

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. Instead, we reconsider Bar-Hillel’s relevance theory, presenting results suggesting that base rates differing in their perceived degree of trustworthiness are correspondingly discounted by people. More generally, we argue that base rates are, intrinsically, less relevant than context-specific data and that base rate neglect may well be a rational strategy.

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 oftentimes 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 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, arguing that people should, as a result, adapt their behavior in the real world would actually hinder performance.

Guided by these points, 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 a series of three experiments 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 ignoring 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 – regardless of 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. Expanding on this, recent work by Krynski and Tenenbaum (2007) looked at the use of causal explanations in base rate neglect problems, showing that Bayesian causal models can predict base rate neglect and, further, 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.

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 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 the 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. More generally, when deciding how much faith to place in a base rate in ecologically-based decisions, at least four environmental factors would appear to be relevant:

- *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 aging” 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.

While these factors seem intuitively reasonable as potential reasons why people might be expected to possess a general, prior bias to neglect base rates to some extent, it is important to demonstrate that people are appropriately sensitive to them 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. 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. The approach is similar to, but more systematic than, the variations considered by Bar-Hillel (1980). It also differs 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.

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). 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. Twenty-four scenarios were created so each participant would see a scenario in each condition.

Procedure. All scenarios were incorporated into a GUI and presented in random order. Participants sat at the computer and read the introductory cover story 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 24 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). Suppose the participant (implicitly) makes the assumption that the observed data reflect some unknown Bernoulli probability θ , and reports the expected value for θ given the data. If n_0 denotes the number of observations that make up this prior 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 the obvious estimator for θ to report is simply

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

This estimator can be justified as a maximum likelihood estimator from a frequentist perspective, and as the Bayesian posterior mean for a standard beta-binomial model under a non-informative Haldane prior (see Jaynes, 2003 for a detailed exposition).

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, then a natural description of the prior is a $\text{Beta}(tx_0, tn_0 - tx_0)$.² Updating in the usual Bayesian manner, we might expect the participant to report the value,

$$E[\theta|r_0, r_1, n_0, n_1, t] = \frac{tr_0n_0 + r_1n_1}{tn_0 + n_1}, \quad (3)$$

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 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: “even the stupidest of men, by himself and without any instruction (which is a remarkable thing), is convinced that 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 to be expected when one applies Bayes’ theorem

to the *subjective* sample values, with some constant “trust” effect expected to arise due to the cover story:

$$E[\theta|r_0, r_1, \tilde{n}_0, \tilde{n}_1, t] = \frac{tr_0\tilde{n}_0 + r_1\tilde{n}_1}{t\tilde{n}_0 + \tilde{n}_1}. \quad (4)$$

In order to fit the data from the 24 conditions, we fit 4 values for t (high and low trust for both base rates), and 4 values for subjective sample size. Assuming that $f(20) = 20$, we estimate the values for \tilde{n} that correspond to $f(4)$, $f(8)$, $f(12)$ and $f(200)$, which are expected to be more-or-less invariant across experimental conditions. 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, it will transpire that $f(\cdot)$ has a very regular form, allowing these parameters to be fixed in a sensible fashion, and leaving only the explicit trust parameters (i.e., the t values) as truly ‘free’.

Results

Since the simplified framework discussed here makes no provision for extrapolation (i.e., participants perceiving a trend and thus estimating a value outside the range dictated by the base and new rates), only those 14 participants whose data show no evidence of extrapolation (i.e., all 24 judgments lie in the range [25, 75]) are considered in this initial investigation, though all 20 participants will be discussed in a later analysis. Figures 1 and 2 show the mean estimates for the underlying probability 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.

