Simplified mental representations as a cause of overprecision

Raúl López-Pérez†  Antonio Rodriguez-Moral‡  Marc Vorsatz§

January 31, 2021

*All materials related to this study (z-tree programs, raw data, Stata-scripts, and Mathematica programs) are publicly available on the Open Science Framework (OSF) under the link https://osf.io/wq5nm/?view_only=2581875617c942f9bdf1c4db02ade4fd.

†Corresponding author. Institute of Public Goods and Policies Spanish National Research Council (CSIC) C/ Albasanz 26-28, Madrid, 28037, Spain. Email: raul.lopez@csic.es. Financial support from the Spanish Ministry of Economy and Competitiveness, through the project ECO2017-82449-P, is gratefully acknowledged.

‡DEcIDE - Interuniversity Doctorate in Economics, Universidad Nacional de Educación a Distancia (UNED). Email: arodrigue2431@alumno.uned.es.

§Departamento de Análisis Económico, Universidad Nacional de Educación a Distancia (UNED), Paseo Senda del Rey 11, 28040 Madrid, Spain. Email: mvorsatz@cee.uned.es. Financial support from the Spanish Ministry of Science and Innovation, through the project PGC2018-096977-B-I00, is gratefully acknowledged.
Abstract

Although no consensus on the issue exists yet, some evidence indicates that people are typically overprecise in their inferences. In particular, subjective confidence intervals are often too narrow when compared with Bayesian ones. This paper uses a quasi-Bayesian theory and lab experiments to explore overprecision when people learn about the empirical frequency $\theta$ of some random event. Motivated by the literature on limited attention, we hypothesize that, when there is a large number of potential values of $\theta$, individuals mentally operate with simplified representations of the objective state space. Their mental models are however sophisticated in that they co-move with the signals observed, focusing on the values of $\theta$ most consistent with the evidence available. As a result, they elaborate accurate point estimates of $\theta$, but also become too confident about them, as they hardly reflect on those values of $\theta$ that would call more into doubt their conclusions. In this line, subjects in our experiment almost exclusively report overly narrow confidence intervals, but also unbiased point estimates of $\theta$ (except when $\theta$ takes extreme values, i.e., close to 0 or 1). Indirect evidence suggests that subjects often consider about 1/5 of the objective state space.

Keywords: beliefs, bias, inference, omission of information, overprecision, probability, simplification.

JEL–Numbers: D01, D80, D83, D91.
1 Introduction

People often have estimates about the frequency or rate of occurrence of repeatable events—finding a good job, suffering a crime, a successful surgery, food poisoning, a tax inspection, winning the lottery, etc. These beliefs take sometimes an exact numerical shape (“the risk of death after heart surgery is 17 percent”), but most often the form of a less precise, although still fairly quantifiable guess (“the crime rate is very low” or “the chances that the president gets re-elected look high”). Whatever its form, economists tend to view these beliefs as one of the triggers for action. Investigating their formation and evolution is therefore an important task.

In this vein, we explore a bias called overprecision, which in the context of this paper is understood as people having overly narrow subjective confidence intervals in inference problems.\(^1\) Like many other cognitive biases, overprecision is important in part because it can lead to behavior that is arguably too risky or simply wrong, at least from a Bayesian perspective. For example, an overprecise firm with scarce data about a rival may be too confident about potential synergies and acquire it for an excessive price (Malmendier and Tate, 2005). Also, overprecision in asset valuation could be a potential explanation for two well-known phenomena: (a) the premium paid in some trading transactions, which might be driven by economic agents with excessive faith in their knowledge about how much the asset is worth, and therefore about how much they should pay for it (see, Daniel, Hirshleifer, and Sabrahmanyam, 2001; Malmendier and Tate, 2005), and (b) the high rate of trading in stock markets (see, Odean 1999).

The existing lab evidence on overprecision is somehow mixed. On one hand, a strand of literature close to the “heuristics and biases” research program pioneered by Daniel Kahneman and Amos Tversky has concluded that people tend to infer too much from small

---

\(^1\)A less precise but common definition is the excessive faith that you know the truth. Overprecision is commonly considered as one of three varieties of overconfidence. The other two are overestimation—thinking that you are better than you are—and overplacement—an exaggeration of the degree to which you think that you are better than others (see for example Svenson, 1981; Moore and Healy, 2008).
samples and report too narrow confidence intervals (see, Tversky and Kahneman, 1974; Alpert and Raiffa, 1982; Lichtenstein et al., 1982; Block and Harper, 1991). For instance, in Alpert and Raiffa (1982) more than 1,000 subjects report the quartiles and median of the posterior distribution of some random variable – e.g., the share of some sample of first-year students who prefer bourbon to scotch. The authors find that the inter-quartile ranges are too tight for a significant majority of subjects, i.e., the actual response to a question falls most often outside the subject’s reported range, thus suggesting overprecision (also see, Russo and Schoemaker, 1992; Soll and Klayman, 2004; Bazerman and Moore, 2013; and the review by Moore et al., 2016). Moore et al. (2016) contend that overprecision is “the most robust and the least understood form of overconfidence” and still “remains in need of a full explanation”. On the other hand, a different strand of the literature on belief revision suggests that agents do not update their beliefs as fully as Bayes’ theorem indicates. In this line, Benjamin (2019) offers an extensive analysis of the literature on bookbag-and-poker-chip experiments and concludes that “the evidence overwhelmingly indicates that people tend to infer too little from signals rather than too much, even from small samples of signals”. In the majority of these experiments, importantly, the set of states includes just two elements and the typical question is how much weight is placed on one state or the other, that is, how strongly an individual updates the mode of her beliefs.

This paper contributes in several ways to the existing debate on overprecision. First, we provide new laboratory evidence of overly narrow confidence intervals, carefully controlling for the knowledge subjects have, a point stressed by Moore et al. (2015) and relatively unattended by prior research approaches. Since we control for the subjects’ priors and the signals they observe, we can verify that overprecision in our experiment is at odds with the Bayesian standard. Second, interval estimations are incentivized in our study, something that has not been common so far. Third, in our experiment the set of states is (much) larger than two. As we will argue later, the size of the state space is a crucial factor in accounting for overprecision. A focus on minimal-size scenarios, as has been usual in the literature on bookbag-and-poker-chip experiments, can be hence misleading, and we believe that the results from those studies should not be directly extrapolated to settings.
with numerous states. Fourth, we propose a simple quasi-Bayesian theory of overprecision based on limited attention that can be tested with the help of our experiment.

The experiment consists of two phases. In the first phase, each subject faces an (individual) urn with 100 balls. Each ball is either blue or red. The subject does not know the actual number of blue balls in her urn, but she is informed that it has been determined by choosing an integer between 1 and 99 (both included) from the uniform distribution. In other words, there are 99 states and the subjects know that all are a priori equally likely. Each subject then observes 20 consecutive random extractions with replacement from the urn and must give a point estimate of the rate or percentage of blue balls \( \theta \) after each pair of draws. In the second phase, the urn is renewed for each subject, that is, \( \theta \) is randomly determined again, and the subject must provide an interval estimation of \( \theta \) after each pair of draws. More precisely, she must indicate two bounds such that \( \theta \) is with 95% confidence within those bounds. Both phases are relevant to test the predictions of our theory.

We find that subjects report intervals that are systematically narrower than the Bayesian ones, thus providing evidence for overprecision. We conjecture that this partly occurs because Bayesian updating becomes cognitively very demanding when there are too many contingencies. In our experiment, there are 99 potential values of \( \theta \), which most likely makes it impossible for most humans to consider them all.\(^2\) The underlying theory that we explore here is that people cope with this type of learning problems as if they operated with a simplified representation of the state space. When making estimations in our experiment, a Bayesian agent’s beliefs are defined on the set of all integers between 1 and 99. A limited agent, in contrast, operates with a subset of that set. To formalize this idea in a simple and tractable manner, we resort to the following as if assumption: the agent assigns zero probability to any omitted rate and renormalizes the objective prior probabilities to the rates actually considered, thus effectively assigning a uniform prior to those considered rates. The idea is further reflected in the notion of mental model, which is the subset of the true state space to which non-zero probabilities are assigned. Importantly, the critical

\(^2\)Abundant evidence suggests that the number of information items – rates, priors, etc. – that the human mind can simultaneously think of is inherently limited; Miller (1956) is a classical reference.
distinction between Bayesian and limited agents refers to the considered rates; otherwise, we assume that both types update their beliefs according to Bayes’ rule.³

Now, how are mental models formed? For parsimony, we assume that the consideration of a rate is heavily influenced by the signals observed. Specifically, consider a subject who has observed that \( k \) out of \( n \) extractions are blue so that the empirical frequency of blue balls is \( f = k/n \). Then we assume that \( f \) is included in the subject’s mental model (in addition, some adjacent rates may be considered as well, depending on how attentive the subject is). This can be conceived as a facet of WYSIATI (What You See Is All There Is), i.e., the idea that humans often and spontaneously construct stories that are as coherent as possible with the information readily available –see Kahneman (2011) and references therein.⁴ For humans, with limited attention, this does not seem a bad cognitive strategy. In fact, mental models become relatively sophisticated in that they co-move with the evidence and focus on the most relevant details of the environment –note that \( f \) is the mode of the posterior distribution and, thus, the likeliest true rate. Agents are limited in that they cannot mentally operate with the whole set of rates, but they are able to adapt their mental models by focusing on the most relevant rates.⁵ We formally prove

³We stress the as-if character of the theory. In real life, people tend to use basic heuristics: in our first phase, for instance, it is very likely that they simply extrapolate \( \theta \) from the sample. What our theory suggests is that these simple rules are (implicitly) derived from simplified representations. In some cases, as in the first phase, these simplifications (and the rules derived from them) do not lead to biases, but in other cases, as in the second phase, systematic biases like overprecision appear.

⁴Our hypothesis on how mental models are formed is sufficient to generate overprecision, but it is surely a simplification of the WYSIATI idea. For instance, if a subject observes a sample where 80% of the balls are blue but at the same time he is asked how likely it is that 10% of the balls are blue, he will certainly consider that 10% contingency. However, a more complex theory of how mental representations are constructed seems unnecessary to account for overprecision, at least in our setting. We discuss further this point in the conclusion.

⁵A parallelism can be made between our theory and the anchoring explanation proposed by Tversky and Kahneman (1974) and Block and Harper (1991). In these studies, people form their beliefs about the confidence intervals by starting with some “best estimate” from which they fail to adjust sufficiently. In our theory, people include in their mental models the mode and some adjacent rates.
that our assumptions imply overinference, showing as well that other assumptions about
how mental models are formed do not necessarily entail it. Intuitively, since limited agents
distribute the probability mass within a set of limited and adjacent rates, they inflate these
posteriors and overprecision follows.

