Running head: TRUST AND BASE RATE NEGLECT

Seeing is believing: Priors, trust, and base rate neglect

Matthew B. Welsh

Australian School of Petroleum

University of Adelaide

Daniel J. Navarro

School of Psychology

University of Adelaide

Abstract

Tversky and Kahneman (1973) described an effect they called 'insensitivity to prior probability of outcomes', later dubbed base rate neglect, which describes people's tendency to underweight prior information in favor of new data. As probability theory requires that prior probabilities be taken into account, via Bayes' theorem, the fact that most people fail to do so has been taken as evidence of human irrationality and, by others, of a mismatch between our cognitive processes and the questions being asked (Cosmides & Tooby, 1996). In contrast to both views, we suggest that simplistic Bayesian updating using base rates is not necessarily rational. To that end, we present experiments in which base rate neglect is often the right strategy, and show that people's base rate usage varies systematically as a function of the extent to which the data that make up a base rate are perceived as trustworthy.

Seeing is believing: Priors, trust, and base rate neglect

In the closing remarks to *A Philosophical Essay on Probabilities*, Laplace (1814/1951) argues that "the theory of probabilities is at bottom only common sense reduced to calculus; it makes us appreciate with exactitude that which exact minds feel by a sort of instinct without being able ofttimes to give a reason for it". Within probability theory, Bayes' rule provides the mechanism by which a set of prior beliefs can be updated in light of evidence, as follows: given a hypothesis, h, which we believe has some prior probability of being correct $P(h)$, if we then observed some data, x, Bayes' theorem tells us how to find $P(h|x)$, the posterior probability that h is true given that we have now seen x,

$$P(h|x)=\frac{P(x|h)P(h)}{P(x)}. \quad (1)$$

As to whether Laplace's claim provides a plausible account of human reasoning, one of the principal sources of discussion is *base rate neglect*, a phenomenon that seems to contradict the assertion that analytic probabilities are merely formalized versions of people's intuitions about chance. The general finding is that, when people are provided with prior information (in the form of a base rate) along with new evidence, they typically weight the evidence provided by the new data far more heavily than that provided by the base rates (Tversky & Kahneman, 1973). This tendency to downgrade the value of the prior relative to the likelihood is taken to imply that: firstly, Bayes' theorem does not provide a complete account of the reasoning employed by people (Villejoubert & Mandel, 2002); and, secondly, that people are therefore suboptimal or biased in their judgments, and may be regarded as acting irrationally. Note, however, that there are two distinct claims here. Clearly, underweighting the base rate information will lead people to make judgments that differ from those provided by a simplistic application of Equation 1. The charge of irrationality, however, is a stronger claim, and somewhat more questionable.

Traditional approaches to the study of human decision making have tended to assume that rational behavior is best operationalized in terms of the strict adherence to some optimal strategy calculated by the researcher in advance (as is the case in most uses of expected utility theory). Any deviations from this researcher-specified strategy are then deemed to be evidence of irrational behavior. The major problem with this approach is that the manner in which these optimal strategies are designed is often extremely impractical – most notably, no consideration is given to the costs associated with time spent and computations performed. As argued by Todd and Gigerenzer (2000), it is by no means clear that a rational actor should, in fact, expend a great deal of time and effort in computing exact solutions to complicated problems, especially when fast and simple approximations are available.

This accords well with observations such as those made by McKenzie (2003), who argues that rational models should, properly, be regarded as *theories* but not *standards* of behavior. This, it is argued, is because apparent errors observed in laboratory tasks actually tend to result from participants' use of strategies that deliver good results in real world tasks. Thus, while such strategies can be regarded as "irrational" within the context of the specific task, it might in fact be the case that the strategies that are optimal in the experiment would in fact hinder performance in real life.

Guided by these ideas, we consider the question of base rate neglect with respect to how people *should* appropriately weight base rates and novel information in order to make predictions in real environments. To do so, we present two experiments (in three parts) that manipulate the quality of different sources of data presented to people. In doing so, we depart somewhat from the classic base rate neglect approach; specifically, we design scenarios that minimize the potential computational problems (as explained below) by making explicit what information needs to be aggregated and, instead, manipulate the

apparent quality of the data. The results suggest that the strength of base rate neglect effects can be systematically manipulated by altering the trustworthiness of the data – while some people display some base rate neglect, the majority of our participants made decisions in a fashion consistent with assigning different levels of trust to different sources of evidence.

Base Rates: Stability and Neglect

The Existence and Mitigation of Base Rate Neglect

The original characterization of base rate neglect (Kahneman & Tversky, 1973) was simply that people tended to rely on a *representativeness* heuristic when making probability judgments while neglecting (i.e., underweighting) base rate information. That is, they would assign a high probability to a person being, for example, an engineer if the description they were provided sounded, stereotypically, like an engineer – while paying little attention to how many or few engineers were in the group of people from which the description was drawn. Later work, however, extended this to include examples such as the now classic "taxi-cab" examples (Bar-Hillel, 1980), where the novel information was less clearly 'representative' in the sense described by Kahneman and Tversky and the effect seemed, instead, to reflect a more general tendency to underweight base rate information when new information is presented. Thus base rate neglect was seemingly established as a cognitive bias to which people were susceptible and which, therefore, needed to be ameliorated.

Since this early work, however, research on base rate neglect has become somewhat polarized, with the 'heuristics and biases' school of thought continuing to argue that base rate neglect is robust – resulting from people's inability to update in a Bayesian manner (Kahneman & Tversky, 1996) – while others have argued that the effect disappears under experimental conditions better suited to human cognition (Cosmides & Tooby, 1996; Gigerenzer, 1996). In particular, it has been suggested that questions phrased in a frequency

format rather than in terms of probabilities are more easily dealt with by people and thus less likely to produce base rate neglect. The reasoning behind this argument is the claim that frequencies of events are easily observed whereas the probability of a single event is intrinsically unobservable (Cosmides & Tooby, 1996). As a result, people might be expected to have cognitive abilities suited to counting events and comparing frequencies rather than computing one-off probabilities. From this alternative point of view, base rate neglect effects are often regarded as artifacts of experimental designs that impose unnecessary computational costs on people.

Initial support for this idea was found by a number of researchers (Cosmides & Tooby, 1996; Gigerenzer & Hoffrage, 1995), with the strength of the base rate neglect effect diminishing when count data were used. However, subsequent research found that relative frequencies, such as percentages, gave an equal or greater reduction in base rate neglect (Harries & Harvey, 2000; Sloman, Over, Slovak, & Stibel, 2003). Indeed, Sloman et al. argued that this mitigation was not really due to a fundamental difference in how people processed probabilities and counts; instead, they suggested that it occurs because the data were presented in a manner that makes clear to participants which values need to be compared, and that the same benefit can be achieved in probability formats if the data are presented equivalently (though clearly this is a variant on the "computational costs" suggestion). Similarly, Bar-Hillel (1980) showed that neglect is reduced when the base rate's relevance to the outcome is made clear; while Ajzen (1977) demonstrated that phrasing questions in such a way as to make the causal connections between the base rate and the outcome clear does reduce base rate neglect. Recent work by Krynski and Tenenbaum (2007) expanded on the use of causal explanations in base rate neglect problems, showing that Bayesian causal models can predict base rate neglect.

The previous discussion notwithstanding, it should be noted that, even when people

are given direct experience of a sample rather than merely provided with summary statistics, base rate neglect persists (e.g., Goodie & Fantino, 1999; Gluck & Bower, 1988) and an analogous effect has been observed in pigeons (Zentall & Clement, 2002). Moreover, the general finding is that the various methods tend to reduce levels of base rate neglect rather than eliminate it. Given this, it appears unlikely that base rate neglect is entirely an artifact of the computational costs imposed by experimental design. Thus, it seems equally unlikely that the effect could be entirely avoided by changing question formats. With that in mind, it may be worth considering the nature of the information available in the problem: specifically, is there some sense in which base rate data tend to be less informative than would be implied by the naive application of Bayes' theorem?

Unstable Base Rates

In some respects, the debate over base rate neglect seems a little confusing. As the proponents of bounded rationality, Gigerenzer and Todd (1999; Todd & Gigerenzer, 2003) have argued that we should seek to understand cognitive processes in light of the environments in which they are designed to operate. Given the preponderance of data illustrating the robustness of the effect (e.g., Kahneman & Tversky, 1996), rather than attempt to force it to disappear, it seems more productive to consider the ecological factors that might produce situations where neglecting base rates is the right thing to do. This is the approach we take in this paper.

The central observation that differentiates our approach from previous approaches is in how we view the base rate. The tendency within much of the literature has been to treat base rates as if they were eternal and unchanging truths given unto people and which, therefore, it is irrational to ignore. This is not, however, a good characterization of the base rates that people are likely to encounter in real life. In particular, the informativeness of base rate data is limited, or bounded, in a number of important ways. Goodie and Fantino

(1999) touch on this, arguing that people need to be sensitive not just to base rates but also to how base rates change. In the classic "taxi-cab" problem they argue that the base rates given for taxi-cab colors and eyewitness reliability are specific to one place at one time and thus subject to change. Indeed, in advocating a relevance-based account for the base rate neglect effect, Bar-Hillel (1980) examined variants of the taxi-cab problem in which the source of information for the base rate is varied, and notes corresponding effects on the strength of the phenomenon. Fiedler (2000) and Krynski and Tenenbaum (2007) also touch on the fact that base rates, as traditionally described, do not reflect prior data as it exists in the real world: base rates require large samples whereas real-world decision making generally involves predictions made based from a limited sample. Nevertheless, Krynski and Tenenbaum (2007) treat base rates as known characteristics of the world in their experiments.

In this paper, we expand on the perceived-relevance view of base rate neglect but argue that, in many real-world situations, people may be correct to treat the base rate as irrelevant. In contrast with a great deal of previous research, we regard base rates not as unchanging characteristics of the real world but rather as *summary representations of previously collected data*. That is, any base rate information that humans have to deal with always originates somewhere and, in most cases, the source of the "base rate" is actually an older, larger data set collected under different conditions to those to which it needs to be applied.[1] With that in mind, we consider the idea that base rate neglect (in part) corresponds to the idea that these preexisting data may be less relevant than newer observations. The approach can be understood by considering the following example, adapted from one commonly used by philosophers and originating in the work of David Hume's consideration of the problem of induction (1739/1898).

