some school degree (9 years of school or more), 53 people (50%) had a University degree (i.e., Bachelor, Master or PhD). Most people (105) indicated that they found the study interesting and responded at least with a three on a Likert scale from one (very boring) to five (very interesting). Only one person responded with a one. We conclude that in general people were engaged in the study as they found it mostly

[4] We preregistered that we planned to collect the data of 130. We expected that we have to exclude up to 30% of the participants, hence our goal was to have 100 valid data. Due to miss communication with the service provider who collected the data, only 118 people participated.

interesting. Almost everyone had at least some experience with booking hotels online: Four people indicated that they have no experience with booking hotels online, 15 people that they only have little experience, 40 people that they have average experience, 31 people that they have much experience and 16 people that have very much experience with booking hotels online.

Participants received a flat fee of 4.5 Euro (approximately $5.24 when we conducted the study) and an additional bonus that was choice dependent ($M = 0.66$ Euro, $SD = 0.09$). To determine this bonus, for five randomly drawn trials, we multiplied the Bayesian means (as predicted by the BU model) of the chosen hotels with .04. The final bonus equalled the sum of these bonuses. Importantly, as we did not want to influence participants' strategies, we only told them that they will receive a bonus payment and asked them to choose as if they were booking a hotel for real. The reason for incentivizing participant's choices in addition to paying a flat fee, was that we wanted to keep participants engaged in the task.

Materials and Procedure

Participants made 45 choices (39 experimental trials and six catch trials) between two hotels. The order of trials was randomized per person. As introduced above, in the experimental trials, the hotels had different numbers of ratings. One hotel has been rated at maximum ten times and the other more than 100 times. We manipulated the difference between the average ratings (0, +.5, -.5) and whether the better rated hotel had been rated more often. In two additional trials, one hotel has not been rated yet and the second hotel has been rated often with 1 (5) stars. The number of items per combination was chosen to ensure that two competing models make different predictions in more than 50% of the trials. In the six catch trials, the more often rated hotel was at least 2 quality points better than the less often rated hotel. These trials ensured that the differences between the average ratings of the hotels showed sufficient variation to keep people engaged in the experiment. Further, we used the trials to determine whether people made meaningful choices. We predefined to exclude data of people who choose the worse hotel in more than three trials (i.e., 50%). Appendix A provides a detailed description of all choice trials and how we created them. We randomly

located the hotels on the screen, ensuring that sometimes the more rated option appeared on the left and sometimes on the right.

Imitating online platforms, we displayed summaries about how previous customers have rated the hotels on a scale from one (worst possible rating) to five (best possible rating). Figure 3 illustrates how a choice trial looked like. After 22 trials, we included a short attention check asking participants whether they paid attention to the ratings. After the choice block, participants answered three questionnaires. The first questionnaire measured people's statistical numeracy (Cokely, Galesic, Schulz, Ghazal, & Garcia-Retamero, 2012). Then, we asked several questions about how people solved our study; as how important they consider the number of ratings, the average quality ratings, and the number of good and bad ratings when they book hotels. We also asked people how experienced they are with booking hotels, which average rating is typically good enough for them to book a hotel and how they expect hotel ratings to be distributed. These questions were exploratory and we did not analyze them in detail. Further people made two more choices between hotels but this time they also provided reasons that speak for choosing each of the hotels. The first choice consisted of a slightly worse hotel (1 star) that has been rated less often (2 times) than a slightly better hotel (1.5 stars) that has been rated extremely often (1044 times). We explored whether thinking about reasons that speak for a hotel brings more people to consider the uncertainty few rating entail and hence to choose according to Bayesian principles. The second choice consisted of a extremely well (5 stars) and rarely (2 times) rated hotel and a slightly worse (4 stars) but very often (644 times) rated hotel. With this trial, we explored whether people will prefer the slightly worse but more often rated hotel even if the average rating difference is slightly larger (1 star) than the difference that we used in the other trials of our study (.5 stars) Last, we assessed demographic information.

Results

For all analyses we used the software R (R Core Team, 2014). All participants indicated that they seriously participated in the study and that we can trust their data. On average, people choose the worse rated option in 0.05% of the catch trials and no one chose it in more

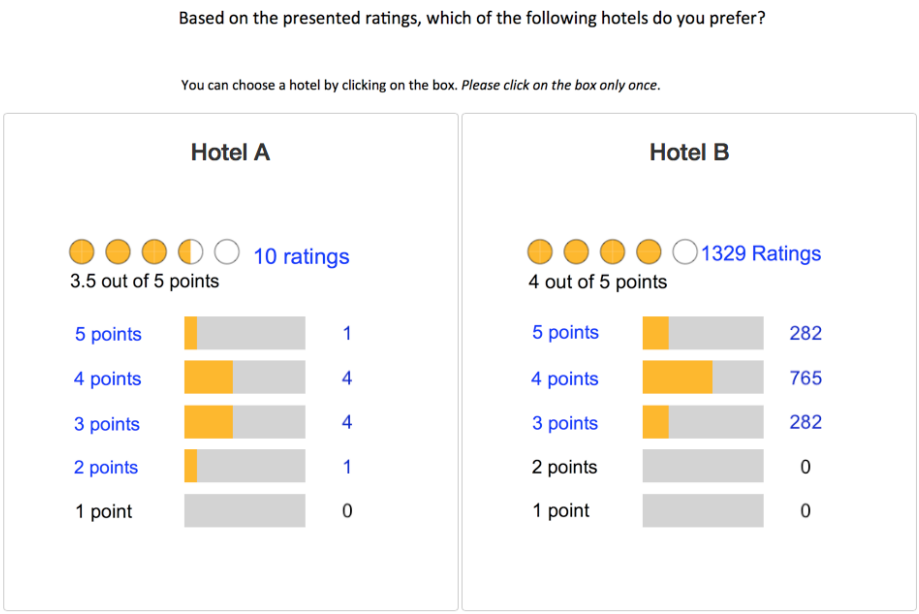

Figure 3. Example of a decision trial.

than 50%. We conclude that all participants paid attentions and made meaningful choices.

Experimental Trials

Figure 4 displays people's choice proportions for choosing the more often rated Hotel A as function of the average rating of this hotel separately for the four choice classes. Qualitatively, the data of choice classes c1, c2 and c4 speak in favour of Bayesian information integration: For hotels with high average ratings, people on average choose the hotel with more ratings. For hotels with poor average ratings, people on average choose the hotel with fewer ratings.

As preregistered we ran separate logistic regressions with random subject intercepts for choice classes c1, c2 and c3. For choice class c4, we could not conduct a logistic regression as all people choose the five-star hotel in the choice between an often rated five star hotel and a never rated hotel.

For all choice classes positive coefficients for the average rating variable indicated that people's preferences for the hotel with more ratings increased as the average rating increased $\beta_{1,c1} = 2.09$, $p_{c1} < .001$; $\beta_{1,c2} = 1.32$, $p_{c2} < .001$; $\beta_{1,c3} = .74$, $p_{c3} = .03$. Generally, the

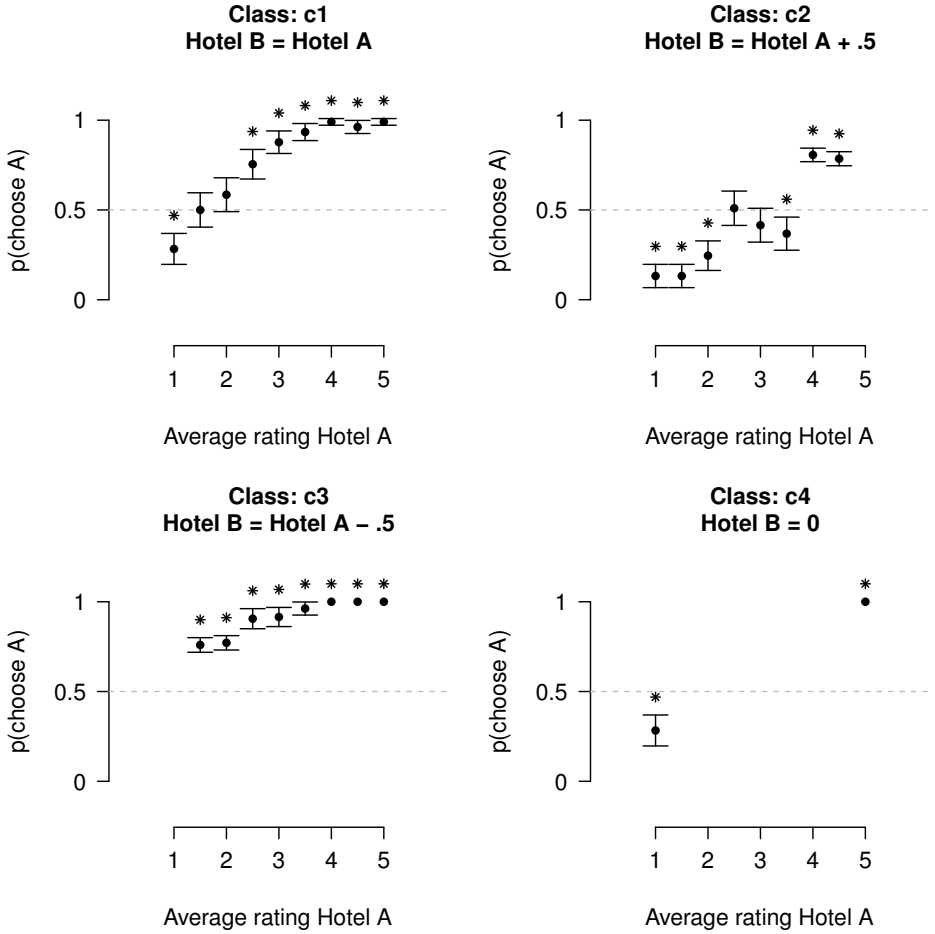

Figure 4. Proportion of people who chose the more rated hotel (Hotel A) for separate choice classes. The title indicates the relationships between average ratings of Hotel B and Hotel A (x-axis). The black dots the average choice proportions. The * indicates whether the choice proportions differed significantly from .5. at p <.001, analysed with Wilcoxon Rank-Sum tests.

significant positive coefficient for the average rating, that is an increase in choice of the hotel with more ratings with higher absolute rating, can only be predicted by the Bayesian theory. However, Bayesian principles also predict a switch in modal preferences. This means on the lower end of the scale, people should choose the less often rated option but on the higher end of the scale, the more often rated option. To analyze whether people's choice proportions significantly differed from .5, we conducted Wilcoxon rank sum tests.

The results are displayed in Figure 4, where all choice proportions marked with a $*$differed significantly from .5 ($p < .001$)[5].

In sum, on the aggregate level, in the majority of the choice classes, we found evidence for Bayesian information integration: When both hotels have been rated poorly, people preferred the option with fewer ratings. However, when both hotels have been rated well, they prefer the more often rated hotel. This finding was especially strong when the more often rated hotel was worse (choice class c2) and when one hotel was not rated yet. When both hotels had the same average rating, people switched already for relatively low average ratings. In principle, this is consistent with Bayesian information integration. However, the fact that already for relatively poor average ratings people prefer the more often rated hotel, may indicate that they give relatively much weight to the number of ratings and treat it as a indicator for quality. Importantly, when the better rated hotel also has received more ratings (choice class c3), people's behaviour clearly deviated from Bayesian principles: There, people *always* preferred the on average better and more often rated hotel in line with the principles of the belief in the law of small numbers and the idea that people treat sample size as indicator for quality.

Model Comparison

We compared the BU model, the BLSN model and the WA model as specified in the Model section with a baseline model that predicts random choices. We fit every model to participants' individual data using maximum likelihood parameter estimation and the L-BFGSB method of the optim function in R (R Core Team, 2014). To account for different numbers of model parameters, we compared the models using Bayesian model weights that we computed from the BICs (Kass & Raftery, 1995; Lewandowsky & Farrell, 2011).

Figure 5 displays for how many participants (x axis) each model was favored and how

[5]We have preregistered to bin the trials based on predictions by the Bayesian model and to base our analysis on these bins. However, we realized that while this analysis would test our BU model specifically, it does not test the general concept of Bayesian information integration. The average rating at which people switch from preferring a less often rated option to a more often rated option, depends on how a model is formalized. The preregistered analysis can be found in the Appendix C. It replicates the most important conclusions.

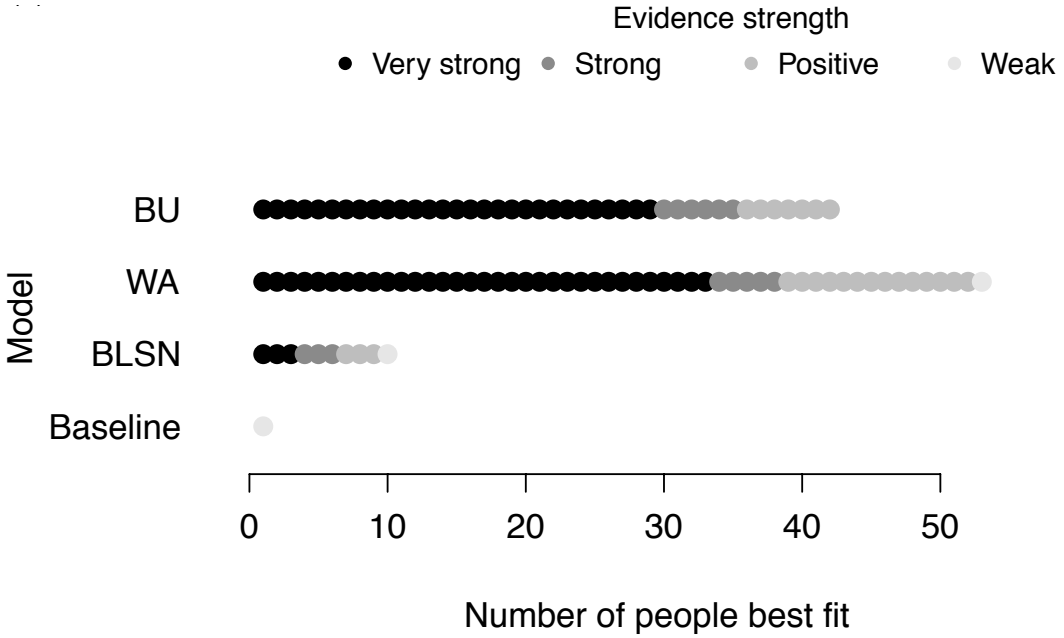

Figure 5. Results of the quantitative model comparison. Number of people (x axis) best fit by every model (y axis). The color indicated the evidence strength.

strong this evidence was (gray-scale). Only one participant was best fit by the Baseline model. 53 people (50%) were best described by the WA model. 42 participants (39.6%) were best described by the BU model and 10 participants (9.4%) were best described by the BLSN model. The high share of people best classified by the WA model is surprising given the overall tendency for Bayesian choice patterns in the aggregate data presented above.

To better understand the discrepancy between aggregate choice data and the results of the individual model comparison, we explored how well the models described the data qualitatively. We plotted the proportion of trials where people choose the more often rated hotel separately for people best fit by each model against the best fitting model's predictions. Figure 6 displays the data (black dots) and the predictions (grey triangle). The Figure reveals that the predictions of the models in some choice situations violate the data: In particular, we noticed a severe mismatch between the WA model's predictions and the data of people best fit by the WA model (column 3): For choice class 2 (row 2) we observe an interaction between average ratings and sample size. On the lower end of the rating scale people chose the hotel with the better average rating even if it has been rated less often. However, on the higher end of the scale people preferred the hotel with the worse average rating that has been rated more

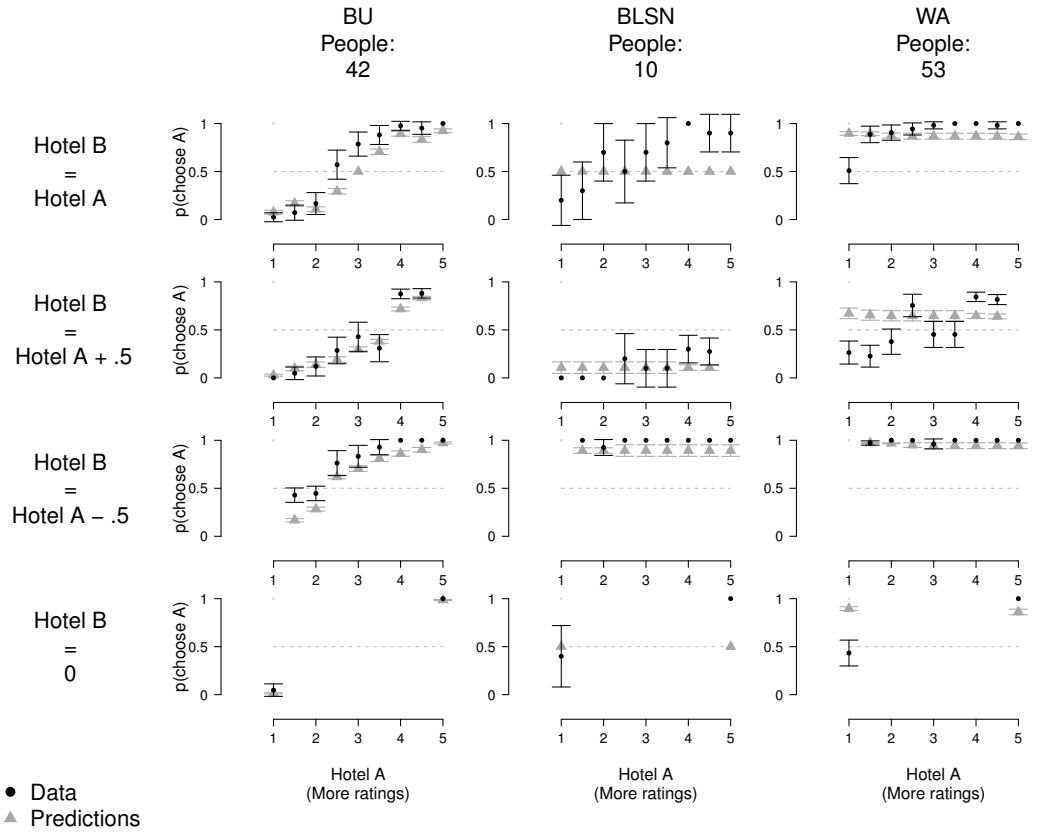

Figure 6. Qualitative model results for different choice classes (rows), separately for people best fit by the different models (columns). Grey dots indicate the model predictions, black dots the average choice proportions. The error bars indicate the 95% confidence intervals.

often. Such an interaction cannot be predicted by the WA model, because the average rating and the number of ratings are considered as independent cues which are given a particular importance. If the average ratings is given the higher importance, so that on the lower end of the scale the hotel with the better average rating is chosen, according to the model also on the higher end of the scale the hotel with the better average rating has to be chosen. However, the data suggest that people trade off average ratings and sample size as predicted by Bayesian principles in this choice class. Similarly, on the lower end of the scale (1 star), when one hotel has not been rated yet (row 4), we observed much variance in the preferences of people best fit by the WA model as indicated by the average choice proportion of 50%. However, the WA

model as we have implemented it predicts that people would prefer the often rated hotel. Also, the data of people following the belief of the law in small numbers deviated from the model predictions in some cases. Yet, as only few people (10) were best fitted by this model, we do not know how reliable these deviations are will not elaborate further on them. The data of people classified as Bayesian Updaters (column 1) were generally well captured by the model qualitatively. Only, when the better rated hotel also has more ratings (row 3), on the lower end of the scale people's choose the more often and better rated hotel slightly more often than predicted by the Bayesian model.

