Belief Distributions, Bayes Rule and Bayesian Overconfidence

by

Glenn W. Harrison and J. Todd Swarthout †

January 2021

ABSTRACT.

We provide evidence that individuals generally behave consistently with Bayes Rule defined over continuous events in the sense that they report subjective belief distributions that are unbiased, but that they behave inconsistently in the sense that they exhibit significant overconfidence compared to the appropriate Bayesian confidence as defined by the variance of the posterior distribution. Only by eliciting distributions are we able to assess the dispersion of an individual’s beliefs, giving us a natural metric for subjective confidence. The evidence involves the elicitation of beliefs with financial incentives, and tests Bayes Rule in a setting in which the posterior distribution is known for each individual and belief elicitation. Previous tests of the consistency of beliefs with Bayes Rule were not designed to independently test bias and overconfidence.

JEL Codes: D84, D81, C11
Keywords: Bayes Rule, Overconfidence, Behavior

† Department of Risk Management & Insurance and Center for the Economic Analysis of Risk, Robinson College of Business, Georgia State University, USA (Harrison); and Department of Economics and Experimental Economic Center, Andrew Young School of Policy Studies, Georgia State University, USA (Swarthout). Harrison is also affiliated with the School of Economics, University of Cape Town. E-mail contacts: gharrison@gsu.edu and swarthout@gsu.edu. We are grateful to Jim Cox for helpful comments.
# Table of Contents

1. Identifying and Measuring Bayesian Overconfidence ........................................... -4-

2. Theory .............................................................................................................................. -6-
   A. Eliciting Subjective Belief Distributions ................................................................. -7-
   B. Calculating the Bayesian Posterior Distribution ..................................................... -9-

3. Experimental Design ....................................................................................................... -10-

4. Results .............................................................................................................................. -12-
   A. Reports about Subjective Beliefs ............................................................................. -13-
   B. The Bias and Confidence of Subjective Beliefs ....................................................... -16-

5. Previous Literature .......................................................................................................... -18-
   A. Early Literature .......................................................................................................... -18-
   B. Classic Designs ......................................................................................................... -19-
   C. Overconfidence ......................................................................................................... -23-
      The Three Faces of Overconfidence ........................................................................ -23-
      Is There Only One Interesting Definition of Overconfidence? ............................. -25-
   D. Additional Issues in the Modern Literature ............................................................. -26-

6. Conclusion ......................................................................................................................... -29-

References ........................................................................................................................... -42-

Appendix A: Instructions and Parameters (Online Working Paper) ................................... -A1-
   A. Risk Aversion Task ..................................................................................................... -A2-
   B. Generic Belief Elicitation Task ................................................................................ -A6-
   C. Urn-Specific Belief Elicitation Task .......................................................................... -A10-
   D. Lotteries for Risk Aversion Task ............................................................................... -A15-

Appendix B: Detailed Statistical Test Results (Online Working Paper) ............................. -A18-

Appendix C: Additional Figures (Online Working Paper) ................................................ -A21-
In general, Bayes Rule generates a posterior probability distribution to characterize subjective inferences about the parameters characterizing a random data-generating process. Based on that posterior distribution, one can define a mean, median or mode as descriptive measures of central tendency, and one can define a variance, equal-tailed credible interval, or highest posterior density credible interval as descriptive measures of dispersion.\(^1\) We provide the first incentivized behavioral test of Bayes Rule in this general sense.

We focus on a canonical case in Bayesian analysis: beliefs about the sole parameter characterizing a Binomial data-generating process for a single random variable. Although the sample is a series of independent draws from a Bernoulli trial, with a binary outcome in each trial, the parameter of the Binomial distribution is continuous. Hence the subjective beliefs predicted by Bayes Rule to be the posterior probability distribution will, in general, be characterized by a non-zero variance that is independent of the mean.

Bayes Rule allows us to define an appropriate level of confidence in beliefs about that continuous parameter. Assuming a diffuse prior at the outset of the sampling process, consider two samples drawn from the same data-generating process. Let one sample be larger than the other. Then it is common sense, and a formal prediction of Bayes Rule, that the individual observing the smaller sample should \textit{appropriately} have a subjective belief about the parameter that exhibits less confidence than the individual observing the larger sample. The formal application of Bayes Rule in this general setting allows us to state precisely what that appropriate confidence is, conditional on the sample realization. And, as informative priors are updated with later samples of varying size and composition, Bayes Rule updates the predicted, appropriate confidence. It is this evolving “appropriate confidence” that we call Bayesian Confidence. And then Bayesian Overconfidence\(^2\) is directly defined as observed

\(^1\) These are, of course, not the only statistics of a posterior distribution, but they are the most popular.

\(^2\) We maintain the term \textit{Bayesian Overconfidence} throughout to explicitly differentiate our definition of overconfidence from the other definitions used across literature in psychology, economics and finance.
or inferred confidence in subjective beliefs for an individual that are more precise\(^3\) than this Bayesian Confidence.

Finally, to provide a meaningful behavioral test of Bayesian Overconfidence, we elicit the latent subjective belief distribution that an individual has for some observable event and compare that elicited subjective belief distribution to the appropriate Bayesian posterior distribution to determine if observed behavior is consistent with Bayes Rule. Perhaps surprisingly, this general distributional test of Bayes Rule, and the appropriate Bayesian Confidence in beliefs, has not been addressed previously.

There are many ways to elicit subjective belief distributions, and we employ the familiar, and strictly proper, Quadratic Scoring Rule defined over continuous events or discrete, non-binary events. We take care to infer latent subjective beliefs from observed reports that are financially motivated. Our behavioral interface does not place heavy demands on subjects, who simply place bets in accordance with their beliefs and see the implied payoff for each event conditional on the event being the true outcome.\(^4\)

We employ an experimental task that allows us to precisely calculate the posterior distribution by applying Bayes Rule to the informative draws. This posterior distribution is used as the basis for identifying bias, defined as a statistically significant difference between the mean elicited belief and the mean of the posterior distribution. This posterior distribution also allows us to directly identify Bayesian Overconfidence or Bayesian Underconfidence, defined as a statistically significant difference between the variance of elicited beliefs and the variance of the posterior distribution.

Remarkably, to the best of our knowledge there are no studies that consider incentivized reports about non-degenerate belief distributions and use a metric for evaluation of bias and confidence.

Moore and Healy [2008] and Merkle and Weber [2011] provide excellent statements of the literature. In §5.C we discuss the relationship of these other definitions to Bayesian Overconfidence.

\(^{3}\) Increased precision is measured by a reduction in the variance of the posterior distribution.

\(^{4}\) See Harrison, Martínez-Correa, Swarthout and Ulm [2017] for more details. As a robustness check, we also recover the latent beliefs of individuals, conditioning our inferences on estimated risk preferences for the individual following Harrison and Ulm [2015].
for a specific individual and event. Many of the pioneering experimental studies of Bayes Rule, such as Grether [1980][1992], used incentivized methods for eliciting subjective probabilities for a binary event, but this does not permit any discussion of confidence with respect to a continuous or discrete, non-binary event; nor was that the goal of those experiments. Indeed, virtually all textbooks on Bayesian inference start off with defining posterior probabilities over binary events, before generalizing, so this was a natural place to start the behavioral evaluation of Bayes Rule. 5

We discuss the subtleties of rigorously identifying Bayesian Overconfidence in section 1, present the theoretical framework motivating our approach in section 2, the experimental design that flows from it in section 3, results in section 4, a detailed discussion of the previous literature in section 5, and concluding remarks in section 6.

Our findings are simple and clear. In general subjects exhibit no statistically significant bias in relation to Bayes Rule, but exhibit statistically significant overconfidence in relation to Bayes Rule. This Bayesian Overconfidence starts with the first samples observed by the subject, and persists as additional sample evidence accumulates. There are also some identifiable demographic effects. The marginal effect of being older, at least within the limited age range of our sample, is to exhibit statistically significant bias as well as Bayesian Overconfidence. The marginal and total effect of being a Business Major is to exhibit statistically significant Bayesian Underconfidence with larger samples, albeit in an unbiased manner.

5 For example, the classic text of Gelman, Carlin, Stern, Dunson, Vehtari and Rubin [2013] starts (p. 8ff.) with an example in which the posterior probability refers to a woman being a carrier of the hemophilia gene given data on the status of her sons. Their next example (p. 9ff.) considers a posterior probability mass function with three possible correct spellings of the typed word “radom.” The bulk of their applications refer to inferences about estimation of parameters of parametric probability density functions, such as the historically important case of estimating the population probability of a sample from a data-generating process that follows a binomial distribution (p. 29ff.). Their first example does not generate Bayesian inferences about the confidence of the posterior probability, but the second and third do. To take another example, the popular textbook of Kruschke [2015] starts in §5.1.2 with an example of someone having a rare disease or not, given some text results, and then moves in §5.3 to the estimation of parameters of the probability of a coin being biased from a sample of toss-outcomes from a data-generating process that follows a Bernoulli distribution. Again, the first example does not generate Bayesian inferences about the confidence of the posterior probability, but the second does.
manner. Females have a statistically significant marginal effect\(^6\) of Bayesian Overconfidence with larger samples, although again in an unbiased manner. Finally, the marginal and total effect of having a small initial sample is associated with a statistically significant Bayesian Underconfidence with no bias, but these beliefs become better calibrated as larger samples accumulate over time.

1. Identifying and Measuring Bayesian Overconfidence

Many applications of Bayes Rule in economics and psychology leave no role for confidence at all.\(^7\) Start with the simplest application of Bayes Rule, derived from the formula for conditional probabilities, as

\[
p(A \mid B) = \frac{p(A, B)}{p(B)} = \frac{p(A) \cdot p(B \mid A)}{p(B \mid A) \cdot p(A) + p(B \mid \lnot A) \cdot p(\lnot A)}
\]

where \(A\) and \(B\) are events, \(P(A \mid B)\) is the conditional probability of \(A\) given that \(B\) has occurred, \(p(A,B)\) is the joint probability of \(A\) and \(B\) occurring, and \(\lnot A\) is the complement of \(A\). To use the mammography problem of Gigerenzer and Hoffrage [1995; p. 685], think of \(A\) as the diagnosis of breast cancer, \(\lnot A\) as the diagnosis of not having breast cancer, and \(B\) as the data provided by a positive mammography reading. Assume that \(p(A) = 0.01\), \(p(B \mid A) = 0.8\) and \(p(B \mid \lnot A) = 0.096\), where \(p(B \mid A)\) is the “true positive” of the medical test and \(p(B \mid \lnot A)\) is the “false negative” of the medical test. Then the posterior probability that a woman with a positive mammography has breast cancer is

\[
(0.01)(0.8)/[(0.01)(0.8) + (0.99)(0.096)] = 0.07763975 = 0.078, \text{ a scalar.}
\]

In order to see how Bayes Rule can be used to derive a posterior distribution that might have a

---

\(^6\) The total effects are close to being statistically significant at the 5% two-tailed level as well.

\(^7\) One can certainly identify distinct values for the mean and variance of distributions defined over binary events, but once the mean is known the variance is known. Let the events be a “heads” or a “tails” in a coin toss, and consider the Bernoulli distribution with probability \(p\) for “heads” and probability \(1-p\) for “tails.” The mean is \(p\) and the variance is \(p(1-p)\). Variance is highest when \(p = \frac{1}{2}\) and smallest when \(p = 0\) or \(p = 1\), but one cannot know if a belief of \(p = 0\), say, reflects overconfidence or just a subjective belief that the outcome “heads” is not likely to happen. In effect, variance must be able to take on values independently of the value of the mean before one can identify confidence (and hence, relative to the Bayesian posterior variance, overconfidence or underconfidence). And there are many distributions, such as the Beta or Normal, which allow that.
variance, extend (1) to the case where there are three mutually exclusive and jointly exhaustive events $A, A'$ and $A''$. Then

$$p(A | B) = \frac{p(B | A) p(A)}{p(B | A) p(A) + p(B | A') p(A') + p(B | A'') p(A'')}.$$  

(2)

Now interpret $A, A'$ and $A''$ as three possible, distinct values of some discrete random variable $x$. Then the data $B$ can be seen to revise the posterior probabilities of the possible values of a discrete random variable, thereby changing the probability mass function for $x$. In general, a probability mass function has a well-defined mean, variance, and indeed higher-order central moments, and these can vary independently. The extension of (2) to the case of a continuous random variable, and a probability distribution function, is immediate: see Winkler [2003; §3.2 and §4.2].

All experiments testing Bayes Rule that we are aware of, and certainly those doing so in an incentivized manner, employ (1). None, to our knowledge, have employed (2) or the extension to a continuous random variable. Hence none of these experiments can test overconfidence, or underconfidence, in the sense that theory predicts some non-zero variance for beliefs that is independent of the predicted mean for beliefs. We elicit subjective belief distributions that test (2), and that therefore have a predicted level of confidence from the Bayesian posterior distribution. We can then directly assess overconfidence and underconfidence relative to that predicted Bayesian posterior distribution.