The hypothesis that people operate with simplified mental models fits well with the
results in our experiment. Indeed, we estimate the median mental model and find that it is
almost always a contraction of the objective state space. Omissions or inattention appear
to be quite significant, because a regression analysis indicates that the (unconditional) size
or cardinality of the median mental model is smaller than 19, whereas the objective state
space contains 99 potential values of \( \theta \). We also find supporting evidence for the hypothesis
that subjects’ mental models are far from naive. First, the estimated median mental model
often includes \( f \). Second, since it can be theoretically shown that \( f \) is the best predictor
of \( \theta \) for any mental model that includes \( f \), we can use the data from the first phase of
the experiment about point predictions to implicitly test the hypothesis of whether the
empirical frequency is an element of the mental model. Our data tends to corroborate
this hypothesis: if subjects accumulate some experience, there is no systematic bias in
their point predictions.\(^6\) In fact, the median error is 12.5 balls, which we find relatively
low. Further, this error gets smaller as the first phase proceeds and for those subjects
who score high in the Cognitive Reflection Test (CRT). The latter possibly means that
some deviations are due to simple mistakes and not the consequence of a different logic of
inference than the one we assume. Third, in any interval estimation of \( \theta \), our assumptions
imply that subjects should co-move the interval endpoints with \( f \); in fact, \( f \) should form
part of the reported interval. In this respect, we find that the unconditional likelihood that
the reported intervals include \( f \) is 0.62. While this number is not very high, the likelihood
increases significantly over the course of the experiment and in the CRT score.

\(^6\)In the classical example of the wisdom-of-the-crowds phenomenon, statistician Francis Galton collected
point estimations of the weight of a slaughtered and dressed ox by 800 people at a 1906 country fair in
Plymouth. Galton observed that the median guess was surprisingly accurate, within 1% of the true weight.
Similarly, the median subject in our experiment made rather accurate point estimations.
We would like to stress that our theory is falsifiable. Indeed, it cannot explain all our experimental results. Two contrarian phenomena occur when \( f \) equals 0 or 1, or some close value of these. On one hand, subjects make point estimations of \( \theta \) that deviate systematically and very significantly from the Bayesian estimate. Specifically, they overestimate when \( f \) is close to 0, and underestimate if \( f \) is close to 1. A potential reason is that subjects are averse to give estimates at the extremes of the response scale, a bias sometimes called “floor and ceiling effects” (see, Phillips and Edwards, 1966). On the other hand, subjects express “due doubt” in their interval estimations, that is, they do not deviate much from what a Bayesian analysis recommends. This might be related to the floor and ceiling effects observed in the first phase and signals that over/underprecision is a multifaceted phenomenon, as some other authors (Moore et al. 2015) have already proposed.

Our study contributes to the literature on quasi-Bayesian inference, including Barberis, Shleifer, and Vishny (1998), Rabin and Schrag (1999), Mullainathan (2002), Rabin (2002), Rabin and Vayanos (2010), and Benjamin et al. (2016). Quasi-Bayesian agents misread the world due to cognitive limitations, but given this misreading, they are then assumed to operate as Bayesians in inference and decision making tasks. In contrast to our model, this literature assumes that agents misread or misremember the signals they observe, but do not have an incorrect model of the state space. In Rabin (2002), for instance, an agent observes a sequence of \( i.i.d. \) binary signals, but (wrongly) believes that the process is generated by random draws \( \text{without} \) replacement from an urn. Our paper is more in the spirit of Gennaioli and Shleifer (2010) and Bordalo et al. (2016), who posit that agents omit elements of the state space when evaluating the probability of a hypothesis given some data.\(^7\) Our study is also related to the literature on inattention and focusing. The assumption of sophisticated, non-fixed mental models links our approach with the idea

---

\(^7\)In their models, there are several uncertain dimensions aside from those specified in the hypothesis and the data. Agents simplify the situation by focusing on the “stereotypical” values of these residual dimensions, that is, those values closely associated with the target hypothesis (and the data), but not with the complementary one. While their models can explain several judgment biases, they cannot account for overprecision in our setting, where there are arguably no residual dimensions.
of rational inattention, which started with the seminal paper by Stigler (1961). Finally, the idea that people have reduced representations of the world relates our article to the behavioral literature on inattention –see Gabaix (2019) for a comprehensive review.

We proceed as follows. Section 2 details the experimental design. Then we introduce our theory and derive the key predictions. Section 4 reports the experimental results, and discusses how they fit within the predictions of our theory and of some alternative theories. We conclude in Section 5. Some statistical results, the translated experimental instructions, some Z-Tree decision screens, and the formal proofs of our predictions are relegated to the (online) appendices.

2 Experimental design and procedures

The experiment consists of two different phases, which all subjects successively complete. In the first phase (point estimation), each subject is assigned a virtual urn with $N = 100$ balls. Each ball in the urn is either blue or red, and the urn contains at least one ball of each color. The actual number of blue balls in the urn is unknown to the subject, but she is informed that a priori all integer numbers of blue balls between 1 and 99 are equally likely. Each subject observes then the realization of $n = 20$ consecutive random draws with replacement and must provide after each pair of draws an estimation $\hat{\theta} \in \{0.01, 0.02, \ldots, 0.99\}$ of the true rate of blue balls $\theta$. The elicitation of $\hat{\theta}$ is incentivized. At the end of the session, one of the 10 elicited estimations is randomly selected for each subject, who gets 15 Euros if the absolute error $|e| = |\hat{\theta} - \theta|$ in the randomly selected estimation is zero. Otherwise, the subject gets no monetary compensation for this task.

In the second phase (interval estimation), subjects face an identical urn, but with a new true rate of blue balls $\theta$, and observe again 20 extractions with replacement. The task of the subjects is now to provide, after each pair of draws, two bounds $\hat{\theta}_l, \hat{\theta}_h \in \{0.01, 0.02, \ldots, 0.99\}$ such that they believe that the interval $[\hat{\theta}_l, \hat{\theta}_h]$ is the shortest one containing $\theta$ with 0.95 probability.\footnote{Since the state space is finite in our design, a subject’s posterior distribution may be such that there} Furthermore, subjects are informed that there is a
probability law that can be used to compute these bounds—to be precise, they are referred to Bayes’ law, although the law is not explained. In this vein, the Bayesian interval \([\theta_l, \theta_h]\) is used to incentivize the interval elicitation: after a subject has consecutively completed the 10 interval estimations, one of them is randomly chosen and the subject gets 15 Euros if the absolute difference between the stated lower bound and the Bayesian lower bound \(|e_l| = |\hat{\theta}_l - \theta_l|\) equals 0.03 at most, and zero Euros otherwise. An analogous procedure determines the payment for the upper bound.\(^9\)

The experiment was programmed within the z–Tree toolbox provided by Fischbacher (2007). In total, 120 undergraduates from various disciplines participated in the experiment, which was run at LINEEX in Valencia, Spain. There were two sessions with 60 subjects each. After being seated at a visually isolated computer terminal, each participant received written instructions that described the inference problem. Subjects could read the instructions at their own pace and their questions were answered in private. The instructions clarified that subjects had to make two different types of independent decisions, although the specific instructions for the second phase were only distributed after the completion of the first one. Instructions were also read aloud after subjects read them privately. The understanding of the rules in each part was checked with a computerized control questionnaire (see the online appendix) that all subjects had to answer correctly before the belief elicitation started. The use of calculators was forbidden during the whole experiment, as our focus was on inference without any kind of computing tools.

The decision screens for each phase showed the color of all balls drawn so far in that phase, that is, the history of signal realizations. To make a point estimation, subjects were provided with a slider that contained as grid points all numbers from 0 up to 1 in is no interval with a probability of exactly 0.95. In this case, we asked subjects to report the interval that gets closest to a probability of 0.95, and the narrowest of them if there are several such intervals.

\(^9\)In an alternative design that aims at reducing the perception that the experiment is about a pure math problem, subjects are simply asked to indicate their “personal” 95\% confidence interval. Our approach has the advantage that by incentivizing subjects to report intervals that are as close as possible to the true interval, we are able to draw conclusions regarding overprecision even if subjects are trying to be as Bayesian as possible.
0.01 steps and a triangular pointer that could be moved over the grid. The extreme values 0 and 1, which could not be chosen, corresponded to the extreme beliefs “no blue balls” and “all blue balls” respectively. When a participant moved the triangular pointer over the grid, the corresponding number was displayed on screen in real time. In the decision screens for the second phase, a similar slider was provided, but instead of one value, two values had to be chosen (the lower and the upper bound of the interval). Again, the corresponding numbers of the pointers were displayed in real time. In each phase, subjects were always provided a history with their prior choices. We show in the online appendix two examples of the interfaces used in the experiment, one for each phase. We also remark that the order of the two phases was not counterbalanced across subjects for two main reasons. First, one can argue that interval estimation is a conceptually more complex task than point estimation. Since they are somehow related, however, our ordering of tasks should facilitate comprehension by subjects. Second, a key goal of the interval estimation task is to explore whether overprecision occurs because subjects operate with simplified mental models, and we cannot find any reason why placing this task second could induce any confound in this respect. If anything, we believe that the effect would actually go in the opposite direction: more experience with the problem should lead to relatively more accurate mental models and hence to beliefs closer to the Bayesian ones. In this respect, our design seems to provide a strict test of our hypothesis.

At the end of the experiment, subjects answered a brief questionnaire that gathered information on personal and socio-demographic characteristics, the Cognitive Reflection Test or CRT (Frederick, 2005), and a risk aversion index partly based on Holt and Laury (2002). Afterwards, subjects were privately paid. Each session lasted approximately 75 minutes. Subjects earned on average 12.25 Euros, including a show-up fee of 7 Euros.

3 Theory and predictions

We here introduce our theory of quasi-Bayesian inference with a simplified state space (S3QB), for which we need the following notation. Consider a binary random variable
with outcomes $B$ (blue) and $R$ (red) and suppose that $\theta$, the theoretical probability that the outcome of the binary random variable is $B$, has been determined upfront according to some known continuous random variable distributed on the unit interval. The realization of $\theta$ is not observable. Let $X = (X_1, \ldots, X_n)$ be a sequence of $n$ i.i.d. realizations of the binary random variable. The sequence of realizations $X$ is observable. To derive a limited agent’s posterior distribution of $\theta$ given $X$, we make the following three key assumptions.

**A1:** *Simplified space state.* People make inferences as if their priors about $\theta$ were defined over some (possibly strict) subset $\Theta \subseteq [0, 1]$, the so-called *mental model*, and not over the true state space $[0, 1]$. Intuitively, due to attention limits, agents mentally operate with simplified representations of the state space.

**A2:** *Adaptation.* The mental model is not fixed, but depends on the signals observed. More formally, $\Theta$ depends on $X$, and we write $\Theta(X)$ in order to make this assumption explicit. Intuitively, people simplify complex problems, but these simplifications are shaped by the available evidence.