The Role of Statistical Numeracy

We analyzed whether people who were better described by the Bayesian model had higher scores in the statistical numeracy test. For this purpose, we identified for every individual the deviance between the Bayesian model and the data as well as the individual statistical numeracy score. Lower deviance indicates better model fit. As predicted, deviance was negatively correlated with statistical numeracy ($r = -.24$, $p = .01$). As lower deviance score indicate better model fit, this results imply that people better fit by the BU model had higher statistical numeracy.

Further Analyses

In the second part of the Study, we asked people to again choose between two poorly rated hotels (1 star with 2 ratings vs. 1.5 stars with 1044 ratings). Importantly, before making the choice, people provided reasons that speak for choosing each hotel. 53 people (50%) choose – in line with Bayesian principles – the worse and less often rated hotel. While this choice proportion is at chance level, it is higher than in the first part of the study. There, only in 24% of similar trials, people choose according to Bayesian principles. In line with the model classifications, 36 people classified by the BU model (i.e., 86% of all people classified by this model) choose the less often rated hotel but only 11 people classified by the WA model (i.e., 21% of all people classified by this model). Of the people best classified by the BLSN model 6 out of 10 people (60%) choose – against the model predictions – the less often rated hotel. However, as only few people were classified as BLSN this result may not be very

reliable. Generally, the choices in this trial do not only support our classification of participants, it also shows that with more elaborated thinking, people may be more likely to choose according to Bayesian principles.

We also asked people to make a choice between two well rated hotels (five stars and two ratings vs. four stars and 644 ratings) after providing reasons that speak for each hotel. In line with our findings from the previous choice phase, most people (85%) choose the worse but more often rated hotel which is in line both both Bayesian principles and the assumption that people treat number of ratings as indicator for quality. This indicates that on the higher end of the scale, a more reliable rating may make up for differences larger than .5 stars.

Interim Discussion

In the first study we tested whether and how people integrate average consumer ratings and number of ratings when choosing between hotels. We defined three models that describe how people may integrate number of ratings. First, according to Bayesian principles; second, ignoring the number of ratings; or third, treating the number of ratings as indicator for quality.

While at first our modeling results suggest that most people are best described by the WA model, this finding was weakened when we explored the data qualitatively split for the best fitting models: Qualitatively, the WA model and the data pattern by participants best described by this model deviate in crucial situations. This finding suggests that people's decision processes deviated substantially from the one suggested by the WA model and illustrates how important careful inspection of data patterns is. Relying only on quantitative modeling results may bring researchers to miss important patterns in the data. Based on the deviations between model predictions and the data, the question rises, why most people were nevertheless assigned to a specific model in the quantitative comparison? An explanation is that for some people none of the models captures the choices comprehensively. To account for this observation, we propose a new model, the *Bayesian decision tree*.

A Descriptive Model of Consumer Choice: The Bayesian Decision–Tree

The Bayesian decision–tree (BDT) describes in which situations people's choices conform with Bayesian principles and in which situations people's choices deviate from

Bayesian principles. The BDT assumes that, when making judgments based on consumer ratings, people pay attention to two cues, the average ratings and the number of ratings where higher numbers indicate better quality. For simplicity, here we assume that people make binary comparisons (0,1) for each cue. If all information is available (i.e., both hotels have been rated before), the subjective value of a hotel corresponds to the sum of the cue values. If one hotel is unambiguously better than the other, people do not elaborate further and choose the hotel with the more positive cue values. In our study, this is the case in choice situations where on average both hotels have been rated equally (choice class c1) and in choice situations where the on average better rated hotel has also been rated more often (choice class c3). Yet, when this decision rule does not lead to an unambiguous choice, people elaborate further about the choice options and make choices in line with the Bayesian model. In our study, this is the case in trials where one product has not been rated yet (choice class c4) and in trials where the cues are conflicting and the better rated hotel has been rated less often than the worse rated hotel (choice class c2). For both situations, the BDT predicts that people engage in more elaborated consideration of the choice options which is approximated with the BU model. The BDT has two free choice sensitivity parameters to account for the fact that the binary comparison and the Bayesian information integration lead to values that are on different scales. Figure 7 illustrates the choice process suggested by the BDT with a decision tree.

We included the the Bayesian decision tree in the set of models described above and tested it with our own data. Indeed the majority of participants (59, 55.6%) were best described by this model. In particular 37 participants previously classified by the WA model, 13 participants previously classified by the BU model, 3 previously classified by the BLSN model, and the person previously classified by the Baseline model are now classified by the BDT. Importantly, the qualitative predictions severely improved when we included the BDT in the strategy set. The full analysis is provided in the Appendix E.

Study 2: Testing the BDT

In the second study, we critically tested the Bayesian decision tree with a different data set. For this purpose, we re-analyzed the data of two experiments recently published by Powell

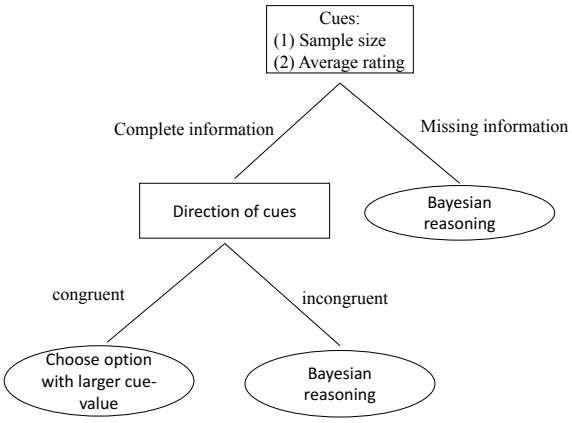

Figure 7. Heuristic Bayesian decision tree. Descriptive model of how people make choices based on customer ratings.

et al. (2017). In these experiments, participants choose between consumer products based on average ratings and the number of ratings. Similar as in our study, one product was always rated more often (150 ratings in Experiment 1 and 26 ratings in Experiment 2) than the other (25 ratings in Experiment 1 and 6 ratings in Experiment 2). The authors manipulated whether both choice options were on average rated equally well, or whether the more often option was rated better (+.1 stars or +.3 stars) or worse (-.1 stars or -.3 stars) than the less often option.

The authors of the study have analyzed the data only on the aggregate level and showed that a model treating sample size as indicator for quality predicts the data better than a Bayesian model. Here, we analyze the data from Powell et al. (2017) on an individual participant level. Based on the findings of Study 1, we expect that people differ in the strategies they use and hence to find more nuanced results, when classifying individual people's strategies as being either best described by the BLSN-, the BU, the WA, or the BDT model. More concretely, we expect a substantial portion of people to be classified best by the Bayesian decision tree. As in their experiments, participants only saw average ratings and the number of ratings (but not the rating distribution), we had to slightly adapt the BU model as described in the Appendix D.

Figure 8 displays how many people have been best fit by each model in Experiment 1

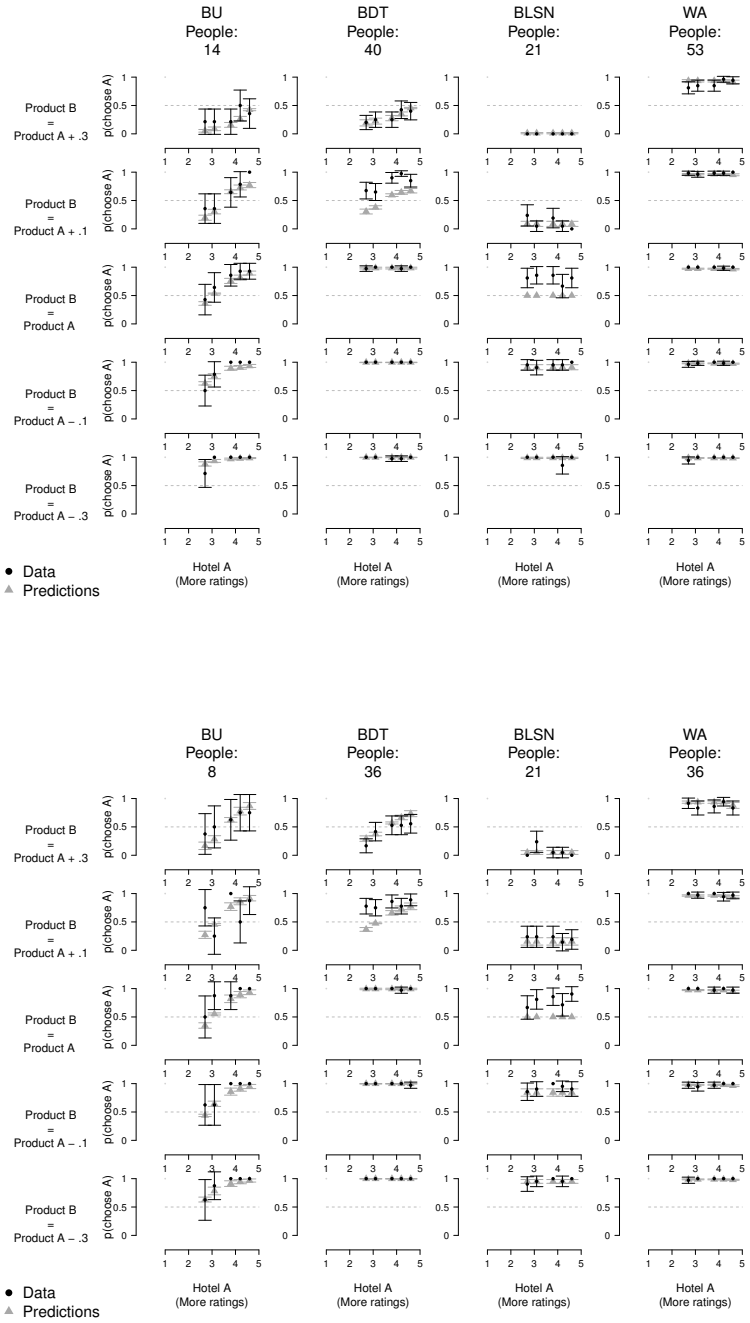

Figure 8. Reanalysis of the data by Powell et al. (2017). Top row = Experiment 1, Bottom row = Experiment 2. Qualitative model results for different choice classes (rows), separately for people best fit by the different models (columns). People best fit by the Baseline model (4 in each experiment) are excluded. The text in the first column of each row indicates how much better/worse Hotel B (hotel with fewer ratings) was in comparison to Hotel A (hotel with more ratings). The number below the model name indicates how many people were best fit by the respective model. Grey dots are the model predictions, black dots the average choice proportions.

(top Figure) and Experiment 2 (bottom Figure) and how well the data and model predictions correspond. In experiment 1 (experiment 2), four (four) people were best fit by the Baseline model, 14 (8) people were best fit by the BU model, 40 (36) people by the BDT model, 21 (21) people by the BLSN model, and 53 (36) people by the WA model. This implies that the WA model best describes the data of most participants in Experiment 1 (40.1%) and of equally many participants as the BDT in Experiment 2 (34.3%). Yet, 30.3% of the participants in Experiment 1 and 34.3% of the participants in Experiment 2 behaved (partly) according to Bayesian principles and were best described by the BDT. Further, a considerable minority of people (10.6% in Experiment 1 and 7.6% in Experiment 1) were classified by the full Bayesian model.

In sum, we showed that the BDT also captures choice patterns in the Powell et al. (2017) study well both on a quantitative and qualitative level. Interestingly, in choice situations where the less often rated option was only slightly (+.1 stars) better than the more often rated option, the choices of people best fitted by the BDT slightly differed from the model predictions. However, when the less often rated option was .3 stars better, people's choices were again perfectly in line with the BDT.

In the first experiment by Powell et al. (2017), the number of ratings of the less often rated product was with approximately 25 ratings already relatively large and the authors did not study choices where the average ratings were on the very low end of the scale. Therefore, for this experiment the BDT does not predict an switch of the modal choice proportions between trials when the less often rated hotel was .3 stars better than the more often rated hotel (first row of upper Figure 8). In experiment 2, where the less often rated hotel has received six ratings, the switch in modal choice proportion appears again. Here, both people's mean choice proportions and the predictions of the Bayesian model switch from preferring the less often rated and better hotel to the more often rated hotel (first row of bottom Figure 8).

General Discussion

Online consumer ratings about products provide valuable information when someone seeks to choose between products as they allow to instantly retrieve the preferences of

thousands of people. To efficiently use rating information, decision makers need to consider the reliability of the ratings. If only few people have rated a product, the average rating may not represent the product's quality as adequately as if many people had rated the product. Hence, when choosing between two products, people should not only consider which product has been on average rated better but also how many people have rated the products. We approached the question of how people integrate rating information, in particular average ratings and the number of ratings, when choosing between hotels. To do so, we contrasted a Bayesian view of information integration where sample size is treated as measure of uncertainty, with two non–Bayesian views on sample size. They were first, the idea that people completely ignore sample size and second that people treat sample size as an additional indicator of quality. We identified that quantitatively most people were best described with a strategy that assumes that people treat the number of ratings and average ratings as indicating quality. Yet, qualitatively people's choice patterns deviated from this predictions in some situations. Based on this observation, we post-hoc developed a Bayesian decision tree that describes, based on the cues *average rating* and *number of ratings*, in which situations people make choices in line with the Bayesian model and in which situations they do not.

More precisely, our BDT assumes that first, people scan if all information is available, i.e., whether both hotels have been rated. If one hotel has not been rated yet, they integrate uncertainty in line with Bayesian principles. If all information is available, people compare the number of ratings and the average ratings between both options. If these cues do not conflict (both are higher for one option or one cue is similar and the other one distinctive), people use a cognitive shortcut and choose the option with higher cue values. Otherwise, if both cues point in opposing directions, people adhere to a more complex reasoning process approximated by Bayesian updating. The idea that people use more cues or elaborate more when cues are not discriminatory, is an often used feature of decision trees and choice heuristics. For example when making decisions between gambles, the priority heuristic assumes that people first consider the minimum outcome that each gamble provides (Brandstätter, Gigerenzer, & Hertwig, 2006). If the values are sufficiently different, people base their choices on this cue. Only when the outcomes are very similar, other cues will be evaluated.

We tested the decision-tree with our data and found it to describe the decisions of the majority of participants best. In particular, it captured the choice patterns of people who were previously quantitatively best described by the WA or the BU model but whose choices qualitatively deviated in several situations from the models' predictions. In the second study, we reanalyzed the data of Powell et al. (2017) by means of individual model classifications including the newly developed BDT. Again, the BDT describes the data of a significant amount of people best.

Interestingly, in one situation where the BDT predicts that people follow Bayesian principles, the data slightly deviate from the model's predictions. This is, when the two choice options hardly differ (the more often rated hotel is only .1 star worse). There, people more strongly prefer the worse but more often rated product than predicted by the model. When the difference gets larger (.3 stars difference), people's behavior conforms much better with the Bayesian predictions. Arguably a small difference of only .1 star is not large enough for people to consider it as a *meaningful* difference. Therefore some people might treat the two options as if they had similar ratings. Assuming that people follow a partial Bayesian strategy as described by the BDT, in a situation where ratings are perceived as equal, people would be more likely to choose the more often rated option. This hypothesis may also explain why in the results of our experiment (Study 1) more people were classified with BDT model than in the second study (reanalysis of the data by Powell et al. (2017)). In our experiment the difference between the average ratings were with .5 stars larger than the differences in the data of the reanalysis. There, the average ratings differed by .3 stars at maximum. Future research should explore which rating differences people perceive as *large enough* and how individuals differ in how sensitively they react to differences in average ratings.

Belief in the Law of Small Numbers

Interestingly, in Study 1, we mostly found evidence that people integrate sample size in their judgments. This contradicts previous findings stating that people mainly ignore sample size (e.g., Griffin & Tversky, 1992; Kutzner et al., 2016; Obrecht et al., 2007). One explanation is that only when decision problems are presented in ecological representative

contexts, such as consumer ratings, people attend to all information. Relating to their experience in the past, people understand well to take ratings based on few samples with a grain of salt. However, in abstract tasks like coin tossing experiments (Griffin & Tversky, 1992), it may be hard for people to estimate the consequences of their actions and/or to correctly interpret the value of different pieces of information. Therefore, in abstract tasks people may be more likely to ignore sample size.