Imagine that you have been calculating the proportion (base rate) of white

swans amongst the general swan population. You have been across Europe and observed 999 swans – all of which were white. You then take a plane to Australia and continue your survey. Your first observation is of a black swan. You have now observed one thousand swans and have a base rate of 99.9% for white swans. As you plan to continue your survey, what is the probability that the next swan you observe will be white?

In a naive use of statistical methods, one would expect the next swan to be white with a 99.9% probability, as the base rate indicates. Rationally, however, human decision-makers are acutely aware of the existence of regional variation, and so will tend to assume that base rates derived from ecological data collected in Europe are less likely to be predictive of new data in Australia (and vice versa). Accordingly, a belief in regional variation provides a strong justification for a decision to neglect the base rate.

In statistical terms, this reflects an oft-overlooked aspect to Bayes' theorem; namely the fact that your prior beliefs about the probability of an event, usually written P(A), should in fact be conditioned on all of your pre-existing background knowledge and beliefs about the situation. If we refer to this knowledge as K, then what Bayes' theorem actually says about the use of base rates is that the decision maker should use $P(A \mid K)$. If it turns out that all of your background knowledge is irrelevant to the inference problem at hand, then $P(A \mid K) = P(A)$, and so it would be reasonable to use the "marginal" base rate P(A). However, to the extent that you have outside knowledge K that is relevant, then $P(A \mid K) \neq P(A)$, and in that case the marginal base rates are not the correct way to specify one's priors. In short, in order to *interpret* base rate neglect findings sensibly, we should be cautious about labeling a person's response as irrational without considering the other background knowledge that they bring to the task.

Understanding this demonstrates a limitation of many base rate neglect experiments,

which commonly assume that people should just accept a probability/base rate given to them without bringing any further information to bear on it. Instead, we would argue that it is impossible to state with certainty, in advance, what probabilities different people may assign to the same event. We can, however, based on our own understanding of the nature of a particular base rate, identify environmental factors that we would *expect* people to feel were relevant. For example:

- *Location.* Even if you genuinely believed that 99.9% was the true, world-wide base rate of white swans, the existence of regional variations implies that the single black swan observed in Australia *should* be more highly weighted. Therefore, when changing from one location to another, a rational person will discount prior observations against current ones – that is, they will neglect the base rate in favor of new data. (In statistics, this is a classic example of a spatial random effect model; see, e.g., Bryk & Raudenbush 1992)

- *Age.* Old data are less likely to be relevant to a new prediction than more current data as base rates change over time. Consider, for example, the proportion of land predators that are dinosaurs. If you were relying on base rates incorporating data collected over the last 170 million years, you might predict a fairly high proportion of observations. A more current analysis, however, would yield a lower figure. While this is a deliberately extreme example, this "information ageing" effect is observed in library curves that track the frequency with which books are borrowed as a function of age, which have a similar shape to human forgetting curves, suggesting that disregarding old information is a rational adaptation to changing environments (Anderson & Schooler, 1991).

- *Source.* In general, people trust the evidence of their own senses to a greater extent than they do that of another person. Thus, a sample that a person has collected themselves is likely to be weighted more heavily than data given to that same person from an outside

source. This is, of course, quite rational in that, with outside data, the degree of certainty over its veracity and how it was collected will tend to be lower than that regarding one's own observations.

• *Quantity.* Sample size must also be considered for both the prior and current samples. This is often ignored in base rate neglect experiments (perhaps due to an implicit assumption that the prior data set is "sufficiently large") but should be considered, as sample size is a determinant of a base rate's reliability. Empirically, a "base rate" can only be discovered via observation: in the real world, base rates are simply older and larger samples. As a consequence, the decision-maker should consider how much data contributes to the base rate itself.

These factors (and others like them) seem intuitively reasonable as an explanation for why people might be expected to possess, at least, a general, prior bias to neglect base rates to some extent. In order for us to argue that base rate neglect occurs as a *result* of probability judgments being conditioned on such factors, however, it is important to demonstrate that people are sensitive to these individual factors in the context of information aggregation. In what follows, we describe a series of experiments designed to illustrate the manner in which people use these cues to determine the trustworthiness of different sources of information and, accordingly, adjust their probability estimates. Throughout, our goal is to minimize any potential "computational" barriers to the integration of old and new information, instead manipulating the strength of base rate neglect via the inherent value of the information itself.

Experiment 1

As an initial examination, we consider the impact of varying the location, age, source and quantity of the data that provides a base rate. Motivated in part by Bar-Hillel's (1980) manipulation of the distinction between 'coincidental' and 'relevant' base rates, we look at

four distinct factors that could, each, affect the perceived degree of relevance of the base rate to the outcome being estimated. In order to focus specifically on these issues, we depart from many traditional base rate neglect studies in that we are, for the purposes of clarity, dealing with base rate information and current data that are completely commensurate, rather than looking at data of differing types. That is, rather than the traditional base rate neglect questions which state the base rate of some event and then provide additional information in the form of eyewitness testimony or some diagnostic test, we are interested in cases where base rate data is being updated with further observations of the same qualitative type in order to predict the future rate of occurrence. In this way, any differences in the salience of the base rate and current data are expected to result from the experimental manipulations rather than being dependent on individual interpretations of relevance/causal structure (Bar-Hillel, 1980; Krynski & Tenenbaum, 2007). Additionally, this characterization of updating maps more clearly onto the forms of updating that people undertake in real-world environments - where previous experience is updated with further information of the same type but from a new location. That is, rather than assuming that the correct prior probabilities are determined by the base rates (a situation that rarely actually happens in real life), we are interested in how people make inferences when the relationship between the prior and the base rate is more complex. By highlighting the fact that base rates are themselves based on data, we are able to investigate these situations.

While we acknowledge this difference between our and the traditional base rate neglect paradigm, we would argue that our method maintains what should be the key aspects of base rate neglect – updating older knowledge using new information – and that, if base rate neglect were to, somehow, fail to generalize to our paradigm, this would be a difficulty not for our experiment but rather for the concept itself. That is, if base rate neglect fails to occur in a situation where base rate data needs to be updated with additional

information of the same type rather than with a qualitatively different type of information, then the effect can hardly be considered to be particularly robust. In fact, if the effect can only be produced when the problem is framed in terms of the aggregation of two qualitatively different sources of information, then one might argue that it is just a framing effect, and not an expression of any fundamentally interesting characteristic of human belief updating. By comparison, if we are right in supposing that base rate neglect occurs due to a general strategy of discounting older data, then one would expect that base rate neglect should be observed in our experimental paradigm – with the strength of the effect being modulated by the extent to which the description of the problem suggests to people that the base rate data are inherently more reliable.

Experiment 1A

Method

Participants. Participants were twenty university students and members of the general public, 10 males and 10 females, with a mean age of 30.4 (SD = 12.1). Each was paid for their participation with a $10 bookstore voucher.

Experimental Design. As noted above, the scenarios used in our experiment were designed to maximize the extent to which people recognize the need to combine both sources of information, by explicitly placing the base rate data on a scale commensurate with a second source of evidence. To do so, both sources of evidence are described as samples of data ("prior" sample and "new" sample) that need to be taken into account. In this experiment we chose to examine the effect of varying sample size, while combining the source, age and location variables into a general cover story. Under the "high trust" cover story, the prior sample was described as recent data, collected by the participant, in the same location. Under the "low trust" cover story, the data was old, collected by someone else, and in a different location. Sample sizes were varied for both the prior data (20 or 200

data points) and for the new data (4, 8 or 12 data points)[2]. Moreover, the implied base rate could be either 25% or 75% (with the new data implying the alternate). With all factors fully crossed, this gave 24 (2x3x2x2) conditions in total.

All of the scenarios used variations on the same cover story: that the participant was part of a survey team exploring an alien planet and reporting on the proportion of some native life form or natural event that met a particular criterion. In every case, the participant was given a prior sample and then told what they had observed. Finally they were asked for an estimate combining both sets of information to be included in their report. For example:

> *You are currently classifying predators according to whether they pose a threat to humans. Your team, working at this location recently collected 200 observations and found that 50 (25%) of them met this criterion. This week, you have made another 4 observations, of which 3 (75%) met the above criterion. What proportion of predators in the area do you estimate pose a threat to humans?*

This example shows a prior sample size of 200 with a base rate of 25%. The current sample has a size of 4 and a rate of 75%. The prior is trustworthy in that it is described as recent, local and self-collected. A total of 32 scenarios were created for Experiment 1 (see Appendix A), 24 of which (selected at random for each participant) were used in Experiment 1A - such that each participant saw one scenario in each of the 24 conditions of this experiment.

Procedure. All scenarios (including the 24 described above and the 8 from Experiment 1B described below) were incorporated into a GUI and presented in random order. Participants sat at the computer and read the introductory cover story (identical to that shown in Appendix A for Experiment 2) before proceeding to the first randomly determined scenario. During each scenario, all of the information remained visible on the

screen until the participant had entered a predicted rate of future occurrence. No time limit was imposed and most participants completed the 32 scenarios within an hour.

Descriptive Model. In order to present an initial analysis of the data, we will adopt a heavily simplified model for how a "rational" decision-maker might solve this kind of induction problem (later in the paper we will introduce a somewhat more careful analysis more suited to handling individual decisions, but for the moment we forbear from doing so in order to avoid complicating the description of the experimental data). Let n_0 denote the number of observations that make up the prior sample and x_0 is the number of those observations that meet the criterion, and n_1 and x_1 are the equivalents for the new example. Then if θ denotes the proportion of items that meet the criterion in the wider population, then a sensible decision procedure is to report $E[\theta|x_0, x_1, n_0, n_1]$ the expected value of θ given the data. This estimate is given by[3]:

$$E[\theta \mid x_0, x_1, n_0, n_1] = \frac{x_0 + x_1}{n_0 + n_1} \qquad (2).$$

The big problem with this model is that it relies on the rather implausible assumption that each datum is equally useful as a predictor of θ. This makes little sense either in our scenarios or in real life. Indeed, our scenarios encourage participants to assume that the prior sample may be less closely related to the quantity of interest θ than the new data. Accordingly, if each prior datum is "worth" only t new data, we might expect the participant to report the value,

$$E[\theta \mid r_0, r_1, n_0, n_1, t] \quad = \quad \frac{tr_0 n_0 + r_1 n_1}{tn_0 + n_1}, \qquad (3)$$

where $r_0 = x_0/n_0$ denotes the base rate, and $r_1 = x_1/n_1$ denotes the sample rate. In this experiment, we vary the way people weight the base rate r_0 against the rate implied by the new sample r_1 in two distinct ways. By altering the description applied to the prior

sample, we expect to see a change in the value of *t*. This is a direct "cover story" manipulation, and is expected to result in some explicit downgrading of the usefulness of the prior sample.