Experiment 2

Method

Experiment 2 aimed to expand on the three factors that contributed to the cover story in Experiment 1. The design of the experiment was the same as for Experiment 1 and was, in fact, conducted simultaneously – using the same 20 participants, with the various conditions intermixed with those used in the first study. In the Experiment 2 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%. 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 1 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:

$$E[\theta|r_0, r_1, n_0, n_1, t_l, t_a, t_s] = \frac{(t_a t_l t_s) r_0 \tilde{n}_0 + r_1 \tilde{n}_1}{(t_a t_l t_s) \tilde{n}_0 + \tilde{n}_1}. \quad (5)$$

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. As shown in Table 2, each manipulation has a substantial effect. Changing the age or source of the data lowers trust to $2/3$, while changing the location lowered trust to $1/3$. 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 1 (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 1. Similarly, the low trust value of 0.23 from Experiment 1 is close to prediction from Experiment 2, 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 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 3

Method

Experiment 3 was conducted to address two issues regarding the design of the previous experiments. The first was that the results of Experiments 1 and 2 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 1; 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.

Procedure. Given the use of only a single question rather than the multiple versions used in Experiments 1 and 2, 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 (triangles) or the logarithmically-scaled subjective sample size (squares) 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 1 and 2. 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 3, in line with previous findings regarding the strength of base rate neglect in between- and within-subjects designs (Birnbau & 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 of the 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. 2 These results, taken in conjunction with those of Experiments 1 and 2, 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 3 are in keeping with previous research (Birnbbaum & 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 the original 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 2 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 $E[\theta] + \text{noise}$, 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 3. 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 effect of the manipulations are 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 1 and 2, since across the two experiments each of the 20 participants provided judgements for 32 different problems. 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 5.³ The results are shown in Figure 7, which plots the model predictions $E[\theta]$ along the horizontal axis and the participant responses on the vertical axis. Each panel shows all 32 responses from a single participant: circles corresponding to the trials that constituted Experiment 2 (75% base rate), while the squares (25% base rate) and triangles (75% base rate) correspond to Experiment 1. 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.

As is clear from inspection (and confirmed by statistical analysis - see Appendix), 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 weak evidence suggesting that this pattern holds for another two (4 and 16). These individuals, referred to as “Skeptics” in Tables 3 and A1, 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 three cases there is either strong (6, 15) or weak (1) evidence suggesting that the participants strongly favor the new sample, but in a fashion that is consistent with a low-trust account, and there is one case (19) in which the old data is favored in a manner that is weakly consistent with a high-trust account. These individuals are labelled “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 12 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 8 participants, however, four (7, 9, 13, 20) consistently favored the old data in a manner that suggests something close to the Bayesian solution as traditionally described (labelled “Naive” in Tables 3 and A1). Of the final four participants, one (2) is not easily characterized (and therefore labelled “Null”), while three (5, 12 and 14) display a pattern that we had not considered a priori: they almost always (about 88% of the time) gave responses of 50% or higher, regardless of what the prior sample and the new sample said (labelled “Big”). In short, 19 participants produced interpretable data, of which 16 corresponded to patterns that one

might have expected a priori. Of those 16, 4 are consistent with naive Bayesian updating, 11 are consistent with the trust model and the remaining 1 (participant 19) is consistent with both the trust model and naive Bayesian updating. 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.

For the 12 participants whose data are consistent with the trust model, Table 3 displays estimates for the three trust parameters. In line with the estimates from the averaged data presented earlier (see Table 2), the general pattern is that data collected elsewhere are discounted most heavily, with both older data and data collected by other people discounted less.

General Discussion

As noted earlier, our base rate neglect paradigm differs significantly from that commonly used and, as such, our results need to be viewed in light of the knowledge that 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 favour of newer (but completely commensurate) information, in a manner that is 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; emphasizing some of the drawbacks of more traditional explanations of base rate neglect, which ignore the nature of base rates and the environments in which people have to operate. 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, largely 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 judgement 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 don't 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. 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 pick 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 isn’t 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 questions from a psychological testing point-of-view, is that they are inherently uncertain. While “decision making under uncertainty” is a common description for psychological decision making research, it is very frequently an inaccurate one. In fact, most such research requires that the researcher know the “true” answer in advance, so that any bias in participants’ responses can be measured. In principle this is a good experimental method, but in practice we suspect that the intuitive statistical models that people rely on are actually much more sophisticated than the naive one applied by the researcher to calculate the “correct” answer, leading to what we might call an illusion of irrationality. For example, in the coin-spinning problem a participant might, quite rationally, decide that the question is unimportant, and make up a random answer, which the experimenter then incorrectly interprets as evidence of a flaw in the participant’s reasoning. With this in mind, we suspect that future research might need to rely more on problems such