Number of Ratings as Quality Indicator

The finding, that people do not choose a slightly worse but less often rated option (choice class c3), is in line with Kutzner et al. (2016) who also reported, that people do not choose a worse option with lower sample size, even if only choosing this option allows people to reach predefined targets. As one explanation, the authors suggest that people disregard sample size completely as it does not contain informative value on its own. This may be different in consumer choice: While sample size and average ratings do not show a linear relationship (Powell et al., 2017), very poorly rated options are almost never rated often. This is because on booking platforms, ratings of hotels are not uniformly distributed but instead are rarely below three stars. Figure 9 displays how 29836 customer ratings are distributed for a set of 620 hotels (these ratings were retrieve from dataworld.com)[6]. The Figure reveals that only 16.5% of the hotels have on average been rated with fewer than three stars and that only 16.1% of all individual customers' ratings were below three stars. Most importantly, if the average rating of a hotel was low, this average was typically based on only few people's opinions: Hotels that have on average been rated with two stars or fewer, rarely (only 12% of the times) received more than ten ratings. In sum, people may treat sample size as having informative value on its own because products that are extremely bad, will not survive on the market and hence may not be available long enough for receiving many ratings. For instance, if a hotel's quality is below a crucial threshold and has huge deficiencies in the hygiene, the hotel may be shut down by authorities.

[6]The full data set that we retrieved consists of 628 hotels and 34181 ratings. However we excluded eight hotels that have been rated more than 300 times. This increases the resolution of the graphs. The conclusions do not change if we leave these hotels in.

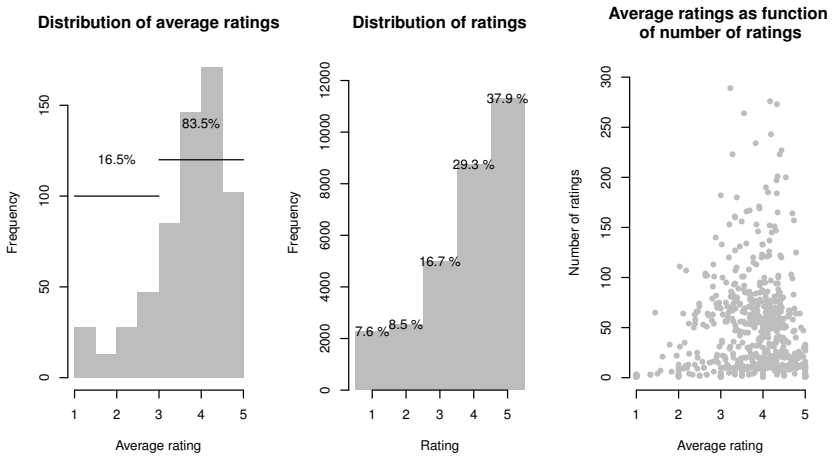

Figure 9. Analysis how 29836 ratings about 620 hotels are distributed. The left graph shows how the average ratings were distributed. The middle graph shows how individual ratings were distributed and the right graph plots how many ratings (y–axis) hotels have received against hoe they have been rated of average (x–axis).

Individual Differences

Essentially, most studies that investigated whether people choose in line with Bayesian principles analyzed data on an aggregate level. This means they aggregated both across participants and across different choice situations and concluded overall that people behaved Bayesian (e.g., De Martino et al., 2017) or did not (e.g., Powell et al., 2017). By investigating the individual strategies that people used in our experiment, we developed a more nuanced picture and show a possibility to reconcile conflicting results. In line with findings showing that many domains people solve problem using considerably different strategies (e.g., Ashby, Maddox, & Lee, 1994; Lewandowsky & Farrell, 2011), we showed that when making consumer choices based on online ratings people apply different strategies. As predicted, when we compared choice behaviour on an individual level, we found that some people choose in principle with Bayesian principles whereas other did not.

But what determines the strategy choice of people? Arguably, strategies differ with respect to how complex they are Payne et al. (1988). Integrating number of ratings and valence in a Bayesian way, presumably is much more complex than ignoring number of ratings or always treating them as cue indicating quality. This is the case as only the Bayesian

strategy, assumes an interaction between the two cues: Many ratings are a bad sign if products are rated poorly but a good sign if products are rated well. Based on the assumption that Bayesian updating requires more cognitive effort, we predicted that people with higher numeracy are more likely to adhere to Bayesian principles. Confirming this hypothesis, we found that people who were better described by the Bayesian model had higher statistical numeracy. Our findings were in line with previous research on reasoning tasks where people who were better in dealing with numerical information, reached higher performance (e.g., Brase & Hill, 2017; Chapman & Liu, 2009).

Alternatively or in addition, people using different strategies may have different goals that they seek to reach. For instance, one hypothesis is that a decision maker who has a *stretch goal* – which means that only a very good quality would satisfy her – may choose more in line with Bayesian principles than a decision maker who is already satisfied with a relatively low quality Kutzner et al. (2016). This is because the person pursuing stretch goals needs to exploit the upside risk associated with few bad ratings to reach her goal. An option that has been rated poorly by many people will very unlikely allow her to reach her goal.

Conclusion

Our findings shed light on an important point that previous research has overlooked: We identified that individual people use different strategies when they make choices based on customer ratings. Some people integrate sample size and average ratings as predicted by Bayesian principles. Others treat sample size as indicating quality. We identified that people's ability to deal with statistical information predicts how well their behaviour can be predicted by a Bayesian model. In addition, we found that some situations seem to trigger less behavior in accordance to Bayesian principles than others. Our findings may explain why previous research has provided conflicting results to the question whether people incorporate the number of ratings in their judgments. Thus, instead of asking whether or not Bayesian principles are taken into account, future studies should explore *who* behaves Bayesian *in which situations*.

References

Ashby, F. G., Maddox, W. T., & Lee, W. W. (1994). On the dangers of averaging across subjects when using multidimensional scaling or the similarity-choice model. *Psychological Science*, *5*, 144–151. doi: https://doi.org/10.1111/j.1467-9280.1994.tb00651.x

Brandstätter, E., Gigerenzer, G., & Hertwig, R. (2006). The priority heuristic: making choices without trade-offs. *Psychological review*, *113*, 409–432. doi: https://doi.org/10.1093/acprof:oso/9780199744282.003.0007

Brase, G. L., & Hill, W. T. (2015). Good fences make for good neighbors but bad science: A review of what improves Bayesian reasoning and why. *Frontiers in Psychology*, 340. doi: http://dx.doi.org/10.3389/fpsyg.2015.00340

Brase, G. L., & Hill, W. T. (2017). Adding up to good Bayesian reasoning: Problem format manipulations and individual skill differences. *Journal of Experimental Psychology: General*, *146*, 577–591. doi: http://dx.doi.org/10.1037/xge0000280

Call, J., & Tomasello, M. (1995). Use of social information in the problem solving of orangutans (pongo pygmaeus) and human children (homo sapiens). *Journal of Comparative Psychology*, *109*, 308-320. doi: https://doi.org/10.1037//0735-7036.109.3.308

Chapman, G. B., & Liu, J. (2009). Numeracy, frequency, and Bayesian reasoning. *Judgment and Decision Making*, *4*, 34–40. Retrieved from `http://www.sjdm.org/journal/8708/jdm8708.pdf`

Chen, Y.-F. (2008). Herd behavior in purchasing books online. *Computers in Human Behavior*, *24*, 1977–1992. doi: http://dx.doi.org/10.1016/j.chb.2007.08.004

Cokely, E. T., Galesic, M., Schulz, E., Ghazal, S., & Garcia-Retamero, R. (2012). Measuring risk literacy: The Berlin numeracy test. *Judgment and Decision Making*, *7*, 25–47.

Courville, A. C., Daw, N. D., & Touretzky, D. S. (2006). Bayesian theories of conditioning in a changing world. *Trends in Cognitive Sciences*, *10*, 294–300. doi: http://dx.doi.org/10.1016/j.tics.2006.05.004

De Martino, B., Bobadilla-Suarez, S., Nouguchi, T., Sharot, T., & Love, B. C. (2017). Social

information is integrated into value and confidence judgments according to its reliability. *Journal of Neuroscience*, *37*, 6066–6074. doi: http://dx.doi.org/10.1523/jneurosci.3880-16.2017

Deutsch, M., & Gerard, H. B. (1955). A study of normative and informational social influences upon individual judgment. *The journal of abnormal and social psychology*, *51*, 629–636. doi: https://doi.org/10.1037/h0046408

Fang, J., Wen, C., George, B., & Prybutok, V. R. (2016). Consumer heterogeneity, perceived value, and repurchase decision-making in online shopping: The role of gender, age, and shopping motives. *Journal of Electronic Commerce Research*, *17*, 116–131.

Fiedler, K., Walther, E., Freytag, P., & Plessner, H. (2002). Judgment biases in a simulated classroom — A cognitive-environmental approach. *Organizational Behavior and Human Decision Processes*, *88*, 527–561. doi: https://doi.org/10.1006/obhd.2001.2981

Garcia-Retamero, R., & Rieskamp, J. (2009). Do people treat missing information adaptively when making inferences? *Quarterly Journal of Experimental Psychology*, *62*, 1991–2013. doi: https://doi.org/10.1080/17470210802602615

Gretzel, U., & Yoo, K. H. (2008). Use and impact of online travel reviews. *Information and Communication Technologies in Tourism 2008*, 35–46. doi: http://dx.doi.org/10.1007/978-3-211-77280-5_4

Griffin, D., & Tversky, A. (1992). The weighing of evidence and the determinants of confidence. *Cognitive Psychology*, *24*, 411–435. doi: http://dx.doi.org/10.1016/0010-0285(92)90013-r

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*, 773–795. doi: https://doi.org/10.2307/2291091

Kersten, D., Mamassian, P., & Yuille, A. (2004). Object perception as Bayesian inference. *Annual Reviews*, *55*, 271–304. doi: http://dx.doi.org/10.1146/annurev.psych.55.090902.142005

Khare, A., Labrecque, L. I., & Asare, A. K. (2011). The assimilative and contrastive effects of word-of-mouth volume: An experimental examination of online consumer ratings. *Journal of Retailing*, *87*, 111–126. doi: https://doi.org/10.1016/j.jretai.2011.01.005

Kutzner, F. L., Read, D., Stewart, N., & Brown, G. (2016). Choosing the devil you don't know: Evidence for limited sensitivity to sample size–based uncertainty when it offers an advantage. *Management Science*, *63*, 1519–1528. doi: http://dx.doi.org/10.1287/mnsc.2015.2394

Lewandowsky, S., & Farrell, S. (2011). *Computational modeling in cognition: Principles and practice.* Thousand Oaks, CA: Sage.

Obrecht, N. A., Chapman, G. B., & Gelman, R. (2007). Intuitive *t* tests: Lay use of statistical information. *Psychonomic Bulletin & Review*, *14*, 1147–1152. doi: http://dx.doi.org/10.3758/bf03193104

Obrecht, N. A., Chapman, G. B., & Suárez, M. T. (2010). Laypeople do use sample variance: The effect of embedding data in a variance-implying story. *Thinking & Reasoning*, *16*, 26–44. doi: https://doi.org/10.1080/13546780903416775

Payne, J. W., Bettman, J. R., & Johnson, E. J. (1988). Adaptive strategy selection in decision making. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*, 534–552. doi: http://dx.doi.org/10.1037/0278-7393.14.3.534

Powell, D., Yu, J., DeWolf, M., & Holyoak, K. J. (2017). The love of large numbers: A popularity bias in consumer choice. *Psychological Science*, *10*, 1432-1442. doi: http://dx.doi.org/10.1177/0956797617711291

R Core Team. (2014). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `http://www.R-project.org/`

Rieskamp, J., & Hoffrage, U. (2008). Inferences under time pressure: How opportunity costs affect strategy selection. *Acta psychologica*, *127*, 258–276. doi: https://doi.org/10.1016/j.actpsy.2007.05.004

Schöbel, M., Rieskamp, J., & Huber, R. (2016). Social influences in sequential decision making. *PloS ONE*, *11*, 1–23. doi: https://doi.org/10.1371/journal.pone.0146536

Tenenbaum, J. B., Griffiths, T. L., & Kemp, C. (2006). Theory-based Bayesian models of inductive learning and reasoning. *Trends in Cognitive Sciences*, *10*, 309–318. doi: http://dx.doi.org/10.1016/j.tics.2006.05.009

Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin*, *76*, 105–110. doi: http://dx.doi.org/10.1037/h0031322

Xu, F., & Tenenbaum, J. B. (2007). Word learning as Bayesian inference. *Psychological Review*, *114*, 245–272. doi: http://dx.doi.org/10.1037/0033-295X.114.2.245

Ye, Q., Law, R., & Gu, B. (2009). The impact of online user reviews on hotel room sales. *International Journal of Hospitality Management*, *28*, 180–182. doi: http://dx.doi.org/10.1016/j.ijhm.2008.06.011

Zhu, F., & Zhang, X. (2010). Impact of online consumer reviews on sales: The moderating role of product and consumer characteristics. *Journal of Marketing*, *74*, 133–148. doi: https://doi.org/10.1509/jmkg.74.2.133

Appendix A

Stimuli and Choice Classes

The items we used differed on two dimensions. First, how much the average ratings differed between the two hotels and second, the number of people who have rated the hotels. Generally, all decision trials consisted of one hotel with a *low* number of ratings and one hotel with a *large* number of ratings. We defined *low* as any integer between 1 and 10 and *large* as any integer above 100. All average ratings were between one and five in .5 steps with two exceptions. Those exceptions were two hotels that have not been rated before. Table B (Appendix B) gives an overview about all decision items.

In the first class (c1), the average ratings of both choice options were similar but the number of ratings differed. We created one trial for each possible average quality rating between 1 and 5 in steps of .5.

In the second class (c2), the hotel with more ratings was .5 quality points worse than the hotel with fewer ratings. We created eight trials involving 4.5 quality points (i.e., 4 vs. 4.5 and 4.5 vs. 5). For all other possible combinations of average ratings, we created one trial.

In the third class (c3), the hotel more ratings was .5 quality points better than the hotel with fewer ratings. We created eight trials involving 1.5 quality points (i.e., 1 vs. 1.5 and 1.5 vs. 2). For all other possible combinations of average ratings, we created one trial.

The fourth class (c4) consisted of two *uncertainty* trials. There, one hotel has not yet been rated and the other hotel has on average been rated with 5 (1) stars by many people.

The fifth class (c5) consisted of six catch trials where the better rated Hotel (at least +2 stars) has also been rated more often.

We determined the number of ratings that was associated with each stimulus as follows: To assign the number of ratings to hotels, we randomly sampled a integer I_1 between one and nine if the number of ratings was low and between 100 and 999 if the number of ratings was large. Then, we randomly sampled a second integer I_2 between 1 and I_1. Afterwards, we repeated this procedure and randomly sampled a third integer I_3 between 1 and I_2. With these integers, we created rating summaries. Importantly, these summaries were unimodally distributed. To create the summaries, we first assigned I_1 to the average rating. To understand

this logic, consider Hotel A in Figure 3. The average rating of this hotel is 4 points. The random integer I_1 is 870. Therefore, we assigned 870 ratings to four quality points. If the average rating was not an integer, we assigned $I_1 \times .5$ (rounded to the next integer) to the two quality points that surround the average ratings. For instance, if the average ratings is 4.5, we assigned $I_1 \times .5$ (rounded to the next integer) to four quality points and $I_1 \times .5$ (rounded to the next integer) to five quality points. In a second step, we assigned $I_2 \times .5$ (rounded to the next integer) to the quality points surrounding the rating(s) that has/have already be assigned. In our example, I_2 was 448. Therefore, we assigned 224 ratings to five points and 224 ratings to three points. If the average rating was 3 points, we repeated this procedure with I_3. This procedure ensured that all ratings were unimodally distributed.

Importantly, in stimulus class 1 and 4 trials below the mean of the rating scale (i.e., 3) exactly mirror trials above the scale. Furthermore, stimuli of class 2 that are below the mean of the rating scale (the mean of the rating scale; above the mean) mirror stimuli of class 3 that are above the mean (the mean of the rating scale; below the mean).