The second manipulation involves sample size, and is somewhat more complex, since sample size is already built into the naive model predictions. By altering the ratio n_0/n_1, we would expect some reweighting of the two estimates. However, in view of the widely studied "insensitivity to sample size" effect (e.g., Tversky & Kahneman, 1974), the subjective "value" of a particular sample size is unlikely to be the same as its actual value. Nevertheless, following Sedlmeier and Gigerenzer (1997), we might reasonably expect that people's behavior will accord with Bernoulli's (1713) statement of the so-called empirical law of large numbers: "the more observations have been made, the less danger there is of wandering from one's goal" (see Stigler, 1986, p.65). For the moment, then, we make the assumption that the subjective value, $\tilde{n}$, is related to the objective value, n, via some unknown monotonic increasing function $\tilde{n} = f(n)$. Given this, we model the participants' judgments by assuming that they will report the value of θ expected when one applies Bayes' theorem to the *subjective* sample values, with some constant "trust" effect expected to arise due to the cover story:

$$E[\theta \mid r_0, r_1, \tilde{n}_0, \tilde{n}_1, t] \quad = \quad \frac{t r_0 \tilde{n}_0 + r_1 \tilde{n}_1}{t\tilde{n}_0 + \tilde{n}_1} \tag{4}$$

In order to fit the data from the 24 conditions, we fit 4 values for t, corresponding to the high trust and low trust values for both base rates. In addition to this, we need to estimate the various subjective sample sizes ñ, which means we need an additional four parameters. The reason it is 4 and not 5 comes from the structure of equation 4: if we multiplied all the ñ values by a constant, it would not change the result. As a consequence, we can fix one of the subjective sample sizes in advance without altering anything of

consequence. To that end, we assumed that the subjective sample size for n=20 observations is ñ = 20, and estimated the values for ñ that correspond to the other four sample sizes, namely 4, 8, 12 and 200. Note that, since 8 parameters are used to fit 24 data points, there is a sense in which this model is more descriptive than explanatory. However, as it turned out (see Figure 3), the subjective sample sizes that we estimated were an almost perfect fit to a logarithmic function ñ = log(n), consistent with previous research on number representation. As a consequence, it is possible to set all of the ñ values to principled values, leaving only the explicit trust parameters (i.e., the t values) as truly "free".

Results

Initial examination of the results revealed a small number of surprising estimates – lying outside the 25 to 75% range implied by the new and old data; that is, either above 75% or below 25%. In total, our 20 participants made 640 estimates (32 each) and, of these, 22 (<3.5%) lay outside the implied range. All 22 'out-of-bounds' estimates were made by 6 of our 20 participants (NB - all of the estimates made by every participant are shown in Figure 7 as part of the 'modeling individual responses' section of the paper).

Traditional base-rate neglect accounts are incapable of explaining such results – except to categorize them as mistakes or as evidence of irrationality. These data also, however, represent results that we did not expect – that is, the four factors that we considered a priori do not explain these results. Our general approach, however, is capable of interpreting these results in a meaningful way – without resorting to claims of irrationality or simply assuming them all to be errors.

Specifically, we failed to consider what might be termed extrapolation – the perception of a trend in the data. For example, a participant might believe that, because the first sample had a low rate and the second sample had a higher one, a hypothetical third sample would have an *even higher* rate; and, thus, estimate a value outside the range

dictated by the base and new rates. While we did not consider this a priori, it is quite consistent with our stand on probability estimation to assume that people will sometimes do exactly this as the existence of trends in data is part of the "given everything else you know" that people bring to bear when assessing probabilities. The simplified, rational model developed for our initial analyses, however, does not have the capability to interpret extrapolation beyond the original range. As a result, those participants whose data showed any evidence of extrapolation, no matter how little, were excluded from the initial investigation, leaving 14 participants. The excision of 30% of our data based on the relatively small number of problematic data points is not ideal, of course, but in the absence of any experimental control on this effect, attempting to incorporate it after the fact would be an exercise in post hoc supposition regarding in which scenarios this is most likely to occur. Therefore, we strictly apply the rule that all of a participant's data points be 'valid' in order to be used (of course, all 20 participants *are* included in our later analysis of individual responses).

Figures 1 and 2 show the mean estimates for the underlying rate given by these participants in all 24 conditions. The triangles show empirical data for the "high trust" cover story, and the circles show data for the "low trust" cover story. The dashed line shows the predictions made by the simplistic Bayesian solution (Equation 2). Overall, there is a clear base rate neglect effect: the empirical predictions tend to be shifted away from the Bayesian solution towards the current rate (i.e., above it in Figure 1 and below it in Figure 2). In total, data for 23 of the 24 conditions are shifted in this direction (one-tailed sign test gives $p \approx 1.5 \times 10^{-6}$). More important, however, is the fact that trustworthiness is having a clear effect. In all 12 cases, the mean predictions made by participants in high trust scenarios are closer to the Bayesian solution than estimates made in otherwise equivalent low trust scenarios (one-tailed sign test gives $p \approx 2.4 \times 10^{-4}$).

A finer grain of analysis is possible by fitting the model described by Equation 4. Parameter estimates for t and $\tilde{n}$ were obtained by minimizing sum squared errors. Figure 3 shows the recovered parameter estimates for the subjective sample size parameters, $\tilde{n}$. Comparison with the solid line makes clear that $\tilde{n} \propto \log n$: in this task, subjective impressions of sample size rise logarithmically with the actual sample size. This logarithmic relationship is in agreement with both the classic Weber-Fechner law and with other data suggesting that the mental representation of magnitude is approximately logarithmic (e.g., Dehaene 2003).

The implied trust statistics t for the cover story, shown in Table 1, are more complex. Most importantly, but not surprisingly, in both the 25% base rate conditions and the 75% base rate conditions, the estimated value for t is much higher when the cover story suggests high trust as opposed to low trust. Parameter estimates for low trust suggest that a prior datum is worth only 1/4 of a new datum, in subjective (i.e., log) terms. When the base rate is 25%, the high trust parameter is approximately 1, suggesting that the only effect in this condition is the logarithmic scaling of subjective sample size effect shown in Figure 3. The inferred value of 1.4 for the 75% base rate and high trust is odd, since it implies that a prior subjective datum is treated as being worth more than one subjective new datum. This observation, and the fact that the corresponding empirical data for these conditions (solid line at the top left of Figure 2) do not show strong evidence of base rate neglect, suggests that this case may be somewhat different to the others. However, as will become clear when we turn to the individual subjects analysis, the effect appears to be due to 3 participants who had a strong tendency to report large percentages regardless of the experimental condition.

Discussion

The results provide a somewhat intriguing view of base rate neglect. To a large extent, the base rates implied by larger samples are weighted more heavily than for small

samples, in keeping with the so-called empirical law of large numbers (Sedlmeier & Gigerenzer, 1997). In that sense, people can be seen to adapt to the trustworthiness of the data in a very sensible fashion. That said, a kind of "insensitivity" to sample size is observed, since the subjective value rises nearly logarithmically with sample size, rather than linearly. Additionally, altering the cover story to devalue the base rate has a large effect on trust, lowering the subjective value of the base rate by three quarters in both the 25% and 75% conditions.

Of course, further discussion is also required regarding the implications of the existence of extrapolation in our data. That is, does the fact that some people gave values outside what we initially considered to be the 'appropriate' range provide evidence against our hypothesis that people are, by displaying base rate neglect, behaving in a rational manner?

As argued above, we believe that it does not. Firstly, the number of responses falling outside this range was small – less than 3.5%, which would not be considered an unusually high error rate for data entry using a novel system (which is what participants were, in essence, doing). Added to this is the fact that almost half of these extrapolations (10 of 22) were observed in a single participant's data, implying an 'error' rate of only ~1.6% for the remaining participants.

Finally, while beyond the scope of the simple model we built as an initial approximation for how we thought a rational person might approach our experimental task, the use of extrapolation is not in any way irrational and could be incorporated in a more complex model of rational behavior. In fact, the presence of extrapolated values serves as an example of our primary hypothesis – that all probabilities are assessed in light of all other information that a person possesses. In this case, that data sometimes displays trends and that future observations can thus be higher or lower than any observations in the

current data set. While not fitting within a narrowly defined base rate neglect paradigm, effects like this do fit within the broader area of updating with new data and can, in theory at least, be accounted for.

Experiment 1B

Method

Experiment 1B aimed to expand on the three factors that contributed to the cover story in Experiment 1A. The design of the experiment was the same as for Experiment 1A and was, in fact, conducted simultaneously – using the same 20 participants, with the conditions intermixed with those used in the first study. In the Experiment 1B scenarios, however, the "base rate" was fixed at 75% using a sample of size 20 (i.e., 15 hits), and the new data always based on a sample size of 4 with a single hit, suggesting a rate of 25%. The conditions here were high and low trustworthiness based on the: age (recent vs old); location (local vs other continent); and source (self vs other) of the base rate data, yielding a 2x2x2 design and 8 conditions in all. The 8 scenarios used for this experiment were selected randomly for each participant from amongst the 32 listed in Appendix A.

An independent effect model takes the same format as Equation 4, but with separate terms for the effect of location t_l , age of data t_a and source of the data t_s. In view of the fact that there are only 8 questions in this experiment, we decided that it would be excessive to fit a high and low trust statistic for each of the three factors. Instead, in light of the results from Experiment 1A we fixed $t = 1$ for the high trust condition in all cases, and as before fixed $f(20) = 20$ for the subjective sample size. As a result, we estimated only the low-t values and $\tilde{n}_1$ from the raw data. In any case, the simple model used to analyze the data relies on the expression:

$$E[\theta \mid r_0, r_1, \tilde{n}_0, \tilde{n}_1, t_a, t_l, t_s] = \frac{(t_a t_l t_s) r_0 \tilde{n}_0 + r_1 \tilde{n}_1}{(t_a t_l t_s) \tilde{n}_0 + \tilde{n}_1}. \quad (5)$$

Note that this is not a new model in any substantive sense: it is just Equation 4 with specific trust effects (t_a, t_l and t_s) for each of the three factors being manipulated.