as the unknown-phone-number problem: although it might require greater sophistication on our part to figure out what counts as a good answer or a bad answer to these, more realistic questions.

On a different topic, an interesting extension of this work would be to consider the role of individual differences in response strategy when choosing what actual numbers people prefer to use (Baird et al, 1970). Given the level of uncertainty associated with any base rate encountered in the real world, one would expect that people's responses to such problems will be less than confident – and the less confidence people become, the more strongly they are likely to round their responses towards “significant numbers.” As a result, a response of 25% may simply indicate the participant's belief that the true number probably lies closer to 25% than to 50%. This conclusion is supported both by the initial modeling referred to earlier – in which an explicit model of number preference seemed to improved the predictive power of the trust model – and by recent work by the authors showing that people *do* use number preference to flag how accurate and precise an estimate should be interpreted as being (Welsh, Navarro & Begg, under review).

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 both ignore potentially misleading base rate data (and we suggest that, almost by definition, every real world base rate is potentially misleading) and respond using numbers that seemingly reflect low confidence in their answers, ought to be interpreted as evidence in *favor* of their rationality rather than as biases to be overcome. In short, to quote the protagonist of the TV series *House*, “everybody lies”: the only time a datum should be completely trustworthy is when you collect it yourself (and probably not even then).

References

- 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.
- Welsh, M.B., Navarro, D.J. & Begg, S.H. (under review). Number preference, data precision and implicit confidence. Manuscript submitted for publication.
- 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

Model selection

As noted in the main text, in the individual subject analysis we estimated values of t_a , t_l and t_s for each participant, using a uniform prior on t and reporting the posterior mode (equivalent to maximum likelihood estimation). Thus, the scatterplots in Figure 7 plot the model prediction for $E[\theta]$ against the response x made by the participant, where each datum is shown in the “75% baserate, 25% new data” format. A simple descriptive measure of agreement is provided by the correlation coefficient between $E[\theta]$ and x (i.e., regression with intercept fixed at 0 and slope 1), but to determine whether the model fit is adequate, we need to be a little more careful. Since, as noted in the text, people have a tendency to “round to the nearest multiple of 5 or 10”, a conservative approach would be to declare that when both x and $E[\theta]$ lie in the interval $[50, 75]$, then the model has correctly predicted that the old sample (base rate) contributes more to the human judgment than the new sample (an “O+” datum). Similarly, if both x and $E[\theta]$ lie in the interval $[25, 50]$ the model has predicted correctly that the new sample is more important (an “N+” datum). Conversely, if x lies in $[75, 50]$ but $E[\theta]$ does not, then the model has failed to predict that the participant relies more on the old sample (an “O–” datum). Vice versa, if x is in $[25, 50]$ and $E[\theta]$ is not, we have an N– datum. If $x = 50$, the participant has weighted the two samples equally (a “50” datum), which could be either model-consistent or model-inconsistent. Finally, if $x < 25$ or $x > 75$, the participant has extrapolated beyond the range of the two samples (an “EX” datum), and the model cannot predict the response. We can then count the frequency of each type of datum for each participant (see Table A1), and use this to quantitatively choose between one of seven different explanations, based on the relative frequency of N+, N–, O+ and O– (for simplicity, we treat 50 and EX data as ambiguous and so exclude those data for the present purposes). The seven accounts are as follows:

- **NULL EFFECT.** In the null effect case, we assume that all four events (N+, N-, O+ and O-) are equally likely with probability 1/4.
- **RANDOM EFFECT.** In the random effect case, we assume that all four events have different probabilities, but with no particular pattern (that is, we assume a uniform prior on the probabilities).
- **NEGLECT.** For the neglect account, we assume that a classic base rate neglect effect occurs, and moreover that this effect is inconsistent with the trust model. Neglect implies that the data are more likely to be N than O, and model inconsistency implies that they are no more likely to be + than -. In short, we assume $P(N+) = P(N-) > P(O+) = P(O-)$.
- **NAIVE.** The logical complement of the neglect account is a naive-Bayesian who sides with the old sample in a manner that that is inconsistent with the trust model That is, $P(O+) = P(O-) > P(N+) = P(N-)$.
- **SKEPTIC.** The skeptic account corresponds to a participant who sometimes sides with the new sample and sometimes sides with the old sample, but tends to do so in a manner that is consistent with the trust model. That is: $P(N+) = P(O+) > P(N-) = P(O-)$.
- **NEW.** An important submodel to consider is the intersection of neglect and skeptic: namely a person who sides with the new sample most of the time, in a manner that is consistent with the trust model. In this case we assume $P(N+) > P(N-) = P(O+) = P(O-)$.
- **OLD.** The last account to consider is the reverse of the last one, in which the participant generally gives responses consistent with the old sample, but also consistent with the model. hat is: $P(O+) > P(N+) = P(N-) = P(O-)$.

Formally, the null effect model assigns probability to an observed data set X as follows:

$$P(X|\text{null}) = 4^{-N} \tag{6}$$

The random effect model is a standard Dirichlet-multinomial model, in which the marginal likelihood is:

$$P(X|\text{random}) = \frac{n_1!n_2!n_3!n_4!}{(N+3)!} \tag{7}$$

where n_1, n_2, n_3 and n_4 denote the numbers of observations in each cell. For the neglect model, the naive model and the skeptic model, note that we have two “high probability” cells and two “low probability cells (so we denote them the “2-2” models). If N_h denotes the number of observations that fall in a high probability category and N_l is the number that have low probability, the marginal probability is:

$$\begin{aligned}
P(X|2-2) &= \int_{1/2}^1 \left(\frac{\phi}{2}\right)^{N_h} \left(\frac{1-\phi}{2}\right)^{N_l} 2 d\phi \\
&= 2^{1-N} \int_0^{1/2} (1-u)^{N_h} u^{N_l} du \\
&= 2^{1-N} \sum_{k=N_l+1}^{N+1} \frac{N_h!N_l!}{k!(N+1-k)!} \frac{1}{2^{N+1}} \\
&= 4^{-N} \sum_{k=N_l+1}^{N+1} \frac{N_h!N_l!}{k!(N+1-k)!}
\end{aligned} \tag{8}$$

For the new and old models, the approach is similar, but in these cases there is one high probability cell and three low probability cells. For these “3-1” models,

$$\begin{aligned}
P(X|3-1) &= \int_{1/4}^1 \phi^{N_h} \left(\frac{1-\phi}{3}\right)^{N_l} (4/3) d\phi \\
&= \frac{4}{3^{N_l+1}} \int_0^{3/4} (1-u)^{N_h} u^{N_l} du \\
&= \frac{4}{3^{N_l+1}} \sum_{k=N_l+1}^{N+1} \frac{N_h!N_l!}{k!(N+1-k)!} \left(\frac{3}{4}\right)^k \left(\frac{1}{4}\right)^{N+1-k} \\
&= 4^{-N} 3^{-N_l+1} \sum_{k=N_l+1}^{N+1} \frac{N_h!N_l!3^k}{k!(N+1-k)!}
\end{aligned} \tag{9}$$