**A3:** *Bayesian updating.* People update their priors over the considered states using Bayes rule. That is, given $X$, the posteriors over the states in $\Theta(X)$ are derived by applying Bayes’ rule.

To assumptions A1 to A3, we add the following three ancillary hypotheses.

**H1:** *Connectedness:* $\Theta(X)$ is a closed interval of $[0, 1]$.

**H2:** *Empirical frequency:* Suppose that $k$ of the $n$ realizations of the binary random variable are $B$. It is assumed that the empirical frequency $f = k/n$ belongs to the mental model; i.e., $f \in \Theta(X)$.

**H3:** *Renormalization:* The agent’s priors over her mental model $\Theta(X)$ are a renormalization of the true, objective prior distribution. For example, if $\theta$ is actually distributed uniformly on $[0, 1]$, then the agent believes that $\theta$ is uniformly distributed on $\Theta(X).^{10}$

---

10Hypothesis H3 has an important implication in dynamic problems. Whereas Bayesian updating is
To derive formal predictions regarding point estimations, let \( \theta, \bar{\theta} \subseteq (0, 1) \) be the closed interval that, by hypothesis \( H_1 \) above, defines the mental model (we omit the dependence on \( X \)). Further, let \( L(\hat{\theta}, \theta) \) be the loss if the subject announces that the rate of blue balls in the urn is \( \hat{\theta} \) given that the actual rate of blue balls is \( \theta \). Since subjects get paid a fixed amount if the estimator coincides with the true rate of blue balls and nothing otherwise, we have \( L(\theta, \theta) = 0 \), whereas \( L(\hat{\theta}, \theta) = \delta > 0 \) for all \( \hat{\theta} \neq \theta \). We say that \( L \) is a Dirac \( \delta \) loss function. Subjects’ incentives are to minimize their expected loss, which implies that they should indicate the mode of the posterior probability distribution function.\(^{11}\) Formally, the posterior is obtained by truncating the beta probability distribution function when there are \( k + 1 \) successes and \( n - k + 1 \) failures to the interval \( [\theta, \bar{\theta}] \). The mode of the unrestricted beta probability distribution function is \( f \). Since \( f \) forms part of the mental model by \( H_2 \), truncating the unit interval to the mental model does not change the mode. Thus, the best estimator \( \hat{\theta}^* \) of \( \theta \) is \( f \), which means that limited agents estimate as Bayesians in the first phase of our experiment. Consult appendix IV for a more detailed proof.\(^{12}\)

**Prediction 1 (point estimation).** For all \( \Theta(X) \), \( \hat{\theta}^* = f \).

In the second phase, subjects are asked to provide an estimate of the shortest confidence interval that contains 95% of the probability mass. To be more precise, they have to indicate two bounds \( \hat{\theta}_l, \hat{\theta}_h \in [0, 1] \) such that they believe that the interval \( [\hat{\theta}_l, \hat{\theta}_h] \) is the shortest one containing \( \theta \) with 0.95 probability – we highlight that the true Bayesian interval \( [\theta_l, \theta_h] \) may not be defined by the 2.5-percentile and the 97.5-percentile. According to our theory, path-independent, meaning that it does not matter whether an agent updates from the prior all at once or signal by signal, this is not true of our model. In particular, if a subject in our experiment estimates given some sequence \( X \) at \( t = 1 \) and then faces \( X' \) in the next estimation at \( t = 2 \), the posterior distribution in \( t = 1 \) does not become the prior distribution in \( t = 2 \).

\(^{11}\)In contrast, if the loss function for point estimates took the form \( \max\{\delta - |\hat{\theta} - \theta|, 0\} \) or \( \max\{\delta - (\hat{\theta} - \theta)^2, 0\} \), the optimal choice would be the median or the mean of the posterior distribution, respectively.

\(^{12}\)Since the experiment works in a discrete setting, the empirical frequency \( f \) is often different from any of the feasible rates \( \{0.01, 0.02, \ldots, 0.99\} \); in this case, the optimal point estimate is the closest feasible rate to \( f \).
the interval reported should include only rates belonging to $\Theta(X)$ –recall that hypotheses H1 and H2 imply that the mental model contains $f$ and some contiguous rates. Since the truncated beta probability distribution function is in our case always log-concave, with $f$ being the unique maximum, it follows that the posterior probabilities get smaller the further one moves away from $f$. Hence, $f$ must necessarily be in the shortest 95% interval.

**Prediction 2 (interval estimation).** For all $\Theta(X)$, $f \in [\hat{\theta}_l, \hat{\theta}_h]$.

We say that a subject is overprecise whenever the declared interval $[\hat{\theta}_l, \hat{\theta}_h]$ is a subset of the corresponding Bayesian interval $[\theta_l, \theta_h]$, while the subject is underprecise if the opposite occurs. Our theory conjectures that overprecision occurs because agents omit elements of the state space, possibly because of attention limits. To clarify, consider a numerical example in which the set of all (objectively) possible rates $\Theta^0$ is finite and consists of all multiples of 0.1 in the interval $(0,1)$, that is, $\Theta^0 = \{0.1, 0.2, \ldots, 0.8, 0.9\}$. Further, the subject starts with a uniform prior and has observed a sample with $n = 10$ and $k = 5$, which implies that $f = 0.5$. Given this data, the first row in Table 1 indicates the posteriors for the true set $\Theta^0$, while the rows below indicate the posteriors for various subsets $\Theta^1, \Theta^2, \ldots, \Theta^6$ of $\Theta^0$ (mental models). A blank space in row $\Theta^i$ indicates that the rate is omitted in $\Theta^i$.

| Rates   | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 |
|---------|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| $\Theta^0$ | 0.002 | 0.029 | 0.113 | 0.221 | 0.27 | 0.221 | 0.113 | 0.029 | 0.002 |
| $\Theta^1$ | 0.449 | 0.551 |
| $\Theta^2$ | 0.310 | 0.380 | 0.310 |
| $\Theta^3$ | 0.121 | 0.235 | 0.288 | 0.235 | 0.121 |
| $\Theta^4$ | 0.295 | 0.705 |
| $\Theta^5$ | 0.052 | 0.204 | 0.487 | 0.204 | 0.052 |
| $\Theta^6$ | 0.003 | 0.226 | 0.542 | 0.226 | 0.003 |

**Table 1:** (Rounded) posterior distributions for various mental models ($n = 10$, $k = 5$).

A first thing to note is that the posterior of any rate is lower for the true set $\Theta^0$ than
for any $\Theta^i$ that considers such rate. This is natural, because the probability mass in $\Theta^i$ is shared among less rates. However, not all $\Theta^i$ generate overprecision. This requires that the rates considered are close to each other. More precisely, observe that the models $\Theta^1$, $\Theta^2$, and $\Theta^3$ are connected — that is, for all $\theta_i < \theta_j < \theta_k$ such that $\{\theta_i, \theta_k\} \subseteq \Theta$, we also have that $\theta_j \in \Theta$. In these models, the posteriors for any interval centered at $f$ are higher than the Bayesian ones. As a result, the confidence intervals are always a contraction of the Bayesian interval. Models $\Theta^4$ to $\Theta^6$, in contrast, are not connected and some intervals receive lower posteriors than the Bayesian ones. For instance, the Bayesian confidence interval at the 95% confidence level is $\{0.3, 0.4, 0.5, 0.6, 0.7\}$, whereas under $\Theta^5$ it is either $\{0.2, 0.3, 0.5, 0.7\}$ or $\{0.3, 0.5, 0.7, 0.8\}$. Underprecision is thus possible when $\Theta^i$ is not connected. Finally, note that the confidence intervals become narrower when the cardinality of the mental model gets smaller. In the extreme case when $\Theta(X) = f$, the confidence interval contains just one rate.

Formally, it can be shown that given two mental models $\Theta, \Sigma$ such that $\Theta \subseteq \Sigma$, the probability under the truncated beta probability distribution function is for all $\theta \in \Theta$ greater for the mental model $\Theta$ than for the mental model $\Sigma$ (the weak inequality is strict whenever the set inclusion is strict). Let $[\hat{\sigma}_l, \hat{\sigma}_h]$ denote the confidence interval with model $\Sigma$. The set inclusion of the 95% confidence intervals follows directly from the log-concavity of the truncated beta probability distribution function and the assumption that $f$ belongs to $\Theta$. Overprecision follows from setting $\Sigma$ equal to the unit interval.

**Prediction 3 (overprecision).** For all $\Theta(X), \Sigma(X)$ such that $\Theta \subseteq \Sigma$, $[\hat{\theta}_l, \hat{\theta}_h] \subseteq [\hat{\sigma}_l, \hat{\sigma}_h]$.

## 4 Results

### 4.1 Evidence on Prediction 1

Since the color of the balls is inessential in our experiment, we can concentrate in our data analysis throughout on the rate of the color that shows more often (instead of the rate of blue balls). Formally, if $k$ out of $n$ drawn balls are blue, the *top frequency* is defined to be
max\{k/n; (n − k)/n\}. So, if in the sample there are at least as many blue as red balls, a subject’s error for a given point estimate \(\hat{\theta}\) is defined as \(\hat{\theta} − f\). And if there is a majority of red balls, then the error is \(1 − \hat{\theta} − (n − k)/n = k/n − \hat{\theta} = f − \hat{\theta}\). Figure 1, which is divided into four panels, analyzes the median error (the two panels to the left) and the standard deviation of the error (the two panels to the right) in the first phase of the experiment. The top-left panel of Figure 1 depicts the median error as a function of the number of extracted balls —recall that subjects make a point estimation after each pair of draws; i.e., when the sample contains 2, 4, 6, . . ., and 20 balls. Since the error is calculated with respect to the top frequency, which necessarily takes values in the interval \([0.5, 1]\), it is rather natural that the median error is always negative. The bottom-left panel disaggregates the data into three subcases according to the top frequency, that is, when such frequency falls into interval (a) \([0.500, 0.525]\) (balanced sample), (b) \([0.525, 0.975]\), or (c) \([0.975, 1.000]\) (single-colored sample). For phase 1, 157 out of the total of 1200 observations fall into subcase (a), 784 in subcase (b), and 259 in subcase (c).

**Figure 1:** Median error (to the left) and standard deviation of the error (to the right).

The left panels reveal that the point estimations track the top frequency well, except in two cases. First, the top-left panel shows that there is some considerable underestimation
when the number of balls drawn is below 10. After that, the median error is rather small. Second, the bottom-left panel reveals that for almost all top frequencies, the median error is very small regardless of the number of balls drawn. The only exception occurs when the top frequency is above 0.975. In this case, subjects underestimate this relative frequency, and this underestimation continues even when the number of balls drawn increases. For an additional remark, as the bottom-left panel shows, if the relative frequency is very close to 0.5, then there is no bias, independently of the number of balls that are extracted from the urn. Finally, we note that the unconditional error equals 0.125.\textsuperscript{13}

We show in the right panels the standard deviation of the error. The top-right panel refers to the aggregated data, whereas the bottom-right panel considers the same three intervals for the top frequency as before. On the aggregate, the standard deviation decreases from about 0.26 when 2 balls are extracted to about 0.17 when 20 balls are extracted. The disaggregated data uncovers that the tendency depends on the top frequency: while the standard deviation seems to increase a bit as more balls are extracted whenever the top frequency falls into the interval [0.5, 0.525], it slightly decreases in the other two cases.