Appendix B

Decision Problems

		Average Rating		Number Ratings		Ratings per Qualitypoint									
Name	Class	A	B	A	B	a1	a2	a3	a4	a5	b1	b2	b3	b4	b5
1	c1	*1*	1	825	5	825	0	0	0	0	5	0	0	0	0
2	c1	*1.5*	1.5	660	8	330	330	0	0	0	4	4	0	0	0
3	c1	*2*	2	145	1	2	141	2	0	0	0	1	0	0	0
4	c1	*2.5*	2.5	596	4	102	196	196	102	0	1	1	1	1	0
5	c1	*3*	3	774	10	18	38	662	38	18	0	2	6	2	0
6	c1	*3.5*	3.5	596	4	0	102	196	196	102	0	1	1	1	1
7	c1	*4*	4	145	1	0	0	2	141	2	0	0	0	1	0
8	c1	*4.5*	4.5	660	8	0	0	0	330	330	0	0	0	4	4
9	c1	*5*	5	825	5	0	0	0	0	825	0	0	0	0	5
10	c2	*1*	**1.5**	579	2	579	0	0	0	0	1	1	0	0	0
11	c2	*1.5*	**2**	532	9	266	266	0	0	0	2	5	2	0	0
12	c2	*2*	**2.5**	1329	10	282	765	282	0	0	1	4	4	1	0
13	c2	*2.5*	**3**	462	3	21	210	210	21	0	0	0	3	0	0
14	c2	*3*	**3.5**	422	6	4	8	398	8	4	0	1	2	2	1
15	c2	*3.5*	**4**	1560	9	0	390	390	390	390	0	0	2	5	2
16	c2	*4*	**4.5**	1230	4	0	0	242	746	242	0	0	0	2	2
17	c2	*4*	**4.5**	686	4	0	0	44	598	44	0	0	0	2	2
18	c2	*4*	**4.5**	933	2	0	0	189	555	189	0	0	0	1	1
19	c2	*4*	**4.5**	1716	8	0	0	368	980	368	0	0	0	4	4
20	c2	*4.5*	**5**	420	2	0	0	0	210	210	0	0	0	0	2
21	c2	*4.5*	**5**	266	3	0	0	0	133	133	0	0	0	0	3
22	c2	*4.5*	**5**	884	7	0	0	0	442	442	0	0	0	0	7
23	c2	*4.5*	**5**	980	5	0	0	0	490	490	0	0	0	0	5
24	c3	***1.5***	1	420	2	210	210	0	0	0	2	0	0	0	0
25	c3	***1.5***	1	266	3	133	133	0	0	0	3	0	0	0	0

26	c3	***1.5***	1	884	7	442	442	0	0	0	7	0	0	0	0
27	c3	***1.5***	1	980	5	490	490	0	0	0	5	0	0	0	0
28	c3	***2***	1.5	1230	4	242	746	242	0	0	2	2	0	0	0
29	c3	***2***	1.5	686	4	44	598	44	0	0	2	2	0	0	0
30	c3	***2***	1.5	933	2	189	555	189	0	0	1	1	0	0	0
31	c3	***2***	1.5	1716	8	368	980	368	0	0	4	4	0	0	0
32	c3	***2.5***	2	1560	9	390	390	390	390	0	2	5	2	0	0
33	c3	***3***	2.5	422	6	4	8	398	8	4	1	2	2	1	0
34	c3	***3.5***	3	462	3	0	21	210	210	21	0	0	3	0	0
35	c3	***4***	3.5	1329	10	0	0	282	765	282	0	1	4	4	1
36	c3	***4.5***	4	532	9	0	0	0	266	266	0	0	2	5	2
37	c3	***5***	4.5	579	2	0	0	0	0	579	0	0	0	1	1
38	c4	*1*	0	318	0	318	0	0	0	0	0	0	0	0	0
39	c4	*5*	0	318	0	0	0	0	0	318	0	0	0	0	0
40	c5	***4***	2	345	4	0	0	64	217	64	1	2	1	0	0
41	c5	***5***	3	610	12	0	0	0	0	610	1	2	6	2	1
42	c5	***4***	1	609	8	0	0	94	421	94	8	0	0	0	0
43	c5	***4***	1	1237	4	0	0	161	915	161	4	0	0	0	0
44	c5	***4.5***	1.5	508	4	0	0	0	254	254	2	2	0	0	0
45	c5	***5***	1.5	769	4	0	0	0	0	769	2	2	0	0	0

Note. Choice trials of the experiment. For each row, the option that is written with bold numbers indicates the choice option predicted by the BLSN model that ignores the number of ratings. The option that is underlined numbers indicates the option predicted by the Bayesian model that predicts an interaction between average ratings and number of ratings. The option that is written in italic indicates the option that is predicted by the WA model that treats number of ratings as quality indicator. If no choice option is marked, the respective model predicts random choices.

Appendix C

Results Preregistered Wilcoxon Rank Sum Tests

As preregistered and described above, we conducted Wilcoxon rank sum tests to identify whether people's preferences within the different classes switched as predicted by our Bayesian model. For each choice class, we identified the trials where the Bayesian model predicted a switch in preference. We clustered trials below and above the switching points (in c1 also the trial where the BU model predicts random choices) and compared whether people's average choice proportions are in line with the modal model predictions (Figure 2). Table C gives an overview about the clusters that we have generated, the modal predictions made by the theories, people's choice proportions and the results of the Wilcoxon tests.

The Bayesian model predicted modal choice proportions in six of the nine comparisons correctly, the number of ratings as quality theory in six of the nine comparisons and the BLSN in only four of the nine comparisons.

			Predictions			Data	
Class	Avg_A	Avg_B	Bayesian Updating p(choose A)	BLSN p(choose A)	SS as Quality p(choose A)	p(choose A)	Wilkoxon Test
c1: A and B rated equally	<3	<3	<.5	**= .5**	>.5	.53	W = 5989
	3	3	= .5	= .5	**>.5**	.88	W = 9858***
	>3	>3	**>.5**	= .5	**>.5**	.97	W = 11077***
c2: A worse than B	<4	<4.5	**<.5**	**<.5**	>.5	.3	W = 2332***
	>3.5	>4	**>.5**	<.5	**>.5**	.8	W = 9699***
c3: A better than B	<2.5	<2	<.5	**>.5**	**>.5**	.76	W = 8692***
	>2	>1.5	**>.5**	**>.5**	**>.5**	.96	W = 11183***
c4: B not yet rated	1	-	**<.5**	.5	>.5	.28	W = 3180***
	5	-	**>.5**	.5	**>.5**	1	W = 11263***

Note. Column 1 indicates the choice class. Columns 2 and 3 relate to the clusters, we have built. Avg_A relates to the average ratings of Hotel A and Avg_B ratings of Hotel B. For instance $\text{Avg}_\text{A} < 3$ and $\text{Avg}_\text{B} < 3$ read: For all trials in which the average ratings of Hotels A and B are smaller than 3. Columns 4 to 6 display the model predictions with regards to choosing the more often rated Hotel A. A value of $< .5$ indicates that a model predicts that people are more likely to choose the less often rated Hotel B, a value of $= .5$ indicates that the model predicts random choices, and a value of $> .5$ that the model predicts that people are more likely to choose the more often rated hotel A. Values in bold indicate that the average choice proportions that we have observed in the experiment (column 7) correspond to these predictions. The Wilkoxon test (column 8) indicates whether people's choices differ significantly from .5. *** significant at $p < .001$.

Appendix D

BU model: Adaptation to Analyze the Data by Powell et al. (2017)

To model people's prior beliefs, we assumed that average ratings were uniformly distributed following a Dirichlet distribution with one parameter ($h_{avg=i}$) for each possible average rating ($Dir(h_{avg=1.0} = 5/41, h_{avg=1.1} = 5/41, ..., h_{avg=5.0} = 5/41)$). We set the strength of the prior belief to 5, as in the original version of the Bayesian model. As 41 average ratings between 1 and 5 are possible, we divided the total strength of the prior belief by 41 to retrieve the strength of a rating i. Again, when receiving rating information, people update their belief distribution by adding the number of times that the hotel has been rated to the parameter representing the average rating of the hotel. People's subjective values are again simplified by the mean of the posterior, i.e., subjective value of Hotel j (V_j) equals $V_j = \frac{\sum_{i=1}^{5} h_i \times i}{\sum_{i=1}^{5} i}$.

Appendix E

Reanalysis Data Of Study 1 Including the BDT

We added the heuristic Bayesian decision tree to the set of models that fitted to the data of the hotel study. For the model comparison, we followed the same protocol as described in the introduction.

Figure E1 illustrates how many people were best described by each model and how strong the evidence was: No one was best described by the Baseline model. Seven people (6.6%) were best described by the BLSN model; 11 people (10.4%) by the weighted additive model; 29 people (27.4%) by the BU model; and 59 people (55.6%) by the BDT. Of these 59 people, three people switched from the BLSN model; 42 people switched from the WA model, 1 person switched from the Baseline model; and 13 people switched from the BU model.

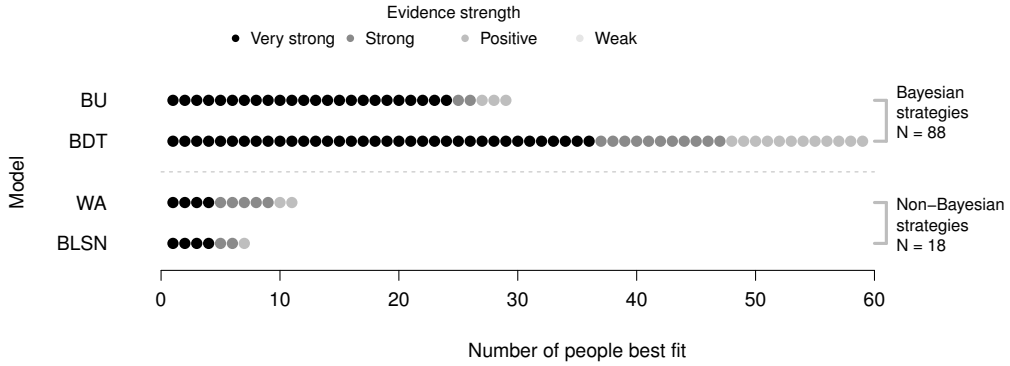

Figure E1. Results of the quantitative model comparison. Number of people (x axis) best fit by every model (y axis). The color indicated the evidence strength.

Again, we explored the data quantitatively. We grouped people best fit by each model and investigated the respective model's predictions based on individual parameter estimates. Figure E2 shows that the models capture the data much better than before.

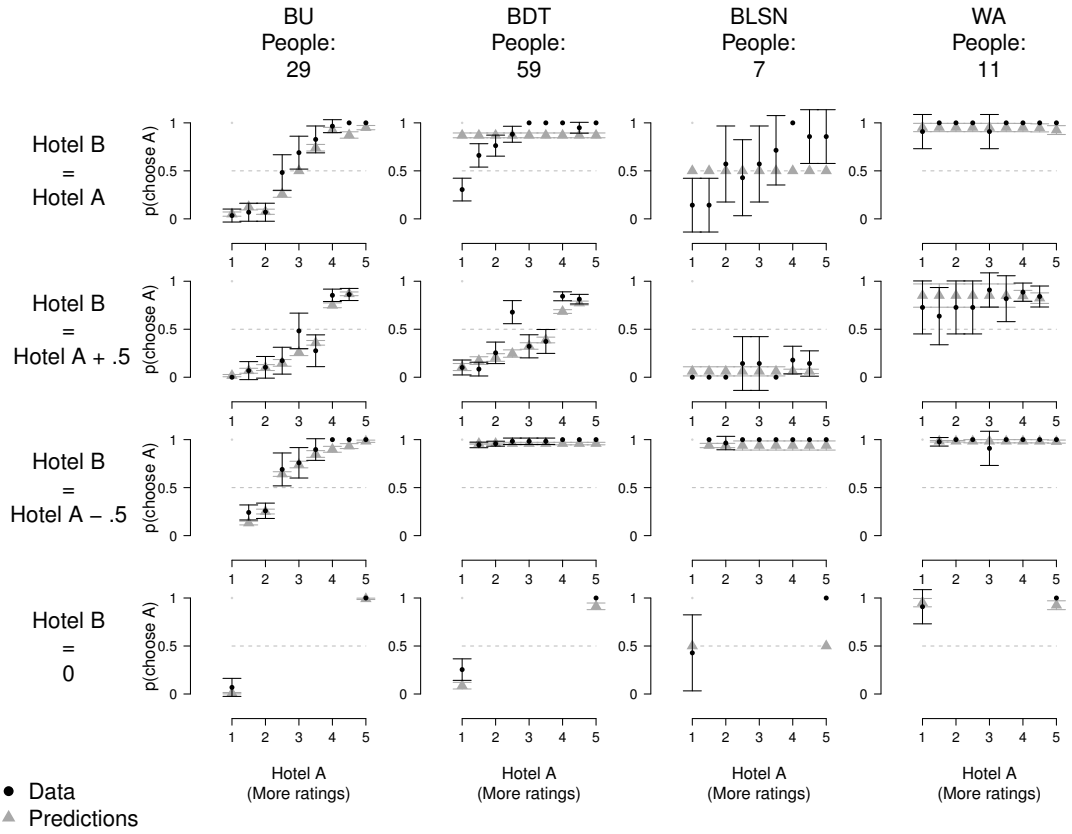

Figure E2. Qualitative model results for different choice classes (rows), separately for people best fit by the different models (columns). Grey dots indicate the model predictions, black dots the average choice proportions.

How environmental regularities affect people's information search in probability judgments from experience

Janine Christin Hoffart

Jörg Rieskamp

Gilles Dutilh

University of Basel

Author Note

This research was supported by a grant (SNF # 143854) from the Swiss National Science Foundation to the second and third author. Correspondence concerning this article should be addressed to Janine Christin Hoffart. University of Basel, Department of Psychology, Missionsstrasse 62a, 4055 Basel, Switzerland. E-mail: janine.hoffart@unibas.ch

Abstract

In everyday life, people encounter smaller rewards with higher probability than larger rewards. Do people expect this reward–probability regularity to hold in experimental settings? To answer this question, we tested whether people's behavior in probability judgment tasks is affected by the correlation between reward size and reward probabilities. In Study 1, we asked people to judge reward probabilities under uncertainty. In line with the ecological reward–probability correlation, people assumed that larger rewards were less likely than smaller rewards. In Study 2, we tested the prediction that people's information search and integration depend on the representativeness of the environment. Participants performed an experience-based probability judgment task in which they sampled outcomes from unknown gambles until they felt confident to estimate the probabilities of the gambles' outcomes. We manipulated the reward–probability relationship of the gambles in three experimental groups. Rewards and reward probabilities were either negatively correlated, positively correlated, or not correlated at all. A negative correlation mimics the ecological reward–probability relationship often present in real life. We analyzed people's search effort and whether they integrated sample-based uncertainty into their judgments. We found that people sampled fewer outcomes in the ecologically representative condition than in the other two conditions. However, people did not integrate sample-based uncertainty in their judgments: In all conditions people treated the observed outcomes as representative of the underlying outcome distribution. People's prior beliefs about regularities in environments provides a potential explanation why people often rely on small sample sizes when making judgments and decisions from experience.

Keywords: probability updating, sample size, probability judgments, decisions from experience, representative design

How environmental regularities affect people's information search in probability judgments from experience

How people search for information, when they stop searching, and how they integrate information are central questions of cognitive psychology (e.g., Busemeyer & Rapoport, 1988). Recently, there has been growing interest in how people make decisions from experience. In experience-based tasks, people sample outcomes from unknown gambles to learn about the gambles' outcomes and outcome probabilities (Hertwig, Barron, Weber, & Erev, 2004). Several factors have been identified that influence information search and integration in these tasks (e.g., Hau, Pleskac, Kiefer, & Hertwig, 2008; Hertwig et al., 2004; Rakow, Demes, & Newell, 2008; Roth, Wänke, & Erev, 2016). Yet in this area past work has mostly neglected how initial beliefs and expectations about the structural regularities of risky gambles influence search effort and behavior (but see Lejarraga, Hertwig, & Gonzalez, 2012; Mehlhorn, Ben-Asher, Dutt, & Gonzalez, 2014). This is surprising when considering that regularities in the structure of experimental stimuli influence behavior in a range of tasks such as causal reasoning and probability judgments. In particular, when people's expectations are congruent with the labeling of stimuli, they learn relationships faster (e.g., Busemeyer, Byun, Delosh, & McDaniel, 1997). If the environment contradicts prior beliefs, for instance, those formed on the basis of real-life experience, people need strong evidence to overcome their beliefs (e.g., Alloy & Tabachnik, 1984).

On the basis of real-life experience, people can, for instance, expect that larger rewards will be less likely than smaller rewards (Pleskac & Hertwig, 2014). Interestingly, specific experimental designs can also lead to regularities between properties of gambles, such as correlations between returns and risk or between rewards and reward–probabilities. In some cases, correlations are representative of real-life correlations, in others not. Generally, researchers may not be aware of how specific regularities of their stimulus material influence people's behavior (Fiedler, 2000). More concretely, it has not been well studied whether regularities of gambles, particularly correlations between rewards and reward probabilities, affect behavior in decision-making experiments. Furthermore, it is not known whether the representativeness of these correlations influences behavior. We address these questions in two

studies. In Study 1, we examined whether people have specific expectations about how rewards and reward probabilities of risky gambles are correlated. In Study 2, we examined whether people search for less information and integrate information differently in ecologically representative environments where rewards and reward probabilities of risky gambles are negatively correlated as compared to unrepresentative environments.

The Structure of the Environment and Ecological Rationality

Natural environments are characterized by certain statistical regularities. In financial decision-making situations, expected returns are typically positively correlated with risk (measured as the variance of returns). For instance, investments in stocks usually offer higher returns than investments in bonds but involve larger risk. This risk–return trade-off forms the basis of Markowitz's (1952) mean–variance model and standard portfolio theory (e.g., Sharpe, 1964). Whereas risks and returns are positively correlated, rewards and reward probabilities are negatively correlated across many situations: Pleskac and Hertwig (2014) analyzed the underlying regularities of several real-world domains including roulette, horse racing, life insurance, artificial insemination, and scientific publications. In essence, they showed that across all domains, larger gains occurred with lower probabilities than smaller gains. For instance, the larger the prize of a national lottery is, the smaller the probability of winning this prize.

Relying on regularities of the environment can simplify judgment processes and help people to make accurate decisions and judgments (e.g., Gigerenzer, Todd, & the ABC Research Group, 1999; Todd & Gigerenzer, 2012). For instance, in environments where different sources provide redundant information, people search for little information and focus on the most valid information when making decisions. But in environments with little information redundancy, people search for more information and integrate all information when making decisions (Dieckmann & Rieskamp, 2007). In sum, people often adapt their behavior to the environment by selecting decision strategies that are appropriate for the structure of the environment (e.g., Mata, Schooler, & Rieskamp, 2007; Pohl, 2006; Rieskamp & Otto, 2006). People's adaptation to natural environments has been recently demonstrated

with the risk–reward heuristic.