A detailed review of previous experiments is provided in section 5. Two examples are sufficient to see the manner in which previous experiments have been conducted. Gigerenzer and Hoffrage [1995] simply asked subjects to state the posterior probability, with no incentives, given data on $p(A)$, $p(B | A)$ and $p(B | \neg A)$ as defined earlier for instances such as the mammography problem. Grether [1980] generated priors by draws from a cage $\chi$: subjects were told that if the number drawn from $\chi$

---

8 The values of these probabilities was presented in percentage form (e.g., the probability of breast cancer is 1%) or in “natural frequency” form (e.g., “10 out of 1,000 women have breast cancer”), and the posterior probability elicited in the same form.
was less than or equal to k, for k = 2, 3 or 4, cage α would be used to generate the observed data, otherwise cage β would be used. Cage α was known to contain 4 balls marked N and 2 balls marked G, and cage β was known to contain 3 balls marked N and 3 balls marked G. Subjects were told a value of k that applied in any instance for the draws from cage χ, so if k=2 the prior probability of cage α was ⅓, if k=3 the prior probability for cage α was ½, and if k=4 the prior probability for cage α was ⅔.

Subjects did not see the outcome of the single draw from case χ, but they did see 6 draws with replacement from cage α or cage β, depending on k and the outcome of the draw from cage χ. So the sample data consisted of one of {0N, 6G; 1N, 5G; 2N, 4G; 3N, 3G; 4N, 2G; 5N, 1G; 6N, 0G}. Subjects were then asked to state whether they believed that cage α had been used or cage β had been used. Under Bayes Rule, a subject would state that cage α had been used if the posterior probability for cage α was greater than 0.5. This design generated Bayesian posterior probabilities for cage α of 0.042, 0.081, 0.149, 0.260, 0.413, 0.584 or 0.737 if k=2 was applied, depending on the sample realization from cage χ, respectively. Hence for a given k and observed data, there was again one Bayesian posterior probability, not a Bayesian posterior distribution.

2. Theory

We want to rigorously characterize the bias and confidence of beliefs relative to the predictions of Bayes Rule. Consider the domain over which beliefs are defined to be the probability of an event occurring, π. Specifically, and without loss of generality, let π be the true fraction of balls of one color in an urn with two colors. Consider a decision maker with a diffuse prior over the composition of the urn. The decision maker observes a series of informative draws with replacement from the urn and then reports a distribution characterizing her belief of π. We classify the belief distribution into one of

---

9 Grether [1992] extended the same general design to elicit a subjective probability, not just a report about which cage was subjectively the most likely.

10 If k=3 the Bayesian posterior probabilities for α were 0.081, 0.149, 0.260, 0.413, 0.584, 0.737 or 0.849. If k=4 the Bayesian posterior probabilities for α were 0.149, 0.260, 0.413, 0.583, 0.737, 0.849 or 0.918.
six categories according to degrees of bias and confidence relative to the Bayesian posterior as shown in Figure 1: beliefs are either biased or unbiased, and beliefs can exhibit overconfidence, underconfidence or appropriate confidence. The Bayesian ideal is to be unbiased with appropriate confidence, as in panel F of Figure 1 where the posterior belief is literally overlaid by the subjective belief. The other panels demonstrate possible deviations from Bayes rule.

For each of the panels shown in Figure 1, we show a Bayesian posterior distribution (solid, black line) and a corresponding subjective belief distribution (dashed, blue line). These panels remind us that Bayesian Overconfidence arises due to an agent subjectively overweighting the probability of some events and, consequently, underweighting the probability of other events in relation to the Bayesian posterior. Hence it is critical when evaluating Bayesian Overconfidence, or Bayesian Underconfidence, to not make an *a priori* judgment about the domain of subjective overweighting. By definition, nobody can overweight the probability of events over the whole range of the distribution: as a logical matter, we simply cannot declare an individual to be overconfident over events without adding some context about what event domain that refers to. It is natural to talk about overconfidence around the mean of the belief distribution, akin to the variance of beliefs viewed as the expectation of a certain deviation around the mean of a distribution, and that is the definition we adopt. Our definition of Bayesian Overconfidence, then, refers to summary statistics\(^{11}\) that pertain to the entire distribution, and not some restricted domain of events.

We review the theory behind the elicitation of the subjective belief distribution and the calculation of the Bayesian posterior distribution.

\[ A. \textit{Eliciting Subjective Belief Distributions} \]

Let the decision maker report subjective beliefs in a discrete version of a Quadratic Scoring

---

\(^{11}\) These are descriptive statistics. They are only sufficient statistics in limited, but popular, cases.
Rule (QSR) for continuous distributions (Matheson and Winkler [1976]). Partition the domain of the continuous distribution into $K$ intervals, and denote as $r_k$ the report of the likelihood that the event falls in interval $k = 1, ..., K$. Assume for the moment that the decision maker is risk neutral, and that the full report consists of a series of reports for each interval, $\{ r_1, r_2, ..., r_k, ..., r_K \}$ such that $r_k \geq 0 \ \forall k$ and $\sum_{i=1,K} (r_i) = 1$.

If $k$ is the interval in which the actual value lies, then the payoff score is defined by Matheson and Winkler [1976; p.1088, equation (6)]: $S = (2 \times r_k) - \sum_{i=1,K} (r_i)^2$. So the reward in the score is a doubling of the report allocated to the true interval, and the penalty depends on how these reports are distributed across the $K$ intervals. The subject is rewarded for accuracy, but if that accuracy misses the true interval the punishment is severe. The punishment includes all possible reports, including the correct one.\(^\text{13}\)

To ensure complete generality, and avoid any decision maker facing losses, allow some endowment, $\alpha$, and scaling of the score, $\beta$. We then get the following scoring rule for each report in interval $k$

\[ \alpha + \beta \left[ (2 \times r_k) - \sum_{i=1,K} (r_i)^2 \right], \]  

where we initially assumed $\alpha=0$ and $\beta=1$. We can assume $\alpha>0$ and $\beta>0$ to get the payoffs to any

\(^{12}\) Alternative scoring rules could be characterized, and Harrison and Ulm [2015] prove that all results generalize to the class of proper scoring rules. The QSR is the most popular scoring rule in practice, and all of the practical issues of recovering beliefs can be directly examined in that context. For instance, Andersen, Fountain, Harrison and Rutström [2014] show that behavior under a Linear Scoring Rule and QSR are behaviorally identical when applied to elicit subjective probabilities for binary events and one undertakes calibration for the different effects of risk aversion and probability weighting on the two types of scoring rules.

\(^{13}\) Take some examples, assuming $K = 4$. What if the subject has very tight subjective beliefs and allocates all of the weight to the correct interval? Then the score is $S = (2 \times 1) - (1^2 + 0^2 + 0^2 + 0^2) = 2 - 1 = 1$, and this is positive. But if the subject has tight subjective beliefs that are wrong, the score is $S = (2 \times 0) - (1^2 + 0^2 + 0^2 + 0^2) = 0 - 1 = -1$, and the score is negative. So we see that this score would have to include some additional “endowment” to ensure that the earnings are positive. Assuming that the subject has very diffuse subjective beliefs and allocates 25% of the weight to each interval, the score is less than 1: $S = (2 \times \frac{1}{4}) - (\frac{1}{4})^2 + (\frac{1}{4})^2 + (\frac{1}{4})^2 + (\frac{1}{4})^2 = \frac{1}{2} - \frac{1}{4} = \frac{1}{4} < 1$. So the tradeoff from the last case is that one can always ensure a score of $\frac{1}{4}$, but there is an incentive to provide less diffuse reports, and that incentive is the possibility of a score of 1.
positive level and units we want. Let $p_k$ represent the underlying, true, latent subjective probability of an individual for an outcome that falls into interval $k$.

Harrison, Martínez-Correa, Swarthout and Ulm [2017] show that the QSR directly elicits “approximately correct” subjective beliefs for EUT individuals with risk attitudes in the observed range in experiments. Essentially, EUT subjects provide reports which are flattened with respect to their true beliefs, so as to reduce variability over possible outcomes with positive subjective probability. This is a second-order adjustment compared to the first-order adjustment in the binary case, where risk averse individuals provide reports closer to $\frac{1}{2}$ so as to reduce variability over the two possible outcomes.

On the other hand, RDU individuals distort probabilities and employ “decision weights” when evaluating ranked payoff outcomes, and exhibit first-order adjustments in both binary and non-binary cases. Harrison and Ulm [2015] demonstrate that one can \textit{exactly} recover subjective beliefs conditional on estimates of an EUT or RDU model of risk preferences. If one assumes a probability mass function for the inferred beliefs of each individual, as we do here, one can indeed infer these beliefs using closed-form systems of equations. For simplicity, we report the results from the reports of subjects, and confirm in Appendix C (online) that recovering beliefs does not change our conclusions.

\textbf{B. Calculating the Bayesian Posterior Distribution}

Assume a data-generating process in period $t$ in which a sample of $N_t$ is drawn with replacement from an urn with two colors. The parameter of interest is $\theta \in [0, 1]$, the fraction of balls of a specific color. Let $y_t$ denote the number of balls of that color in the sample $N_t$. An appropriate sampling model for $y_t$ is a binomial model, so that we have $y_t | \theta \sim \text{Binomial}(N_t, \theta)$.

A “natural conjugate” family of prior and posterior distributions for beliefs about $\theta$ is the Beta distribution, with non-negative parameters $a$ and $b$. A diffuse prior is traditionally defined with the Beta distribution by setting $a = b = 1$, although alternative characterizations of an “uninformative” Beta
prior have been proposed. The advantage of this natural conjugate family is that the posterior
distribution is immediately characterized once we know the prior parameters and the sample realization
\( y_t \); if \( a \) and \( b \) are the prior parameters, whether or not they reflect a diffuse prior, then the posterior
parameters are \( a' = a + y_t \) and \( b' = b + N_t - y_t \). These parameters \( a' \) and \( b' \) characterize the posterior
distribution in period \( t \), then become the parameters characterizing the prior distribution for period
\( t+1 \), the posterior parameters for period \( t+1 \) are then \( a'' = a' + y_{t+1} \) and \( b'' = b' + N_{t+1} - y_{t+1} \), and so
on.\(^{15}\)

3. Experimental Design

Subjects were recruited from the undergraduate population at Georgia State University,
spanning several colleges. Subjects received no information about the task or expected earnings in
advance of the session. Each subject received $5 for participating, in addition to earnings from the
incentivized tasks.\(^{16}\) We presented instructions to subjects in both print and video formats. Before each
task, the printed instructions were distributed and then the video instructions were played on each
participant workstation. The video instructions included an audio narration of the printed text. By
using video instructions we minimize the speaking required by an experimenter during a session, and
thus abate a possible source of variation across sessions. Appendix A (online) contains all instructions.
A total of 122 subjects participated.

The primary stimulus in the beliefs task was a sample of draws from an urn consisting of 100
balls, colored blue or orange. Four independent samples were drawn with replacement for each subject,

\(^{14}\) Haldane [1932] proposed \( a = b = 0 \) and Jeffreys [1946] proposed \( a = b = \frac{1}{2} \).

\(^{15}\) One advantage of modern Bayesian computational methods is that analysts need not restrict
themselves to natural conjugate families, where the sufficient statistics get updated in such a direct manner
with sample evidence.

\(^{16}\) Apart from the belief task that is the focus here, each subject initially completed a task consisting
of 50 binary lottery choices and one of those choices was selected at random for payment. Earnings from the
selected lottery choice were realized prior to the belief elicitation task, and subjects were paid for both tasks.
with beliefs elicited after each sample. Each subject observed a total of 40 draws from the urn. Subjects were randomly assigned to be one of three types, differentiated by the sample sizes drawn in each of the four periods. Subject type A received 28, 4, 4 and 4 draws in period 1 through 4, respectively, so their posterior distribution after period 1 would be relatively informative. Subject type C received 4, 12, 12 and 12 draws in periods 1 through 4, respectively, so their posterior distribution after period 1 would be relatively uninformative. Subject type B received 16, 8, 8 and 8 draws in periods 1 through 4, respectively, so their posterior distribution after period 1 would be intermediately informative relative to the other two types. Each session consisted of a roughly even mix of subjects of each type. The true fraction of blue balls was 0.77 in session 1, 0.23 = 1-0.77 in session 2, 0.63 in session 3, 0.37 = 1-0.63 in session 4, and 0.15 in session 5. There were 24, 20, 26, 22 and 20 subjects in sessions 1 through 5, respectively.