\textbf{Figure 2:} Density of estimation error for 2 and 20 balls drawn

To further investigate how the distribution of the error changes with the number of

\textsuperscript{13} Table 5 in the appendix details the median error for all possible extraction histories.
extractions, we compare in Figure 2 the density of the estimation error distributions for the cases of $n = 2$ and $n = 20$ extracted balls. We find for $n = 2$ not only that the median error is -0.25 instead of 0.00 but also that the distribution is wider in comparison with $n = 20$. This suggests again that as the experiments goes on, the estimation error and its associated uncertainty (variance) decreases. Next, we estimate two linear mixed models. Let $e_{i,t}$ be the size of the error, that is the absolute value of the error, of subject $i$ in the $t$-th extraction round. The model is

$$
e_{i,t} = \alpha + \beta \cdot X_{i,t} + \gamma \cdot t + \delta \cdot Z_i + u_i + \varepsilon_{i,t},$$

where $\alpha$ is a constant, $X_{i,t}$ are the individual and time specific determinants of the subject’s error like the top frequency, $Z_i$ are the individual characteristics obtained from the ex-post questionnaire, $u_i$ is the individual random effect, and $\varepsilon_{i,t}$ the idiosyncratic error for subject $i$ in round $t$.

Table 2 presents the estimation results. We can see that the error is increasing in the top frequency ($\beta_1$), particularly for extreme values of it ($\beta_2$), which confirms our insights from Figure 1. Also, the error decreases as the experiment goes on ($\gamma$), perhaps due to learning. In the second column, we also consider the variables extracted in our ex-post questionnaire. We find that female subjects perform better than their male peers ($\delta_3$). In fact, the error of female subjects is roughly 5 percentage points lower than that of male subjects. Subjects who are relatively more willing to engage in reflection, as measured by the CRT score, have a smaller error as well ($\delta_2$). Risk aversion, on the other hand, is not correlated with the error ($\delta_1$).

**Result 1.** *Point estimations track the top frequency fairly well unless the evidence points to a single-color urn or the number of extractions is relatively low. The point estimations of female and more reflective subjects are better than those of their peers.*

One possible interpretation of the signs of the estimated coefficients ($\gamma$) and ($\delta_2$) is that the biases observed in Table 2 are in part due to random mistakes, which disappear after subjects become more familiar with the inference problem or that more attentive subjects
commit less. In line with this, the bias observed in Figure 1 and in the first columns of 
Table 5 is systematically negative, *i.e.*, biased towards 1/2, which is natural given that the 
top frequency is higher than 1/2 and the error is random –there are more rates below the 
top frequency than above. If this interpretation were true, therefore, the biases for small 
sample sizes would not fundamentally contradict our theory. When one of the colors is 
very rare, nevertheless, this account appears to be totally wrong. Subjects seem to follow 
a different logic in this contingency, something that our later analysis will tend to confirm.

### 4.2 Evidence on Prediction 2

In the second half of our experiment, subjects have to provide an estimate $[\hat{\theta}_l, \hat{\theta}_h]$ of the 
shortest confidence interval $[\theta_l, \theta_h]$ at the 95% confidence level. According to our theory,
the interval reported should include only rates belonging to $\Theta(X)$, which according to Hypotheses H1 and H2 in Section 3, contains the empirical frequency of blue balls $f$ and some contiguous rates.

![Graph](image)

**Figure 3**: Likelihood that the reported confidence interval contains the empirical frequency

Figure 3 shows for each extraction round the likelihood that the empirical frequency is contained in the reported confidence interval.\(^{14}\) The lower panel disaggregates the data in the same way as in Figure 1. In phase 2, 140 out of the total of 1200 observations fall into the interval $[0.500, 0.525]$, 787 in the interval $(0.525, 0.975)$, and 273 in the interval $(0.975, 1.000]$. The upper panel shows that subjects make better estimations as the experiment goes on. Yet much of this improvement happens after the first round (2 balls), where the sample can be only balanced or single-colored. This is relevant because the lower panel suggests that subjects perform substantially worse in the extreme case when all extracted balls are of the same color. When comparing the two panels, the reader should take also into account that, as the experiment proceeds and the sample size increases, more and more subjects observe frequencies in the interval $(0.525, 0.975)$.

\(^{14}\)Table 6 in the appendix details the likelihood for all possible extraction histories.
To offer further insight, we run two logit estimations with random effects. In particular, if for subject $i$ in round $t$ it is the case that $f \in [\hat{\theta}_l, \hat{\theta}_h]$, then $y_{i,t} = 1$. Otherwise, $y_{i,t} = 0$. We then estimate the equation

$$y_{i,t} = \alpha + \beta \cdot X_{i,t} + \gamma \cdot t + \delta \cdot Z_i + u_i + \varepsilon_{i,t}.$$  

Observe that the determinants are the same as before in Prediction 1. The results, presented in Table 3, confirm our descriptive insights ($\beta_2$ is negative and $\gamma$ is positive). Moreover, similar to what we have already seen in our analysis of Prediction 1, more reflective subjects make better decisions. The subjects’ degree of risk aversion and their gender, on the other hand, do not correlate with the likelihood that the empirical frequency is in the reported confidence interval.

|                      | Coefficient | Standard Error |
|----------------------|-------------|----------------|
| Constant ($\alpha$)  | -0.6107     | (0.6410)       |
|                      | -2.0638     | (2.5137)       |
| Top frequency ($\beta_1$) | 1.3049       | (0.8526)       |
|                      | 1.1760      | (0.8558)       |
| Top frequency equals 1 ($\beta_2$) | -0.9600** | (0.3460) |
|                      | -1.0057**   | (0.0297)       |
| Round ($\gamma$)     | 0.1074***   | (0.0297)       |
|                      | 0.1041***   | (0.0013)       |
| Risk aversion – Holt and Laury ($\delta_1$) | -0.0022   | (0.0112)       |
| CRT Score ($\delta_2$) | 0.3869*     | (0.1805)       |
| Gender – Female ($\delta_3$) | 0.6007      | (0.4126)       |

Table 3: Logit random effects estimation of the likelihood that the reported confidence interval contains the empirical frequency. Standard deviations are in parenthesis. We control for the subject’s age, major, and statistics knowledge (neither of these variables is significant). *** indicates significance at $p = 0.001$, ** at $p = 0.01$, and * at $p = 0.05$ (all two-sided).
We summarize our findings as follows.

**Result 2.** The likelihood that the reported confidence interval contains the empirical frequency increases over the course of the experiment, but it is smaller if all extracted balls are of the same color. More reflective subjects present a higher likelihood than their peers.

Before analyzing our prediction regarding overprecision, we explore in more detail why some subjects fail to set confidence intervals containing the empirical frequency. In particular, our hypothesis is that such “anomaly” occurs less frequently among subjects with relatively precise point estimates in the first phase. That is, one might expect that the quality of the estimations between the two phases are positively correlated. We address this question with the help of a panel data estimation with subject random effects. Formally,

\[ y_{i,t} = \alpha + \beta \cdot e_{i,t} + \gamma \cdot x_{i,t} + \varepsilon_{i,t}, \]

where, as before, \( y_{i,t} \) indicates whether the confidence interval contains the empirical frequency in phase 2 and \( e_{i,t} \) is the size of the error when estimating the empirical frequency in phase 1. We include as a control the top frequency of the second phase, \( x_{i,t} \). Additional controls for the top frequency in the first phase do not turn out to be significant at the 5% level and are therefore not considered in the final specification.

We find that \( \hat{\alpha} = 0.1003, \hat{\beta} = -0.0931, \) and \( \hat{\gamma} = -0.1032 \). Since the corresponding standard deviations are 0.0083 for the intercept, 0.0187 for the error in phase 1, and 0.0092 for the top frequency in phase 2, it is not surprising to see that all parameter estimates are significantly different from zero at \( p = 0.0001 \). Most importantly, a 1 percentage point increase of the error in phase 1 decreases the likelihood that the empirical frequency is contained in the indicated confidence interval by more than 9 percentage points. This indicates that at least part of the anomaly is due to failures in the point estimates.

### 4.3 Evidence on Prediction 3

Figure 4 presents the median lower and upper bounds of the reported intervals in the second phase of the experiment (in dark triangles) together with the median Bayesian bounds
(in light triangles). As before, the upper panel shows the development over the course of the experiment, while the lower panel disaggregates the data according to top frequency.

**Figure 4:** Difference between the estimated and Bayesian bounds of confidence intervals

The figure provides clear evidence on overprecision. On one hand, the reported upper bounds are in all instances lower than the Bayesian ones. On the other hand, as the lower panel shows, the reported lower bounds tend to be higher than the Bayesian ones, the only exceptions being situations when both the number of extracted balls is large and when the top frequency is extremely high, *e.g.* the subjects observe many balls that are all of the same color. Since we rarely find underprecision, there is support for our $H_1$ that mental
models are connected.\footnote{Table 7 in the appendix details the bounds for all extraction histories.}

Further evidence comes from a regression analysis that follows the same structure as before. We focus for the moment on the first column in Table 4, which presents the results of a linear mixed model estimation of the error $|\hat{\theta}_h - \theta_h| + |\hat{\theta}_l - \theta_l|$. We observe several correlations that are in fact also perceivable in Figure 4. First, the error decreases with the top frequency, particularly when all balls drawn are of one color. As we already observed with point estimations, therefore, inference presents some particular characteristics when one event is rare (and thus the other very frequent). Second, the error also decreases over the course of the experiment. This could be a learning effect, but another possibility is simply that subjects do not anticipate well that the Bayesian confidence interval narrows down as the sample size increases. Since their reported intervals are always of similar size (as we discuss later), this statistical effect decreases the error. Third, the error does not depend on individual characteristics like risk aversion, statistical knowledge, or gender. For readers who suspected that men are less cautious or moderate in their assertions than women (in the sense of having narrower confidence intervals given the same evidence), we hence note that, although our data points in that direction, the effect is not significant. In contrast, the error significantly decreases with the CRT score: more reflective subjects have a smaller error.

**Result 3.** The reported confidence intervals are systematically narrower than the Bayesian ones. The error is smaller the more reflective a subject is. It decreases as well in the number of rounds and when the top frequency observed is close to 1.