The Risk–Reward Heuristic

Work on the risk–reward heuristic describes how people estimate reward probabilities from potential costs and benefits of risky options (Pleskac & Hertwig, 2014). Pleskac and Hertwig (2014) offered participants an opportunity to gamble at the cost of $2. Participants had a chance to win a monetary reward whose magnitude was known and varied between participants. Importantly, participants had to estimate the unknown reward probability. In line with the ecological negative correlation between rewards and reward probabilities, people estimated the chances of winning smaller rewards as being higher than the chances of winning larger rewards. Pleskac and Hertwig (2014) concluded that under complete uncertainty, people heuristically inferred reward probabilities from the ratio of the reward magnitude and the costs of gambling. Yet in many experimental situations, people have some information about the probabilities of possible outcomes. For instance, in experience-based tasks, people sample outcomes to estimate outcome probabilities. Such tasks involve what Knight (1921) called *statistical probabilities*, which lie on the continuum between complete uncertainty and risk (Camilleri & Newell, 2013; Hau, Pleskac, & Hertwig, 2010; Hertwig & Erev, 2009).

Rewards and Reward Probabilities Are Correlated in Experience-Based Tasks

Interestingly, the payoff structure of the most frequently used gambles in decisions-from-experience (dfe) studies (e.g., Hau et al., 2008; Hertwig et al., 2004; Rakow et al., 2008) resembles the negative reward–reward probability correlation observed outside the laboratory: Larger rewards are less likely than smaller rewards. Figure 1 displays gains and probabilities of gambles used in dfe studies (e.g., Hau et al., 2008; Hertwig et al., 2004; Rakow et al., 2008). Rewards and reward probabilities are negatively linked across competing gambles (gray lines link competing gambles) and across all gambles (black line). While the negative correlation across gambles is particular to the gamble pairs we analyzed, a similar correlation as for the competing gambles has been reported elsewhere (Pleskac & Hertwig, 2014). A reason for this correlation can be deduced from the way researchers design experiments and specially the pairs of gambles they use: For choices to be informative for

researchers, competing gambles typically have similar expected values (Rieskamp, 2008). Consider a pair of gambles where each gamble offers a reward with some probability and zero otherwise. When the first gamble's reward size is larger than the second gamble's reward size, the first gamble's reward probability needs to be smaller to lead to similar expected values of both gambles. Using gambles with similar expected values is required to avoid trivial, noninformative choices. However, the observation that reward sizes and reward probabilities are negatively correlated in experience-based tasks may help to explain unresolved questions about how people search for and integrate information in experienced-based tasks.

In experience-based tasks, people often search less than would be required for an unbiased representation of the gambles' outcome distribution (e.g., Hau et al., 2008; Hertwig et al., 2004; Rakow et al., 2008). Reasons why people rely on small sample sizes range from sampling costs such as time and opportunity costs (Hau et al., 2008) to memory constraints that make it difficult to deal with large sample sizes (e.g., Hertwig et al., 2004; Rakow et al., 2008, but see Plonsky, Teodorescu, and Erev, 2015). Lejarraga et al. (2012) proposed that people in an experience-based task learn over the course of the experiment how gambles are structured and exploit this knowledge by sampling less. Here, we extend the hypothesis that learning influences behaviour in experience-based tasks and investigate how the structure of gambles and people's expectations influence search effort.

Outline of the Studies and Predictions

In Study 1 we investigated whether people expect a negative correlation between rewards and reward probabilities in decision-making tasks. This study was an extension of Pleskac and Hertwig's (2014) work on the risk–reward heuristic. There, participants had two cues (i.e., reward probability and costs of gambling) to inform probability judgments. In choice studies, typically no *explicit costs* are associated with gambling. Therefore, in Study 1, we assessed people's probability expectations for different reward amounts when no costs were associated with gambling.

In Study 2 we investigated people's search effort and accuracy in an experience-based probability judgment task. Briefly, we asked participants to repeatedly judge reward

probabilities for two-outcome gambles (rewards of varying magnitudes or zero outcome). To make their judgments, participants first sampled outcomes from a gamble's outcome distribution. Using a between-subjects design, we manipulated the reward–reward probability correlation. We contrasted a representative condition in which rewards and reward probabilities were negatively correlated (representative: negative correlation, RNC) with two nonrepresentative conditions in which rewards and reward probabilities either were not correlated (nonrepresentative: no correlation, NRNC) or were positively correlated (nonrepresentative: positive correlation, NRPC). We hypothesized that a correlation between rewards and reward probabilities affects search effort and/or information integration.

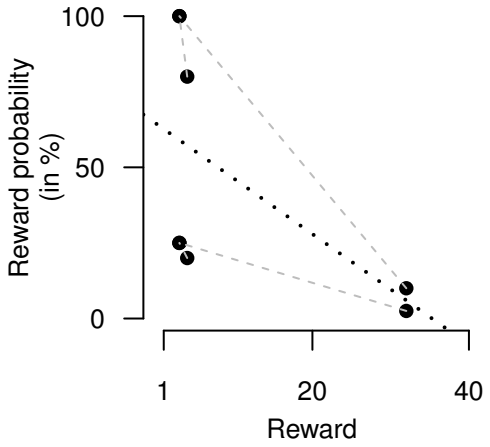

Figure 1. Relationship between reward probabilities and gains of gambles that have been used in many dfe studies. Each dot represents one gamble that consists of two outcomes (reward or zero outcome). Gray lines link competing gambles. Gambles that are linked with more than one other gamble were used in more than one choice trial. The black line indicates the regression line between rewards and probabilities across all gambles.

How Can a Correlation Influence Learning and Search Effort in Experience-Based Tasks?

As Lejarraga et al. (2012) has proposed, learning can influence search effort. To test this hypothesis, we compared how much people sampled between conditions. If general learning

influences sample size, participants should sample less when they have learned the relationship between rewards and reward probabilities. To determine if participants learned the correlation, we prompted their probability expectations at the end of the experiment for different rewards *without* allowing them to draw any outcomes. If participants have learned how rewards and reward probabilities are correlated, their probability estimates in this test block should resemble the correlation experienced during the previous phases. Generally, people can learn the relationships in both correlation conditions, implying that people should draw fewer outcomes in both correlated conditions. However, previous research has shown that people learn relationships between cues and criteria faster when the relationship is congruent with their expectations (e.g., Busemeyer et al., 1997). If people expect rewards and probabilities to be negatively correlated, they may learn the reward–probability correlation in the representative condition faster. In this case, we expect that sample sizes will be smaller *and* people's probability estimates in the test block will more closely resemble the previously learned correlation in the representative condition than in the nonrepresentative condition.

We further analyzed the impact of learning by comparing how sample sizes evolved over trials. If learning impacts search effort and the control block reveals that people learned the relationships in both correlated conditions, sample sizes should decrease more in the two conditions with correlated rewards and reward probabilities than in the condition where rewards and reward probabilities are not correlated.

Alternatively, it could be that it is not general learning but familiarity with the search environment that influences search effort. In line with the notion that human cognition is adapted to natural environments (e.g., Brunswik, 1947, 1955; Gigerenzer, Hoffrage, & Kleinbolting, 1991; Gigerenzer & Hug, 1992), we hypothesized that people are adapted to navigating in environments where smaller rewards occur with higher probabilities than larger rewards. If participants' search effort is selectively influenced by representative correlations, we would expect first that they will draw fewer outcomes in the representative condition than in both nonrepresentative conditions and second that in both conditions where rewards and reward probabilities are correlated, people's responses in the control block will resemble the reward–probability structure observed in the previous blocks.

How Can a Correlation Between Rewards and Reward Probabilities Influence Accuracy in Experience-Based Tasks?

We propose that in correlated environments, people can exploit their knowledge about correlations and make reasonably accurate judgments based on relatively small sample sizes. In our experiment, we showed participants the outcomes (i.e., zero and a positive reward) *before* they started sampling. This implies that in the correlated conditions, participants could form prior beliefs about the reward probabilities based on the reward magnitudes.

However, previous research showed that people often integrate sample-based uncertainty insufficiently. Instead, they simply rely on the relative observed outcome frequency when they make judgments (e.g., Griffin & Tversky, 1992; Tversky & Kahneman, 1971). The "natural-mean" heuristic describes such a strategy and has been successfully applied in modeling decisions from experience (Hertwig & Pleskac, 2008).

We contrasted the two hypotheses stated above by comparing two models: The Bayesian updating model (M_B) assumes that people's probability judgments are based on their prior belief about the reward probability and a Bayesian updating process that relies on the sampled information. The model makes the assumption that people's prior beliefs follow the objective correlation present in the experimental environment. The second model is a variant of the natural-mean heuristic and captures the idea that people attend only to the observed relative outcome frequencies when judging probabilities (M_{NM}). The models are described in detail in Appendix A.

Study 1: Do People Believe That Rewards and Probabilities Are Correlated in the Laboratory?

To investigate whether people believe that reward magnitudes and probabilities are correlated across gambles in experimental tasks, we asked people in two experiments to infer reward probabilities from reward magnitudes when no costs were associated with gambling.

Method Experiment 1

Fifty-seven undergraduates from the University of Basel volunteered to participate in a paper-and-pencil survey (49 women, 8 men, M_{age} = 22.21 years, $SD = 3.67$, range = 18-35 years). Participants estimated reward probabilities for two-outcome gambles. The outcomes were zero or a specific known reward (reward magnitudes: CHF 2.40, 4.00, 4.70, 12.00, 16.00, 28.00). These rewards were taken from past studies conducted at the Center for Economic Psychology and thus represented realistic outcomes of psychological experiments. The amounts were displayed randomly in increasing or decreasing order. For all analyses in this manuscript, we used the software R (R Core Team, 2014).

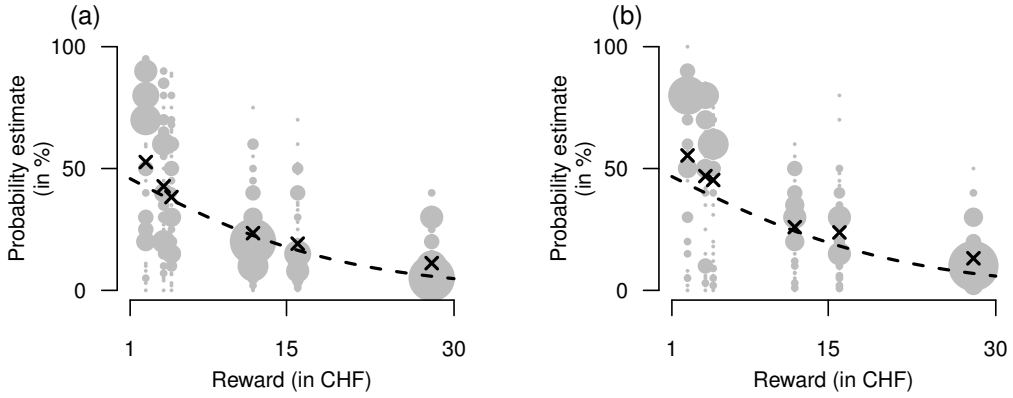

Figure 2. Probability estimates as a function of rewards in Experiment 1 (a) and Experiment 2 (b). The dots display individual responses; the dot size indicates response frequency. The crosses display the mean probability estimates. The dashed line shows the regression line based on the median posterior group estimates of the coefficients of the beta regression.

Results and Discussion of Experiment 1

Participants' probability estimates varied as a function of reward amount: Participants assigned higher probabilities to smaller rewards than to larger rewards (see Figure 2). This finding is supported by a Bayesian hierarchical beta regression (see Appendix B for model description). With vague priors, the median posterior estimate for the overarching slope coefficient was $\beta_1 = -0.1$ ($\text{HDI}_{95\%}$: -0.11 to -0.09). Described in terms of odds of winning,

the results imply that for every 10 CHF decrease of reward amount, participants believed they were 2.72 ($HDI_{95\%}$: 2.6 to 3) times more likely to win. The 95% highest density interval (HDI) of the slope coefficient excluded zero, indicating a reliable negative correlation between participants' probability estimates and rewards. The results were similar for both orders of reward presentation, slope: $\beta_{1,increasing} = -0.1$, $HDI_{95\%}$: -0.12 to -0.08 and $\beta_{1,decreasing} = -0.09$, $HDI_{95\%}$: -0.11 to -0.08; intercept (converted to the probability scale): $\beta_{0,increasing} = 0.47$, $HDI_{95\%}$: 0.33 to 0.6 and $\beta_{0,decreasing} = 0.51$, $HDI_{95\%}$: 0.4 to 0.63.

Experiment 1 supports the findings of Pleskac and Hertwig (2014). Participants assumed larger rewards were less likely to occur than smaller rewards. However, the fact that we always presented rewards in either increasing or decreasing order may be a potential confound. To control for the possibility that our results are an artifact of the presentation order of rewards, we replicated Experiment 1 with a fully randomized order of rewards.

Experiment 2

In the second experiment, we followed the protocol of Experiment 1 with one exception: We randomized the order in which we presented the rewards for every participant.

Method Experiment 2. We predefined the number of subjects with a power analysis using a simulation approach. We searched for the minimum number of required subjects for finding a reliably negative slope coefficient when the data-generating slope coefficient was half as large as the median coefficient of Experiment 1 in 90% of the simulation. The power analysis revealed that we needed to test 42 subjects to reach the desired power. We collected data of 44 psychology students (38 women, 4 men, 2 not reported, $M_{age} = 25.3$ years, $SD = 5.2$, range = 20–53 years) at the University of Basel.

Results of Experiment 2. Similar to in Experiment 1, participants' probability estimates varied as a function of reward magnitude: Participants estimated higher probabilities for smaller rewards than for larger rewards (see Figure 2). A beta regression revealed that the median posterior estimate for the overarching slope coefficient was $\beta = -.09$ ($HDI_{95\%}$: $-.11$ to $-.08$). This means that for every 10 CHF decrease of reward amount, participants believed they were 2.54 ($HDI_{95\%}$: 2.16 to 3) times more likely to win. Again, the 95% HDI of the slope

coefficient excluded zero, indicating a reliable effect of the reward magnitudes on participants' probability estimates.

Discussion of Study 1

Experiment 2 confirmed the findings of Experiment 1: People judged larger rewards as less likely than smaller rewards also when the rewards were presented in random order. In both experiments, we put our participants in a situation of total uncertainty: They had to judge reward probabilities without any context or guidance on how they should solve the task. Participants used the reward magnitude as the only available cue to inform their probability judgments. Thus, participants seemed—under uncertainty—to treat rewards and reward probabilities as correlated variables.

Study 2: Does a Correlation Between Rewards and Reward Probabilities Influence Behavior in an Experienced-Based Probability Judgment Task?

With Study 2, we explored whether a correlation between rewards and reward probabilities, and in particular the direction of correlation, influences behavior in a probability judgment task.

Method

We asked people to repeatedly judge from experience the probabilities for two-outcome (zero or positive outcome) gambles. Between subjects, we manipulated the mapping between rewards and reward probabilities. The range of possible rewards over the gambles was the same in each condition, however in the RNC condition, rewards and reward probabilities were negatively correlated. In the NRPC condition they were positively correlated. In the NRNC condition they were not correlated. Figure 3 graphically represents these correlations.

Participants. Ninety subjects from the student pool of the University of Basel participated in the study (49 women, 41 men, M_{age} = 26 years; $SD = 5$; range = $19 - 45$ years). The experiment was approved by the ethics committee of the University of Basel and all participants signed informed consent. Participants decided between a show-up fee of CHF 15 (roughly $15 at the time of the experiment) or course credit. Additionally, they received a

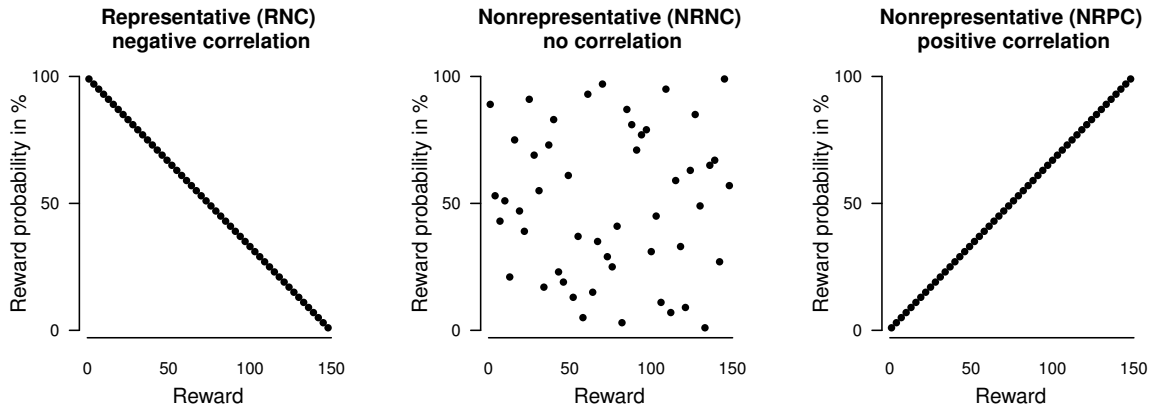

Figure 3. The relationship between the reward probabilities and the rewards in the test block separately for the three conditions. Zero occurs with counter probability. Each gamble is illustrated by a dot.

performance-dependent bonus payment (CHF 0 – CHF 10). To determine the bonus, in each trial of the test block, participants were endowed with CHF 0.20. From this endowment, we subtracted the squared deviation between response and objective reward probability, expressed as a percentage. If the deviation was larger than 9.99, the participant did not get any bonus for this trial. Participants received the summed bonuses of all trials of the test block ($M =$ CHF 4.85, $SD = 1.06$).