We carefully integrated the sampling process and belief elicitation process in the same computerized interface, illustrated in Figures 2 through 6. Figure 2 shows the initial display of sample draws in period 1, with a virtual hand drawing a blue ball from the urn. Figure 3 shows the end of sample draws in period 1, if there had been 10 draws: subjects see that there were 6 blue balls and 4 orange balls drawn. Figure 4 displays the initial belief interface for period 1. Each subject has 100 tokens to allocate across 10 bins, each labeled with an interval of percentages of blue balls. Bins 1 through 10 are labeled 0% to 10%, 11% to 20%, 21% to 30%, 31% to 40%, 41% to 50%, 51% to 60%, 61% to 70%, 71% to 80%, 81% to 90%, and 91% to 100%, respectively. The implied payoffs from the token allocation are displayed in dollars and cents above each histogram bar, and varied in real time as

There is a long tradition in the psychology literature on Bayesian updating when subjects experience samples whose size they choose: see Hertwig, Barron, Weber and Erev [2004] for a review and further results. We discuss this experimental paradigm later, but wanted to avoid it here because it complicates analysis by adding a search process that brings with it potential behavioral biases of its own. Indeed, one of the key findings of this literature is that subjects seem to select very small samples, leading to biases in their inferences about the underlying population. The issue we want to focus on is whether beliefs are consistent with Bayesian inferences conditional on the objective sample.
the subject adjusted the token allocation. Figure 5 shows a typical display of the belief elicitation when an individual is finished re-allocating tokens. A button to the left of the sliders allow the subject to initiate a uniform distribution or clear all tokens to zero, to reduce the burden of moving to these default allocations. After period 4, Figure 6 shows the accumulated samples in a table on the right hand side of the screen, and reveals the true distribution on the left side of the screen. The results for this subject, #4 in the example, are displayed, as well as results for all other subjects, so that the credibility of the data-generating process is enhanced. Earnings from each period are displayed, as well as accumulated earnings, in this case $30.29.

From a Bayesian perspective this is a Binomial distribution defined over the true fraction of blue balls. Conjugate families of prior and posterior distributions are the Beta distribution. If we assume a diffuse prior at the outset, it is a simple matter to generate the posterior distribution that each subject faced at the end of each period. This posterior distribution is path-dependent, in the sense that it reflects the previous independent samples observed by each subject, since those affect the priors in the current period. We also consider prior beliefs that reflect varying degrees of “rational expectations” about the true fraction of blue balls, but a diffuse prior is the most natural for this setting.

4. Results

It is useful to convey the general patterns of behavior that we observe in two ways. The first is to look at reports by individuals, which allows us to see how the relevant Bayesian posterior prediction varied over the four periods of updating, as well as how the realized samples affected subjective beliefs. The second way of seeing the patterns of behavior builds on these insights from individual behavior by statistically evaluating the overall size and significance of Bayesian Bias and Bayesian Overconfidence.
A. Reports about Subjective Beliefs

Figures 7 through 10 display detailed results for each of 10 subjects. In each case the four periods are stacked on top of each other to make it easy to see the sequential evolution of the task. In the top right corner the sample data are presented, showing the number of realized blue balls out of the total number of draws from the urn, and the implied sample frequency. The true population probability is shown in black, with a probability of 1, and of course remains constant over all periods. Assuming a diffuse prior distribution, the blue bars show the implied Bayesian posterior distribution. Although the actual posterior is a continuous distribution, we computed the mass inside each elicitation interval and report that, to make it easier to compare to reports and beliefs. The red bars show the reports of the subject, normalized by the 100 tokens assigned to each bin. We consider observed reports rather than recovered beliefs, without loss of generality.\textsuperscript{18}

Subject #36 in Figure 7 illustrates several important, general features of our design. First, with small samples, and realizations that have average frequencies that differ from the population probability, there is no presumption at all that the Bayesian posterior distribution be close to the population probability. One would expect that with a sufficient sample size it would converge to the population probability, but even with samples of 40 as we have here this is not always the case. In period 1 this subject had a small sample of 4, and 3 blue balls were drawn for a sample mean of 0.75, well above the population probability. In periods 2, 3 and 4 the sample sizes were each 12, so by period 4 the posterior distribution is much closer to the population probability.

A key feature of this example is to see that the posterior for each subject would typically change from period to period, and in some striking ways. These changes might be the posterior mean moving

\textsuperscript{18} As a theoretical matter we know that for any EUT or RDU risk attitude, a report of zero tokens in a bin is associated with a true latent belief of zero probability for that bin. So recovered beliefs are always on the positive support of the observed allocations. Moreover, EUT decisions makers have recovered beliefs that are closely approximated by the observed reports, and less flat than the observed reports for risk averse individuals.
towards the population probability as here, but there are also changes in the standard deviation of the posterior, in this case getting much smaller after period 1. An obvious, but critical, implication is that behavioral evidence of bias and inappropriate confidence has to be defined relative to the relevant posterior distribution facing each subject in each period. It should not be defined, in small samples like this at least, relative to the population probability.

Substantively, we observed subject #36 tracking the posterior distribution quite well, particularly given the “misleading” lure of the sample realization in period 1. To be sure, beliefs do not adjust enough in periods 3 and 4, but were clearly qualitatively consistent with Bayesian updating.

Subject #39 in Figure 7 was a good Bayesian. Period 1 started with a small sample that had an unfortunately biased realization from a sample frequency of 0 blue balls, but the reported beliefs allowed for the fact that this was just from a sample of 4. Then in periods 2, 3 and 4 she received larger samples of 12 that were more representative of the population parameter. We then see evolution of the posterior towards the population parameter, and beliefs reflecting that movement.

Period 1 for subject #36 and subject #39 display a planned feature of our design: a subject received a sample signal that was not representative of the population parameter, but was from such a small sample that it should have led to an appropriate lack of confidence in reported beliefs. And then the larger samples of periods 2, 3 and 4, along with the application of Bayes Rule, serve to generate beliefs that are less biased with respect to the population parameter and more confident: again, “more confident” means appropriately confident, not overconfident compared to Bayes Rule.

In Figure 8 subject #49 and subject #50 are examples where the sample realizations, as measured by the sample frequency, are representative of the population probability. So the issue here is primarily whether the subjects reported the appropriate levels of belief relative to the Bayesian posterior as the evidence accumulated. Subject #49 steadily tightens her confidence, as she should according to Bayes Rule. Subject #50, on the other hand, moved in the right direction with respect to
confidence but too slowly from the perspective of Bayes Rule, particularly given the relatively large samples in periods 2, 3 and 4.

Subject #53 in Figure 9 is an example of a good Bayesian with a terrible initial realization. From a data-generating population process with probability 0.63, the small sample has a frequency of only 0.25. The reported beliefs are unbiased, but far too confident given the tiny sample. Then larger samples arrive, the sample frequency is closer to the population probability, and the reported beliefs track the posterior remarkably. So this is an instance of an individual who was overconfident when faced with a tiny sample, but who calibrated to become a good Bayesian as the sample got larger and more representative. Nonetheless, it is worth noting again that even at the end of period 4, the posterior itself is not assigning modal weight to the correct bin.

Subject #112 in Figure 9 is an example of someone who starts out being well-calibrated to the Bayesian posterior distribution, but then exhibits too much confidence compared to that posterior as the samples accumulate. This is also a case where the samples are representative of the population probability.

Subject #116 in Figure 10 is another example of a Good Bayesian who starts with a small sample and an extreme sample realization, but reports an appropriate lack of confidence. As the samples accumulate, and in a more representative manner, she tracks the evolution of the posterior distribution well.

Finally, subject #9 in Figure 10 starts off as a well-calibrated Bayesian dealt a potentially cruel sampling hand: small sample of 4, with a sample frequency of only 1 blue ball. So far so good, and discounting the information from small samples is good from a Bayesian perspective. But then as more representative samples arrive in periods 2 and 3, and with much larger samples than in period 1, the adjustments towards the evolving posterior are far too muted. And the slightest hint of confirmation of those biases resulting from the sampling frequency of 0.92 in period 4 generates a surge in the beliefs
reported in period 4 back toward the initial biased beliefs reported in period 1.

**B. The Bias and Confidence of Subjective Beliefs**

Figures 11 and 12 display the main results for bias and confidence, pooled over all subjects. Bias is measured by the ratio of the estimated mean of the subjective belief distribution\(^{19}\) divided by the posterior mean for each subject and period. Hence a value of 1 implies no bias, and values less (greater) than 1 imply a mean belief less (greater) than the appropriate posterior mean under Bayes Rule. We display point estimates by a solid circle and a 95% confidence interval around that point estimate. Confidence is measured similarly, by the ratio of the estimated variance of the subjective belief distribution divided by the posterior variance for each subject and period. Hence a value of 1 implies appropriate Bayesian confidence, and values less (greater) than 1 imply over (under) confidence relative to appropriate Bayesian confidence. We again display point estimates and 95% confidence intervals.\(^{20}\)

Our primary conclusion is that subjects exhibit no statistically significant bias with respect to Bayes Rule, but do exhibit persistent and statistically significant excess confidence with respect to Bayes Rule.

Figures 13, 14 and 15 show the effects of demographics on bias and confidence, and Figure 16 shows the effect of the back-loading treatment. In each case we display both marginal effects for each period as well as total effects for each period; each type of estimated effect is an answer to a different inferential question, and both are of interest.

\(^{19}\) Comparable displays using the recovered beliefs, rather than observed reports, can be found in Appendix C (online). The primary, qualitative conclusions are the same.

\(^{20}\) Since we use rely on this type of display to easily visualize results, it is worth recalling the relationship between confidence interval bounds and statistical significance. When the null hypothesis falls wholly outside the confidence intervals, as in Figure 12, one can correctly infer from the visual display that the null hypothesis is rejected. But when the null hypothesis falls within the confidence intervals, as in Figure 11, the converse is not automatically correct, despite the temptation to believe that the null hypothesis is not rejected. As it happens, each of the null hypotheses implied by the results for bias in Figure 11 are not rejected at conventional significance levels. Unless otherwise noted, the same is true for comparable displays of marginal and total effects in Figures 13 through 16. Appendix B (online) lists all \(p\)-values for all hypothesis tests of bias, confidence and the joint hypothesis of bias and confidence.
Figure 13 shows that age has a statistically significant marginal effect on bias, and a statistically significant marginal and total effect on confidence after the first sample. Older subjects exhibit more Bayesian Overconfidence. Figure 14 shows that gender does not lead to any significant bias, but is associated with increased Bayesian Overconfidence as samples accumulate in the last two periods. Women become overconfident with larger samples, in terms of marginal and total effects. Figure 15 shows that having a business major was associated with Bayesian Underconfidence in later periods after the first sample, whether measured by marginal or total effects.

Finally, Figure 16 shows that back-loading the sample has the *a priori* expected behavioral effect on confidence: subjects with smaller samples in the first period exhibit Bayesian Underconfidence, in terms of both marginal and total effects. We stress that these inferences are relative to the appropriately less precise posterior variance that Bayes Rule predicts in earlier periods, due to the smaller sample. Hence this underconfidence is relative to a prediction from Bayes Rule of a high variance posterior to start with. The joint hypothesis test of consistency with Bayes Rule, reflecting both bias and confidence, can be rejected with a $p$-value of 0.006. It is intriguing to also note that the same back-loaded subjects exhibited Bayesian Overconfidence in period 3, over-adjusting for the asymmetry of sample sizes; in this case consistency with Bayes Rule can be rejected with a $p$-value of 0.043.

The preceding analysis assumes, arguably plausibly, a diffuse prior at the outset of the sampling process. Do the conclusions change if we assume some degree of “rational expectations,” leading to non-diffuse, informative priors that are centered on the true population fraction? In short, the answer is that the conclusions do not change significantly, providing we adopt priors that are not degenerate in the sense of assigning essentially all mass to the true probability. Such a prior, of course, leads to limited updating with new data, by definition of Bayes Rule.\textsuperscript{21}

\textsuperscript{21}If the prior assigns zero probability to some outcome from a univariate distribution, then no amount of data accumulated at that outcome can generate a posterior with positive probability at that outcome. This is the reason that many Bayesians recommend assigning some (small) positive probability to all possible outcomes rather than literally assigning zero probability. Detailed results with these informative
Our analysis has employed the direct reports of subjects, relying on the proof in Harrison, Martínez-Correa, Swarthout and Ulm [2017] that the QSR directly elicits “approximately correct” subjective beliefs for EUT individuals with risk attitudes in the observed range in experiments. As a robustness check, we also recovered the latent beliefs of individuals, conditioning our inferences on estimated risk preferences for the individual, and come to the same qualitative conclusions. Figures C11 through C16 in Appendix C (online) use these recovered beliefs, and mimic Figures 11 through 16.

5. Previous Literature

A. Early Literature

There have been two broad strands in the earlier behavioral literature examining Bayes Rule across psychology and economics. One approach was defined by Edwards [1954][1961a][1961b] and involves subjects experiencing a series of sample draws from some population distribution, and reporting beliefs after each sample. This is very much the approach we have adopted, and allows the sequential application of Bayes Rule as a dynamic updating rule for beliefs. Petersen and Beach [1967] reviewed a massive literature in psychology generated by this approach. By and large the evidence supported the view that agents followed the Bayesian statistical models reasonably well, but certainly not perfectly.