Results

The basic pattern of results is shown in Figure 4. As more reasons to distrust the prior data (distant location, old data, collected by someone else) are added to the cover story, participants' estimates move away from the base rate (75%) and closer to the new data (25%). Moreover, a model that assumes that each manipulation has a constant effect on trust (with no interaction effects) provides a very close fit to the data. Each manipulation has a substantial effect. Changing the age or source of the data lowers trust to 0.63 and 0.62, respectively, while changing the location lowered trust to 0.34. Fitting the subjective value of the new data, we obtained $\tilde{n} = 4.72$.

The overall pattern of results is highly consistent with results from corresponding conditions in Experiment 1A (i.e., those with 75% base rate, prior sample of 20 and current of 4): if all parameter values are multiplied by 1.41 (the high trust value found for these conditions in Experiment 1), we obtain $\tilde{n} = 6.61$ for the subjective sample size, which is fairly close to the value of 7.82 found in Experiment 1A. Similarly, the low trust value of 0.23 from Experiment 1A is close to prediction from Experiment 1B, which would be $1.41 \times 0.34 \times 0.63 \times 0.62 = 0.18$. In other words, although we have analyzed the two parts of the data set separately since we conceptualized them as somewhat distinct experiments, it is clear that the two are highly consistent with each other.

Discussion

It is clear that all three elements of the cover story affect the trust that people assign to data in reasonable ways. For example, location has a stronger effect on beliefs about ecological phenomena than time or source of the data, corresponding with natural expectations – specifically, that a change in continent will, in most cases, alter the value of

an ecological dataset more than it being 100 years old or collected by someone else.

Experiment 2

Method

Experiment 2 was conducted to address two issues regarding the design of the previous experiment. The first was that the results of Experiments 1A and 1B might have been affected by the use of a within-subjects design, which previous research has shown to increase the perceived salience of base rates and thus reduce the level of base rate neglect (see, e.g., Birnbaum & Mellers, 1983; Stolarz-Fantino, Fantino & van Borst, 2006). The second was to compare scenarios containing base rates explicitly indicated to be trustworthy or untrustworthy with scenarios where the trustworthiness was never specifically stated.

Participants. Participants were 80 University of Adelaide undergraduate students, 28 male and 52 female, with a mean age of 19.3 (SD = 1.7). Participants completed the task either for course credit or a $10 book voucher.

Experimental Design. A between-subjects design was used, with participants divided into four groups of 20 at random – each of which saw one of four versions of a single base-rate question. The question, in all cases, was based on the example question described in Experiment 1 (proportion of predators posing a threat to humans), with a base rate of 25%, calculated from a sample of 200 observations, and a current rate of 75%, calculated from a sample of 4 observations. The difference between the versions was in the description of how the base rate had been collected. Versions 1 and 2 replicated, respectively, the 'low trust' and 'high trust' conditions from Experiment 1A; with the base rate described as being derived from a sample collected by another team, on another continent, and 100 years ago (low trust) or by the participant's own team, locally and recently (high trust). Version 3 (unstated) included no markers of trustworthiness, containing no indication of when, where or by whom the data had been collected. Version 4 (unknown), similarly,

included no information regarding the source of the base rate data but, in this case, the absence of this information was specifically pointed out. The basic structure of the base rate question and the specific wording of each condition are given in Appendix A.

Procedure. Given the use of only a single question rather than the multiple versions used in Experiments 1A and 1B, participants completed the task in a pencil-and-paper format as part of a battery of psychological tasks unrelated to this experiment. Participants were given a sheet of paper containing the base rate question and asked to read the question and respond by writing their estimate of the future rate of occurrence – that is the proportion of predators in the area they believe pose a threat to humans – in a provided space.

Results

Figure 5 shows the mean estimated rate under each of the trust conditions. Looking at this figure, one sees that some base rate neglect is observed in all conditions, with all of the estimates lying above the naive statistical solution, irrespective of whether the objective sample size (squares) or the logarithmically-scaled subjective sample size (triangles) was used in the calculation of this solution. The degree of base rate neglect, however, varies as a function of the trust condition, as was the case in Experiments 1A and 1B. Comparing the results for the *low* and *high* trust conditions back to the equivalent points in Figure 1, one sees that a stronger base rate neglect effect is observed in Experiment 2, in line with previous findings regarding the strength of base rate neglect in between- and within-subjects designs (Birnbaum & Mellers, 1983; Storlaz-Fantino et al., 2006). Nevertheless, the pattern of results remains the same as in our earlier studies with highly trustworthy base rate data being incorporated more fully than less trustworthy data.

For three of the four conditions, the trust levels are both obvious and interpretable. Naturally, when the prior data are described as highly trustworthy, they are assigned more weight by participants than if the source seems less trustworthy. Moreover, when the

source of the data is explicitly described as unknown, the implied level of trust assigned to that data lies in between the high and low conditions, as one might expect. These results, taken in conjunction with those of Experiments 1A and 1B, suggested that participants were making use of the base rate and current rate in a sensible fashion, with the base rate being awarded more weight in situations where it was deserving of more trust.

The most interesting aspect to the data, however, lies in the fourth condition. While the *high* > *unknown* > *low* ordering is obvious, people's decisions when all details about trustworthiness are omitted (the *unstated* case) require further explanation. In particular, looking at the results in Figure 5, it seems clear that, when no details are given, people trust the base rate data even more than they do if the data are explicitly described as highly trustworthy. That is, by default, the pragmatic assumption made by people is that the base rates *are* fairly trustworthy and, in fact, that the act of including explicit statements suggesting that the data are trustworthy acts to call attention to the possibility that the data *could* be untrustworthy.

Discussion

The general increase in neglect levels in Experiment 2 are in keeping with previous research (Birnbaum & Mellers, 1983; Stolarz-Fantino et al, 2006), showing that between-subjects designs elicit higher levels of base rate neglect. Moreover, the most interesting aspect to the data, namely that the *unstated* condition (which most closely mimics more traditional studies) actually produced the lowest levels of neglect, has a natural, pragmatic interpretation. Viewed in terms of Grice's (1975) communicative maxims, the decision-maker assumes that the language used is as minimal as required to convey the correct intuition: one would only go to the effort of specifying that the data should be trusted if there were some reason to think that they might not (a kind of "lady doth protest

too much" effect). That is, as one might expect, the general effect of including explicit trust markers is to *increase* the extent of base rate neglect. Along similar lines, an implied-relevance explanation suggests that even in low-trust scenarios, the base rate should not be entirely discounted, because if there were *literally* no reason to trust the base rates, the speaker should not have included them at all. In fact, once we make the logarithmic-adjustment of sample size, so as to correspond to the standard (e.g., Dehaene 2003) view of human number representation, the *unstated* condition led to a comparatively modest base rate neglect effect. That is, after logarithmic scaling to compensate for sample size effects, the trust statistics for the conditions in this experiment correspond to an implied belief that an old datum is worth .08, .17 and .29 new data, respectively, in the three conditions where a statement of trustworthiness is made – *low*, *unknown* and *high* – rising to .54 in the *unstated* condition.

While it is difficult to make accurate comparisons with the levels of base rate neglect observed in Experiment 1B due to the differences in experimental design, we suspect that this weighting of data in the *unstated* condition corresponds to something close to the "one-reason to distrust" cases in that data set. Essentially, even in the unstated condition when the base rates are most highly trusted, people remain sensitive to the fact that base rate data are older, and probably less useful than the new data. In short, while the experimental manipulations have not eliminated base rate neglect in the between-subject design (where people only get to see one scenario and are less likely to be made aware of the potential usefulness of the older, less trustworthy source), the trust-related manipulations alter the magnitude of the effect in sensible ways, lending at least some credence to the idea that base rate neglect corresponds to something like the "background" or "default" level of distrust for the kinds of old, context-general data that generally make up a base rate.

Modeling Individual Responses

The analyses presented in the previous sections focused primarily on how presenting various different markers of trust (data age, location, source and quantity) systematically alters the *average* of the judgements given by participants. Implicitly, this analysis relies on the assumption that individual participants all use the same strategy: if everyone reports a number described by $E[\theta]$ + noise, then averaging helps to remove the noise. As an initial analysis of the data, this approach is useful, since it illustrates a general pattern that is sensible, and shows that people's decisions are heavily influenced by these markers. However, much is lost by examining data only at this level of generality: focusing only on average responses and expected values tends to oversimplify the way in which individual decisions are made, because it fails to consider the full distribution over responses, or accommodate individual differences among participants. To illustrate the point, Figure 6 shows the empirical distribution of responses for all four conditions in Experiment 2. The individual responses tend to match the base rate (25%), the likelihood (75%) or the average of the two (50%), although some proportion of responses cannot be described in this fashion. Clearly, while the *averages* change quite remarkably across conditions, and in a fashion that we think is consistent with a rational decision-making strategy (i.e., distrust untrustworthy data!), the individual response categories (i.e., 25, 50, 75) tend not to change. This effect seems indicative of a general tendency for participants to respond with certain numbers preferentially – in particular, with multiples of 5 and 10 as predicted by research on human number preference (Baird, Lewis & Romer, 1970), combined with a possible anchoring effect (Tversky & Kahneman, 1974) resulting from those numbers presented as part of the problem. In short, the majority of the manipulation's effect is observed to result from alterations to the *proportion* of responses belonging to each category.

Following up on this observation, a finer grained analysis of individual responses was conducted for Experiments 1A and 1B, since across the two parts of the

experiment each of the 20 participants provided judgements for 32 different problems. Note that this includes those participants previously excluded for having estimates falling outside the expected bounds (also that an examination of Figure 7 or Table A1 confirms that, while six individuals did extrapolate on some trials, overall this was a rare occurrence).