To provide further evidence on overprecision and test Prediction 3, we compute for each extraction history the model $\Theta(X)$ that rationalizes the reported confidence intervals best. To do so, we proceed as follows. First, given any history of extractions, we calculate the posterior distribution for all possible $\Theta$. Note that we do not restrict them to satisfy our frequency hypothesis $H2$, $\{f\} \subseteq \Theta(X)$, thus allowing for an additional check of this assumption. For any such posterior distribution, we then compute the probability mass
Table 4: Linear mixed model estimation on the determinants of the error, the bounds, and the size of $\Theta^*$. Standard deviations are in parenthesis. We control for the subjects’ age, major, and statistical knowledge (neither of these variables is significant). For the bounds, we first estimate the lower bound and then include the fitted values as an explanatory variable in the estimation of the upper bound. Symbol $^{***}$ indicates significance at $p = 0.001$, $^{**}$ at $p = 0.01$, and $^*$ at $p = 0.05$ (all two-sided).

|                        | Error       | Lower bound | Upper bound | Size        |
|------------------------|-------------|-------------|-------------|-------------|
| Constant ($\alpha$)    | $0.7618^{***}$ | $13.8517$   | $52.0984^{***}$ | $18.6785$   |
|                        | $(0.1639)$  | $(22.1028)$ | $(11.0949)$  | $(20.3410)$ |
| Top frequency ($\beta_1$) | $-0.2596^{***}$ | $66.5051^{***}$ | $65.7911^{***}$ | $-4.3340$   |
|                        | $(0.0455)$  | $(6.4180)$  | $(5.0070)$   | $(5.9848)$  |
| Top frequency equals 1 ($\beta_2$) | $-0.0158^{***}$ | $-18.6342^{***}$ | $-10.7149^{***}$ | $8.9231^{***}$ |
|                        | $(0.0172)$  | $(2.4385)$  | $(1.6808)$   | $(2.2747)$  |
| Round ($\gamma$)       | $-0.0219^{***}$ | $0.2445$    | $0.4283^{***}$ | $0.1696$    |
|                        | $(0.0014)$  | $(0.1987)$  | $(0.1119)$   | $(0.1854)$  |
| Risk aversion – Holt and Laury ($\delta_1$) | $0.0001$ | $-0.04311$ | $-0.0436$ | $0.0039$ |
|                        | $(0.0007)$  | $(0.1036)$  | $(0.0549)$   | $(0.0960)$  |
| CRT score ($\delta_2$) | $-0.0274^{**}$ | $-0.3142$  | $1.2840$    | $1.6097$    |
|                        | $(0.0122)$  | $(1.6500)$  | $(0.8850)$   | $(1.5294)$  |
| Gender – Female ($\delta_3$) | $-0.0364$ | $8.2756^*$ | $4.3465^*$ | $-4.4258$ |
|                        | $(0.0287)$  | $(3.8774)$  | $(2.1811)$   | $(3.5939)$  |

Observations 1200 1200 1200 1200

The only exceptions appear when at least 12 balls have been extracted.
from the urn and all are of the same color. The median state space is then very close to the Bayesian one. We also note that, for almost any number of balls extracted, the median mental model satisfies fairly well the frequency condition. The only exceptions occur when the top frequency is 1 or almost 1. Figure 5 also shows that, for this case, the fraction of subjects that report an estimated interval that contains the empirical frequency is much lower (up to three times lower) than in the other top frequency intervals. On the other hand, when the top frequency is close to 0.5, the frequency condition is satisfied by a much higher fraction of subjects.

**Result 4.** The optimal mental models $\Theta^*$ are almost always substantially smaller than the Bayesian one and of a constant size. The only exceptions appear when the top frequency is close to 1.

increase the upper bound and gender effects for both bounds (both bounds are larger for females).
4.4 Potential causes of overprecision: a discussion

Our results fit the three predictions of Section 3 fairly well. Although one could conjecture other explanations for overprecision, we find them less convincing. To start, one might hypothesize that (some) subjects do not have uniform priors over the considered rates.\textsuperscript{17} In particular, it could be suggested that subjects assign a large prior to $\theta = 0.5$ and a lower, uniform prior to the remaining rates. We first note in this respect that our experimental instructions explicitly indicated that all rates are equally likely; hence we tend to doubt that subjects misinterpreted this statement in a systematic, non-random way. Regardless, such a model can account for two facts: (a) the median error in the point estimations is zero when $f = 0.5$ and (b) the systematic underestimation of $\theta$ that we observe in the first phase of our experiment when the sample size is small; see Table 5 in the Appendix. As we have noted, however, our model can also account for those phenomena if we assume that people commit initially some random errors. Our interpretation is also consistent with the fact that the errors committed are often of a very similar (and reduced) size, independently of the actual $f$, which is at odds with this alternative theory. Surely, the median error is rather high when $f$ is very close to 1, but it is always so, \textit{i.e.}, even when the sample size is large. A Bayesian with wrong, non-uniform priors should converge to the Bayesian prediction with accurate priors as the evidence accumulates.

The hypothesis that subjects assign (substantially) more probability to $\theta = 0.5$ also fails to explain overprecision in the second phase, at least when $f$ is sufficiently different from 0.5. Suppose for instance that $f = 0.75$. If the sample size was sufficiently large, the mode of the posterior distribution would be 0.75, but the posterior for that rate would be lower than the Bayesian one (since the posterior of 0.5 is higher). As a result, the interval around 0.75 with probability mass of 0.95 should be larger than the Bayesian one. Therefore, people with wrong, non-uniform priors seem to spread their posteriors “too much” when the evidence is at odds with their priors, thus leading to underprecision. As

\textsuperscript{17}More generally, if subjects rank the considered rates by their prior probability, their ordering is not the same as that of a fully rational agent.
we have seen, however, the median subject is almost never underprecise. Although we have obviously not checked all potential specifications of the non-uniform priors explanation, the relatively minor errors observed in the subjects’ point estimations, together with the systematic overprecision, suggest that subjects do not wrongly assign a (substantially) higher prior to some rates.

There are other theories of biased inference but, again, our evidence seems hardly explainable by them. Motivated reasoning (e.g., Benabou and Tirole, 2016) states that people have biased beliefs due to their preferences. However, we designed our experiment on purpose so that emotions arguably played a very minor role (if any): subjects are unlikely to have any strong preference on a particular state of the world. Similarly, the availability heuristic of Kahneman and Tversky (1973) should play little role, because subjects are always given feedback of the previous extractions and do not need to memorize any. Social learning seems also irrelevant because subjects did not know others’ choices. Base-rate neglect as in Grether (1980) is neutralized because priors are uniform. It also seems unlikely to us that subjects did not understand that extractions were made with replacement, as our control questions explicitly checked this point. Finally, with respect to recent theories of focusing and relative thinking like Bordalo et al. (2012) or Köszegi and Szeidl (2013), where actors over-weight salient attributes of the options available, we note first that these theories share insights with our explanation. Indeed, we implicitly assume that agents over-weight some elements of the state space, under-weighting or directly omitting others. However, these theories deal with choice problems of a substantially different nature than our inference problems. In particular, their notion of salience of an attribute, based on how it compares across options, does not seem directly applicable to our setting.

\[\text{18}\] This is something relevant in the model by Rabin (2002). A full test of this model is however out of the scope of this paper
5 Concluding discussion

This paper makes three key contributions. The first one is an analysis of belief formation or inference when agents simplify the state space due to limited cognition. The theoretical model is based on three fundamental assumptions. First, limited agents mentally operate with a reduced state space in that they omit some (objective) elements and focus on others. Second, these simplified mental representations are nevertheless flexible, possibly co-moving with the data – e.g., they might include the most likely rates, given the available evidence, as in our frequency hypothesis. Third, agents have prior beliefs over the rates included in their models, which they update by means of Bayes’ law if they receive new evidence. The theory, although extremely parsimonious, makes testable predictions.

Second, overprecision, which takes the form of overly-narrow confidence intervals and inflated posteriors, is a consequence in this theory. But our theory can account for other biases as well. One is the fallacy that Tversky and Kahneman (1971) denominate the “law of small numbers”, which can be defined as the tendency to exaggerate the similarities between the empirical distribution of a random sample and the population distribution, particularly for small samples. In our theory, this bias can be represented for instance by the difference in the posterior probability that the actual rate \( \theta \) equals the empirical frequency \( f \). This difference is predicted to be positive and decreasing (due to the law of large numbers) in the sample size \( n \). Further, the bias increases as the cardinality of the mental model decreases. If the mental model only contains \( f \), for instance, the limited agent will be positive that the empirical frequency \( f \) equals the actual rate \( \theta \) regardless of the size of the sample. If the model includes more rates, in turn, she may not believe for sure that \( \theta = f \), but the difference may be large if the mental model includes only a few rates, as the probability mass is shared among few rates (Table 1 illustrates this point).

An additional corollary from our theory is that limited agents will misperceive the phenomenon of “regression to the mean”. Since a short sequence of identical realizations of the signal is misinterpreted as indicating that such realization has very likely a large rate, a limited agent will be surprised if further observations deviate from the observed...
pattern. Tversky and Kahneman (1974) tell the story of some flight-training instructors who observed that performance usually deteriorated after praising pilots for smooth landings, whereas it improved after criticizing them for poor landings. Based possibly on erroneous statistical reasoning, the instructors applied carrot-and-stick incentives that underestimated the role that mere chance might play in the success of a pilots’ landings, and overestimated the effect of approval and disapproval.

As a third contribution, we offer new experimental data largely in line with our model. The experiment here reported allows to control for aspects that are from our point of view essential, like the complexity of the objective state space, the evidence observed, the loss function, or the subject’s priors. To our knowledge, we are the first to elicit in an incentivized manner both point and interval estimations, exploring how simplified mental representations affect them. Except for extreme frequencies, we observe no systematic biases in point estimations after some rounds. Interval estimations, in contrast, are often contractions of the Bayesian ones. In exploring the mechanism behind this overprecision, we offer evidence consistent with our hypothesis that people simplify the state space, although in a sophisticated manner, \( i.e. \), changing the model with the available evidence.\(^{19}\) In effect, people tend to co-move the stated intervals with the empirical frequency, particularly as the experiment proceeds.