The other approach was defined by Kahneman and Tversky [1972][1973] and involves static descriptions of Bayesian inferential problems, and a request for the subject to state a posterior probability given that description. The facts in the description contained information on a prior and a sampling process, usually presented in terms of rates of false-positives and false-negatives. These descriptions provided a “one shot” approach to testing Bayes Rule, in contrast to the sequential, 

---

22 Extending the earlier reviews by Edwards [1954][1961a].
“multiple shot” approach. The consensus from these static evaluations is that agents violated the Bayesian statistical model in numerous ways.\textsuperscript{23}

These can be formally viewed as complementary approaches, with the static approach referring to what we call the period 1 elicitation\textsuperscript{24} and the sequential approach referring to what we call the elicitation in periods 2, 3 and 4. So from a formal, abstract Bayesian perspective our design incorporates both. In fact, our results shed some light on why these two approaches might find different conclusions. As shown in Figure 16, we find that subjects with smaller samples in the first period periods exhibit Bayesian Underconfidence, and hence violate Bayes Rule. This result will be important for treatments considered in later literature.

In general, all of the experiments in this early literature were unincentivized. There were some exceptions,\textsuperscript{25} and occasional references to robustness checks with real rewards (e.g., Edwards [1962; p.121]).

\textbf{B. Classic Designs}

Studies by Grether [1979][1992], Griffin and Tversky [1992] and Gigerenzer and Hoffrage [1995] stand out as classic designs to test Bayes Rule that have been emulated by many.

Grether [1979] extended a design that had been used in many of the earlier experimental studies of sequential updating with Bayes Rule. Peterson and Beach [1967; p.32] summarized the design used

\begin{itemize}
\item \textsuperscript{23} The contrast was particularly stark with respect to the alleged “base rate fallacy.” Evidence of overweighting of the base rate, called “conservatism,” was stressed by Edwards [1968] and surveyed by Petersen and Beach [1967]. And of course neglect of the base rate was one of the primary conclusions of Kahneman and Tversky [1972].
\item \textsuperscript{24} By and large the descriptions in the static approach contained considerable context, with field referents to things such as diseases and medical tests. And those field referents should not be viewed as unimportant, for reasons stressed by Harrison and List [2004].
\item \textsuperscript{25} For example, Edwards [1955] rewarded subjects for binary choices between a certain amount of money and a risky lottery that depended on objective or subjective probabilities. In order to infer subjective probabilities, he acknowledged that the “... method used in this experiment [...] is based on a sweeping, convenient assumption [...] that what the economists call the marginal utility function for bets [...] is a constant” (p.203).
\end{itemize}
Imagine yourself in the following experiment. Two urns are filled with a large number of poker chips. The first urn contains 70% red chips and 30% blue. The second contains 70% blue chips and 30% red. The experimenter flips a fair coin to select one of the two urns, so the prior probability for each urn is .50. He then draws a succession of chips from the selected urn. Suppose that the sample contains eight red and four blue chips. What is your revised probability that the selected urn is the predominantly red one? If your answer is greater than .50, you favor the same urn that is favored by most subjects and by statistical man. If your probability for the red urn is about .75, your revision agrees with that given by most subjects. However, that revised estimate is very conservative when compared to the statistical man’s revised probability of .97. That is, when statistical man and subjects start with the same prior probabilities for two population proportions, subjects revise their probabilities in the same direction but not as much as statistical man does.

Of course the expression “statistical man” refers to a good Bayesian. As explained earlier, Grether [1979] used draws from a physical cage, rather than an urn, and the subject was given priors for one cage of $\frac{2}{5}$, $\frac{1}{5}$ or $\frac{1}{5}$, and 6 sample draws with replacement. They were asked to simply state which of the two cages they believed was generating the sample. Ignoring the non-salient sessions, subjects were paid $5 if their prediction was incorrect and $15 if it was correct.\footnote{It is not obvious if any given subject made more than one prediction after the instructional “run 0.” The instructions (p. 556) only refer to one decision, but the text notes (p. 542) that “the procedure was repeated again (with possibly a different prior)” and (p. 545) that one of the decisions was selected for payment.} Although salient, these incentives only allows inferences that the posterior probability for the selected cage was greater than 0.5 or less than 0.5, indifference aside.

Grether [1992] extended the experiment in Grether [1979], with a wider range of induced priors and sample sizes, to use an incentive-compatible elicitation mechanism for a subjective probability report. The mechanism is a version of the Becker, DeGroot and Marschak [1964] procedure, adapted from eliciting the certainty equivalent of a risky lottery to eliciting a probability.\footnote{A modern version of these experiments is provided by Holt and Smith [2009].} Since these are eliciting the probability of a binary event, they do not consider confidence by design.\footnote{Several recent studies of “overconfidence” in the psychology literature have clearly recognized the need for a metric that is specific to the individual and event for which confidence intervals of belief distributions are elicited, such as Glaser and Weber [2007], Glaser, Langer and Weber [2007][2013] and}
The alternative hypothesis tested by Grether [1979][1992] is the Representativeness Heuristic of Kahneman and Tversky [1972]. That heuristic is that subjects focus of what we call the sample frequency of blue balls, and ignore other aspects relevant to the Bayesian prediction (the prior, and the sample size).

Griffin and Tversky [1992] developed a remarkable design intended to shed light on the determinants of confidence in subjective judgments about binary events. They posited that confidence was an interaction between the “strength” and “weight” of some evidence: in our context, strength is just the sample frequency of blue balls, and weight is just the sample size. For a binary event, strength is lowest when the sample frequency is 0.5, and higher when it is closer to 0 or 1. Our discussion of individual results displayed in Figure 7 through 10 referred repeatedly to these two factors. Griffin and Tversky [1992] hypothesize that when strength is high and weight is low, again for a binary event, overconfidence will result; and when strength is low and weight is high, underconfidence will result. Their primary design chose combinations of strength and weight that generated the same posterior probability of the binary event, allowing a direct test of their hypotheses about these determinants of confidence.29 Again, every example they present involved a subjective probability judgment about a binary event. And all elicitations were unincentivized.

The issue with their analysis is that the word confidence is actually used to refer to the elicited subjective probability, and “overconfidence” refers what we called the positive bias of that subjective probability in relation to the posterior probability.30 This might certainly be a colloquial meaning of the word confident, in the sense that “she is confident that it will rain today,” but it is not confidence in the sense that we have used the term, to mean the precision of one’s beliefs. In one instance they elicited a

---

29 A modern version of these experiments is provided by Antoniou, Harrison, Lau and Read [2015][2017].
30 To be explicit, consider their first example. On page 415 in the text they refer to the “probability judgment” and in the corresponding Table 1 this is labeled “confidence.”
subjective probability of some event, and then also asked subjects to indicate “their confidence in each prediction (on a scale from 50% to 100%)” (p. 424). Although this is conceptually what we mean by confidence, this hypothetical scale is not incentive-compatible.

Hence Griffin and Tversky [1992] can be viewed as presenting an alternative theory to Bayes Rule in order to make predictions about what we call the bias. And also to be presenting a theory which generalizes the Representativeness Heuristic of Kahneman and Tversky [1972], which was about bias, but just looked at “strength” and posited that “weight” was ignored.

Gigerenzer and Hoffrage [1995] argue that much of the evidence against Bayes Rule for subjective beliefs about the outcomes of binary events arises because of the artefactual manner in which humans are provided the key information about priors and sampling variability (in the form of diagnostic errors of tests). In many experiments this information is literally provided using probabilities. Their hypothesis is that subjects would better approximate Bayes Rule if the information is presented in the form of “natural frequencies,” which are frequencies that arise from sequentially updating event frequencies through “natural sampling” processes. The manner in which we draw blue or orange balls in our experiment was designed to mimic just this process: how one might naturally sample one ball at a time from a physical urn. This is not the same as telling someone that an event occurs a certain number of times out of a 100, which is just another way of stating a probability.

In their experiments Gigerenzer and Hoffrage [1995] presented subjects with information that is formally sufficient to infer a posterior probability a binary event. The information has a standard exposition: there is some test that has a certain false positive rate and a certain false negative rate (e.g., a medical test), there is a base rate of the condition (e.g., a medical condition), and there is someone that has tested positive (e.g., been told that the test says that they have the medical condition). The posterior probability then refers to the probability that the person has the condition given all this information. Subjects were given the information in probability formats or in natural frequency formats. They are
asked to state the probability that the person has the condition, or to state the X and Y that completes the expression “how many of a representative sample of people that had tested positive have the condition? X out of Y.”

Set aside the manner in which Gigerenzer and Hoffrage [1995] score consistency with Bayes Rule, in claiming that use of the natural frequency format dramatically improves consistency. And set aside the lack of any incentives for these reports of subjective judgment. For our purposes the central point is that they claim reduced bias in the reports of subjective probabilities, where bias is correctly measured relative to the outcome that has posterior probability greater than $\frac{1}{2}$. Their design was not intended to address confidence, nor does it, since it focused on subjective probabilities over binary events.

C. Overconfidence

When we say that an individual exhibits too much confidence in their subjective beliefs about an event, what do we mean? Some have defined overconfidence to mean overestimation of one’s ability or performance; others have defined overconfidence to mean overplacement of one’s self relative to others; and yet others have defined overconfidence to mean excessive certainty about the accuracy of one’s belief.31 Our focus has been on the last of these definitions, the notion of overprecision of beliefs, and we choose to name it Bayesian Overconfidence to stress the theoretical basis for the notion of excess confidence.

The Three Faces of Overconfidence

One definition of overconfidence considers overestimation of one's absolute ability or

31 Moore and Healy [2008] and Merkle and Weber [2011] carefully tease these three notions apart in the literature.
performance on a task. For example, someone might take an exam scored out of 100, and be asked what they believe their score is. If the reported belief exceeds the actual score, ignoring perfect scores for the moment, overconfidence in this sense is said to exist. To assess statistical significance, the same person is asked for several such scores across several domains, and the systematic overestimation then assessed over that sample. But someone might be overconfident in some domain but underconfident in other domains, and average out to being well calibrated in terms of confidence.

Another definition considers overplacement of one’s relative ability or performance on a task, also known as the “better than average” effect or, more deliciously, the “Lake Wobegon effect,” where everyone is smarter than average. The classic example is from Svenson [1981], where a majority of drivers in a well-defined group viewed themselves as more skillful and less risky than the average driver in that group.

A third definition is the one that we employ, and refers to overprecision of one’s subjective beliefs about some event. We are careful to define overprecision relative to some appropriate level of precision, recognizing that sometimes we should have less precise beliefs about some event. The simple example used in our experiment provides a standard example. Consider draws from an urn with a number of blue or orange balls, and the belief over the fraction of blue balls. One person observes a small sample drawn from the urn with replacement, and another person observes a large sample. Assuming that both start with the same diffuse prior, the person with the small sample will, in general, have a Bayesian posterior distribution over the fraction of blue balls that has a higher variance than the person with the large sample. So what one means by being “overconfident” in the sense of overprecision depends on which sample was observed.

---

32 We set aside for the moment the manner in which this belief is elicited, since that is not critical to understanding these definitions.
Is There Only One Interesting Definition of Overconfidence?

Apart from overprecision of beliefs being interesting by itself, it is not clear how one can make inferences about statistically significant overestimation or overplacement without knowing how much subjective precision the individual has about those things. Beliefs about overestimation and overplacement are only ever elicited in an incentive-compatible way as point estimates, so sufficient subjective imprecision in beliefs makes these biases in point estimates less interesting. For example, someone might have a modal belief that they performed better than the average or median, but have sufficient imprecision in those beliefs such that any reasonable confidence interval or credible interval includes the belief that they performed no better than the average or median.

To be explicit, consider a task in which someone has just completed a popular test of “fluid intelligence,” the Raven Advanced Progressive Matrices (RAPM) test documented by Raven, Raven and Court [1998]. The primary component of this test consists of 36 problems illustrated by a block of 9 images, arrayed as a 3×3 matrix, with one image deleted. The subject is presented with 8 possible solutions to fill in the deleted image, and should select the correct image. Assume that a sample of 55 subjects gets an average score of 18 correct, with only a $5 show-up fee as reward; this is a fact.