Materials and procedure. The experiment had a learning block, a test block, and a control block.

Learning block. In the learning block, participants estimated reward probabilities for nine randomly presented two-outcome (reward or zero) gambles (see Table 1). In each trial, participants first saw the reward amount and then drew 15 random outcomes from the gamble's outcome distribution by clicking a button on a computer keyboard. We chose the sample size of 15 for two reasons: First, it ensured that participants experiences a range of probabilities in the learning block, which is only possible when the sample size is not too small. Second, a sample size of 15 was observed as the median sample size per trial in Hertwig et al. (2004). Rewards were displayed in white numbers in the middle of a black circle. Zeros were displayed in black numbers in the middle of a white circle. After participants had drawn all 15 required outcomes, we prompted their reward–probability

estimate. Participants then got feedback about the objective reward probability and how much their estimate deviated from this probability. The purpose of the learning block was twofold. First, it allowed people to adapt to the reward–probability relationship in this study. Importantly, this means that prior beliefs about possible correlations could be corrected in condition NRNC. Second, it ensured that participants understood that the draws were random.

Table 1

The Nine Gambles of the Learning Block

Gamble	Reward	Reward probability per condition		
		Representative negative correlation	Nonrepresentative no correlation	Nonrepresentative positive correlation
1	2	.99	.52	.03
2	6	.96	.96	.06
3	11	.93	.03	.09
4	68	.55	.93	.49
5	72	.52	.55	.52
6	77	.49	.06	.55
7	137	.09	.09	.93
8	141	.06	.49	.96
9	146	.03	.99	.99

Note. All three conditions had the same rewards and probabilities.

Test block. In the test block, participants estimated probabilities for 50 gambles (see Figure 3 for an overview of the gambles). The test block trials followed the same procedure as used in the learning block with two exceptions: Participants decided freely how many outcomes they wanted to draw, and they did not receive performance feedback.

Control block. In the control block, participants estimated probabilities for nine gambles: Three gambles provided small rewards $(3; 8; 12)$; three provided medium rewards $(69; 74; 78)$; and three provided large rewards $(138; 143; 147)$. Participants did not get to sample any outcomes. They saw a gamble's reward and estimated the reward probability

based on knowledge acquired in the learning and test blocks without receiving feedback. The purpose of this block was to examine whether participants learned the correlations between rewards and probabilities during the previous two blocks.

Results

Learning block. We removed trials in which participants' probability estimates deviated by more than 50 percentage points from observed relative outcome frequencies (18 of 810 responses). We identified this outlier criterion *before* running the study on the basis of a pilot study in which some participants reported for a few trials that they erroneously judged the probability of the zero outcome instead of the positive reward. On average, participants' observed relative reward frequencies deviated by 6.3 ($SD = 6$) percentage points from the objective reward probabilities (M_{RNC} = 6.7, SD_{RNC} = 6.1; M_{NRNC} = 5.7, SD_{NRNC} = 5.5; M_{NRPC} = 6.5, SD_{NRPC} = 6.2). Figure 4 displays probability estimates as a function of objective reward probability and observed relative reward frequencies. As expected, participants' probability estimates increased as a function of increasing objective probabilities and relative observed reward frequencies. This was confirmed by a Bayesian hierarchical beta regression that regressed probability estimates on relative observed reward frequencies. The 95% HDIs of condition-dependent posterior slope and intercept estimates overlapped (Table 2), indicating no reliable difference in accuracy between conditions. As the regression line in Figure 4b and the intercepts reveal, participants tended to slightly overestimate small probabilities in all conditions.

Test block. Before we analyzed the data, we log-transformed the number of outcomes that participants drew to approximate normally distributed data. We then removed trials in which participants sampled more than $+/-2SD$ of the condition mean (163 of 4,500 responses). Furthermore, we removed trials in which participants' probability estimates deviated by more than 50 percentage points from observed relative outcome frequencies (52 of 4,337 responses).

Table 2

Beta Regression: Probability Estimates as a Function of Relative Observed Reward Frequencies

Block	Condition	Coefficient	Median	95% HDI
Learning	RNC	Intercept	0.06 (-2.7)	[0.05 (-2.9) to 0.07 (-2.6)]
		Slope	1.06 (0.06)	[1.05 (0.05) to 1.06 (0.06)]
	NRNC	Intercept	0.06 (-2.8)	[0.05 (-3.0) to 0.07 (-2.7)]
		Slope	1.06 (0.06)	[1.06 (0.05) to 1.06 (0.06)]
	NRPC	Intercept	0.06 (-2.7)	[0.06 (-2.9) to 0.07 (-2.6)]
		Slope	1.06 (0.06)	[1.06 (0.05) to 1.06 (0.06)]
Test	RNC	Intercept	0.07 (-2.7)	[0.06 (-2.8) to 0.07 (-2.6)]
		Slope	1.05 (0.05)	[1.05 (0.05) to 1.06 (0.06)]
	NRNC	Intercept	0.06 (-2.8)	[0.06 (-2.8) to 0.07 (-2.7)]
		Slope	1.06 (0.05)	[1.05 (0.05) to 1.06 (0.06)]
	NRPC	Intercept	0.07 (-2.6)	[0.06 (-2.7) to 0.08 (-2.5)]
		Slope	1.05 (0.05)	[1.05 (0.05) to 1.06 (0.06)]

Note. Coefficients of the hierarchical beta regression where we regressed probability estimates on the relative observed reward frequencies. Similar estimates for the median and the HDI bounds result from rounding. The estimates describe the posterior group estimates of the condition-dependent slope and intercept estimates converted back from the logit scale. A slope coefficient of x can be interpreted in terms of the odds of participants' probability estimates: For each percentage point increase in the relative observed reward frequency, the odds of the probability estimates increased by a factor of x. The numbers in parentheses describe the median posterior estimates on the logit scale. RNC = Representative, negative correlation; NRNC = Nonrepresentative, no correlation; NRPC = Nonrepresentative, positive correlation.

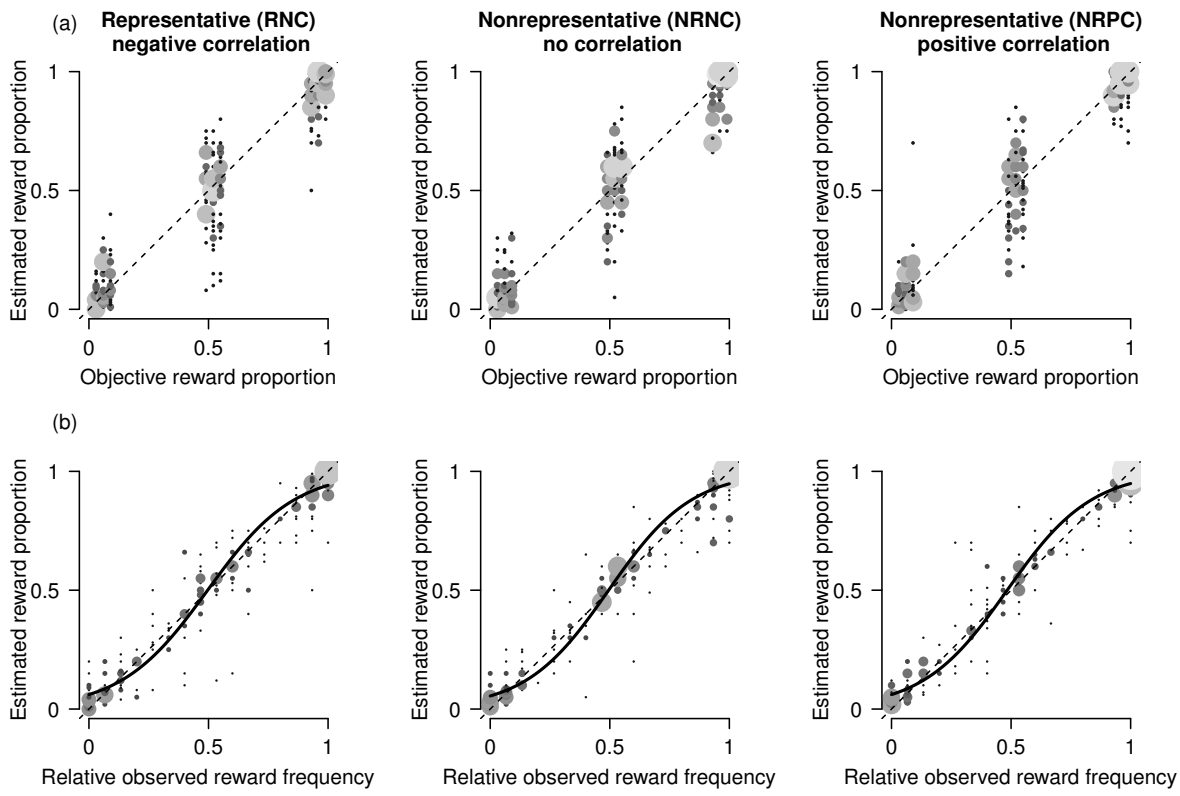

Figure 4. Estimated reward proportions in the learning block as a function of (a) objective reward proportion and (b) relative observed reward frequencies. The sizes of the dots illustrate how often each response was given. The dashed lines represent the $45°$ identity line. The solid lines represent the regression line based on the median posterior group estimates of the coefficients of the beta regression.

Sample size. Across conditions, participants drew on average 21.1 ($SD = 9.5$) outcomes (M_{RNC} = 17.3, SD_{RNC} = 6.5; M_{NRNC} = 22.8, SD_{NRNC} = 8.8; M_{NRPC} = 23.2, SD_{NRPC} = 11.3).

Did participants sample less when rewards and probabilities were (negatively) correlated? We compared search effort between conditions with a Bayesian hierarchical model. We modeled the sample size y that a participant j drew in a given trial i as being drawn from a normal distribution with a specific participant mean μ_{j} and a participant precision τ_{j}. The participant means were modeled as draws from overarching condition-dependent normal distributions with mean $\mu_{\text{Condition}}$ and group precision $\tau_{\text{Condition}}$.

The condition-specific parameter $\mu_{\text{Condition}}$ measures the posterior estimates for sample

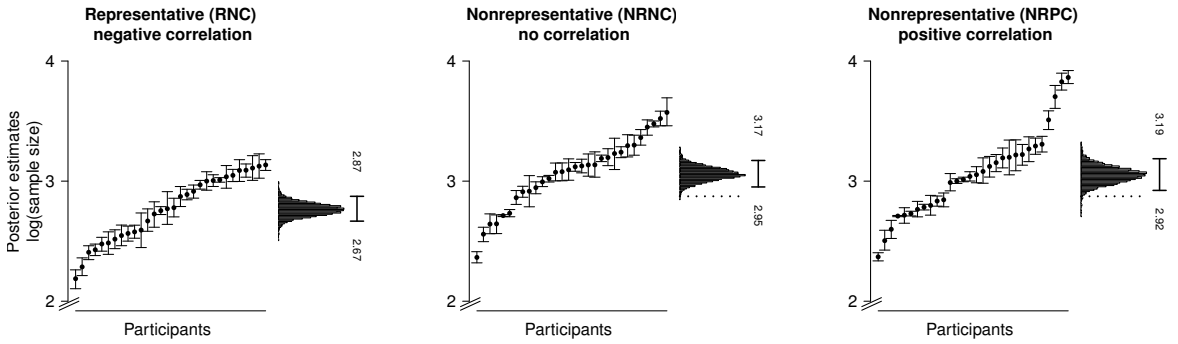

Figure 5. Posterior estimates of the group means (histograms) and posterior individual subject means (dots). The error bars indicate the 95% highest density intervals (HDIs). The dotted lines indicate the upper bound of the 95% HDI of μ_{RNC}.

sizes separately for each condition. The posterior densities of these overarching $\mu_{Condition}$ are shown in Figure 5 alongside the posterior densities of individual mean sample-size estimates (μ_j). The upper bound of the 95% HDI of μ_{RNC} does not overlap with the lower bounds of the HDIs of both other conditions. This indicates that participants reliably sampled less in the representative condition as compared to both nonrepresentative conditions[1].

Figure 5 shows that the individual parameters μ_j are slightly more dispersed in both nonrepresentative conditions, suggesting larger between-subject variability of mean sample sizes than in the representative condition. However, the 95% HDI of the posterior group estimates for the precision parameters $\tau_{Condition}$ of all conditions overlap, which indicates that there were no reliable group differences in precision (τ_{RNC} = 12.5, $HDI_{95\%}$: 6.6 to 19.9; τ_{NRNC} = 10.9, $HDI_{95\%}$: 5.6 to 17; τ_{NRPC} = 7.3, $HDI_{95\%}$: 3.8 to 11.5).

Did sample size decrease over the course of the experiment? Figure 6 displays how sample size evolved over trials in the different conditions. We analyzed whether sample size decreased with a Bayesian hierarchical linear regression with trial number as a predictor. Only for the representative condition did sample size reliably decrease over the course of the experiment, as indicated by the reliably negative slope parameter $\beta_{1,RNC}$ ($Mdn\ \beta_{1,RNC}$ =

[1]This finding held (and was even accentuated) when we applied a stricter outlier criterion and removed all trials in which sample sizes deviated by more than one standard deviation from the condition mean as well as when we did not exclude any data.

-0.003; $\text{HDI}_{95\%}$: -0.006 to -0.001). For both nonrepresentative conditions, the 95% HDIs of the condition-specific slope parameter include 0, suggesting there were no reliable changes of sample sizes over trials ($Mdn\beta_{1,\text{NRNC}} = -0.001$, $\text{HDI}_{95\%}$: $-.004$ to $.001$; $Mdn\ \beta_{1,\text{NRPC}} = .001$, $\text{HDI}_{95\%}$: $-.004$ to $.001$).

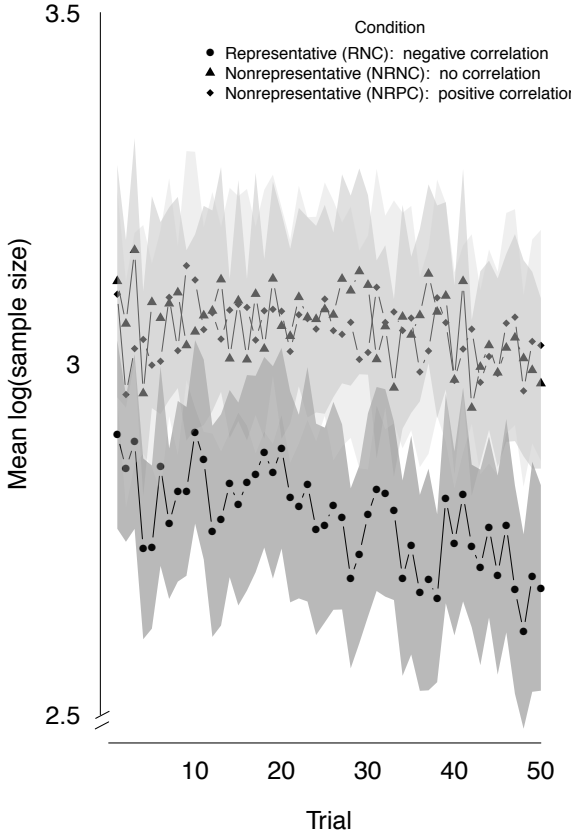

Figure 6. Mean of log sample sizes as a function of trial separately for the three conditions. The shaded area indicates 95% confidence intervals.

Probability judgments. On average, participants' observed relative reward frequencies deviated by 7.3 percentage points (pp) ($SD = 6.2$) from the objective reward probabilities ($M_{\text{RNC}} = 8$ pp , $SD_{\text{RNC}} = 6.9$; $M_{\text{NRNC}} = 6.7$ pp, $SD_{\text{NRNC}} = 5.8$; $M_{\text{NRPC}} = 7.1$ pp, $SD_{\text{NRPC}} = 6$). Figure 7 illustrates participants' probability estimates as a function of objective reward probabilities and relative observed outcome frequencies. Similar to in the learning block, a hierarchical Bayesian beta regression revealed that the 95% HDIs of the posterior slope- and intercept coefficients overlapped in all conditions (Table 2), indicating that accuracy did not differ between conditions.

We further analyzed participants' probability estimates with the Bayesian updating model M_B and the natural-mean heuristic M_{NM} (Appendix A). We estimated both models with participants' individual data by applying maximum likelihood methods. To compute the likelihood, we assumed participants' probability estimates would follow a truncated normal distribution (0 – 1) around the model's predicted probability estimate. We estimated the standard deviation of the truncated normal distribution as a free parameter for each participant. With a grid-search approach we identified the set of parameter values that minimized the deviance (negative log likelihood) between predictions and responses. We searched the parameter space for the strength parameter *N* between 2 and 50 in steps of 1 and for the standard deviation of the truncated normal distribution between 0.00001 and 0.3 in steps of 0.01. We compared the models based on the Akaike information criterion (AIC) that accounts for the number of free parameters (P; $AIC = -2 \times LogL + 2 \times P$; Burnham & Anderson, 2002).

Our model comparison revealed that most participants' data were best fit by the natural-mean heuristic. Only for 4 of the 90 participants did the Bayesian updating model describe the data better (RNC: 1 of 30; NRNC: 0 of 30; NRPC: 3 of 30)[2].