Our results contrast somehow with those coming from studies of belief updating where individuals are asked to report univariate posteriors after receiving new information – see Benjamin et al. (2016) and Benjamin (2019) for comprehensive reviews. In most of these experiments, there are only two possible values of \( \theta \), the priors are uniform, and subjects’ responses are not incentivized. Subjects observe a sample and their posterior beliefs are elicited afterwards. Interestingly, most of these studies find that people assign

\(^{19}\)In this respect, we note that Peterson and Phillips (1966) report no biases (or some slight underprecision) in the 33% confidence intervals. Yet this is somehow anticipated by our theory, which predicts less overprecision (possibly none at all) when the confidence level is low. Since the Bayesian confidence interval is already quite tight and the subjects’ mental models include the most likely rates, the Bayesian and the reported intervals may have the same length.
a probability to the mode that is lower than the Bayesian posterior. Since people do not update their beliefs as fully as Bayes’ theorem suggests, a natural conclusion is that agents do not overinfer generally, but in fact underinfer. As we have explained in Section 4.3, in contrast, underprecision is not a widespread phenomenon in our experiment. While not the goal of our paper, we can at least conjecture a number of reasons for this difference with our results. In most prior experiments, first, the number of potential values of $\theta$ is just two. In other words, the state space is the simplest possible. Further, subjects in most studies have to indicate their posterior of each rate. Under these two conditions, it seems extremely implausible that subjects operate using simplified mental models, as there is not much to simplify and moreover subjects are asked to consider all rates.\footnote{In our design, we do not elicit the subjects’ posterior probabilities because it would be extremely time-consuming. By forcing subjects to make computations for all rates, however, we find likely that overprecision would be reduced.} \footnote{Nonetheless Grether (1980) finds some degree of overprecision in inference experiments in which $\theta$ takes any of two values and priors are not always uniform. The same is true in Camerer (1987), where subjects must gamble based on their beliefs about the actual value. We finally stress that many of the experiments in this literature elicit probabilities and not point or interval estimations. Maybe the response mode affects the prevalence of underinference, a point to be considered in further research.} If subjects do not simplify the state space, however, our account coincides with the standard Bayesian one. That is, these prior studies might have found no overprecision because their designs were simpler compared to ours.\footnote{Nonetheless Grether (1980) finds some degree of overprecision in inference experiments in which $\theta$ takes any of two values and priors are not always uniform. The same is true in Camerer (1987), where subjects must gamble based on their beliefs about the actual value. We finally stress that many of the experiments in this literature elicit probabilities and not point or interval estimations. Maybe the response mode affects the prevalence of underinference, a point to be considered in further research.} This of course cannot explain why subjects in those studies happen to underinfer, but possibly different factors play a role here; Benjamin et al. (2016) discuss some of them. We note finally that two results from this literature seem in line with our findings. First, we find that overprecision decreases when the sample size grows (recall Figure 4), whereas previous studies find increasing underprecision. Second, prior research suggests greater underprecision when the two potential values of $\theta$ are further apart from 0.5. This seems again consistent with our data, as we observe that overprecision is significantly attenuated only if the top frequency is close to 1, that is, when one of the events seems a rare one – e.g., when all balls observed are blue, a subject in our experiment guesses a rate of 0.8 instead of 1.
In further research, we plan to incorporate into our model a more complex account of how mental models are formed, explore then additional insights on how simplification affects inference about event frequencies, and test our predictions in the lab and, if possible, also in the field. In addition to the signals observed, for example, it seems obvious that the context has a great effect in shaping our mental representations. This can for instance shed some light on the psychology of tail events. In a tentative summary of the available evidence, Barberis (2013, p. 611) reckons that “when asked to estimate the probability of a tail event, people tend to overestimate this probability”\textsuperscript{22}. On the other hand, Barberis (2013) also notes that people sometimes underestimate the likelihood of rare events. Our idea that people infer based on simplified representations of the state space might explain part of this paradox. When experimental subjects are asked the probability of some (rare) event, naturally that event becomes part of their own mental model. Since that model is limited, however, this probability is overestimated. When people are not directly asked about the rare events or somehow made aware of them, in contrast, we tend to agree with Taleb’s (2007) Black Swan theory in that people are often blind to them; they are omitted and hence their probability is largely underestimated. Hopefully these avenues of future exploration can reveal some additional insights into the mechanisms by which humans form beliefs.

\textsuperscript{22}Incidentally, this is related to the overweighting of tail events in the probability weighting function of Tversky and Kahneman’s (1992) cumulative prospect theory model, or in its more general form as proposed by Wakker (2010).
Bibliography

Alpert, M. and H. Raiffa (1982). A progress report on the training of probability assessors. In D. Kahneman, P. Slovic and A. Tversky (eds.): *Judgment under uncertainty: heuristics and biases*, 294–305, Cambridge University Press.

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

Barberis, N. (2013). The psychology of tail events: progress and challenges. *American Economic Review: Papers and Proceedings* 103(3): 611–616.

Bazerman, M. and D.A. Moore (2013). *Judgment in managerial decision making* (8th ed.). New York: Wiley.

Bénabou, R. and J. Tirole (2016). Mindful economics: the production, consumption, and value of beliefs. *Journal of Economic Perspectives* 30(3): 141–64.

Benjamin, D., M. Rabin, and C. Raymond (2016). A model of nonbelief in the law of large numbers. *Journal of the European Economic Association* 14: 515–544.

Benjamin, Daniel J. (2019). Errors in Probabilistic Reasoning and Judgment Biases. *Handbook of Behavioral Economics* (edited by Doug Bernheim, Stefano DellaVigna, and David Laibson). Elsevier Press.

Block, R.A., and D. Harper (1991). Overconfidence in estimation: Testing the anchoring-and-adjustment hypothesis. *Organizational Behavior and Human Decision Processes*, 49(2): 188–207.

Bordalo, P.K., N. Coffman, N. Gennaioli, and A. Shleifer (2016). Stereotypes. *The Quarterly Journal of Economics* 131(4): 1753–1794.

Bordalo, P., N. Gennaioli, and A. Shleifer (2012). Salience theory of choice under risk. *The Quarterly Journal of Economics* 127(3): 1243–1285.

Camerer, C.F. (1987). Do biases in probability judgments matter in markets? Experimental evidence. *American Economic Review* 77: 981–997.
Daniel, Kent D., D. Hirshleifer, and A. Subrahmanyam (2001). Overconfidence, Arbitrage, and Equilibrium Asset Pricing. *Journal of Finance* 56(3): 921-965.

Fischbacher, U. (2007). z-Tree: Zurich toolbox for ready-made economic experiments. *Experimental Economics* 10(2): 171–178.

Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives* 19: 25–42.

Gabaix, X. (2019). Behavioral inattention, in *Handbook of Behavioral Economics*, Vol. 2, eds. D. Bernheim, S. DellaVigna, and D. Laibson, Elsevier.

Gennaioli, N. and A. Shleifer (2010). What comes to mind. *The Quarterly Journal of Economics* 125(4): 1399–1433.

Grether, D.M. (1980). Bayes’ rule as a descriptive model: the representativeness heuristic. *The Quarterly Journal of Economics* 95: 537-557.

Grether, D.M. (1992). Testing Bayes’ rule and the representativeness heuristic. *Journal of Economic Behavior and Organization* 17: 31–57.

Holt, C.A. and K. Laury (2002). Risk aversion and incentive effects. *American Economic Review* 92: 1644–1655.

Juslin, P., A. Winman, and P. Hansson (2007). The naïve intuitive statistician: a naïve sampling model of intuitive confidence intervals. *Psychological Review* 114(3): 678–703.

Kahneman, D. (2011). Thinking, fast and slow. Farrar, Straus and Giroux.

Köszegi, B. and A. Szeidl (2013). A Model of focusing in economic choice. *The Quarterly Journal of Economics* 128(1): 53–104.

Lichtenstein, S., B. Fischhoff, and L.D. Phillips (1982). Calibration of probabilities: the state of the art to 1980. In D. Kahneman, P. Slovic and A. Tversky (Eds.): *Judgment Under Uncertainty: Heuristics and Biases*, Cambridge University Press, 306–334.
Malmendier, U. and G. Tate (2005). CEO Overconfidence and Corporate Investment. 
*Journal of Finance* 60(6): 2661–2700.

Miller, G.A. (1956). The magical number seven, plus or minus two: some limits on our 
capacity for processing information. *Psychological Review* 63: 81–97.

Moore, D. A. and P. J. Healy (2008). The trouble with overconfidence. *Psychological 
Review* 115(2): 502–517.

Moore, D.A., A. Carter, and J. Yang (2015). Wide of the mark: evidence on the unider-
lying causes of overprecision in judgment. *Organizational Behavior and Human Decision 
Processes* 131: 110–120.

Moore, D.A., E. R. Tenney, and U. Haran (2016). Overprecision in judgment. In G. Keren 
and G. Wu (eds.): *Handbook of Judgment and Decision Making*, 182–212, New York: Wiley.

Mullainathan, S. (2002). A memory-based model of bounded rationality. *The Quarterly 
Journal of Economics* 117(3): 735–774.

Odean, T. (1999). Do investors trade too much? *American Economic Review* 89(5): 1279– 
1298.

Peterson, C.R. and L.D. Phillips (1966). Revision of continuous subjective probability 
distributions. *IEEE Transactions on Human Factors in Electronics* HFE-7(1): 19–22.

Phillips, L.D., Edwards, W., 1966. Conservatism in a Simple Probability Inference Task. 
*Journal of Experimental Psychology*, 72 (3), 346-354.

Rabin, M. and J.L. Schrag (1999). First impressions matter: A model of confirmatory 
bias. *The Quarterly Journal of Economics* 114(1): 37–82.

Rabin, M. (2002). Inference by believers in the law of small numbers. *The Quarterly 
Journal of Economics* 117: 775–816.

Rabin, M. and D. Vayanos (2010). The gambler’s and hot-hand fallacies: theory and ap-
plications. *The Review of Economic Studies* 77(2): 730–778.
Russo, J. and P. Schoemaker (1992). Managing overconfidence. *Sloan Management Review* 33(2): 7–17.

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

Stigler, G.J. (1961). The economics of information. *Journal of Political Economy* 69(3): 213–225.

Svenson, O. (1981). Are we all less risky and more skillful than our fellow drivers? *Acta Psychologica* 47(2): 143–148.

Taleb, N.N. (2007). *The black swan: the impact of the highly improbable*. New York: Random House.

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

Tversky, A. and D. Kahneman (1973). Availability: a heuristic for judging frequency and probability. *Cognitive Psychology* 5: 207–232.

Tversky, A. and D. Kahneman (1974). Judgement under uncertainty: heuristics and biases. *Science* 185: 1124–1131.

Tversky, A. and D. Kahneman (1992). Advances in prospect theory: cumulative representation of uncertainty. *Journal of Risk and Uncertainty* 5(4): 297–323.