For simplicity, we fit individual subjects' data separately, rather than constructing a full individual difference model (e.g., Lee & Webb 2005; Navarro, Griffiths, Steyvers & Lee 2006). In each case, we fixed the subjective sample size function to a logarithmic form $\tilde{n} = f(n) \propto \log(n)$ with $f(20) = 20$, and again fixed the high trust value to 1 regardless of whether the factor in question was location, age or source. Thus, for each participant we estimate a distinct low-trust value for location (t_l), age (t_a) and source (t_s), by minimizing sum squared error between their responses and the predictions of Equation 5.[4] The results are shown in Figure 7, which plots the model predictions $E[\theta]$ along the horizontal axis and the participant estimates on the vertical axis. Each panel shows all 32 responses from a single participant (participant number is given in the bottom, left corner). In order to facilitate the visual display of the data, however, all trials are shown *as if* the base rate was always 75% and the new data suggested 25%. That is, in the case of the squares, the value plotted is actually 100 − x on the vertical and $100(1 - E[\theta])$ on the horizontal.

Looking at the individual subplots, the expectation is that, if the model is accurately predicting human responses, then the data points will correlate positively, falling along the diagonal or, at least, primarily in the bottom left and top right quadrants. By comparison, base rate neglect would predict values falling within the bottom left and bottom right quadrants and the Bayesian solution predicts values falling in the top left and top right quadrants. Broadly speaking, a datum was taken as evidence for a particular approach to the base rate neglect task if it fell within the predicted quadrants and against it if it fell into non-predicted quadrants. The fits of each possible model to the total data set of 32

observations per participant were then compared to determine the best fitting model.[5]

As is clear from inspection of the posterior model probabilities in Figure 7 (representing the probability that a given model is the best explanation - of the seven we tested - for a participant's behavior), we can be highly confident that six of the participants (3, 8, 10, 11, 17 and 18) switched between a reliance on the old data to a reliance on the new data in a fashion similar to that predicted by the trust model (i.e., Equation 5), with weaker evidence suggesting that this pattern holds for another three (4, 14 and 16). These individuals, referred to as "Skeptics" in Table 2, made estimates that accorded with model predictions. That is, in situations where the base rate seemed untrustworthy, they discounted it, whereas when the base rate seemed trustworthy they relied on it more.

In four cases there is either strong (6, 15) or weak (1, 12) evidence suggesting that the participants favored the new sample, but in a fashion that is consistent with a variable-trust account, and there is one case (19) in which the old data is favored in a manner that is weakly consistent with a variable-trust account. These individuals are labeled "New" and "Old" in the tables and can be regarded as having a bias towards relying, preferentially, on either the base rate (Old) or current data (New) but still being somewhat affected by the experimental manipulation of the trustworthiness of the base rate.

That is, for 14 of the 20 participants, the best explanation for their behavior is that they trust or distrust the base rate data in a manner consistent with the model. Of the remaining 6 participants, however, four (7, 9, 13, 20) consistently favored the old data in a manner suggesting something close to the Bayesian solution as traditionally described (labeled "Naive" in Table 2), with a further one doing the same but less consistently (5). The final participant (2) produced no clear relationship between their estimates and either the old or new data and is, therefore, labeled "Random".

In short, 19 participants produced interpretable data, of which 5 are best described as

naive Bayesian updating while 14 are consistent with the trust model.[6] That is, despite the observation of what seemed to be base rate neglect in both experiments, none of the individual participants' responses are best characterized by a simple "base rate neglect" explanation – which would be the mirror image of the "Naïve" pattern of results in Figure 7.

For the 14 participants whose data are consistent with the trust model, Table 2 displays estimates for the three trust parameters and correlations between the model and participants estimates. In line with the estimates from the averaged data presented earlier, the general pattern is that data collected elsewhere are discounted most heavily, with both older data and data collected by other people discounted less.

Note, however, that the weight assigned to the three trust parameters varies markedly between individuals. For example, looking at Table 2, one sees that participant 8 assigned data collected at a distant location only .09 the weight of local data, whereas participant 4 weighed local and distant data equally. This suggest that participants held different prior beliefs regarding the usefulness of these parameters in predicting future data – in line with our argument regarding differences in prior knowledge affecting probability judgments. That is, the K portion of P(A|K) differs between individuals and, as a result, given the same event, A, each person can, rationally, make a different estimate.

General Discussion

As noted earlier, our base rate neglect paradigm differs significantly from that commonly used and, as a result, effects commonly argued to affect the magnitude of base rate neglect – such as the salience and representativeness of novel information – have been (deliberately) restricted. However, even in the absence of these factors, the base rate tended to be neglected in favor of newer (but completely commensurate) information, in a manner consistent with our view of base rate neglect as a natural and rational strategy. Our analyses also point to a greater complexity in the problem of base rate neglect than is sometimes

assumed. While, in all of our experiments, there is evidence of what would, classically, be called base rate neglect, the effect has been shown to be strongly influenced by what seem to be rational rules for information-updating in real environments. A secondary observation is that the behavior of groups, which display base rate neglect in the manner described by previous research, differs significantly from that of individuals, whose responses are, in our data, often constrained to a limited number of response categories and are better characterized by alternate explanations.

To expand on the main point somewhat: the general perspective we have adopted relies on the idea that people are drawing inductive inferences that are constrained by the way base rates operate in real life. In essence, handling multiple sources of evidence with quite different pedigrees is a kind of data analysis problem, and so we have aimed to show that people (to some extent) respond to exactly the kinds of pressures that apply in real world data analysis. We know of no data analyst who believes that all data sets are equally valuable, nor have we ever found an old data set that precisely matches our needs in any novel context. As a result, it seems wise to be skeptical when trying to apply the base rates they imply to any new context.

From this "inference as data analysis" perspective, however, the basic framework used to construct judgment and decision-making problems can seem somewhat unsatisfying. Consider, for example, this decision-making problem taken from Griffin & Tversky (1974):

> *Imagine that you are spinning a coin, and recording how often the coin lands heads and how often the coin lands tails. Unlike tossing, which (on average) yields an equal number of heads and tails, spinning a coin leads to a bias favoring one side or the other because of slight imperfections on the rim of the coin (and an uneven distribution of mass). Now imagine that you know this bias*

is 3/5. It tends to land on one side 3 out of 5 times. But you do not know if this bias is in favor of heads or in favor of tails. You then spin the coin 10 times and see 6 heads. Does the bias favor heads? How confident are you?

Suppose this actually happened in real life. You might want to know *who* told you that the bias (which, much like the base rates in the taxicab problem, acts as a way of characterizing background knowledge about the problem) is exactly 3/5, and *why* they do not know its direction. In fact, the very set-up is suspicious – the fact that you apparently know the magnitude of the bias *exactly* but have no knowledge at all about its direction suggests that something funny is going on. Moreover, the fact that a psychologist is asking the question doubtless adds to the suspicion that the game is rigged. While the taxi-cab problem is more realistic than the coin-spinning one, it still seems odd to treat the base rate as inherently fixed, relevant and trustworthy – ignoring its nature as a prior sample. By comparison with these idealized, experimental base rates, every real-world database the authors have had to analyze has been riddled with coding errors, missing data and erroneous information (while the agencies that collect and disseminate statistics have a good motivation to exaggerate their fidelity).

Further, even if the base rate is, by some statistical quirk, completely accurate, there are innumerable ways in which it might still be inapplicable to the specific incident described in the taxi-cab problem. For example, the problem gives no background on the location of the accident, which allows for the possibility (noted by Goodie & Fantino, 1999) that it could have occurred right outside one of the company headquarters and that this information has simply been omitted (hardly a stretch given that legal proceedings are involved – not situations renowned for producing unbiased data). Thinking in these terms, it seems clear that any statement of a base rate is, at best, a *generalization*, which must be adapted by the decision maker to suit the current context before being incorporated into any

prediction. That is, when thinking about the probability of an event 'A' – P(A) - we must incorporate the often implicit "given everything else we know" and consider P(A|K).

Compounding these problems, a lot of real world situations involve base rates that are not directly accessible to people, but instead have to be estimated by the decision maker from their own past experience – making it susceptible to memory effects, and cognitive biases such as anchoring and availability (Tversky & Kahneman, 1974), leading to an even less reliable estimate.

Once again, however, data analysts in the real world are entirely aware of this, and adapt their methods accordingly. For instance, regional variations in the rate of a phenomenon are ubiquitous, and routinely handled by statisticians via hierarchical (i.e., random effect) models that are highly sensitive to this variation (Bryk & Raudenbush 1992). Similarly, dealing with outliers and missing data are a routine part of the "data cleaning" process that precedes any professional statistical analysis, because data analysts have learned that it is *irrational* to assume that every datum is equally helpful. To illustrate the difference in mindset that this perspective induces, compare the reasoning problem presented as a "psychological" experiment (i.e., the Griffin & Tversky example above) to a sample problem taken from a statistics textbook (Mackay 2003):

> *You move into a new house; the phone is connected, and you're pretty sure that the phone number is 740511, but not as sure as you'd like to be. As an experiment, you* p*ick up the phone and dial 740511; you obtain a 'busy' signal. Are you now more sure of your phone number? If so, how much?*

Clearly, the statistics question is much richer, and more closely relates to the kinds of knowledge and environments in which people really operate. The number of possible phone numbers is not stated, the proportion of phones engaged at any given time is not clear, and the loss function for an incorrect decision is not obvious either. To answer such questions,

people need to bring relevant prior knowledge to the task; trying, for example, to recall the proportion of times when ringing other numbers resulted in a busy signal.

Of course, the problem with using these sorts of inference questions, from a psychological testing point-of-view, is that they are inherently uncertain; making it difficult for the researcher to be sure what the "correct" answer is. Even so, the problems that humans face in real life actually have this characteristic; and while "decision making under uncertainty" is a common description for psychological decision making research, it is very frequently an inaccurate one. Most such research requires that the researcher know the "true" answer in advance, so that any bias in participants' responses can be measured. Given this, the fact that we did not specify, in advance, what a rational person *should* do and what their answer *should* be given any set of inputs in our experiments may seem unusual to those familiar with common practice in the literature.

In principle, this is a good experimental method but, far too frequently, the statistical model used for calculating the "correct" answer is so naive that no statistician would ever apply it to anything except a fictitious data set. However, if human cognition is adapted to dealing with real world data analysis problems, it seems very unlikely that people would give answers that agree with statistical models that are overly simple. In other words, there is a danger that an over-reliance on naive statistical tools can produce what we might call an illusion of irrationality. With this in mind, we suspect that future research might need to rely more on problems such as the unknown-phone-number problem and our base rate neglect paradigm which allow a person's pre-existing knowledge to affect what the correct answer *should* be and thus allow for the fact that different people can, based on their different knowledge, make *different* but *equally rational* assessments of probabilities. Although it requires greater sophistication on our part to figure out what counts as a good answer or a bad answer to these, more realistic questions - given that we can not specify a *single* right

answer in advance - we feel that such data might be much more helpful in explaining everyday human decision making.