Control block. In the control block, participants estimated reward probabilities without sampling. Probability estimates clearly depended on the magnitude of the reward (see Figure 8) in both correlated conditions. This finding was supported by a Bayesian hierarchical beta regression that used reward magnitudes to predict probability estimates. In the RNC condition, probability estimates decreased as a function of reward magnitudes as shown by the median slope parameter $b_{1,\text{ Condition}}$ ($b_{1,\text{ RNC}} = -.023$, $\text{HDI}_{95\%}$: $-.03$ to $-.014$). In the NRPC condition probability estimates increased as a function of reward amount ($b_{1,\text{ NRPC}} = .016$, $\text{HDI}_{95\%}$: $.009$ to $.022$). In the NRNC condition there was no reliable systematic relation between probability estimates and reward magnitudes ($b_{1,\text{ NRNC}} = -.001$, $\text{HDI}_{95\%}$: $-.006$ to $.004$). The absolute values of the 95% HDIs' slope parameters of the two correlated conditions overlapped, suggesting that participants in both conditions learned the contingencies equally

[2]We also tested variants of the Bayesian updating model where we estimated prior probability beliefs on the individual level from the data or using the regression coefficients of the control block to estimate prior probability beliefs. None of the models outperformed the natural-mean heuristic.

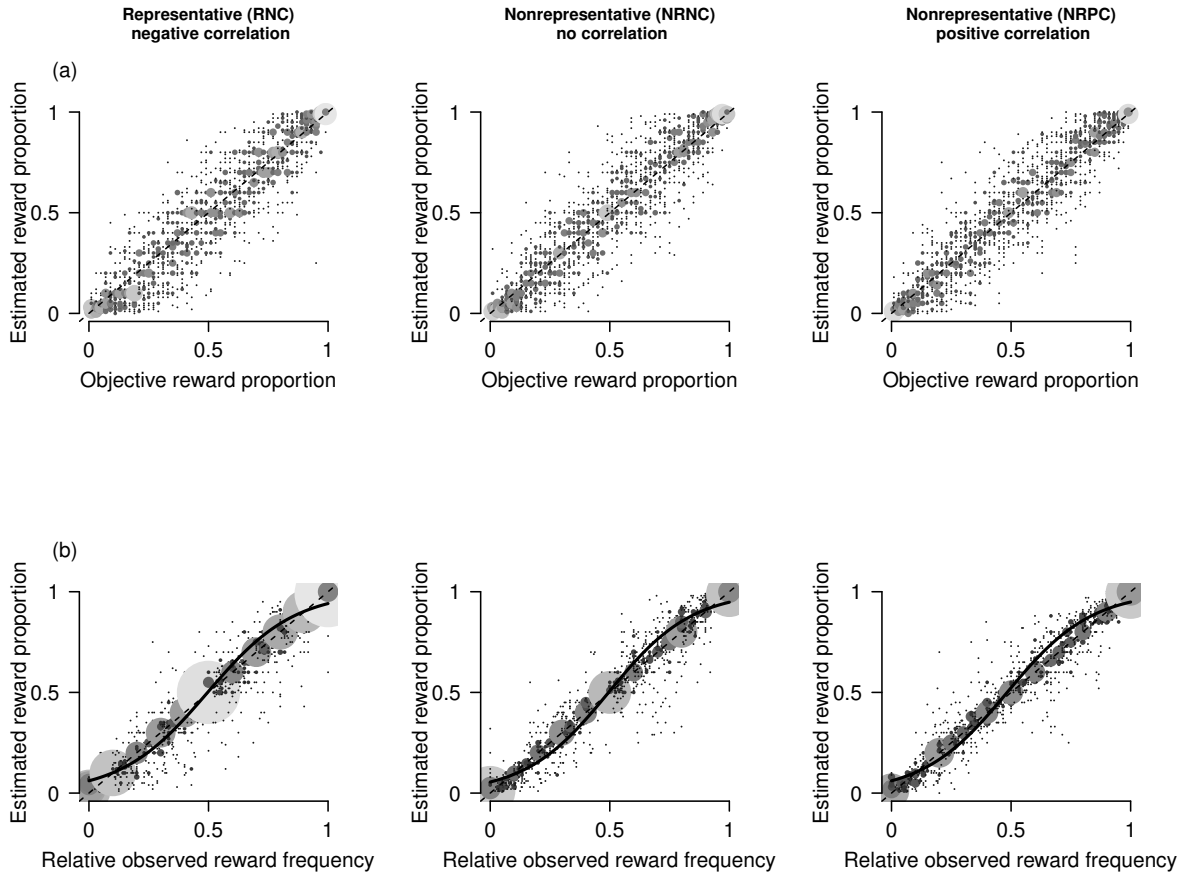

Figure 7. Estimated reward proportions in the testblock as a function of (a) objective reward proportion and (b) relative observed reward frequencies. The sizes of the dots illustrate how often each response was given. The dashed lines indicate the 45° identity line. The solid lines indicate the regression line based on the median posterior group estimates of the coefficients of the beta regression.

well.

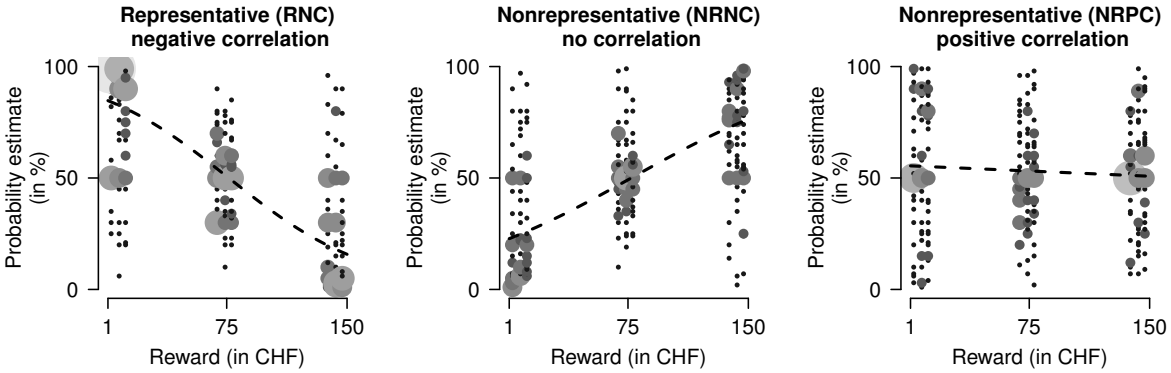

Figure 8. Probability estimates as a function of reward amount. Sizes of the dots indicate how often individual responses were given. The dotted lines represent the regression line based on the median estimates of the posterior intercept and slope distributions.

Discussion

The present work shows that people generally expect that the size of rewards and the probabilities with which these rewards occur in gambles are negatively correlated. Furthermore, people search for less information under uncertainty when this expectation is met by the stimuli. In Study 1, we asked participants to estimate reward probabilities for two-outcome gambles under complete uncertainty conditions. Their probability estimates were guided by their expectations and varied as a function of reward magnitude: Smaller rewards were assumed to be more likely than larger rewards. With Study 2, we tested whether encountering a task environment that conforms with people's prior beliefs affects their information search and probability judgments. In this study, we asked participants to estimate reward probabilities of two-outcome gambles. Participants sampled outcomes from the gambles. Across participants, we manipulated the relationship between rewards and reward probabilities. In a representative condition, the correlation mimicked the ecological reward and reward probability relationship such that larger rewards were less likely than smaller rewards. In two nonrepresentative conditions, rewards and reward probabilities were either not correlated or positively correlated. We hypothesized that a correlation between reward and reward probabilities could influence people's search effort and how they integrate the acquired

information for their final probability judgment.

In sum, participants searched for less information when rewards and reward probabilities were negatively correlated as compared to situations in which they were positively correlated or not correlated. But the way participants integrated the acquired information, that is, their judgment strategy, did not depend on different reward and reward probability relationships: They treated the outcomes that they drew as if it was representative of the true outcome distribution.

These findings contradict a full Bayesian approach, which would suggest that people integrate knowledge about how rewards and probabilities are correlated to determine how much to sample *and* when to estimate probabilities. However, our data suggest that participants heuristically applied a two-stage strategy: First, they sample X outcomes. Second, they estimated the relative reward frequency of the outcomes that they observed. But what is the mechanism that influenced the sample size X (i.e., the number of outcomes participants drew) and led to lower sample sizes in the representative condition than in both nonrepresentative conditions?

Does Awareness of Correlation Influence Sample Size?

Generally, people's search effort could be influenced by any correlation between rewards and probabilities. Lejarraga et al. (2012) argued that in decisions from experience, people learn how gambles are structured. Consequently, people's search effort is directly influenced by their knowledge. In our experiment, participants in both correlated conditions learned the relationship between rewards and reward probabilities. This becomes evident from the results of the control block where participants estimated reward probabilities *without* sampling. We found that in all conditions, their responses resembled the contingencies between rewards and reward probabilities that they had observed in the previous blocks (see Figure 8). We conclude that a general awareness that a correlation exists does not influence sample size. We argue instead that familiarity or belief in the plausibility of a correlation between rewards and probabilities influences sample size.

Belief in the Plausibility of Correlation

Arguably, participants learned the correlation in the nonrepresentative condition as well as in the ecologically representative condition but apparently did not *trust* this correlation as much. The notion that larger rewards are more likely than smaller rewards could appear so counter-intuitive that in each trial, participants felt that they needed to sample more outcomes to verify that the correlation still existed. This hypothesis receives support from the result that only in the representative condition did sample size decrease over the course of trials. Future research should test this hypothesis, for instance, by assessing confidence in judgments.

Conclusion and Outlook

Our study provides evidence that people exploit representative reward–probability regularities of the environment when they search for information in experience-based judgment tasks. This finding is crucial for the well-studied domain of decisions from experience (Hertwig et al., 2004). There, the expected values of the two competing gambles that participants choose between are often matched. This matching of expected values creates an environment where rewards and reward probabilities of competing gambles are negatively correlated (see Figure 1). People potentially exploit this correlation by inferring properties of one gamble from knowledge about the other gamble. Suppose a participant draws a few outcomes from Gamble A, showing small rewards, and a few outcomes from Gamble B, showing a number of zero outcomes. The person expects that rewards and probabilities are negatively correlated and correctly guesses from her observations that the nonobserved reward value in Gamble B must be relatively high. Hence, the subjective representation of the gambles may be more accurate than would be assumed based on the observed outcome frequencies which potentially influences a decision makers search effort and/or decisions.

More generally, our study shows an aspect of human cognition that is often overlooked: People behave differently depending on whether the research environment (e.g., stimuli) corresponds to the structure of everyday situations outside the laboratory. When one ignores people's prior beliefs about the structure of the world, one might fail to observe the ecological rationality of human cognition.

References

Alloy, L. B., & Tabachnik, N. (1984). Assessment of covariation by humans and animals: The joint influence of prior expectations and current situational information. *Psychological review*, *91*, 112–149.

Brunswik, E. (1947). *Systematic and representative design of psychological experiments. With results in physical and social perception.* Berkeley, CA: University of California Press.

Brunswik, E. (1955). Representative design and probabilistic theory in a functional psychology. *Psychological Review*, *62*, 193–217.

Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference: A practical information-theoretic approach*. New York, NY: Springer.

Busemeyer, J. R., Byun, E., Delosh, E. L., & McDaniel, M. A. (1997). Learning functional relations based on experience with input-output pairs by humans and artificial neural networks. In K. Lamberts & D. Shanks (Eds.), *Knowledge concepts and categories* (p. 405-437). New York: Psychology Press.

Busemeyer, J. R., & Rapoport, A. (1988). Psychological models of deferred decision making. *Journal of Mathematical Psychology*, *32*, 91–134.

Camilleri, A. R., & Newell, B. R. (2013). Mind the gap? Description, experience, and the continuum of uncertainty in risky choice. In N. Srinivasan & C. Pammi (Eds.), *Progress in brain research: Vol. 202. Decision making: Neural and behavioural approaches.* (pp. 55–71). Oxford, UK: Elsevier.

Dieckmann, A., & Rieskamp, J. (2007). The influence of information redundancy on probabilistic inferences. *Memory & Cognition*, *35*, 1801–1813.

Fiedler, K. (2000). Beware of samples! A cognitive-ecological sampling approach to judgment biases. *Psychological review*, *107*, 659-676.

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

Gigerenzer, G., & Hug, K. (1992). Domain-specific reasoning: Social contracts, cheating, and perspective change. *Cognition*, *43*, 127–171.

Gigerenzer, G., Todd, P. M., & the ABC Research Group. (1999). *Simple heuristics that make*

us smart. New York, NY: Oxford University Press.

Griffin, D., & Tversky, A. (1992). The weighing of evidence and the determinants of confidence. *Cognitive Psychology*, *24*, 411–435.

Hau, R., Pleskac, T., & Hertwig, R. (2010). Decisions from experience and statistical probabilities: Why they trigger different choices than a priori probabilities. *Journal of Behavioral Decision Making*, *68*, 48–68.

Hau, R., Pleskac, T. J., Kiefer, J., & Hertwig, R. (2008). The description–experience gap in risky choice: The role of sample size and experienced probabilities. *Journal of Behavioral Decision Making*, *21*, 493–518.

Hertwig, R., Barron, G., Weber, E. U., & Erev, I. (2004). Decisions from experience and the effect of rare events in risky choice. *Psychological Science*, *15*, 534–539.

Hertwig, R., & Erev, I. (2009). The description-experience gap in risky choice. *Trends in Cognitive Sciences*, *13*, 517–523.

Hertwig, R., & Pleskac, T. J. (2008). The game of life: How small samples render choice simpler. In N. Chater & M. Oaksford (Eds.), *The probabilistic mind: Prospects for rational models of cognition* (p. 209–236). Oxford, UK: Oxford University Press.

Knight, F. H. (1921). *Risk, uncertainty, and profit.* New York, NY: Sentry Press.

Lejarraga, T., Hertwig, R., & Gonzalez, C. (2012). How choice ecology influences search in decisions from experience. *Cognition*, *124*, 334–342.

Markowitz, H. (1952). Portfolio selection. *Journal of Finance*, *7*, 77-91.

Mata, R., Schooler, L. J., & Rieskamp, J. (2007). The aging decision maker: Cognitive aging and the adaptive selection of decision strategies. *Psychology and Aging*, *22*, 796–810.

Mehlhorn, K., Ben-Asher, N., Dutt, V., & Gonzalez, C. (2014). Observed variability and values matter: Toward a better understanding of information search and decisions from experience. *Journal of Behavioral Decision Making*, *27*, 328–339.

Pleskac, T. J., & Hertwig, R. (2014). Ecologically rational choice and the structure of the environment. *Journal of Experimental Psychology: General*, *143*, 2000–2019.

Plonsky, O., Teodorescu, K., & Erev, I. (2015). Reliance on small samples, the wavy recency effect, and similarity-based learning. *Psychological review*, *122*, 621–647.

Pohl, R. F. (2006). Empirical tests of the recognition heuristic. *Journal of Behavioral Decision Making*, *19*, 251–271.

R Core Team. (2014). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `http://www.R-project.org/`

Rakow, T., Demes, K. A., & Newell, B. R. (2008). Biased samples not mode of presentation: Re-examining the apparent underweighting of rare events in experience-based choice. *Organizational Behavior and Human Decision Processes*, *106*, 168–179.

Rieskamp, J. (2008). The probabilistic nature of preferential choice. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*, 1446-1465.

Rieskamp, J., & Otto, P. E. (2006). SSL: A theory of how people learn to select strategies. *Journal of Experimental Psychology: General*, *135*, 207-236.

Roth, Y., Wänke, M., & Erev, I. (2016). Click or Skip: The Role of Experience in Easy-Click Checking Decisions. *Journal of Consumer Research*, *43*, 583–597.

Sharpe, W. F. (1964). Capital asset prices: A theory of market equilibrium under conditions of risk. *The Journal of Finance*, *19*, 425–442.

Todd, P. M., & Gigerenzer, G. (2012). *Ecological rationality: Intelligence in the world*. New York, NY: Oxford University Press.

Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin*, *76*, 105-110.

Appendix A

Bayesian Updating Model

The Bayesian updating model describes how people can make relatively accurate probability judgments based on a few observed outcomes. Basically, the model suggests that people learn the underlying patterns of correlation between rewards and probabilities and integrate this information into their judgments. The model makes two assumptions: First, in every trial people start sampling with a prior belief about the reward probability of this trial t $p_{t=0}$. Second, people update this belief as they sample.

We assumed that people's prior belief depends on the objective correlation between rewards and probabilities and can be described as $p_{t=0} \sim Beta(a, b)$. To estimate a and b, we assumed that in both correlated conditions the mode of people's prior belief distribution is approximated by the objective reward probability. In the uncorrelated condition, we used .5 as approximation of the mode of people's prior belief distribution. With these assumptions, parameter $b = N \times -p + N + 2 \times p - 1$. Here N is a free parameter that describes the strength of the prior belief. The larger N is, the more influence the prior belief has relative to the sampled information. N has to equal or be greater than 2. Parameter $a = N - b$. We assumed that people integrate new information as they sample. Mathematically, this can be described by adding 1 to parameter a of the beta-distributed probability belief p_t every time a person samples a reward and by adding 1 to parameter b of the beta-distributed probability belief p_t every time a person samples a zero. In a third step, people deduct an estimate of the reward probability ($p_{M_B,reward,t}$). After t outcomes this estimate equals the mean of the updated Beta distribution ($p_{M_B,reward,t} = \frac{a_t}{a_t+b_t}$).

Natural-Mean Heuristic

We compared the Bayesian model described above to a variant of the natural-mean heuristic M_{NM}. There, people's probability estimates are derived from the relative frequency of reward and zero outcomes. This M_{NM} model implies that people treat the observed outcomes as if they describe the underlying outcome probabilities comprehensively.

Mathematically, the probability judgment about the reward probability $p_{M_{NM},reward}$ after

t observations is defined as $p_{M_{NM},reward,t} = \frac{1}{t} \times \sum_{t=1}^{t} f(t)$ where $f(t)$ is a sign function that equals 1 when the observed outcome was a reward and zero otherwise.