Now consider a question that might be posed to a new subject who has just completed the same task with the same rewards, but not yet been told the fraction of questions she got correct: “Compared to 55 other students from this university, how well do you think you did in relation to the average score they achieved?” Imagine two ways that the response to this question might be elicited. In one variant the subject is asked to pick one of 9 possibilities, and in another variant the subject is asked to allocate 100 tokens over the 9 possibilities. The 9 possibilities are: over 3 less, 3 less, 2 less, 1 less, the same, 1 more, 2 more, 3 more, and over 3 more. Assume the subject is incentivized to respond truthfully, and does: in the first variant she gets $25 if she is correct and $0 otherwise, and in the second variant she receives comparable payoffs generated by a QSR and is known to be risk-neutral. In
the first variant she answers “1 more.” In the second variant she allocates 12 tokens to the “1 more” possibility and 11 tokens to each of the others, so the modal response is “1 more” than the average. In the first case the subject rationally selected her modal belief. Knowing her almost-completely-diffuse subjective belief distribution from the second variant, we would hopefully be loathe to brand her as overestimating her score in any statistically meaningful manner.

Now extend the example by allowing 100 new subjects to respond in exactly the same manner. Assume that the researcher only asks the first variant, and observes 100 data points declaring that they all scored better than the average. Without the benefit of the responses to the second variant, the researcher concludes that the sample overestimates their fluid intelligence and is overconfident by that measure. The problem, we suggest, is that the researcher has ignored the subjective confidence of the individual, and hence the confidence of the sample of 100. Without some information on the confidence of the subjective beliefs of the individual, responses that only elicit modal subjective beliefs are not sufficiently informative to allow one to draw meaningful conclusions. Obviously the same general point holds if we had asked the subject to rank herself compared to the previous sample, rather than estimate her score compared to theirs.

D. Additional Issues in the Modern Literature

One focus of modern literature has been the distinction between “experienced” samples and “described” samples. An experienced sample is one that accumulates through a natural, sequential sampling of event outcomes, as in our experiment. A described sample is one that is presented in the form of a probability, or an artificial frequency. Furthermore, the literature has focused on experienced samples where the sample size is selected by the individual: see Hertwig, Barron, Weber and Erev [2004] and Hertwig and Pleskac [2010] for reviews. The stylized puzzle that this literature contends with is why individual seem to choose samples that are “too small.” To an economist, this is an optimal
sampling problem familiar from job search models. One can rigorously distinguish whether an
individual is making the right sequential, dynamic decision about sampling and sample size from
whether they draw valid inferences from the sample at hand (e.g., Harrison and Morgan [1990]). As
important as endogenous sampling is, it confounds the testing of the consistency of behavior with
Bayesian predictions.

Another issue addressed by the modern literature on behavior consistent with Bayes Rule is the
use of incentivized and incentive-compatible elicitation procedures for subjective probabilities.33 With
two exceptions, none of these studies consider confidence of beliefs by an individual, for which one
needs comparable methods to elicit subjective probability distributions.34 The exceptions are Moore and
Healy [2008] and Merkle and Weber [2011], who also make the correct point that identifying
overconfidence in the sense used here requires eliciting a subjective probability distribution. Each study
had subjects complete a quiz or series of quizzes about some topic (e.g., science or geography), and
elicited beliefs about performance.

Moore and Healy [2008; p.508] used a QSR to elicit beliefs about which of 11 integer scores
between 0 and 10 the subject would get prior to taking the quiz, but after knowing the topic.35 No
adjustments for risk attitudes are made to the reported beliefs, so subjects are assumed to behave as if
they are risk neutral. Subjects were explicitly recruited to an experiment dealing with “trivia quizzes.”
For each of 6 topics, there were 3 banks of 10 questions rated easy, medium or hard in terms of

33 For example, Holt and Smith [2009], Offerman et al. [2009] and Antoniou et al. [2015][2017].
34 There is a large literature on the significance of disagreement across elicited point forecasts of
different individuals as a measure of uncertainty in the forecast, which might be viewed as a measure of
confidence. These are different things, forced together solely because elicited distributions have not been
available, as explained well by Zarnowitz and Lambros [1987] and Engelberg, Manski and Williams [2009].
And the context for this literature has not been Bayesian updating.
35 Since their subjects were incentivized in both the belief elicitation task and the subsequent quiz,
there is an obvious concern of a portfolio effect (e.g., report that I am stupid and will get a score of only 0, 1
or 2, to get a high payoff in the belief elicitation, then pick the clearly incorrect option in the quiz to ensure
that my score is as close to this low range as possible). The rewards for the quiz were orders of magnitude
higher than the maximal rewards for the belief elicitation, and this “sandbagging” strategy appears not to
have been used from ex post inspection of individual responses (footnote 8, p. 508).
difficulty. The order of the 18 quizzes was random, although each 3-round block contained one quiz of each difficulty level. Coming into the belief elicitation stage in each round, the subject knew neither the topic nor the difficulty level, and had to form a belief over their likely score. This is a challenging task, even if the subject was aware that there could be 6 topics and 3 levels of difficulty.\footnote{At the first level the subject has to determine what these levels of difficulty mean, a legitimate issue of semantics. At the second level the subject also has to determine what the topic means: is “geography” country-specific or global? At the third level, “do I know much about these things if asked a lot of representative questions given my understanding of the topic and level of difficulty?” And then the final level, “how representative of my understanding will a sample of 10 questions be?” Once these determinations have been made for a given round, there is an issue of potential updating over the 16 rounds. If the subject was told the design (6 topics and 3 levels of difficulty per topic), or developed some sense of it from the initial rounds, would they have revised their semantic interpretations in some way?} Their interest was in relating evidence of overestimation, overplacement and overprecision to each other. Overprecision was measured by taking the integer scores for an individual in each round that best approximated their reported 90% confidence interval, and comparing that to their ex post accuracy in that round after taking the quiz: this is how the previous literature had measured overprecision, but by just using incentivized judgements about the subject’s 90% confidence interval instead of incentivized belief reports. They find that the actual score fell within this 90% confidence interval range 73\% of the time, and deem this to be evidence of overprecision on average. In our design this is akin to comparing the reported belief to the \textit{population probability}, rather than the posterior probability given the \textit{sample evidence}.

Merkle and Weber [2011; p. 268] employed a similar design, and elicited beliefs using a QSR over the ranking that each subject expected to receive in a quiz in groups of 10. Although subjects were incentivized in the QSR, no documentation of the rewards or instructions are provided. Again, no adjustments for risk attitudes are made to the reported beliefs, so subjects are assumed to behave as if they are risk neutral. In contrast to Moore and Healy [2008], these subjects had already completed the unincentivized quiz, presumably knew their individual score, and were just forming beliefs about their rank.

Overprecision is measured in relation to “the uniform distribution implied by Bayesian...
updating” (p.269). Although it is not clear how this is related to Bayesian updating, this prediction arises because someone has to be in each rank, so if one has a diffuse prior about the group one is in, the correct response would have been to state that there is a 10% chance of each rank.\textsuperscript{37} In fact, to take average reports for the Intelligence quiz as an example (Table 4, p. 268), probabilities from lowest to best rank were 1.4\%, 2.0\%, 4.5\%, 7.3\%, 12.0\%, 14.0\%, 18.9\%, 18.5\%, 12.9\% and 8.5\%, respectively. Hence subjects were, on average and at the individual level, more precise than a uniform distribution, so deemed to be overconfident.

6. Conclusion

We provide the first evidence of overconfidence in the application of Bayes Rule, defined as having subjective belief distributions that understate the variance of the chance of an outcome relative to the Bayesian posterior variance. This conclusion is the result of carefully controlling for the appropriate level of precision in beliefs at the level of the individual as predicted by Bayes Rule. We find that subjects tend to exhibit no statistically significant evidence of bias, in the sense that the mean of the reported belief is close to the mean of the Bayesian posterior distribution, but do clearly tend to exhibit Bayesian Overconfidence, in the sense that the variance of reported beliefs are much smaller than the variance of the Bayesian posterior distribution.

\textsuperscript{37} This logic is explained (p. 264/5) as follows: “The experimental setting proposed here is to ask subjects directly for the probabilities with which they would place themselves in the different quantiles (e.g., deciles). This avoids the complication of different aggregation methods and preserves the additional information coming from people’s distributional beliefs. The setup imposes clear restrictions on what is possible under rational Bayesian updating. Posterior probabilities calculated by Bayes’ law, weighted by their occurrence, must add up to the relative frequencies within the population. In a quantile framework these real probabilities are simply defined by the chosen partition of the scale: for instance, for any decile, there are 10\% of the population who belong to that decile.” In incentivized experiment using a QSR deciles were replaced by ranks, which are equivalent in groups of 10. So stated, the prediction seems to be a matter of definition rather than any updating process.
Figure 1: Bias and Confidence

A. Unbiased Excess Confidence

B. Unbiased Insufficient Confidence

C. Biased Excess Confidence

D. Biased Insufficient Confidence

E. Biased Appropriate Confidence

F. Unbiased Appropriate Confidence
Figure 2: Initial Display of Sample Draws in Period 1

Figure 3: Final Display of Sample Draws in Period 1
Figure 4: Initial Display of Belief Elicitation in Period 1

Figure 5: Final Display of Belief Elicitation in Period 1
Figure 6: Final Display of Earnings after Period 4

The urn is composed of 50 Blue balls and 50 Orange balls. So the percentage of Blue balls in the urn is 50.0%.

Player 1: [Blue balls] 65.0%
Player 2: [Orange balls] 42.6%
Player 3: [Orange balls] 37.5%
Player 4: [Blue balls] 50.0%
Player 5: [Orange balls] 50.0%

The observed percentage of Blue balls across everyone is 49.0%.

The experiment is over. You have earned $30.29 in this task.
Figure 7: Results for Subject #36 and Subject #39
Figure 8: Results for Subject #49 and Subject #50
Figure 9: Results for Subject #53 and Subject #112

Subject 53

Period 1
1/4 = 0.25

Period 2
5/12 = 0.42

Period 3
9/12 = 0.75

Period 4
8/12 = 0.67

Subject 112

Period 1
3/16 = 0.19

Period 2
1/8 = 0.13

Period 3
0/8 = 0.00

Period 4
1/8 = 0.13
Figure 10: Results for Subject #116 and Subject #9

Subject 116

Period 1
0/4 = 0.00

Period 2
3/12 = 0.25

Period 3
2/12 = 0.17

Period 4
2/12 = 0.17

Subject 9

Period 1
4/4 = 1.00

Period 2
7/12 = 0.58

Period 3
8/12 = 0.67

Period 4
11/12 = 0.92
Figure 11: Estimates of Bias

Average subjective report divided by the posterior mean for that period and subject

Posterior based on a Diffuse prior in period 1

Estimated coefficients from interval regression model
Figure 12: Estimates of Confidence

Subjective report $\sigma$ compared to the posterior $\sigma$ for that period and subject
Posterior based on a Diffuse prior in period 1
Estimated coefficients from interval regression model
Figure 15: Effects of Having a Business Major on Bias and Confidence

<table>
<thead>
<tr>
<th>Type of Effect</th>
<th>Bias</th>
<th>Confidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Marginal effect of Having a Business Major in period 1</td>
<td>-0.1</td>
<td>-0.1</td>
</tr>
<tr>
<td>Marginal effect of Having a Business Major in period 2</td>
<td>0.05</td>
<td>0.05</td>
</tr>
<tr>
<td>Marginal effect of Having a Business Major in period 3</td>
<td>0.1</td>
<td>0.1</td>
</tr>
<tr>
<td>Marginal effect of Having a Business Major in period 4</td>
<td>-0.05</td>
<td>-0.05</td>
</tr>
<tr>
<td>Total effect of Having a Business Major in period 1</td>
<td>-0.2</td>
<td>-0.2</td>
</tr>
<tr>
<td>Total effect of Having a Business Major in period 2</td>
<td>0.2</td>
<td>0.2</td>
</tr>
<tr>
<td>Total effect of Having a Business Major in period 3</td>
<td>-0.5</td>
<td>-0.5</td>
</tr>
<tr>
<td>Total effect of Having a Business Major in period 4</td>
<td>0.4</td>
<td>0.4</td>
</tr>
</tbody>
</table>

Effect and 95% Confidence Interval

Figure 16: Effects of Back-Loaded Sample on Bias and Confidence

<table>
<thead>
<tr>
<th>Type of Effect</th>
<th>Bias</th>
<th>Confidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Marginal effect of Back-Loaded Sample in period 1</td>
<td>-0.15</td>
<td>-0.15</td>
</tr>
<tr>
<td>Marginal effect of Back-Loaded Sample in period 2</td>
<td>0.1</td>
<td>0.1</td>
</tr>
<tr>
<td>Marginal effect of Back-Loaded Sample in period 3</td>
<td>-0.2</td>
<td>-0.2</td>
</tr>
<tr>
<td>Marginal effect of Back-Loaded Sample in period 4</td>
<td>0.15</td>
<td>0.15</td>
</tr>
<tr>
<td>Total effect of Back-Loaded Sample in period 1</td>
<td>-0.3</td>
<td>-0.3</td>
</tr>
<tr>
<td>Total effect of Back-Loaded Sample in period 2</td>
<td>0.3</td>
<td>0.3</td>
</tr>
<tr>
<td>Total effect of Back-Loaded Sample in period 3</td>
<td>-0.4</td>
<td>-0.4</td>
</tr>
<tr>
<td>Total effect of Back-Loaded Sample in period 4</td>
<td>0.45</td>
<td>0.45</td>
</tr>
</tbody>
</table>

Effect and 95% Confidence Interval
References


Glaser, Markus; Langer, Thomas, and Weber, Martin, “True Overconfidence in Interval Estimates:
Evidence Based on a New Measure of Miscalibration,” *Journal of Behavioral Decision Making*, 26, 2013, 405-417.