Wakker, P. (2010). *Prospect theory for risk and ambiguity*. New York: Cambridge University Press.
Online appendix I: fully disaggregated data

| Top frequency | 2  | 4  | 6  | 8  | 10 | 12 | 14 | 16 | 18 | 20 |
|---------------|----|----|----|----|----|----|----|----|----|----|
| (0.975,1.000)| -0.35 | -0.30 | -0.26 | -0.20 | -0.17 | -0.15 | -0.18 | -0.13 | -0.23 | -0.15 |
| (0.925,0.975)| -0.06* | -0.05* | -0.07* | -0.03 | -0.02 | -0.05* | 0.00 | 0.00 | 0.00 | 0.00 |
| (0.875,0.925)| -0.07* | -0.05* | -0.03* | -0.03 | -0.02 | -0.05* | 0.00 | 0.01 | 0.02 | 0.00 |
| (0.825,0.875)| -0.04* | -0.07* | 0.00 | 0.00 | 0.00 | 0.00 | 0.01 | 0.01 | 0.03 | 0.01 |
| (0.775,0.825)| 0.00 | 0.02 | -0.01 | 0.03 | -0.01 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| (0.725,0.775)| -0.05* | -0.07* | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| (0.675,0.725)| -0.03* | -0.01 | 0.01 | 0.01 | 0.03 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| (0.625,0.675)| 0.01 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| (0.575,0.625)| 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| (0.525,0.575)| 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| [0.500,0.525]| 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |

Table 5: Median difference between the actual estimation $\hat{\theta}$ and the optimal one $\hat{\theta}^* = f$. The symbol $^*$ indicates situations when the null hypothesis that the median difference is zero is rejected at the 95% confidence interval (two-sided Wilcoxon signed-rank test).

| Top frequency | 2  | 4  | 6  | 8  | 10 | 12 | 14 | 16 | 18 | 20 |
|---------------|----|----|----|----|----|----|----|----|----|----|
| (0.975,1.000)| 0.16 | 0.21 | 0.14 | 0.26 | 0.35 | 0.21 | 0.25 | 0.27 | 0.27 | 0.27 |
| (0.925,0.975)| 0.67 | 0.69 | 0.56 | 0.56 | 0.56 | 0.56 | 0.56 | 0.56 | 0.56 | 0.56 |
| (0.875,0.925)| 0.74 | 0.68 | 0.65 | 0.65 | 0.65 | 0.65 | 0.65 | 0.65 | 0.65 | 0.65 |
| (0.825,0.875)| 0.69 | 0.50 | 0.42 | 0.80 | 0.85 | 0.85 | 0.85 | 0.85 | 0.85 | 0.85 |
| (0.775,0.825)| 0.69 | 0.69 | 0.69 | 0.69 | 0.69 | 0.69 | 0.69 | 0.69 | 0.69 | 0.69 |
| (0.725,0.775)| 0.76 | 0.78 | 0.65 | 0.63 | 0.71 | 0.71 | 0.71 | 0.71 | 0.71 | 0.71 |
| (0.675,0.725)| 0.85 | 0.68 | 0.81 | 0.57 | 0.72 | 0.72 | 0.72 | 0.72 | 0.72 | 0.72 |
| (0.625,0.675)| 0.79 | 0.58 | 0.81 | 0.61 | 0.69 | 0.69 | 0.69 | 0.69 | 0.69 | 0.69 |
| (0.575,0.625)| 0.63 | 0.95 | 0.95 | 0.73 | 0.67 | 0.92 | 0.92 | 0.92 | 0.92 | 0.92 |
| (0.525,0.575)| 0.77 | 0.80 | 1.00 | 0.70 | 0.70 | 0.70 | 0.70 | 0.70 | 0.70 | 0.70 |
| [0.500,0.525]| 0.90 | 0.88 | 0.92 | 0.82 | 0.84 | 0.67 | 1.00 | 1.00 | 0.71 | 0.83 |

Table 6: Likelihood that the reported confidence interval contains the empirical frequency.
| Top Frequency       | 2  | 4  | 6  | 8  | 10 | 12 | 14 | 16 | 18 | 20 |
|--------------------|----|----|----|----|----|----|----|----|----|----|
| (0.975, 1.000)     | -0.19 | -0.09 | -0.09 | -0.07 | -0.03 | -0.05 | -0.03 | -0.03 | -0.03 | -0.03 |
| (0.925, 0.975)     | -0.04 | -0.03 | -0.04 | -0.04 | -0.04 | -0.04 | -0.04 | -0.04 | -0.04 | -0.04 |
| (0.875, 0.925)     | -0.09 | -0.04 | -0.06 | -0.07 | -0.08 | -0.05 | -0.08 | -0.08 | -0.08 | -0.08 |
| (0.825, 0.875)     | -0.10 | -0.13 | -0.12 | -0.04 | -0.04 | -0.07 | -0.07 | -0.07 | -0.07 | -0.07 |
| (0.775, 0.825)     | -0.11 | -0.06 | -0.07 | -0.07 | -0.07 | -0.07 | -0.07 | -0.07 | -0.07 | -0.07 |
| (0.725, 0.775)     | -0.19 | -0.13 | -0.06 | -0.09 | -0.08 | -0.08 | -0.08 | -0.08 | -0.08 | -0.08 |
| (0.675, 0.725)     | -0.14 | -0.08 | -0.07 | -0.08 | -0.08 | -0.08 | -0.08 | -0.08 | -0.08 | -0.08 |
| (0.625, 0.675)     | -0.17 | -0.12 | -0.06 | -0.07 | -0.13 | -0.13 | -0.13 | -0.13 | -0.13 | -0.13 |
| (0.575, 0.625)     | -0.23 | -0.18 | -0.15 | -0.12 | -0.16 | -0.08 | -0.16 | -0.16 | -0.16 | -0.16 |
| (0.525, 0.575)     | -0.23 | -0.18 | -0.15 | -0.12 | -0.16 | -0.08 | -0.16 | -0.16 | -0.16 | -0.16 |
| [0.500, 0.525]     | -0.33 | -0.24 | -0.25 | -0.19 | -0.14 | -0.16 | -0.13 | -0.13 | -0.13 | -0.13 |

Table 7: On the top (at the bottom): Difference between the median upper (lower) bound of the reported and the Bayesian confidence intervals, for each history of extractions.
| Top frequency | 2  | 4  | 6  | 8  | 10 | 12 | 14 | 16 | 18 | 20 |
|---------------|----|----|----|----|----|----|----|----|----|----|
| (0.975, 1.000)| [47,81]| [56,90]| [64,91]| [67,92]| [69,96]| [35,95]| [1,96]| [1,96]| [2,96]| [1,96]| |
| (0.925, 0.975)|           | [81,95]| [83,96]| [84,95]| [84,95]|           |           |           |           |           | |
| (0.875, 0.925)|           |           | [78,95]| [78,94]| [71,91]| [68,90]| [64,94]|           |           |           | |
| (0.825, 0.875)|           |           | [64,89]| [64,85]| [77,92]| [73,89]|           |           |           |           | |
| (0.775, 0.825)|           |           |           | [62,85]| [74,87]| [65,87]| [68,87]|           |           |           | |
| (0.725, 0.775)|           |           |           |           | [62,82]| [67,88]| [63,82]| [61,83]|           |           | |
| (0.675, 0.725)|           |           |           |           |           | [58,77]| [59,82]| [59,82]| [56,81]| [64,81]| |
| (0.625, 0.675)|           |           |           |           |           |           | [59,76]| [58,79]| [62,79]| [53,69]| |
| (0.575, 0.625)|           |           |           |           |           |           |           | [46,67]| [49,72]| [44,64]| [55,72]| |
| (0.525, 0.575)|           |           |           |           |           |           |           |           | [50,64]| [46,60]| [52,64]| [45,62]| |
| [0.500, 0.525]|           |           |           |           |           |           |           |           |           | [42,57]| [42,61]| |

Table 8: Median $\Theta^*$, expressed as intervals where the lower (upper) bound is the lowest (highest) number of blue balls considered by the median agent.

Online appendix II: experimental instructions and decision screens

Welcome

Thank you for participating in this experiment. At the end of the experiment, you will be paid some money. The precise amount will depend on chance and your decisions. All decisions are anonymous, that is, the other participants will not be informed about your decisions, and you will not receive any information about the other participants’ decisions.

You will make decisions with the help of your computer. Read the on-screen instructions carefully before making any decision. There is no hurry to decide. As usual in Experimental Economics, all information provided in the instructions is true, and therefore there is no deceit in them.

Please, do not talk to any other participant. This point is fundamental for our research. If you do not follow this rule, we will have to exclude you from the experiment without any monetary compensation. If you have questions, raise your hand and we will assist you.
Note that the use of calculators is not permitted throughout the whole experiment. Please, switch off your cell phone.

**Description of the experiment**

During the experiment you must perform two types of estimations (type 1 and 2). We first describe the type 1 estimations.

**Type 1 estimations**

There is a “virtual urn” with 100 blue and red balls. You do not know the actual number of blue balls and therefore neither the actual number of red balls. You only know that the number of blue balls, called $Z$ henceforth, has been determined by the computer randomly selecting with equal probability an integer between 1 and 99, both included. **That is, the probability that there are, say, 30 blue balls in the urn is 1/99. The probability that there are 78 blue balls in the urn is the same, 1/99, and so on for any integer from 1 to 99.**

From this urn with $Z$ blue balls, the computer will randomly extract, one by one, a total of 20 balls. You will sequentially observe the extractions, which will be made with replacement. In other words, each ball drawn is introduced again in the urn and can hence be drawn out again in later extractions.

After each pair of extractions, you must perform an estimation of type 1, *i.e.*, you must estimate $Z$ taking into account all the extractions observed so far. That is, you must indicate the number of blue balls that you believe the urn contains. You can indicate any integer between 1 and 99, both included. Note that you will perform a total of 10 type 1 estimations, since 20 balls are extracted in total and you estimate after each pair of extractions.

Once you have made these 10 type 1 estimations, we will describe to you the type 2 Estimations and you will pass on to perform them. Afterwards you will complete a short and anonymous questionnaire. Finally, you will be paid privately in cash. You will receive
7 Euros for participating in this experiment as well as an additional payoff that will depend on the accuracy of your type 1 and type 2 estimations.

To compute your payoff for the type 1 estimations, the computer will randomly select one of them at the end of the experiment –let us call it $\hat{Z}$. You will earn 15 Euros if your estimation $\hat{Z}$ coincides exactly with the actual value $Z$, and zero Euros otherwise. Your payoff for the type 2 estimations will be explained afterwards, but it will not depend on the type 1 estimations that you make.

In summary, the experiment has several stages, in chronological order: (A) You observe random extractions with replacement from the urn and make a type 1 estimation after each pair of extractions, (B) the type 2 estimations are then introduced and you make your choices, (C) you answer a post-experimental questionnaire, (D) the monetary payoffs are computed and you are privately paid.

If you have any questions, raise your hand and we will attend you.

Type 2 estimations

There is a second virtual urn with 100 blue and red balls. The computer has determined the number $Z$ of blue balls again by selecting with equal probability an integer from 1 to 99. Observe therefore that the new $Z$ is very likely to be different than the one you faced in the type 1 estimations. As before, you do not know $Z$ but observe 20 consecutive random extractions, made with replacement from the urn.

You will make a type 2 estimation after each pair of extractions. The idea is to indicate two integers between 1 and 99, both included, so that you are “almost sure” that the actual number $Z$ of blue balls is in the interval formed by those two integers. More precisely, you must believe that the probability that $Z$ is in that interval is 0.95. In other words, the probability that $Z$ is not in the interval will be small, i.e., 0.05. For a hypothetical example, if someone indicates the interval $[19, 43]$, it means that, based on the extractions observed, he/she is 95% sure that $Z$ is between 19 and 43, both included. You will perform a total of 10 of these estimations, since 20 balls are extracted in total and you estimate
after each pair of extractions.