To conclude, in real world data analysis, one rarely if ever comes into contact with a situation in which base rate data are anywhere near as trustworthy as data collected in a context-appropriate fashion – that is, data collected as it is needed for a specific purpose. Thus, we are drawn to the same basic conclusion as McKenzie (2003): demonstrations that people tend to ignore potentially misleading base rate data (and we suggest that, almost by definition, every real world base rate is potentially misleading) ought to be interpreted as evidence in *favor* of their rationality rather than as a bias to be overcome.

References

Ajzen, I. (1977). Intuitive theories of events and the effects of base-rate information on prediction. *Journal of Personality and Social Psychology, 35*(5), 303-314.

Anderson, J. R. & Schooler, L. J. (1991). Reflections of the environment in memory. *Psychological Science, 2*, 396-408.

Baird, J.C., Lewis, C. & Romer, D. (1970). Relative frequencies of numerical responses in ratio estimation. *Perception and Psychophysics, 6*, 78-80.

Bar-Hillel, M. (1980). The base-rate fallacy in probability judgments. *Acta Psychologica, 44*, 211-233.

Bernoulli, J. (1713). *Ars Conjectandi*, Basilea: Thurnisius.

Birnbaum, M. H., & Mellers, B. A. (1983). Bayesian inference: Combining base rates with opinions of sources who vary in credibility. *Journal of Personality and Social Psychology, 45*, 792-804.

Bryk, A. S. & Raudenbush, S. W. (1992). *Hierarchical linear models*. Newbury Park, CA: Sage.

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

Dehaene, S. (2003). The neural basis of the Weber-Fechner law: a logarithmic mental number line. *Trends in Cognitive Sciences, 7*(4),145-147.

Fiedler, K. (2000). Beware of samples! A cognitive-ecological sampling approach to judgment biases. *Psychological Review, 107*(4), 659-676.

Gigerenzer, G. (1996). On narrow norms and vague heuristics: A reply to Kahneman and Tversky (1996). *Psychological Review, 103*(3), 592-596.

Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning

without instruction: Frequency formats. *Psychological Review, 102*(4), 684-704.

Gigerenzer, G., & Todd, P. M. (1999). *Simple Heuristics That Make Us Smart*. New York: Oxford University Press.

Gluck, M. A. & Bower, G. H. (1988). From conditioning to category learning: An adaptive network model. *Journal of Experimental Psychology: General, 117*, 227-247.

Goodie, A. S., & Fantino, E. (1999). What does and does not alleviate base-rate neglect under direct experience. *Journal of Behavioral Decision Making, 12*, 307-335.

Grice, P. (1975). Logic and conversation. In P. Cole & J. L. Morgan (Eds), *Studies in Syntax*, vol 3 (pp. 41-58). New York: Academic Press.

Harries, C. & Harvey, N. (2000). Are absolute frequencies, relative frequencies, or both effective in reducing cognitive biases. *Journal of Behavioral Decision Making, 13*, 431-444.

Hume, D. (1739/1898). *A Treatise of Human Nature*. London: Ward Lock.

Jaynes, E. T. (2003). *Probability Theory: The Logic of Science*. Cambridge: Cambridge University Press

Kahneman, D., Slovic, P. & Tversky, A. (1982). *Judgment Under Uncertainty*. Cambridge, UK: Cambridge University Press.

Kahneman, D. & Tversky, A. (1973). On the psychology of prediction. *Psychological Review, 80*(4), 237-251.

Kahneman, D. & Tversky, A. (1996). On the reality of cognitive illusions. *Psychological Review, 103*(3), 582-591.

Krynski, T. & Tenenbaum, J. (2007). The role of causality in judgment under uncertainty. *Journal of Experimental Psychology: General, 136* (3), 430-

450.

Laplace, P. S. (1814/1951). *Essai Philosophique sur les Probabilites* (F. W. Truscott & F. L. Emory, Trans.). New York: Dover Publications.

Lee, M. D. & Webb, M. R. (2005). Modeling individual differences in cognition. *Psychonomic Bulletin & Review, 12*, 605-621.

McKenzie, C.R.M. (2003). Rational models as theories - not standards - of behavior. *Trends in Cognitive Sciences, 7*, 403-406.

Navarro, D. J., Griffiths, T. L., Steyvers, M. & Lee, M. D. (2006). Modeling individual differences using Dirichlet processes. *Journal of Mathematical Psychology, 50,* 101-122.

Sedlmeier, P. & Gigerenzer, G. (1997). Intuitions about sample size: The empirical law of large numbers. *Journal of Behavioral Decision Making, 10*, 33-51.

Simon, H. A. (1956). Rational choice and the structure of environments. *Psychological Review, 63*, 129-138.

Sloman, S. A., Over, D., Slovak, L. & Stibel, J. M. (2003). Frequency illusions and other fallacies. *Organizational Behavior and Human Decision Processes, 91*, 296-309.

Stigler, S. M. (1986) *The History of Statistics* Cambridge, MA: Harvard University Press.

Stolarz-Fantino, S., Fantino, E. & van Borst, N. (2006). Use of base rates and case cue information in making likelihood estimates. *Memory and Cognition, 34*, 603-618.

Todd, P. M. & Gigerenzer, G. (2000). Simple heuristics that make us smart. *Behavioral and Brain Sciences, 23*(5), 727-741.

Todd, P. M. & Gigerenzer, G. (2003). Bounding rationality to the world. *Journal of*

Economic Psychology, 24, 143-165.

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

Tversky, A. & Kahneman, D. (1982). Evidential impact of base rates. In D. Kahneman, P. Slovic & A. Tversky (Eds.), *Judgment Under Uncertainty: Heuristics and Biases*. Cambridge, UK: Cambridge University Press.

Villejoubert, G. & Mandel, D. R. (2002). The inverse fallacy: an account of deviations from Bayes Theorem and the additivity principle. *Memory and Cognition, 30*(2), 171-178.

Zentall, T. R., & Clement, T. S. (2002). Memory mechanisms in pigeons: evidence of base-rate neglect. *Journal of Experimental Psychology: Animal Behavior Processes, 28*(1), 111-115.

Appendix A

Experimental Scenarios

Thirty-two scenarios were written for Experiment 1 and four for Experiment 2. These are all given below. In Experiment 1, a participant saw all 32 of the listed scenarios - 24 during Experiment 1A and 8 during Experiment 1B. Which scenarios were assigned to which part of the Experiment was randomly determined for each participant.

In Experiment 2, by comparison, each participant saw only one of the 4 scenarios, determined by the group into which they were randomly assigned.

Experiment 1 Scenarios

All scenarios began with the phrase: "You are currently classifying/estimating…" and then gave the class of objects being categorized and the criterion for making the category distinction. Table A1, shows the specific classification task required in each question.

Experiment 1A: High Trust Base Rate.

Your team, working at this location recently, collected n_0 observations and found that x_0 (x_0*100/n_0 %) of them met this criterion. This week, you have made another n_1 observations, of which x_1 (x_1*100/n_1 %) met the above criterion. What proportion of <insert object>s in the area do you estimate <meet stated criterion>?

Experiment 1A: Low Trust Base Rate.

Team Alpha, working on the northern continent more than a century ago, collected n_0 observations and found that x_0 (x_0*100/n_0 %) of them met this criterion. This week, you have made another n_1 observations, of which x_1 (x_1*100/n_1 %) met the above criterion. What proportion of <insert object>s in the area do you estimate <meet stated criterion>?

Experiment 1A allowed: n_0 values of 20 or 200; n_1 values of 4, 8 or 12; base rates of 25% or 75%; and corresponding rates in the new data of 100-(base rate). This gave a 2x2x3x2 design with 24 conditions.

Experiment 1B used only: n_0=20; n_1=4; base rate=75%; and new rate=25%. Here, the trustworthiness of the base rate was manipulated in three ways, yielding a 2x2x2 design with 8 conditions:

By changing *who* collected the data - "your team" vs "Team Alpha".

By changing *where* the data was collected - "at this location" vs "on the northern continent".

By changing *when* the data was collected - "recently" vs "more than a century ago".

Experiment 2 Scenarios

General Question Format..

You are the leader of Team Beta, one of two survey teams recently landed on Epsilon Eridani's fourth planet (EE4) - the other being Team Alpha. Your job is the categorisation of native life forms and conditions according to various criteria the Exploration company has set. Both teams have been to the planet previously but, due to the time taken to travel between star systems, it is more than a century since you were last here (most of that time having been spent in cold-sleep hibernation in transit). You are assessing the site chosen for the first permanent settlement on EE4 - nicknamed Star City - which lies on the southern continent.

Your current task is to estimate the proportion of predators in the local area that pose a threat to humans. You have access to two sources of data to aid you in making your estimate:

Specific base rate information (see below) inserted here.

Second, your team has observed 12 predators in the last week in the immediate Star City area and of these 9 (75%) posed a threat to humans.

What proportion of predators in the Star City area will you report as posing a threat to humans?

P =

Specific Base Rate Information.

High Trust: First, a previous survey, observed 200 predators and found that 50 (25%) of these posed a threat to humans. This survey was undertaken by your team, recently, in an area adjacent to Star City.

Low Trust: First, a previous survey observed 200 predators and found that 50 (25%) of these posed a threat to humans. This survey was undertaken by Team Alpha on the northern continent during the previous survey period, 100 years ago.

Unknown Trust: First, a previous survey, observed 200 predators and found that 50 (25%) of these posed a threat to humans. When, where and by whom this data was collected, however, is not recorded.

Unstated Trust: First, a previous survey, observed 200 predators and found that 50 (25%) of these posed a threat to humans.

Author Note

Correspondence concerning this article should be addressed to Matthew Welsh, Australian School of Petroleum, University of Adelaide (matthew.welsh@adelaide.edu.au).
MBW was supported by ExxonMobil and Santos, through the CIBP at the Australian School of Petroleum. DJN was supported by an Australian Research Fellowship (ARC grant DP-0773794). Part of this research was presented at the 2007 *Cognitive Science* conference. We thank Anastasia Ejova and Ben Schultz for their assistance in collecting the data and Steve Begg, Nancy Briggs, John Dunn, Amy Perfors and Carolyn Semmler for comments on earlier drafts of this manuscript. Finally we wish to thank William P. Bottom and three anonymous reviewers for their comments and suggestions.