Appendix A: Instructions and Parameters (Online Working Paper)

Clean, formatted versions of these instructions are available on request.

These instructions were presented in the order shown here.

Videos for each instruction were presented to subjects, to ensure that session-specific effects were minimized. Links to these MP4 files are here, in order:

https://www.dropbox.com/s/bleg1fargxhu5bd/Instructions%20for%20Risky%20Lottery%20Choices%20with%20Compound%20Lotteries.mp4?dl=0

https://www.dropbox.com/s/9ucegs8kj9e5jkj/Belief%20Instructions%20--%20Generic.mp4?dl=0

https://www.dropbox.com/s/ppj4t4nu882q566/Urns%20Instructions%20and%20Live%20Demo.mp4?dl=0
A. Risk Aversion Task

**Choices Over Risky Prospects**

This is a task where you will choose between prospects with varying prizes and chances of winning each prize. You will be presented with a series of pairs of prospects where you will choose one of them. For each pair of prospects, you should choose the prospect you prefer. You will actually get the chance to play one of these prospects for earnings, and you will be paid according to the outcome of that prospect, so you should think carefully about which prospect you prefer on each decision screen.

Here is an example of what the computer display of such a pair of prospects will look like.

The outcome of the prospects will be determined by the draw of a random number between 1 and 100. Each number between, and including, 1 and 100 is equally likely to occur. In fact, you will be able to draw the number yourself using two 10-sided dice.

You might be told your cash endowment for each decision at the top of the screen. In this example it is $35, so any earnings would be added to or subtracted from this endowment. The endowment may change from choice to choice, so be sure to pay attention to it. The endowment you are shown only applies for that choice.

In this example the left prospect pays twenty-five dollars ($25) if the number drawn is between 1 and 5, pays negative five dollars ($-5) if the number is between 6 and 55, and pays negative thirty-five dollars ($-35) if the number is between 56 and 100. The blue color in the pie chart corresponds to 5% of the area and illustrates the chances that the number drawn will be between 1 and 5 and your prize will be $25. The orange area in the pie chart corresponds to 50% of the area and illustrates the chances that
the number drawn will be between 6 and 55 and your prize will be $-5. The green area in the pie chart corresponds to 45% of the area and illustrates the chances that the number drawn will be between 56 and 100. When you select the decision screen to be played out the computer will confirm the die rolls that correspond to the different prizes.

Now look at the pie on the right. It pays twenty-five dollars ($25) if the number drawn is between 1 and 15, negative five dollars ($-5) if the number is between 16 and 25, and negative thirty-five dollars ($-35) if the number is between 26 and 100. As with the prospect on the left, the pie slices represent the fraction of the possible numbers which yield each payoff. For example, the size of the $25 pie slice is 15% of the total pie.

Even though the screen says that you might win a negative amount, this is actually a loss to be deducted from your endowment. So if you win $-5, your earnings would be $30 = $35 - $5.

Each pair of prospects is shown on a separate screen on the computer. On each screen, you should indicate which prospect you prefer to play by clicking on one of the buttons beneath the prospects.

Some decision screens could also have a pair of prospects in which one of the prospects will give you the chance for “Double or Nothing.” For instance, the right prospect in this screen image pays “Double or Nothing” if the Green area is selected, which happens if the number drawn is between 51 and 100. The right pie chart indicates that if the number is between 1 and 50 you get $10. However, if the number is between 51 and 100 we will flip a coin with you to determine if you get either double the amount or $0. In this example, if it comes up Heads you get $40, otherwise you get nothing. The prizes listed underneath each pie refer to the amounts before any “Double or Nothing” coin toss.

After you have worked through all of the pairs of prospects, please wait quietly until further
instructions. When it is time to play this task out for earnings, you will then roll two 10-sided dice until a number comes up to determine which pair of prospects will be played out. If there are 40 pairs we will roll the dice until a number between 1 and 40 comes up, if there are 80 pairs we will roll until a number between 1 and 80 comes up, and so on. Since there is a chance that any of your choices could be played out for real, you should approach each pair of prospects as if it is the one that you will play out. Finally, you will roll the two ten-sided dice to determine the outcome of the prospect you chose, and if necessary we will then toss a coin to determine if you get “Double or Nothing.”

Here is an example: suppose your first roll was 81. We would then pull up the 81st decision that you made and look at which prospect you chose – either the left one or the right one. Let’s say that the 81st lottery was the same as the last example, and you chose the left prospect. If the random number from your second roll was 37, you would win $0; if it was 93, you would get $20.

If you picked the prospect on the right and drew the number 37, you would get $10; if it was 93, we would have to toss a coin to determine if you get “Double or Nothing.” If the coin comes up Heads then you would get $40. However, if it comes up Tails you would get nothing from your chosen prospect.

It is also possible that you will be given a prospect in which there is a “Double or Nothing” option no matter what the outcome of the random number. This screen image illustrates this possibility.

In summary, your payoff is determined by five things:

• by your endowment, if there is one, shown at the top of the screen;
• by which prospect you selected, the left or the right, for each of these pairs;
• by which prospect pair is chosen to be played out in the series of pairs using the two 10-sided
dice;
• by the outcome of that prospect when you roll the two 10-sided dice; and
• by the outcome of a coin toss if the chosen prospect outcome is of the “Double or Nothing” type.

Which prospects you prefer is a matter of personal choice. The people next to you may be presented with different prospects, and may have different preferences, so their responses should not matter to you or influence your decisions. Please work silently, and make your choices by thinking carefully about each prospect.

All payoffs are in cash, and are in addition to the $5 show-up fee that you receive just for being here, as well as any other earnings in other tasks from the session today.
B. Generic Belief Elicitation Task

Instructions

This is a task where you will be paid according to how accurate your beliefs are about certain things. You will be presented with some questions and asked to place some bets on your beliefs about the answers to each question. You will be rewarded for your answer to one of these questions, so you should think carefully about your answer to each question. The question that is chosen for payment will be determined after you have made all decisions, and that process is explained below.

Here is an example of what the computer display of a question might look like. We pick a question that is not going to be asked of you, just for illustration.

The display on your computer will be larger and easier to read. You have 10 sliders to adjust, shown at the bottom of the screen, and you have 100 tokens to allocate across the sliders. Each slider allows you to allocate tokens to reflect your belief about the answer to this question. You must allocate all 100 tokens, and in this example we start with 0 tokens allocated to each slider. As you allocate tokens, by adjusting sliders, the payoffs displayed on the screen will change. Your earnings are based on the payoffs that are displayed after you have allocated all 100 tokens.

You can earn up to $50 in this task.

Where you position each slider depends on your beliefs about the correct answer to the
question. Note that the bars above each slider correspond to that particular slider. In our example, the tokens you allocate to each bar will naturally reflect your beliefs about the proportion of left-handed Presidents. The first bar corresponds to your belief that the proportion is between 0% and 9%. The second bar corresponds to your belief that the proportion is between 10% and 19%, and so on. Each bar shows the amount of money you could earn if the true proportion is in the interval shown under the bar.

To illustrate how you use these sliders, suppose you think there is a fair chance the true answer is just under 50%. Then you might allocate the 100 tokens in the following way: 50 tokens to the interval 40% to 49%, 40 tokens to the interval 30% to 39%, and 10 tokens to the interval 20% to 29%. So you can see in this picture that if indeed the proportion of left-handed Presidents is between 40% and 49% you would earn $39.50. You would earn less than $39.50 for any other outcome. You would earn $34.50 if the proportion of left-handed Presidents is between 30% and 39%, $19.50 if it is between 20% and 29%, and for any other proportion you would earn $14.50.

You can adjust the allocation as much as you want to best reflect your personal beliefs about the proportion of left-handed Presidents.

Your earnings depend on your reported beliefs and, of course, the true answer. For instance, suppose you allocated your tokens as just shown. The true answer is that there are 8 left-handed Presidents out of 44, so the true proportion is 18.2%, and we would round that to 18%. So if you had reported these beliefs, you would have earned $14.50.
Suppose you had put all of your eggs in one basket, and allocated all 100 tokens to the interval corresponding to a proportion between 10% and 19%. Then you would have faced the earnings outcomes shown here:

Note the “good news” and “bad news” here. Since the proportion of left-handed Presidents is indeed between 10% and 19%, you earn the maximum payoff, shown here as $50. But if the true proportion had been 20%, you would have earned nothing in this task.

It is up to you to balance the strength of your personal beliefs with the possibility of them being wrong. There are three important points for you to keep in mind when making your decisions:

• First, your belief about the correct answer to each question is a personal judgment that depends on the information you have about the topic of the question.
• Second, depending on your choices and the correct answer you can earn up to $50.
• Third, your choices might also depend on your willingness to take risks or to gamble.

The decisions you make are a matter of personal choice. Please work silently, and make your choices by thinking carefully about the questions you are presented with.

When you are satisfied with your decisions, you should click on the Submit button and confirm your choices. When you are finished we will roll dice to determine which question will be played out.
The experimenter will record your earnings according to the correct answer and the choices you made.

All payoffs are in cash, and are in addition to the show-up fee that you receive just for being here, as well as any other earnings in the session today.

We will now have a video demonstration of how you make decisions in this task, using the same hypothetical example. You can then raise your hand if you need more explanation, or replay these instructions. The actual questions we will ask you to state your beliefs about will be presented after we explain the next task. Here is the video demonstration...
C. Urn-Specific Belief Elicitation Task

Your Beliefs About Draws from an Urn

This is a task where you will observe the computer making some draws of colored balls from an urn. The urn contains 100 balls, and each ball is either blue or orange. You do not know the percentage of blue and orange balls in the urn, but each draw you observe gives you information about the percentages. You will be asked to state your beliefs about the true percentage of blue balls in the urn. The true percentage does not change during your session.

You will be shown some initial draws from the urn, and then report your beliefs. Then we will provide you with three more draws from the urn, and in each case ask you to update your beliefs about the true fraction of one of the colors. Your reports will earn you money, which will be determined at the end of this task when we tell you the true fraction. You will earn money for all four of your reports about beliefs. Each of your reports will pay between $0.00 and $12.50, so your earnings in this task will range between $0.00 and $50.00. Your earnings in this task will be added to any other earnings from any other tasks, as well as your participation payment.

At the end of these instructions we will provide a video demonstration of how the task proceeds.

There will be four Periods. In Period 1 the first draw from the urn will look like this:

You see that the urn has balls with two colors: blue and orange. The first draw is blue, as shown here. The blue ball will then be dropped back into the urn, since these are draws with replacement of the ball that is drawn.
Each ball will be drawn one at a time, and a running total of each color shown on your screen. Once we have drawn all of the balls you will be told that “Draws complete” and the total number of blue and orange balls. In this case there were 10 draws, but this might vary from person to person. The total number of draws of each color in Period 1 is also shown in the table on the right of the screen.

At this point you can click, when you are ready, to “Start beliefs.” This is where you report your beliefs about the true fraction of blue balls in the urn. This stage, where you report your beliefs and get rewarded, has already been explained to you.
For example, you might have reported the following beliefs. When all 100 tokens are allocated, you are asked to “Submit” them, and then “Confirm” them. You then proceed to Period 2 and when you click the “Start draws” button, you will watch a new set of draws.

In Period 2 there are draws just as in Period 1, and then you are asked to update your beliefs. This will be the same in Period 3 and Period 4. The number of draws in these Periods may or may not be the same as Period 1, and they may also change from Period to Period. This display shows the draws that this subject observed over all four Periods, along with a total of the draws over All Periods.
So in this case the draws in Period 2 were 3 blue and 7 orange, in Period 3 they were 8 blue and 2 orange, and in Period 4 they were 3 blue and 7 orange. Over all four periods there were 20 blue balls and 20 orange balls drawn.

Once you finish all periods, you must wait for everyone else to finish. When everyone is finished you are shown the draws that everyone else observed, as well as your own, in a display like this:

![Display of draws](image)

Your draws are shown in the thin black box in the top left corner. The correct composition of the urn is displayed at the top. In this example we had an urn with 50 blue balls and 50 orange balls, so the true percentage of blue balls was 50%. Underneath the display of everyone’s draws, we tell you the observed percentage of blue balls across everyone’s draws, which in this case was 49%.

In the front of the room on the whiteboard we have an envelope. This envelope contains the true percentage of blue balls used for the urn. At the end of the session we will open this envelope and show you the number, so that you can verify it matches the true percentage shown at the top of the payment screen. We do this so that you know we are not changing the true percentage during the session.