Appendix B

Description of Bayesian Models

The following code describes the hierarchical beta regression that we used in Studies 1 and 2.

```
model {
for(i in 1:Ndata) {
y[i] ~dbeta(alpha[i], beta[i])
alpha[i] <- mu[i] * tau[subj[i]]
beta[i] <- (1-mu[i]) * tau[subj[i]]
mu[i] <- 1/(1+ exp(-1* (b0[subj[i]]+inprod(b1[subj[i]],x[i]))))
}
for(vp in 1:Nsubj){
tau[vp] ~dgamma(.001,.001)
b0[vp] ~dnorm(mub0[Csubj[vp]], taub0[Csubj[vp]])
b1[vp] ~dnorm(mub1[Csubj[vp]], taub1[Csubj[vp]])
}
for(c in 1:cond){
mub1[c] ~dnorm(0,.0001)
taub1[c] ~dgamma(.001,.001)
mub0[c] ~dnorm(.5,.0001)
taub0[c] ~dgamma(.001,.001)}
}
```

- $Ndata$: Total number of data points
- y: Probability estimate
- $subj$: Identifies to which subject each individual data point belongs
- x: Predictor
- $Nsubj$: Number of participants
- $Csubj$: Identifies in which condition a subject is
- $cond$: Number of between-subject conditions

The following code describes the hierarchical model of the sample-size analysis.

```
model {
for( i in 1 : Ndata ) {
y[i] ~dnorm( mu[ subj[i] ] , tau[ subj[i] ] ) T(0,)
}
for ( j in 1 : Nsubj ) {
mu[j] ~dnorm( muG[ Csubj[j] ] , tauG[Csubj[j]]) T(0,)
tau[j] ~dgamma(.001,.001)
}
for(c in 1:cond){
muG[c] ~dnorm( 2.95, .0001 )T(0,)
tauG[c] ~dgamma( .001,.001 )
}
dif1 <- muG[2] - muG[1]
dif2 <- muG[3] - muG[1]
}
```

- $Ndata$: Total number of data points
- y: Data point (Sample size)
- $subj$: Identifies to which subject each individual data point belongs
- $Nsubj$: Number of participants
- $Csubj$: Identifies in which condition a subject is
- $cond$: Number of between-subject conditions

The following code describes the regression of sample size on trial number.

```
model {
for(i in 1:Ndata){
y[i] ~dnorm(mu[i], tau[subj[i]])
mu[i] <- b0[subj[i]]+b1[subj[i]]*trial[i]
}
for(vp in 1:Nsubj){
tau[vp] ~dgamma(.001,.001)
b0[vp] ~dnorm(mub0[Csubj[vp]], taub0[Csubj[vp]])
b1[vp] ~dnorm(mub1[Csubj[vp]], taub1[Csubj[vp]])
}
for(c in 1:cond){
mub1[c] ~dnorm(0,.0001)
taub1[c] ~dgamma(.01,.01)
mub0[c] ~dnorm(0,.0001) T(0,)
taub0[c] ~dgamma(.01,.01)}
}
```

- $Ndata$: Total number of data points
- y: Data point (Sample size)
- $trial$: Predictor (Trial number)
- $subj$: Identifies to which subject each individual data point belongs
- $Nsubj$: Number of participants
- $Csubj$: Identifies in which condition a subject is
- $cond$: Number of between-subject conditions

Rewards and Perceived Risks in Clinical Trials: Monetary Incentives are a Cue for Side-Effects

Janine Hoffart[1], Benjamin Scheibehenne[2]

University of Basel[1], University of Geneve[2]

Author Note

We report no conflict of interest. Correspondence concerning this article should be addressed to Janine Christin Hoffart. University of Basel, Department of Psychology, Missionsstrasse 62a, 4055 Basel, Switzerland. E-mail: janine.hoffart@unibas.ch

Abstract

In clinical trials, incentivizing human research subjects with too much money is often considered unethical as it may coerce people to participate. This argument implies that people perceive rewards (i.e., incentives) independently of risks (i.e., probability of side-effects) or assume that smaller rewards are associated with higher risks than larger rewards. However, research on risk perception indicates that people associate higher rewards with higher risks. Here, we investigated whether people expect that the magnitude of incentives payed for participation in clinical trials foreshadows the trials' riskiness. In an experiment, participants estimated how many people participating in a hypothetical clinical trials they expect to experience side effects. Between subjects, we manipulated the incentives offered for participation in the trial ($500 vs. $10'000). Participants expected more side effects in the high-incentive condition. This result suggests that paying large rewards may not necessarily be coercive as people implicitly associate them with higher risk.

Keywords: Incentives, Coercion

Rewards and Perceived Risks in Clinical Trials: Monetary Incentives are a Cue for Side-Effects

In medical research, it is an open question how participants shall be compensated (Grady, 2005). The Council for International Organizations of Medical Sciences (CIOMS, 2016) suggests that large incentives serve as undue inducements as they entrap people to participate in studies against their better judgments. Following this argumentation, ethicists and researchers worry that large incentives coerce people to participate in clinical trials (e.g., Macklin, 1981; McNeill, 1997).

As ethical guidelines explicitely state that "the level of compensation should not be related to the level of risk that participants agree to undertake" (CIOMS, 2016; pp. 53 – 54), it is implicitely or explicitely assumed that also subjects perceive incentives and riskiness as independent (Ambuehl et al., 2015). In contrast, research on risk perception shows that people often expect a correlation between risks and benefits. Both evidence for a negative and a positive correlation have been reported.

The first stream of research suggests that people expect risks and benefits to be negatively correlated. For instance, in an experiment people judged safer activities and technologies as more beneficial than riskier activities and technologies (Alhakami & Slovic, 1994). Likewise, probabilities of attractive outcomes are often overestimated (Irwin, 1953). In line with ethical concerns, from this perspective, high incentives may lower participants' risk-estimates of clinical trials and unethically lure them to participate.

In contrast, Edwards (1962), conjectured that "our world is so constructed that the more desirable objects are harder to get" (p.49). Pleskac and Hertwig (2014) provided empirical evidence for this claim and found that in many environments higher payoffs occur with lower probabilities than smaller payoffs. Further, when probability information is missing, people

intuitively infer reward-probabilities from reward-magnitudes and expect larger rewards to be less likely than smaller rewards (Hoffart, Rieskamp, & Dutilh, in press; Pleskac & Hertwig, 2014). This risk-reward heuritic dovetails with the idea that people believe in *fair bets* and believe that expected values are similar across comparable situations (Osherson, 1995). Applied to clinical trials, the risk-reward heuristic predicts that people expect larger risks when incentives are higher. Consistently, Cryder, London, Volpp and Loewenstein (2009) reported that participants in a behavioral experiment perceived research trials as riskier when incentives were larger. Consequently, if higher incentives signal greater risk, increased attractiveness of clinical trials due to better payments may be attenuated: With growing payments also perceived riskiness increases which makes coerced participation less likely than when incentives and risks are perceived as uncorrelated or even negatively correlated.

Given these diverse theoretical predictions outlined above, it is important to understand whether people perceive risks and rewards of clinical trials as independent (ethical guidelines hypothesis), or alternatively assume that higher compensation will either lead to a decrease (desirability hypothesis) or an increase (risk-reward hypothesis) in expected riskiness.

Method

To empirically test the three hypotheses described above, we conducted an online experiment on Amazon Mechanical Turk. The study design and statistical analysis were preregistered on the open science framework (https://osf.io/b4wtr/?view_only=db1bc2ea30cc4cd28126cbecfe010fd6). Participants of our experiment read a hypothetical advertisement for a clinical trial that aimed to test a new vaccine against Ebola for women, adapted from an experiment by Ambuehl et al. (2015). The text mentioned that side effects may occur, yet it did not provide precise information about the

likelihood of occurrence. We manipulated between subjects whether hypothetical participants of the vaccine-trial would be reimbursed with $500 or $10,000. As previous research showed that not comprehensive definition of the term *risk* exists (Slovic, 1987), we clearly defined risk as the proportion of women participating in the trial who will suffer from side effects. After reading the text and passing an attention check, participants estimated how many out of 1,000 women participating in the clinical trial would suffer from *mild* and from *very severe* side effects. We separately asked for mild and severe side effects as a *Trial X* may be judged as riskier than *Trial Y* because a) overall more side effects occur in *Trial X* or b) the absolute number of side effects is similar across trials but relatively more severe (compared to mild) side effects occur in *Trial X*.

In addition, participants stated how ethical they perceived different compensation schemes (i.e., no money, $500, and $10'000) and whether they themselves would participate in such a trial for a) a reimbursement of $500 and b) a reimbursement of $10'000. We asked these questions as we planned to analyze whether people who believe it is more ethical to reimburse participants of medical trials with little money (referred to as “Ethicists”, Ambuehl et al., 2015) expect a greater increase of side effects when incentives are larger than “Economists” who believe that it is more ethical to reimburse participants of medical trials with much money. Further, we assessed whether participants would approve the trial if they were part of an ethical committee on a scale form 1 (definitively reject) to 7 (definitively approve). We asked this question to explore which factors (i.e., incentive and/or personal expectations about side effects) influence approval decision.

In total, we collected valid data from 483 participants (223 women, 258 men, and 3 did not respond, collected in two batches) who were remunerated with $1 for their participation. We collected as many participants as our budget permitted. As preregistered, we controlled for

possible outliers in people's risk estimate by excluding 20% of the most extreme data points (i.e.: 10% of the lowest data points and 10% of the highest data points) within both experimental conditions. On average participants of the final sample (N= 371, 169 women, 201 men, and 1 did not respond) were 36 years old (*SD* = 10.23, range = 18 - 72). To analyze the data, we ran negative binomial regressions for count data (estimated number of side effects) and ordinal regressions for Likert scale data (approval decisions) in R. We based our inferences about whether an effect was significant by comparing the Bayesfactors (calculated from the models' BICs, Kass & Raftery, 1995) of a model including a predictor of interest and the simpler model without this predictor.

Results

We first report the results on people's side effect judgments and second, the results on people's ethical approval judgments.

Side Effect Estimates

People's expectations for all side effect (sum of mild and very severe) differed between conditions. In line with the risk—reward hypothesis, in the high incentive condition people expected that more women would suffer from side effects (*M* = 237.26, *SD* =200.63) than in the low incentive condition (*M* = 170.15, *SD* = 142.91). This data was better described by a regression model assuming that the means differ between both conditions as compared to a model assuming similar means. The corresponding Bayes factor (BF) is 65, indicating strong evidence for the risk-reward hypothesis.

Figure 1 illustrates for both mild and severe side effects participants in the high incentive condition expected that more women will suffer from side effects than in the low incentive

condition. The previously described regression model, conducted separately for mild and severe side effect judgments, confirmed these results (BF_{mild} = 31.49; BF_{severe} = 11.16), providing evidence for the risk-reward hypothesis for both dependent variables. A model including only the main effects of severity and incentives was preferred over a model also including the interaction ($BF_{nointeraction}$ = 24.11).

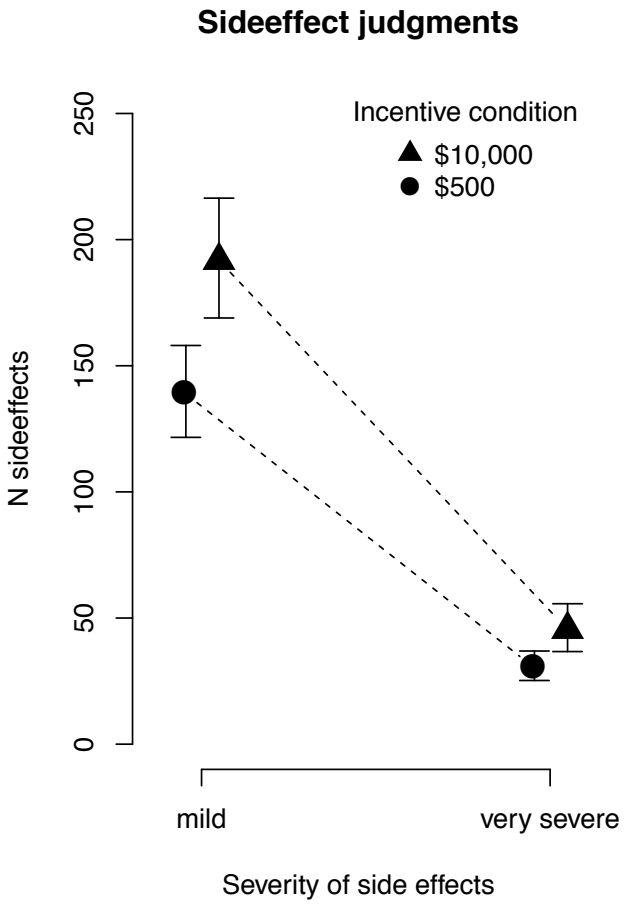

Figure 1. Judgments (y-axis) of mild and severe side effects (x-axis) per incentive condition. The error bars represent the bootstrapped 95% confidence intervals of the mean judgments.

We also planned to separately compare judgments for ethicists (believe that lower incentives are more ethical than larger incentives) and economists (believe that larger incentives are more ethical than lower incentives). 248 people (98 low incentive condition; 150 high incentive condition) were classified as Economists and 40 people (31 low incentive condition; 9 high incentive condition) as Ethicists. The remaining 83 people (54 low incentive condition; 29 high incentive condition) found it equally ethical to reimburse people with $500 and $10,000. The fact that only few people were classified as Ethicists impeded further statistical analyses.

Do side effect judgments influence ethical approval judgments?

In the high incentive condition, people were slightly more likely (*Mdn* = 6, on a scale from 1 to 7) to approve the clinical trial than in the low incentive condition (*Mdn* = 5; $BF_{condition}$ = 6.34). Crucially, higher approval ratings were associated with lower estimates for the number of women who will suffer from side effects (r = -0.13, n = 371, p = 0.01). The Bayesfactor analysis provided more evidence for an ordinal regression predicting approval ratings as function of incentive condition and side effect judgments than a regression only including the incentive condition ($BF_{bothmaineffects}$ = 7.12) or personal side effect judgments ($BF_{bothmaineffects}$ = 33.87). This finding indicates that approval ratings depend on both monetary incentives and individual expectations of side effects.

Discussion

Here, we experimentally studied whether people expect that the magnitudes of rewards payed for participation in research trials foreshadows how risky the clinical trial is. To do so, we contrasted three hypotheses: First, the ethical guidelines hypothesis that predicts that side effect estimates are independent of the incentives' magnitudes. Second, the desirability hypothesis, that

predicts that side effect estimates decrease with increasing incentives. And third, the risk-reward hypothesis that predicts that side effect estimates increase with increasing incentives. Consistently with the risk-reward hypothesis and previous findings (Cryder et al., 2009), people expected that more women will suffer from both mild and very severe side effects when participation in the trial was incentivized with $10,000 instead of $500. This finding contrasts with ethical guidelines that explicitely state that monetary incentives must strictly not be used to compensate for risks (CIOMS, 2016). Clinicians and researchers should consider that people infer riskiness from incentives when explaining potential consequences of participation in clinical trials to patients.

People in the high incentive condition were more likely to approve the trial than people in the low incentive condition although they expect risks to be larger. In addition, approval judgments were higher for people who expect fewer side effects than for people who expect more side effects. In sum, these findings indicate that incentives and personal risk expectations contribute individually to the approval decision and complex interactions may be involved in the way people perceive the riskiness of clinical trials.

To conclude, in the discussion about how subjects of clinical trials shall be compensated, it has mostly been attended that large incentives may harm decision making. However, people treat incentives as signal for how risky trials are. This expectation needs to be considered when determining payment schemes.

References

Alhakami, A. S., & Slovic, P. (1994). A psychological study of the inverse relationship between perceived risk and perceived benefit. *Risk analysis, 14*, 1085-1096.

Ambuehl, S., Niederle, M., & Roth, E. A. (2015). More money, more problems? Can high pay be coerciveand repugnant? *American Economic Review: Papers Proceedings, 105*, 357–360.

Council for International Organizations of Medical Sciences (CIOMS). (2016). *International ethical guidelines for health-related research involving humans.* Geneva, Switzerland: CIOMS.

Cryder, C. E., London, A. J., Volpp, K. G., & Loewenstein, G. (2009). Informative inducement: Study payment as a signal of risk. *Social Science & Medicine, 70*, 455-464.

Edwards, W. (1962). Utility, subjective probability, their interaction, and variance preferences. *Journal of Conflict Resolution. 62*, 42-51.

Grady, C. (2005). Payment of clinical research subjects. *Journal of Clinical Investigation, 115*, 1681-1687.

Hoffart, J. C., Rieskamp, J., & Dutilh, G. (in press). How environmental regularities affect people's information search in probability judgments from experience. *Journal of Experimental Psychology. Learning, Memory, and Cognition*.

Irwin, F. W. (1953). Stated expectations as functions of probability and desirability of outcomes. *Journal of Personality. 21*, 329-335.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistic Association, 90, 773-795.*

Macklin, R. (1981). 'Due' and 'Undue' inducements: On paying money to research subjefcts. *IRB: Ethics & Human Research, 3*, 1-6.

McNeill, P. (1997). Paying people to participate in research: Why not? *Bioethics, 11*, 390-396.

Osherson, D. N. (1995). Probability judgment. In E. E. Smith, & D. N. Osherson (Eds.), *Thinking* (pp. 35–75). Cambridge, MA: MIT Press.

Pleskac, T. J., & Hertwig, R. (2014). Ecologically rational choice and the structure of the environment. *Journal of Experimental Psychology: General, 143*, 2000-2019.

Slovic, P. (1987). Perception of risk. *Psychological Science, 236*, 280-285.

Curriculum Vitæ

Personal Details

Name	Janine Hoffart
Date of birth	18.11.1987 in Velbert, Germany