Footnotes

1. More generally, the base rate may correspond to an inductive inference or theory constructed from such data. Even so, the validity of the inferred base rate relies on the validity of the data used to construct it.

2. Our decision to make the new sample smaller than the older (base rate) sample was informed by ecological considerations whereby pre-existing data sets tend to be larger newly collected data sets and that base rates, almost by definition, are defined from large samples. It also matches more closely the standard base rate paradigm where a base rate of unknown, but presumably large, size is given and then a single, new observation is used as the new data.

3. All of the following has a natural, Bayesian interpretation. For those interested in the details of the derivations of the in-text formulae, please see the supplementary materials available online at: www.compcogscilab.com/dan/publications.

4. We also fitted a model where participants were assumed to have some tendency to prefer round or "natural" numbers (e.g., 25, 40, 50, etc) in line with number preference (e.g., Baird, Lewis & Romer, 1970) and the pattern observed in Figure 6. Doing so improved model performance somewhat, but the model predictions are less easily visualized and, as the general pattern of results is much the same, we restrict discussion to the simpler model.

5. Full details of the models and their fits are available in the supplementary on-line materials available at: www.compcogscilab.com/dan/publications.

6. NB - the models listed in text are the best fitting of those considered *a priori*. However, three participants (5, 12 and 14) display a pattern that we did not consider a priori: they commonly (about 88% of the time) gave responses of 50% or higher, regardless of what the prior sample and the new sample said. This - if interpreted as indicating that participants have informed, prior beliefs about the problem - could be captured by the trust model but, given only 3 cases, we avoid introducing this post hoc extension.

Table 1.

Estimated trust statistics for the low and high trustworthiness conditions, by underlying base rate, from Experiment 1A.

	high trust story	low trust story
25% base rate	0.94	0.25
75% base rate	1.41	0.23

Table 2.

Individual trust parameter estimates for the 14 participants for whom some version of the 'distrust-of-base-rates' model is the preferred account. Correlations between model predictions and human estimates are given in the 'r' column.

ID	location	age	source	r	model
3	0.62	0.17	0.91	.56	Skeptic
4	1.00	0.12	1.00	.35	Skeptic
8	0.09	0.71	0.71	.75	Skeptic
10	0.47	0.39	0.94	.63	Skeptic
11	0.09	1.00	1.00	.81	Skeptic
14	1.00	0.39	0.94	.20	Skeptic
16	0.12	1.00	0.51	.59	Skeptic
17	0.11	0.77	0.75	.71	Skeptic
18	0.39	0.21	1.00	.71	Skeptic
1	1.00	0.35	0.77	.40	New
6	0.76	1.00	0.76	.71	New
12	0.73	0.87	0.87	.46	New
15	0.69	1.00	1.00	.67	New
19	0.16	0.72	0.26	.56	Old
M	0.51	0.68	0.78		

Table A1.

Experiment 1 Classification Scenarios.

	Object	Criterion
1	Predators.	Threat to humans.
2	Animals.	Can be domesticated.
3	Parasites.	Parasitize humans.
4	Grasses.	Digested by Earth herbivores.
5	Fruits .	Poisonous to humans.
6	Soil samples.	>0.5 parts/thousand gold.
7	Soil samples.	Sufficient nitrogen fixing bacteria.
8	Water samples.	Exceed safe levels of heavy metals.
9	Predators.	Nocturnal.
10	Animals.	Scavengers.
11	Herbivores.	> 100kg weight.
12	Plants .	Wind-pollinating.
13	The proportion of days.	Humidity > 90%.
14	The proportion of days.	Humidity < 20%.
15	The proportion of days.	UV index > 11.
16	The proportion of days.	Atmospheric CO_2 > 5%.
17	The proportion of days.	Atmospheric O_2 < 15%.
18	The proportion of days.	Minimum Temperature < 0 C.
19	The proportion of days.	Maximum Temperature > 40 C.
20	The proportion of days.	Pollen count > 200/cubic metre.
21	The proportion of days.	Rainfall > 25mm.

22	The proportion of days.	Sunshine > 9 hours.
23	Birds.	Carnivorous.
24	Weed plants.	Subject to chemical control.
25	Microbes.	Capable of carrying cholera.
26	Water sources.	Iodine < 2 micrograms/litre
27	Volcanoes.	Active.
28	Soil samples.	> 15% humus (organic material).
29	Mushrooms.	Poisonous to humans.
30	Water sources.	Potable.
31	Coelentrates (jellyfish and sea anemone).	Poisonous to humans.
32	Plants.	Evergreen.

Figure Captions

Figure 1. Participants' mean estimated rates (with SE) for the 25% base rate conditions in Experiment 1A.

Figure 2. Participants' mean estimated rates (with SE) for the 75% base rate conditions in Experiment 1A.

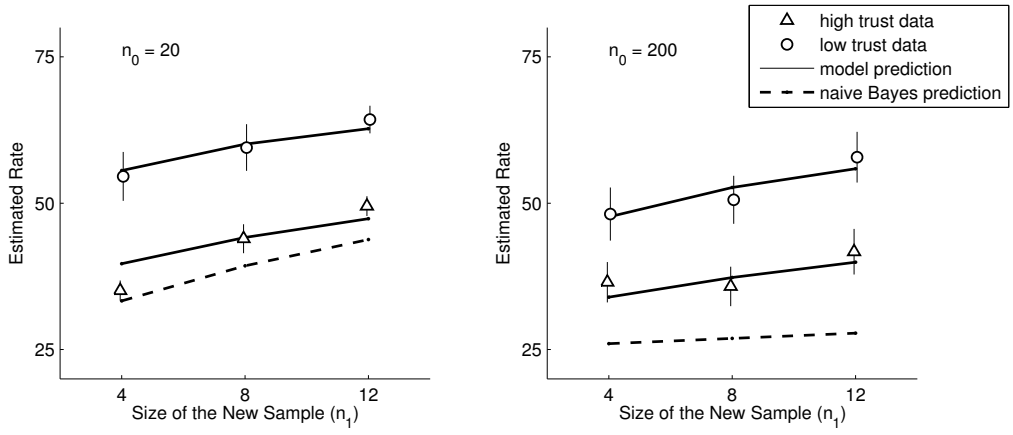
Figure 3. Subjective sample sizes inferred from participants' probability judgments. Note that these follow an approximately logarithmic function – shown by the solid line.

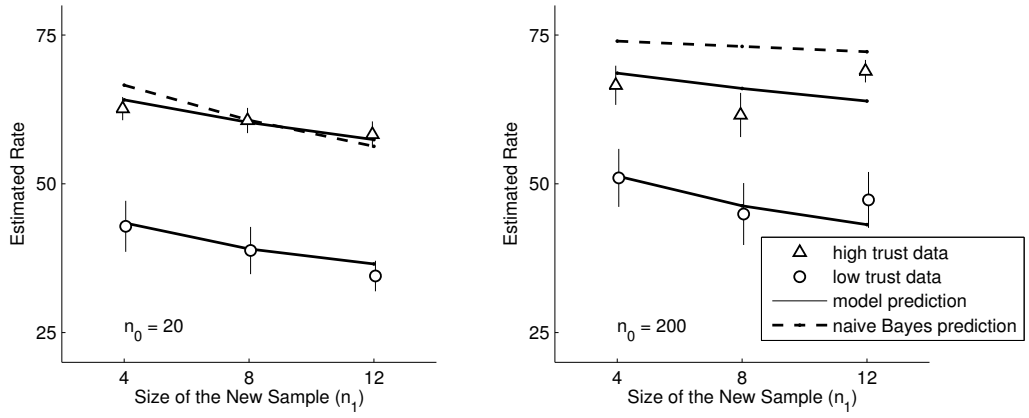
Figure 4. Actual (circles with SE bars) and predicted values (crosses) for participants' estimates of the underlying rate in Experiment 1B. Model predictions are based on the estimated trust effects given in text: Location 0.34; Age 0.63; and Source 0.62.

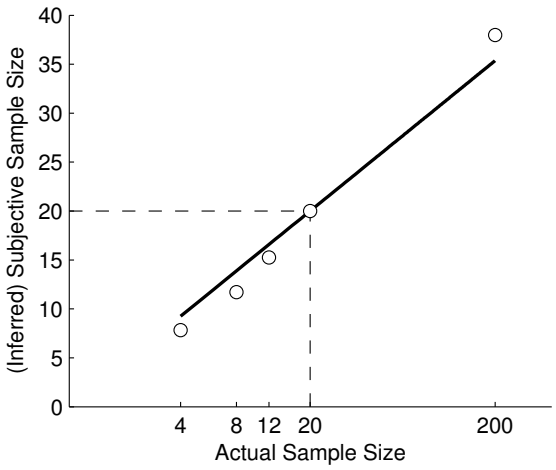
Figure 5. Mean estimated rate (with 95% CI) by trust condition (i.e., trustworthiness of the base rate) in Experiment 2. All scenarios involved a base rate of 25% and a current rate of 75%, with sample sizes of 200 and 4 observations respectively. The dashed lines indicate the Bayesian solutions (i.e., complete-trust; t = 1) based on objective and logarithmically-scaled, subjective sample sizes.

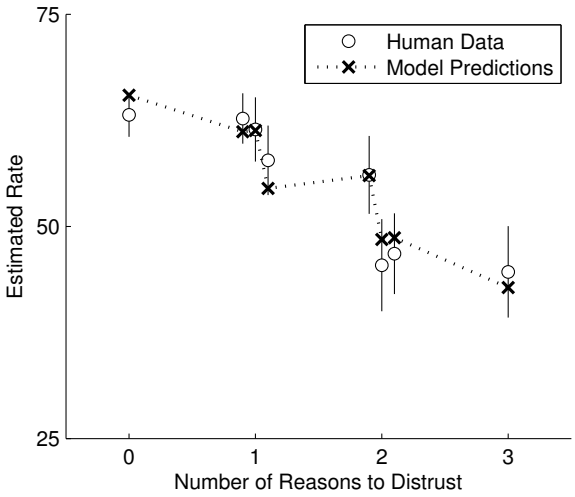
Figure 6. Histogram of estimated rates for all four trust conditions in Experiment 2. Note that individual responses tend to match the base rate (25%), the likelihood (75%) or the simple average of the two (50%) and, thus, the majority of the effect of the manipulation lies in the *proportion* of responses belonging to each category.