On the right of the screen you are then told your earnings from each of the beliefs you reported. Recall that you are rewarded by the earnings for the bin that contains the true fraction. In this instance the subject earned a total of $30.29 for this task.

In summary, your payoff in this task is determined by two things:

- which beliefs you reported in each of the four Periods, based on the draws you observed; and
- the true percentage of blue balls in the urn.

Your beliefs may be different than those of the other people in the room, since you may have different draws than they have, and you might form different beliefs from the draws even if they were
the same. You are not competing against anyone else in the room for money. We will vary the true percentage from session to session, so the true percentage used in one session will not be the same true percentage in the prior session or the next session.

Please work silently, and make your choices by thinking carefully about your beliefs. We now provide a brief demonstration of the flow of the task.
D. Lotteries for Risk Aversion Task

Each subject was asked to make choices for each of 70 pairs of lotteries in the gain domain, designed to provide evidence of risk aversion as well as the tendency to make decisions consistently with EUT or RDU models. The battery is based on designs from Loomes and Sugden [1998] to test the Compound Independence Axiom (CIA), designs from Harrison, Martínez-Correa and Swarthout [2015] to test the ROCL axiom, and designs from Harrison, Lau, Ross and Swarthout [2017], based on Cox and Sadiraj [2008], to test the premiss that subjects exhibit “small stakes risk aversion.” Each subject faced an individually randomized sequence of choices from this 70. Table A1 lists the names of the lottery pairs and the prizes and probabilities. All prizes are stated in U.S. dollars.

To evaluate RDU preferences for individuals we estimate an RDU model for each individual, following procedures explained in Harrison and Rutström [2008]. The utility of income \( x \) in an elicitation is assumed to be defined by the CRRA function \( U(x) = \frac{x^{1-s}}{1-s} \), where \( s \neq 1 \) is a parameter to be estimated. The assumed probability weighting function is a general functional form proposed by Prelec [1998] that exhibits considerable flexibility. This function is \( \omega(p) = \exp\{-\eta\cdot(-\ln p)^\varphi\} \), and is defined for \( 0<p\leq1, \eta>0 \) and \( \varphi>0 \). When \( \varphi=1 \) this function collapses to the Power function \( \omega(p) = p^\eta \).

Additional References


### Table A1: Parameters for Risk Aversion Lottery Battery

See text for explanation of lottery names and variables

<table>
<thead>
<tr>
<th>qid</th>
<th>prize1L</th>
<th>prob1L</th>
<th>prize2L</th>
<th>prob2L</th>
<th>prize3L</th>
<th>prob3L</th>
<th>prize4L</th>
<th>prob4L</th>
<th>prize1R</th>
<th>prob1R</th>
<th>prize2R</th>
<th>prob2R</th>
<th>prize3R</th>
<th>prob3R</th>
<th>prize4R</th>
<th>prob4R</th>
</tr>
</thead>
<tbody>
<tr>
<td>ls10_lr</td>
<td>0</td>
<td>0</td>
<td>10</td>
<td>.5</td>
<td>30</td>
<td>0</td>
<td>50</td>
<td>.5</td>
<td>0</td>
<td>0</td>
<td>10</td>
<td>.1</td>
<td>30</td>
<td>.8</td>
<td>50</td>
<td>.1</td>
</tr>
<tr>
<td>ls13_r1</td>
<td>0</td>
<td>0</td>
<td>10</td>
<td>.5</td>
<td>30</td>
<td>.4</td>
<td>50</td>
<td>.1</td>
<td>0</td>
<td>0</td>
<td>10</td>
<td>.7</td>
<td>30</td>
<td>0</td>
<td>50</td>
<td>.3</td>
</tr>
<tr>
<td>ls13_lr</td>
<td>0</td>
<td>0</td>
<td>10</td>
<td>.65</td>
<td>30</td>
<td>.1</td>
<td>50</td>
<td>.25</td>
<td>0</td>
<td>0</td>
<td>10</td>
<td>.55</td>
<td>30</td>
<td>.3</td>
<td>50</td>
<td>.15</td>
</tr>
<tr>
<td>ls15_r1</td>
<td>0</td>
<td>0</td>
<td>10</td>
<td>.4</td>
<td>30</td>
<td>.6</td>
<td>50</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>10</td>
<td>.5</td>
<td>30</td>
<td>.4</td>
<td>50</td>
<td>.1</td>
</tr>
<tr>
<td>ls161_lr</td>
<td>0</td>
<td>0</td>
<td>10</td>
<td>.88</td>
<td>30</td>
<td>.04</td>
<td>50</td>
<td>.08</td>
<td>0</td>
<td>0</td>
<td>10</td>
<td>.83</td>
<td>30</td>
<td>.14</td>
<td>50</td>
<td>.03</td>
</tr>
<tr>
<td>ls17_lr</td>
<td>0</td>
<td>0</td>
<td>10</td>
<td>.1</td>
<td>30</td>
<td>.6</td>
<td>50</td>
<td>.9</td>
<td>0</td>
<td>0</td>
<td>10</td>
<td>0</td>
<td>30</td>
<td>.25</td>
<td>50</td>
<td>.75</td>
</tr>
<tr>
<td>ls171_r1</td>
<td>0</td>
<td>0</td>
<td>10</td>
<td>.04</td>
<td>30</td>
<td>.15</td>
<td>50</td>
<td>.81</td>
<td>0</td>
<td>0</td>
<td>10</td>
<td>.08</td>
<td>30</td>
<td>.05</td>
<td>50</td>
<td>.87</td>
</tr>
<tr>
<td>ls18_r1</td>
<td>0</td>
<td>0</td>
<td>10</td>
<td>.1</td>
<td>30</td>
<td>.75</td>
<td>50</td>
<td>.15</td>
<td>0</td>
<td>0</td>
<td>10</td>
<td>.4</td>
<td>30</td>
<td>0</td>
<td>50</td>
<td>.6</td>
</tr>
<tr>
<td>ls181_r1</td>
<td>0</td>
<td>0</td>
<td>10</td>
<td>.14</td>
<td>30</td>
<td>.65</td>
<td>50</td>
<td>.21</td>
<td>0</td>
<td>0</td>
<td>10</td>
<td>.38</td>
<td>30</td>
<td>.05</td>
<td>50</td>
<td>.57</td>
</tr>
<tr>
<td>ls21_lr</td>
<td>0</td>
<td>0</td>
<td>10</td>
<td>.12</td>
<td>30</td>
<td>.05</td>
<td>50</td>
<td>.83</td>
<td>0</td>
<td>0</td>
<td>10</td>
<td>.03</td>
<td>30</td>
<td>.2</td>
<td>50</td>
<td>.77</td>
</tr>
<tr>
<td>ls211_r1</td>
<td>0</td>
<td>0</td>
<td>10</td>
<td>.7</td>
<td>30</td>
<td>0</td>
<td>50</td>
<td>.3</td>
<td>0</td>
<td>0</td>
<td>10</td>
<td>.6</td>
<td>30</td>
<td>.25</td>
<td>50</td>
<td>.15</td>
</tr>
<tr>
<td>ls221_lr</td>
<td>0</td>
<td>0</td>
<td>10</td>
<td>.66</td>
<td>30</td>
<td>.1</td>
<td>50</td>
<td>.24</td>
<td>0</td>
<td>0</td>
<td>10</td>
<td>.54</td>
<td>30</td>
<td>.4</td>
<td>50</td>
<td>.06</td>
</tr>
<tr>
<td>ls26_r1</td>
<td>0</td>
<td>0</td>
<td>10</td>
<td>.2</td>
<td>30</td>
<td>.6</td>
<td>50</td>
<td>.2</td>
<td>0</td>
<td>0</td>
<td>10</td>
<td>.4</td>
<td>30</td>
<td>.05</td>
<td>50</td>
<td>.6</td>
</tr>
<tr>
<td>ls281_r1</td>
<td>0</td>
<td>0</td>
<td>10</td>
<td>.12</td>
<td>30</td>
<td>.84</td>
<td>50</td>
<td>.04</td>
<td>0</td>
<td>0</td>
<td>10</td>
<td>.18</td>
<td>30</td>
<td>.66</td>
<td>50</td>
<td>.16</td>
</tr>
<tr>
<td>ls29_r1</td>
<td>0</td>
<td>0</td>
<td>10</td>
<td>.5</td>
<td>30</td>
<td>.3</td>
<td>50</td>
<td>.2</td>
<td>0</td>
<td>0</td>
<td>10</td>
<td>.6</td>
<td>30</td>
<td>0</td>
<td>50</td>
<td>.4</td>
</tr>
<tr>
<td>ls301_r1</td>
<td>0</td>
<td>0</td>
<td>10</td>
<td>.45</td>
<td>30</td>
<td>.45</td>
<td>50</td>
<td>.1</td>
<td>0</td>
<td>0</td>
<td>10</td>
<td>.55</td>
<td>30</td>
<td>.15</td>
<td>50</td>
<td>.3</td>
</tr>
<tr>
<td>ls311_lr</td>
<td>0</td>
<td>0</td>
<td>10</td>
<td>.48</td>
<td>30</td>
<td>.36</td>
<td>50</td>
<td>.16</td>
<td>0</td>
<td>0</td>
<td>10</td>
<td>.42</td>
<td>30</td>
<td>.54</td>
<td>50</td>
<td>.04</td>
</tr>
<tr>
<td>ls32_r1</td>
<td>0</td>
<td>0</td>
<td>10</td>
<td>.7</td>
<td>30</td>
<td>.3</td>
<td>50</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>10</td>
<td>.8</td>
<td>30</td>
<td>0</td>
<td>50</td>
<td>.2</td>
</tr>
<tr>
<td>ls33_r1</td>
<td>0</td>
<td>0</td>
<td>10</td>
<td>.25</td>
<td>30</td>
<td>0</td>
<td>50</td>
<td>.25</td>
<td>0</td>
<td>0</td>
<td>10</td>
<td>.25</td>
<td>30</td>
<td>.3</td>
<td>50</td>
<td>.75</td>
</tr>
<tr>
<td>ls35_r1</td>
<td>0</td>
<td>0</td>
<td>10</td>
<td>0</td>
<td>30</td>
<td>1</td>
<td>50</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>10</td>
<td>.25</td>
<td>30</td>
<td>0</td>
<td>50</td>
<td>.75</td>
</tr>
<tr>
<td>ls351_lr</td>
<td>0</td>
<td>0</td>
<td>10</td>
<td>.2</td>
<td>30</td>
<td>.2</td>
<td>50</td>
<td>.6</td>
<td>0</td>
<td>0</td>
<td>10</td>
<td>.1</td>
<td>30</td>
<td>.6</td>
<td>50</td>
<td>.3</td>
</tr>
<tr>
<td>ls361_lr</td>
<td>0</td>
<td>0</td>
<td>10</td>
<td>.04</td>
<td>30</td>
<td>.92</td>
<td>50</td>
<td>.06</td>
<td>0</td>
<td>0</td>
<td>10</td>
<td>.08</td>
<td>30</td>
<td>.68</td>
<td>50</td>
<td>.3</td>
</tr>
<tr>
<td>ls371_lr</td>
<td>0</td>
<td>0</td>
<td>10</td>
<td>.48</td>
<td>30</td>
<td>.28</td>
<td>50</td>
<td>.24</td>
<td>0</td>
<td>0</td>
<td>10</td>
<td>.44</td>
<td>30</td>
<td>.44</td>
<td>50</td>
<td>.12</td>
</tr>
<tr>
<td>ls39_r1</td>
<td>0</td>
<td>0</td>
<td>10</td>
<td>.5</td>
<td>30</td>
<td>.2</td>
<td>50</td>
<td>.3</td>
<td>0</td>
<td>0</td>
<td>10</td>
<td>.55</td>
<td>30</td>
<td>0</td>
<td>50</td>
<td>.45</td>
</tr>
<tr>
<td>ls411_r1</td>
<td>0</td>
<td>0</td>
<td>10</td>
<td>.24</td>
<td>30</td>
<td>.05</td>
<td>50</td>
<td>.68</td>
<td>0</td>
<td>0</td>
<td>10</td>
<td>.03</td>
<td>30</td>
<td>.45</td>
<td>50</td>
<td>.52</td>
</tr>
<tr>
<td>ls41_lr</td>
<td>0</td>
<td>0</td>
<td>10</td>
<td>.6</td>
<td>30</td>
<td>0</td>
<td>50</td>
<td>.41</td>
<td>0</td>
<td>0</td>
<td>10</td>
<td>0</td>
<td>30</td>
<td>1</td>
<td>50</td>
<td>0</td>
</tr>
<tr>
<td>ls77_lr</td>
<td>0</td>
<td>0</td>
<td>10</td>
<td>.6</td>
<td>30</td>
<td>0</td>
<td>50</td>
<td>.4</td>
<td>0</td>
<td>0</td>
<td>10</td>
<td>.15</td>
<td>30</td>
<td>.75</td>
<td>50</td>
<td>.1</td>
</tr>
<tr>
<td>ls71_lr</td>
<td>0</td>
<td>0</td>
<td>10</td>
<td>.54</td>
<td>30</td>
<td>.1</td>
<td>50</td>
<td>.36</td>
<td>0</td>
<td>0</td>
<td>10</td>
<td>.18</td>
<td>30</td>
<td>.7</td>
<td>50</td>
<td>.12</td>
</tr>
<tr>
<td>ls811_r1</td>
<td>0</td>
<td>0</td>
<td>10</td>
<td>.06</td>
<td>30</td>
<td>.04</td>
<td>50</td>
<td>.86</td>
<td>0</td>
<td>0</td>
<td>10</td>
<td>.86</td>
<td>30</td>
<td>.05</td>
<td>50</td>
<td>.05</td>
</tr>
<tr>
<td>rae1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>.5</td>
<td>10</td>
<td>.5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>.75</td>
<td>10</td>
</tr>
<tr>
<td>rae10</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>35</td>
<td>.75</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>.5</td>
<td>35</td>
</tr>
<tr>
<td>rae11</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>35</td>
<td>.75</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>.25</td>
<td>20</td>
</tr>
<tr>
<td>rae12</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>35</td>
<td>.75</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>.25</td>
<td>30</td>
</tr>
<tr>
<td>rae13</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>.25</td>
<td>10</td>
<td>.75</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>.75</td>
<td>10</td>
</tr>
<tr>
<td>rae14</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>20</td>
<td>.75</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>.25</td>
<td>20</td>
</tr>
<tr>
<td>ss8</td>
<td>20</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>------</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
</tr>
<tr>
<td>ss6</td>
<td>25</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>ss5</td>
<td>25</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>ss30</td>
<td>120</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>ss3</td>
<td>20</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Appendix B: Detailed Statistical Test Results (Online Working Paper)