Applying Bayes’ rule gives the posterior probabilities associated with the i th model,

$$P(M_i|X) = \frac{P(X|M_i)P(M_i)}{\sum_{j=1}^7 P(X|M_j)P(M_j)} \tag{10}$$

where we assume $P(M_j) = 1/7$ for all j . (It should be noted that since the model fitting process attempts to move the data into the two + cells, and there are three free parameters that are fit in the process, an optimal selection procedure should correct for this by penalizing those accounts that allow $N+$ or $O+$ to be higher probability – this is trivial to do via methods such as BIC, but such methods assume that the fit is optimized with respect to the

model in question, which is not the case here: all seven of these models make use of the trust model estimates, which will tend to push data into the $N+$ and $O+$ sectors, but are not optimized for that precise task. While it is quite possible to extend the analysis to do this sensibly, we are loathe to introduce even more statistical analysis into this paper).

Setting aside participants 5, 12 and 14 for the moment (see below), when we apply this analysis to the data shown in Table A1, it turns out that there is an extremely clear winner for 12 of the 17 remaining participants (posterior probability $> .94$ for one of the seven models), and in only one case (participant 19) is there a near-tie. Overall, there are 8 participants for whom the “skeptical” explanation is best, with a further 3 given the “new” label, implying that one of the model-consistent explanations is best in 11 cases. In 4 cases, we observe a model-inconsistent “naive” result, plus one case of a “null” effect and one case where the odds are close to even between the “old” and “naive” explanations. Overall, there are clear but sensible individual differences, with the majority of participants behaving in accordance with the trust model, but a non-trivial minority of people behaving in a manner according with a naive application of Bayes Theorem.

To understand why participants 5, 12 and 14 were excluded, note that in Figure 7, for all three of these participants the circles and triangles are almost always in the N sector (i.e., 50+), and the squares are in the O sector (i.e., 50-). Moreover, in order to draw the data in the “75% is always the old sample” format, only the trials corresponding to squares needed to be recoded (since those were the ones with the 25% base rate). In short, these three people almost always gave answers of 50% or higher, regardless of how the trial was structured (27/32 trials for participant 5, 28/32 for participant 12, and 29/32 for participant 14). Obviously, this pattern of behavior could be captured by the trust model by allowing for the possibility that participants have some informed prior beliefs about the problem, but since there are only three cases to work with, we avoid introducing this post hoc extension in this case.

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 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, as well as the reviewers of the *Cognitive Science* conference paper that this manuscript extends.

Footnotes

¹More 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.

²Technically, the statistical model in this case involves a trivial generalization of the Binomial distribution to handle the fact that t can vary continuously. However, the derivation is both straightforward and of no real interest for the current purposes, so we omit it in the interests of conciseness.

³We also fitted a model in which participants were assumed to have some tendency to prefer round or “natural” numbers (e.g., 25, 40, 50, etc) in line with theories of number preference (e.g., Baird, Lewis & Romer, 1970) and the pattern observed in Figure 6. Doing so improved model performance to some extent, but the model predictions are less easily visualized and, since the general pattern of results is much the same, we restrict the discussion to the simpler model.

Table 1

Estimated trust statistics for the low and high trustworthiness conditions, as a function of the underlying base rate.

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

Table 2

Estimated effect on trust of each element of the cover story. Values reported correspond to the low-trust case, with the high-trust value fixed at 1 for this analysis. Not surprisingly, varying the location of the prior data decreases trust in the base rate by the largest amount.

	location	age	source
<i>t</i>	0.34	0.63	0.62

Table 3

Individual subject parameter estimates for the 12 participants for whom some version of the ‘distrust-of-base-rates’ model is the preferred account. Details for the model selection process can be found in the appendix.

ID	location	age	source	r	model
4	1.00	0.12	1.00	.19	skeptic?
1	1.00	0.35	0.77	.40	new?
19	0.16	0.72	0.26	.47	old? naive?
16	0.12	1.00	0.51	.54	skeptic?
3	0.62	0.17	0.91	.56	skeptic
15	0.69	1.00	1.00	.59	new
10	0.47	0.39	0.94	.62	skeptic
6	0.76	1.00	0.76	.64	new
17	0.11	0.77	0.75	.70	skeptic
18	0.39	0.21	1.00	.72	skeptic
8	0.09	0.71	0.71	.76	skeptic
11	0.09	1.00	1.00	.82	skeptic
M	0.46	0.62	0.80		

Table A1

Data, preferred models and correlations. Numbers in parentheses denote the posterior probability associated with each model.