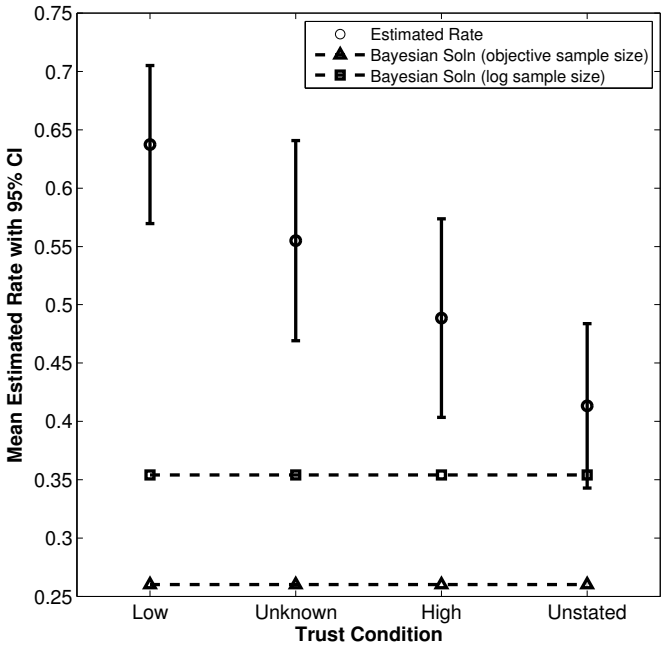
Figure 7. Scatterplots of Experiment 1 rate estimates versus model predictions incorporating base rate trustworthiness. All trials are displayed as if the base rate were 75% (i.e., values from 25% base rate scenarios – squares - are plotted as [100-rate]). Each subplot shows one participant's data from all 32 conditions of Experiment 1 and is labeled with the model best describing their responses and the posterior probability of that model. Grey regions indicate the quadrants where the named model predicts responses will fall, to aid in visual inspection.

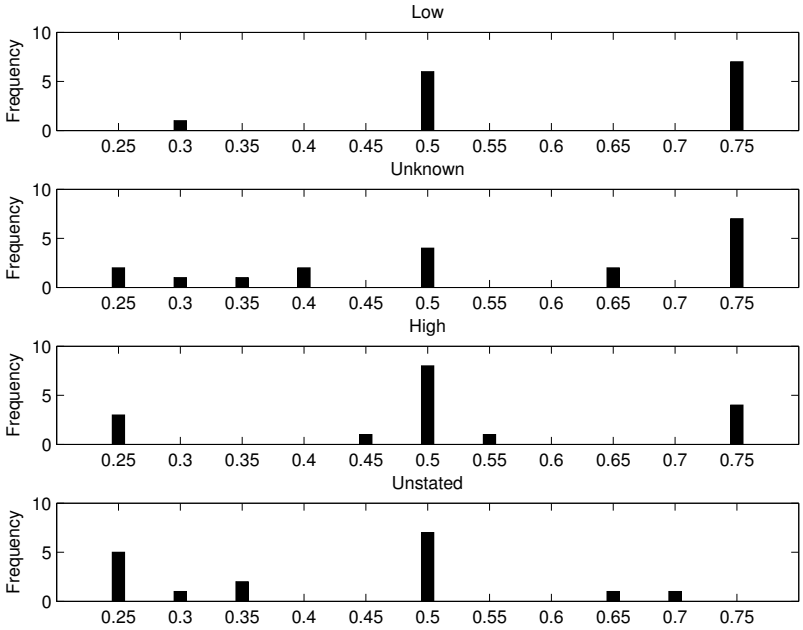

n_0 = 20
n_0 = 200
Estimated Rate
75
50
25
4
8
12
Size of the New Sample (n_1)
high trust data
low trust data
model prediction
naive Bayes prediction

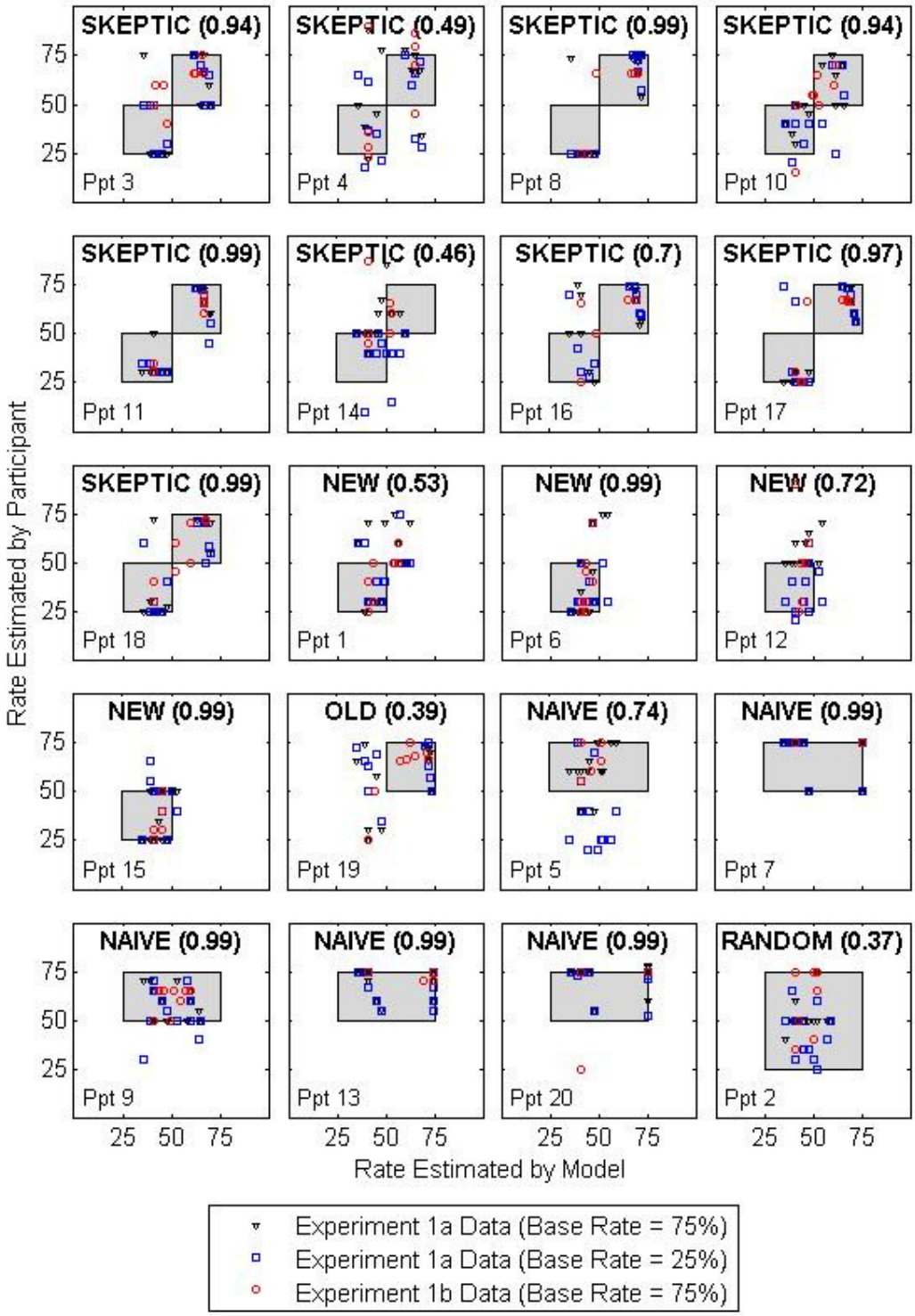

75
50
25
Estimated Rate
n_0 = 20
4
8
12
Size of the New Sample (n_1)
n_0 = 200
high trust data
low trust data
model prediction
naive Bayes prediction


(Inferred) Subjective Sample Size
40
35
30
25
20
15
10
5
0
4
8
12
20
200
Actual Sample Size


75
50
25
Estimated Rate
0
1
2
3
Number of Reasons to Distrust
Human Data
Model Predictions


Estimated Rate
Bayesian Soln (objective sample size)
Bayesian Soln (log sample size)
Mean Estimated Rate with 95% CI
0.75
0.7
0.65
0.6
0.55
0.5
0.45
0.4
0.35
0.3
0.25
Low
Unknown
High
Unstated
Trust Condition


Low
Frequency
10
5
0
0.25 0.3 0.35 0.4 0.45 0.5 0.55 0.6 0.65 0.7 0.75
Unknown
Frequency
10
5
0
0.25 0.3 0.35 0.4 0.45 0.5 0.55 0.6 0.65 0.7 0.75
High
Frequency
10
5
0
0.25 0.3 0.35 0.4 0.45 0.5 0.55 0.6 0.65 0.7 0.75
Unstated
Frequency
10
5
0
0.25 0.3 0.35 0.4 0.45 0.5 0.55 0.6 0.65 0.7 0.75


SKEPTIC (0.94)
Ppt 3
SKEPTIC (0.49)
Ppt 4
SKEPTIC (0.99)
Ppt 8
SKEPTIC (0.94)
Ppt 10
SKEPTIC (0.99)
Ppt 11
SKEPTIC (0.46)
Ppt 14
SKEPTIC (0.7)
Ppt 16
SKEPTIC (0.97)
Ppt 17
SKEPTIC (0.99)
Ppt 18
NEW (0.53)
Ppt 1
NEW (0.99)
Ppt 6
NEW (0.72)
Ppt 12
NEW (0.99)
Ppt 15
OLD (0.39)
Ppt 19
NAIVE (0.74)
Ppt 5
NAIVE (0.99)
Ppt 7
NAIVE (0.99)
Ppt 9
NAIVE (0.99)
Ppt 13
NAIVE (0.99)
Ppt 20
RANDOM (0.37)
Ppt 2
25
50
75
Rate Estimated by Participant
Rate Estimated by Model
Experiment 1a Data (Base Rate = 75%)
Experiment 1a Data (Base Rate = 25%)
Experiment 1b Data (Base Rate = 75%)