This appendix tabulates detailed $p$-values for hypothesis tests to complement the displays in Figures 8 through 13, as well as providing results for all demographics. The first $p$-value in each line tests the hypothesis that the bias value is 1, the second $p$-value in parentheses tests the hypothesis that the confidence value is 1, and the third $p$-value in square brackets tests the hypothesis that both values are 1. As explained in the text, a value of 1 in each case means that the bias or confidence (variance) of the subjective belief distribution for each subject and period is equal to the Bayesian posterior distribution mean or variance. A $p$-value of 0.000 means that the true $p$-value is below 0.001.

- A18 -
p-value for total effect of business in period 3, assuming prior diffuse in period 1, is 0.558 (0.000) [0.000]
p-value for total effect of business in period 4, assuming prior diffuse in period 1, is 0.662 (0.042) [0.009]

p-value for marginal effect of junior in period 1, assuming prior diffuse in period 1, is 0.949 (0.604) [0.007]
p-value for marginal effect of junior in period 2, assuming prior diffuse in period 1, is 0.931 (0.774) [0.094]
p-value for marginal effect of junior in period 3, assuming prior diffuse in period 1, is 0.212 (0.389) [0.037]
p-value for marginal effect of junior in period 4, assuming prior diffuse in period 1, is 0.390 (0.093) [0.078]
p-value for total effect of junior in period 1, assuming prior diffuse in period 1, is 0.772 (0.997) [0.028]
p-value for total effect of junior in period 2, assuming prior diffuse in period 1, is 0.638 (0.301) [0.384]
p-value for total effect of junior in period 3, assuming prior diffuse in period 1, is 0.711 (0.538) [0.762]
p-value for total effect of junior in period 4, assuming prior diffuse in period 1, is 0.574 (0.032) [0.079]

p-value for marginal effect of sophomore in period 1, assuming prior diffuse in period 1, is 0.632 (0.678) [0.862]
p-value for marginal effect of sophomore in period 2, assuming prior diffuse in period 1, is 0.219 (0.763) [0.468]
p-value for marginal effect of sophomore in period 3, assuming prior diffuse in period 1, is 0.128 (0.110) [0.111]
p-value for marginal effect of sophomore in period 4, assuming prior diffuse in period 1, is 0.058 (0.028) [0.014]
p-value for total effect of sophomore in period 1, assuming prior diffuse in period 1, is 0.272 (0.510) [0.536]
p-value for total effect of sophomore in period 2, assuming prior diffuse in period 1, is 0.657 (0.858) [0.827]
p-value for total effect of sophomore in period 3, assuming prior diffuse in period 1, is 0.749 (0.918) [0.940]
p-value for total effect of sophomore in period 4, assuming prior diffuse in period 1, is 0.986 (0.991) [0.997]

p-value for marginal effect of senior in period 1, assuming prior diffuse in period 1, is 0.514 (0.122) [0.152]
p-value for marginal effect of senior in period 2, assuming prior diffuse in period 1, is 0.058 (0.976) [0.161]
p-value for marginal effect of senior in period 3, assuming prior diffuse in period 1, is 0.010 (0.364) [0.027]
p-value for marginal effect of senior in period 4, assuming prior diffuse in period 1, is 0.008 (0.085) [0.011]
p-value for total effect of senior in period 1, assuming prior diffuse in period 1, is 0.263 (0.002) [0.010]
p-value for total effect of senior in period 2, assuming prior diffuse in period 1, is 0.511 (0.412) [0.531]
p-value for total effect of senior in period 3, assuming prior diffuse in period 1, is 0.520 (0.065) [0.134]
p-value for total effect of senior in period 4, assuming prior diffuse in period 1, is 0.593 (0.064) [0.126]

p-value for marginal effect of gpaHI in period 1, assuming prior diffuse in period 1, is 0.980 (0.357) [0.520]
p-value for marginal effect of gpaHI in period 2, assuming prior diffuse in period 1, is 0.124 (0.033) [0.086]
p-value for marginal effect of gpaHI in period 3, assuming prior diffuse in period 1, is 0.312 (0.054) [0.095]
p-value for marginal effect of gpaHI in period 4, assuming prior diffuse in period 1, is 0.642 (0.667) [0.740]
p-value for total effect of gpaHI in period 1, assuming prior diffuse in period 1, is 0.478 (0.545) [0.334]
p-value for total effect of gpaHI in period 2, assuming prior diffuse in period 1, is 0.746 (0.890) [0.855]
p-value for total effect of gpaHI in period 3, assuming prior diffuse in period 1, is 0.087 (0.371) [0.231]
p-value for total effect of gpaHI in period 4, assuming prior diffuse in period 1, is 0.577 (0.155) [0.364]

p-value for marginal effect of gpaLO in period 1, assuming prior diffuse in period 1, is 0.305 (0.005) [0.009]
p-value for marginal effect of gpaLO in period 2, assuming prior diffuse in period 1, is 0.817 (0.533) [0.788]
p-value for marginal effect of gpaLO in period 3, assuming prior diffuse in period 1, is 0.782 (0.957) [0.957]
p-value for marginal effect of gpaLO in period 4, assuming prior diffuse in period 1, is 0.298 (0.248) [0.309]
p-value for total effect of gpaLO in period 1, assuming prior diffuse in period 1, is 0.350 (0.415) [0.226]
p-value for total effect of gpaLO in period 2, assuming prior diffuse in period 1, is 0.825 (0.471) [0.603]
p-value for total effect of gpaLO in period 3, assuming prior diffuse in period 1, is 0.656 (0.580) [0.828]
p-value for total effect of gpaLO in period 4, assuming prior diffuse in period 1, is 0.283 (0.508) [0.552]

p-value for marginal effect of work in period 1, assuming prior diffuse in period 1, is 0.331 (0.402) [0.443]
p-value for marginal effect of work in period 2, assuming prior diffuse in period 1, is 0.985 (0.222) [0.146]
p-value for marginal effect of work in period 3, assuming prior diffuse in period 1, is 0.195 (0.144) [0.097]
p-value for marginal effect of work in period 4, assuming prior diffuse in period 1, is 0.959 (0.017) [0.058]
p-value for total effect of work in period 1, assuming prior diffuse in period 1, is 0.298 (0.714) [0.333]
p-value for total effect of work in period 2, assuming prior diffuse in period 1, is 0.461 (0.490) [0.685]
p-value for total effect of work in period 3, assuming prior diffuse in period 1, is 0.356 (0.480) [0.492]
<table>
<thead>
<tr>
<th>Parameter Description</th>
<th>Period</th>
<th>Assumption</th>
<th>p-value</th>
<th>Standard Error</th>
<th>Lower CI</th>
<th>Upper CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total effect of work in period 4</td>
<td>Assuming prior diffuse in period 1</td>
<td>0.901 (0.520)</td>
<td>0.805</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Marginal effect of christian in period 1</td>
<td>Assuming prior diffuse in period 1</td>
<td>0.690 (0.129)</td>
<td>0.316</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Marginal effect of christian in period 2</td>
<td>Assuming prior diffuse in period 1</td>
<td>0.088 (0.108)</td>
<td>0.114</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Marginal effect of christian in period 3</td>
<td>Assuming prior diffuse in period 1</td>
<td>0.882 (0.089)</td>
<td>0.235</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Marginal effect of christian in period 4</td>
<td>Assuming prior diffuse in period 1</td>
<td>0.277 (0.287)</td>
<td>0.191</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total effect of christian in period 1</td>
<td>Assuming prior diffuse in period 1</td>
<td>0.476 (0.608)</td>
<td>0.767</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total effect of christian in period 2</td>
<td>Assuming prior diffuse in period 1</td>
<td>0.124 (0.604)</td>
<td>0.304</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total effect of christian in period 3</td>
<td>Assuming prior diffuse in period 1</td>
<td>0.179 (0.864)</td>
<td>0.362</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total effect of christian in period 4</td>
<td>Assuming prior diffuse in period 1</td>
<td>0.188 (0.479)</td>
<td>0.256</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Marginal effect of front in period 1</td>
<td>Assuming prior diffuse in period 1</td>
<td>0.606 (0.522)</td>
<td>0.803</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Marginal effect of front in period 2</td>
<td>Assuming prior diffuse in period 1</td>
<td>0.110 (0.521)</td>
<td>0.164</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Marginal effect of front in period 3</td>
<td>Assuming prior diffuse in period 1</td>
<td>0.913 (0.369)</td>
<td>0.636</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Marginal effect of front in period 4</td>
<td>Assuming prior diffuse in period 1</td>
<td>0.677 (0.044)</td>
<td>0.117</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total effect of front in period 1</td>
<td>Assuming prior diffuse in period 1</td>
<td>0.388 (0.856)</td>
<td>0.659</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total effect of front in period 2</td>
<td>Assuming prior diffuse in period 1</td>
<td>0.665 (0.815)</td>
<td>0.852</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total effect of front in period 3</td>
<td>Assuming prior diffuse in period 1</td>
<td>0.859 (0.987)</td>
<td>0.984</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total effect of front in period 4</td>
<td>Assuming prior diffuse in period 1</td>
<td>0.627 (0.902)</td>
<td>0.876</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Marginal effect of back in period 1</td>
<td>Assuming prior diffuse in period 1</td>
<td>0.153 (0.002)</td>
<td>0.006</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Marginal effect of back in period 2</td>
<td>Assuming prior diffuse in period 1</td>
<td>0.885 (0.532)</td>
<td>0.774</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Marginal effect of back in period 3</td>
<td>Assuming prior diffuse in period 1</td>
<td>0.219 (0.582)</td>
<td>0.414</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Marginal effect of back in period 4</td>
<td>Assuming prior diffuse in period 1</td>
<td>0.975 (0.014)</td>
<td>0.043</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total effect of back in period 1</td>
<td>Assuming prior diffuse in period 1</td>
<td>0.164 (0.010)</td>
<td>0.026</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total effect of back in period 2</td>
<td>Assuming prior diffuse in period 1</td>
<td>0.795 (0.653)</td>
<td>0.903</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total effect of back in period 3</td>
<td>Assuming prior diffuse in period 1</td>
<td>0.919 (0.160)</td>
<td>0.340</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total effect of back in period 4</td>
<td>Assuming prior diffuse in period 1</td>
<td>0.869 (0.178)</td>
<td>0.399</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Appendix C: Additional Figures (Online Working Paper)

**Figure C11: Estimates of Bias Using Recovered Beliefs**
Average subjective belief divided by the posterior mean for that period and subject
Posterior based on a Diffuse prior in period 1
Estimated coefficients from interval regression model

**Figure C12: Estimates of Confidence Using Recovered Beliefs**
Subjective belief $\sigma$ compared to the posterior $\sigma$ for that period and subject
Posterior based on a Diffuse prior in period 1
Estimated coefficients from interval regression model

Excess confidence
Figure C13: Effects of Age in Years on Bias and Confidence

Figure C14: Effects of Gender on Bias and Confidence