Note: If you think for any type 2 estimation that no interval formed by integers has a level of confidence of exactly 95%, please indicate the interval that gives you a confidence closest to 95%. If on the contrary you think that there are several possible intervals, indicate the shortest one of them.

There is a statistically accepted method to compute the above mentioned interval depending on prior extractions, called Bayes’ rule. This is relevant because the computer will randomly select at the end of the session one of your type 2 estimations and your payoff for these estimations will depend on how similar that interval is to the corresponding statistical interval (computed by the computer according to Bayes’ rule). For the goals of our experiment, it is not necessary that you know or can apply Bayes’ rule. You must simply indicate at any moment the type 2 estimation that you consider best given the extractions observed so far. If this estimation is sufficiently good because it approximates the statistically accepted one, you will earn money.

Description of the payment: Let us call \([\hat{Z}_1, \hat{Z}_2]\) your type 2 estimation randomly selected by the computer, where \(\hat{Z}_1\) and \(\hat{Z}_2\) are two integers between 1 and 99, both included. Let \([Z_1, Z_2]\) be the corresponding statistical interval, calculated by the computer. You will earn 15 Euros if the absolute difference between \(\hat{Z}_1\) and \(Z_1\) is smaller or equal to 3, and zero Euros if it is larger than 3. Similarly, you will earn 15 Euros if the absolute difference between \(\hat{Z}_2\) and \(Z_2\) is smaller or equal to 3. Hence, you will earn a total of 30 Euros if your selected type 2 estimation is similar enough to the corresponding statistical interval.

To see this in an example, suppose that in the randomly selected type 2 Estimation, you indicated the interval \([19, 43]\). Suppose as well that the corresponding statistical interval is \([25, 40]\). As your error in computing the inferior endpoint is larger than 3 (25 - 19 = 6), you would earn 0 Euros for your estimation of the inferior endpoint. On the other hand, your error in estimating the upper endpoint is not larger than 3 (the distance from 40 to 43 is just 3), so that you would earn 15 Euros with your estimation of the upper endpoint. In total, therefore, your payoff for the type 2 Estimations would be 15 Euros.

If you have any questions, raise your hand and we will attend you.
Estimation Type 1: 6 out of 10.

Remember: There is a total of 100 balls in a virtual urn. Each ball is either blue or red. You do not know the number of blue balls in the urn, only that the number has been determined by the computer, selecting each integer from 1 to 99 with equal probability, both included. The number of blue balls in the urn remains the same in all estimations of type 1.

So far, 12 balls have been extracted from the urn with replacement.

The following graph shows the color of all extracted balls.

Total number of blue balls drawn: 7.
Total number of red balls drawn: 5.

Given the color of the extracted balls and in your opinion, how many blue balls are in the urn? (move the triangle over the bar and drop it to indicate your answer)

Your answer is: 56

Summary

| Estimation type 1 | Total number of blue balls | Total number of red balls | Your answer |
|-------------------|---------------------------|---------------------------|-------------|
| 1 of 10           | 1                         | 1                         | 49          |
| 2 of 10           | 2                         | 2                         | 48          |
| 3 of 10           | 3                         | 3                         | 52          |
| 4 of 10           | 5                         | 3                         | 61          |
| 5 of 10           | 5                         | 5                         | 40          |
| 6 of 10           | 7                         | 5                         |             |

Figure 6: Estimation type 1
**Estimation Type 2: 6 out of 10.**

Remember: There is a total of 100 balls in a virtual urn. Each ball is either blue or red. You do not know the number of blue balls in the urn, only that the number has been determined by the computer, selecting each integer from 1 to 99 with equal probability, both included. The number of blue balls in the urn remains the same in all estimations of type 2. This number of blue balls is determined independently of the number of blue balls in the estimations of type 1.

So far, 12 balls have been extracted from the urn with replacement.

The following graph shows the color of all extracted balls

- Total number of blue balls drawn: 6.
- Total number of red balls drawn: 6.

Moving the triangles over the bar, indicate the lower and the upper bound of the interval that you believe to contain the actual number of blue balls with 95% confidence.

The lower bound is: 33
The upper bound is: 68

**Summary**

| Estimation type 2 | Total number of blue balls | Total number of red balls | Your answer |
|-------------------|---------------------------|---------------------------|-------------|
| 1 of 10           | 1                         | 1                         | [30,76]     |
| 2 of 10           | 2                         | 2                         | [29,76]     |
| 3 of 10           | 3                         | 3                         | [31,71]     |
| 4 of 10           | 4                         | 4                         | [35,64]     |
| 5 of 10           | 5                         | 5                         | [38,84]     |
| 6 of 10           | 6                         | 6                         |             |

**Figure 7: Estimation type 2**
Online appendix III: control questions

Point estimations (type 1 estimations).

1. How many balls are there in the urn?

2. What is the maximal number of blue balls in the urn?

3. How many type 1 estimations are you going to make throughout the experiment?

4. Some of these type 1 estimations are randomly chosen at the end of the experiment for payoff. How many?

5. Remember that after extracting a ball it is re-introduced into the urn. Hence, the extraction of balls from the urn is: with or without replacement?

6. Does the number of blue balls in the urn change during the type 1 estimation phase?

7. Suppose there are 70 blue balls in the urn. How many red balls are in the urn?

8. Suppose that after the first two extractions of balls from the urn, a participant indicates a type 1 estimation of 75. If this estimation is selected for payoff and the true number of blue balls in the urn is 70, what is the subject’s payoff for this task?

Interval estimations (type 2 estimations).

1. How many balls are there in the urn?

2. What is the minimal number of blue balls in the urn?

3. How many type 2 estimations are you going to make throughout the experiment?

4. Some of these type 2 estimations are randomly chosen at the end of the experiment for payoff. How many?

5. Remember that after extracting a ball it is re-introduced into the urn. Hence, the extraction of balls from the urn is: with or without replacement?
6. Does the number of blue balls in the urn change during the type 2 estimation phase?

7. Suppose there are 71 red balls in the urn. How many blue balls are in the urn?

8. Suppose a participant indicates a type 2 estimation of 23 for the lower bound and 59 for the upper bound. If this estimation is selected for payoff and the lower bound of the true statistical interval is 20 and the upper bound of the true statistical interval is 48, what is the subject’s payoff for this task?

Online appendix IV: formal proofs.

Proposition 1. The following statements hold for the Dirac $\delta$ loss function:

1. For all $\Theta(X)$, $\hat{\theta}^* = f$.

2. For all $\Theta(X)$, $f \in [\hat{\theta}_l, \hat{\theta}_h]$.

3. For all $\Theta(X), \Sigma(X)$ such that $\Theta \subseteq \Sigma \subseteq [0, 1], [\hat{\sigma}_l, \hat{\sigma}_h] \subseteq [\hat{\sigma}_l, \hat{\sigma}_h]$.

Proof of Proposition 1.

1. Given any $\Theta = [\theta, \bar{\theta}] \subseteq [0, 1]$, minimizing the expected loss under the Dirac $\delta$ loss function is the same as maximizing the posterior probability that the error is zero. Hence, we only have to show that the mode of the posterior probability distribution function is equal to $f$. So, let $p_{\Theta}(\theta|k, n)$ be the posterior probability when $k$ out of $n$ trials are successes and the subject consider the mental model $\Theta$.

The Bayesian posterior upon observing $k$ successes out of $n$ trials is the beta probability distribution function

$$
\beta(\theta|k+1, n-k+1) = \theta^k \cdot (1-\theta)^{n-k}.
$$

For simplicity, we write $b(\theta|k, n)$ instead of $\beta(\theta|k+1, n-k+1)$. The total probability mass of all elements of $\Theta$ is therefore

$$
\int_\Theta \theta^k \cdot (1-\theta)^{n-k} d\theta = B(\bar{\theta}|k, n) - B(\theta|k, n),
$$

46
where $B(\cdot)$ denotes the beta cumulative probability distribution function. The maximization problem to be solved is thus:

$$\arg\max_{\theta} \frac{b(\theta|k, n)}{B(\theta|k, n) - B(\bar{\theta}|k, n)}.$$ 

This is equivalent to

$$\arg\max_{\theta} k \cdot \ln \theta + (n - k) \cdot \ln(1 - \theta) - \ln \left( B(\theta|k, n) - B(\bar{\theta}|k, n) \right).$$

As $k$ and $n$ are known, the last term in this optimization problem is a constant. The first order condition is thus $k/\theta - (n - k)/(1 - \theta) = 0$. Hence, $\hat{\theta}^* = k/n$. Since the second derivative $-k/\theta^2 - (n - k)/(1 - \theta)^2$ is strictly smaller than zero for all $\theta \in [0, 1]$, the necessary condition is also sufficient. \hfill $\Box$

2. We have essentially used in the first part that the beta probability distribution function is log-concave on our domain of $k$ and $n$. We can thus conclude that (1) for all $\theta_1 < \theta_2 \leq f$, $b(\theta_1|k, n) < b(\theta_2|k, n)$ and (2) for all $f \leq \theta_1 < \theta_2$, $b(\theta_1|k, n) > b(\theta_2|k, n)$. Consequently, the shortest 95% confidence interval must necessarily include $f$.

3. Since $\Theta \subseteq \Sigma$, it must be the case that for all $\theta \in \Theta$,

$$p_{\Theta}(\theta|k, n) = \frac{b(\theta|k, n)}{B(\bar{\theta}|k, n) - B(\bar{\sigma}|k, n)} = \frac{b(\theta|k, n)}{B(\bar{\sigma}|k, n) - B(\bar{\sigma}|k, n)} \cdot \frac{B(\bar{\sigma}|k, n) - B(\bar{\sigma}|k, n)}{B(\bar{\theta}|k, n) - B(\bar{\theta}|k, n)} = p_{\Sigma}(\theta|k, n) \cdot \frac{B(\bar{\sigma}|k, n) - B(\bar{\sigma}|k, n)}{B(\bar{\theta}|k, n) - B(\bar{\theta}|k, n)}.$$ 

Since the rescaling factor is larger than 1, $p_{\Theta}(\theta|k, n) > p_{\Sigma}(\theta|k, n)$. That is, the posterior probability of all $\theta \in \Theta$ is larger under $\Theta$ than under $\Sigma$. Since $f \in [\hat{\sigma}_l, \hat{\sigma}_h]$ and $f \in [\hat{\sigma}_l, \hat{\sigma}_h]$ by the second part of this proof and since the beta probability distribution function is log-concave on our domain of $k$ and $n$, we can conclude that $\hat{\sigma}_l \leq \hat{\theta}_l$ and $\hat{\sigma}_h \leq \hat{\theta}_h$. \hfill $\Box$