ID	N+	O+	O-	N-	50	EX	model	correl.
7	17	0	0	11	4	0	naive (.99)	
9	10	1	1	10	10	0	naive (.99)	
13	19	0	0	13	0	0	naive (.99)	
20	16	1	0	13	0	2	naive (.99)	
19	16	5	0	7	4	0	old? (.39) naive? (.36)	.47
6	2	22	1	3	4	0	new (.99)	.64
15	0	18	1	2	11	0	new (.98)	.59
1	5	12	0	5	10	0	new? (.53)	.40
3	12	9	0	3	8	0	skeptic (.94)	.56
8	16	14	0	2	0	0	skeptic (.99)	.76
10	12	8	2	1	7	2	skeptic (.94)	.62
11	15	15	1	0	1	0	skeptic (.99)	.82
17	16	13	0	3	0	0	skeptic (.97)	.70
18	15	12	1	2	2	0	skeptic (.99)	.72
4	9	6	4	2	1	10	skeptic? (.49)	.19
16	16	8	0	4	4	0	skeptic? (.70)	.54
2	5	5	4	3	15	0	null? (.37)	
5	6	5	4	15	0	2	big? (-)	
12	1	10	2	6	11	2	big? (-)	
14	6	6	3	2	11	4	big? (-)	

Figure Captions

Figure 1. Participants estimated rates for the various conditions in which the base rate was 25%. Triangles denote data from conditions involving the “high trust” cover story, and the circles show the “low trust” condition data. The thin lines are standard error bars. The dashed lines show the predictions of the naive Bayesian model (Equation 2), while the solid lines show the predictions made by the descriptive model.

Figure 2. Participants estimated rates for the various conditions in which the base rate was 75%. The format of the plot is the same as for Figure 1.

Figure 3. Subjective sample sizes inferred from participants’ probability judgments follow an approximately logarithmic function.

Figure 4. Actual and predicted values for participants’ estimates for the underlying rate in experiment 2. Empirical values are shown by white circles with standard error bars shown. Model predictions based on the estimated trust effects in Table 2 are shown with crosses.

Figure 5. Mean estimated rate (with 95% CI) by trust condition in experiment 3. 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 (complete-trust; i.e., $t = 1$) Bayesian solutions based on objective sample sizes (triangles; 26.0%) and logarithmically scaled estimates of subjective sample sizes (squares; 35.4%). Only in the “unstated” condition do the confidence intervals include either of the Bayesian solutions, and then only barely.

Figure 6. Empirical distribution of estimated rates for all four trust conditions in Experiment 3. Clearly, individual responses tend to match the base rate (25%), the likelihood (75%) or the average of the two (50%). The majority of the effect of the manipulation lies in the *proportion* of responses belonging to each category.

Figure 7. Scatterplots of individual participants' estimates from Experiments 1 and 2 versus predictions from individual models incorporating the degree of trustworthiness of the base rate. Circles are data from Exp 2 (base rate = 75%), squares and triangles are data from Exp 1 (base rate = 25% and 75%, respectively) with all trials displayed as if the base rate were 75% (i.e., estimates and model predictions shown as squares are plotted as $[100 - \text{rate}]$). The top row shows participant results conforming to naive Bayesian solution – tending to be close to the rate in the old data (75%) regardless of what the model predicts. The third and fourth rows show individuals whose estimates conform to the model predictions – high when the model predicts high and low when it predicts low. The more complex patterns seen in rows 2 and 5 are discussed in the main text and appendix. Correlations between model predictions and participant estimates are shown for the 12 cases where some variant of the model is the best explanation for an individual's estimates.













