Optimal Predictions in Everyday Cognition: The Wisdom of Individuals or Crowds?

Michael C. Mozer
Department of Computer Science and Institute of Cognitive Science
University of Colorado, Boulder, CO 80309–0430 USA
mozer@colorado.edu

Harold Pashler
Department of Psychology, UC San Diego
La Jolla, CA 92093 USA
hpashler@ucsd.edu

Hadjar Homaei
Department of Computer Science and Institute of Cognitive Science
University of Colorado, Boulder, CO 80309-0430
hadjar@gmail.edu

Abstract

Griffiths and Tenenbaum (2006) asked people to make predictions about the duration or extent of everyday events (e.g., human lifespans), and reported that predictions were optimal, employing Bayesian inference based on veridical prior distributions. Although the predictions conformed strikingly to statistics of the world, they reflect averages over many individuals. We constructed a simple, heuristic approximation to the Bayesian model, one of whose assumptions is that individuals have access merely to a sample of k instances drawn from the relevant distribution. We investigated how large k would have to be to explain the magnitude and variability of the group response reported by Griffiths and Tenenbaum. Two instances proved sufficient. Thus, the accuracy of the group response is achieved by averaging across individuals—a *wisdom of crowds* effect (Galton, 1907)—and is consistent with individuals having crude and impoverished internal models and simple reasoning heuristics.

Keywords: Bayesian models, prediction, wisdom of crowds, optimality, probabilistic models, heuristics

Introduction

In 1906, Francis Galton was impressed with an event in which visitors to the West of England Fat Stock and Poultry Exhibition were each asked to write down their individual estimates of the weight of a certain ox. Obtaining the original responses, Galton noted that the group average (1197 pounds) was strikingly close to the measured weight of the ox (1198 pounds). This effect, ultimately a reflection of the statistical law of large numbers, has come to be commonly referred to as the *Wisdom of Crowds* effect. (See Surowiecki, 2004, for a highly readable review.) The present article points out that this phenomenon can lead to an inflated estimate of the amount of information individuals possess about real world distributions.

Griffiths and Tenenbaum (2006; henceforth abbreviated G&T) evaluated individuals' ability to make conditional estimates regarding "everyday" domains with which they would have had some first- or second-hand experience. Some were commonplace, such as human lifespans and move run times; others were less so, such as cake baking times and the reigns of pharaohs. In their study, G&T asked individuals questions such as

> If you were assessing an insurance case for an 18-year-old man, what would you predict for his life span?
>
> If your friend read you her favorite line of poetry, and told you it was line 5 of a poem, what would you predict for the total length of the poem?
>
> If you opened a book about the history of ancient Egypt to a page listing the reigns of the pharaohs, and noticed that at 4000 BC a particular pharaoh had been ruling for 11 years, what would you predict for the total duration of his reign?
>
> If you were calling a telephone box office to book tickets and had been on hold for 3 minutes, what would you predict for the total time you would be on hold?

The average responses revealed what G&T termed a "close correspondence between peoples implicit probabilistic models and the statistics of the world." To elaborate, G&T constructed a normative prediction based on Bayesian inference and a veridical prior distribution over the domains in question, which G&T were able to obtain from various sources on the web (e.g., mortality statistics by age).

The normative model yielded an excellent fit to the human predictions, suggesting that the computations underlying higher-level kinds of judgment and reasoning may have a statistical sophistication that has often been assumed to be absent from the domain of higher-order cognition (even though it is often believed to be present in perceptual inference). We now describe the G&T analysis in more detail, and then propose an alternative account, which suggests quite different conclusions about the nature of higher-level judgment and reasoning.

The G&T Analysis

Consider a prediction query of the form, "If a person has lived to age t_{cur}, what age t_{total} are they likely to live to?" G&T modeled human predictions with a theory based on four key claims:

1. **Optimal (Bayesian) inference**: Individuals make a prediction for t_{total} in accordance with Bayes rule, which specifies the posterior distribution for t_{total} as:

$$p(t_{total}|t_{cur}) = \frac{p(t_{cur}|t_{total})p(t_{total})}{\int_{\tau} p(t_{cur}|\tau)p(\tau)}.$$

2. **Prior distribution**: Past real-world experience provides individuals with a veridical prior distribution over the domain in question, $p(t_{total})$. For example, in the case of predicting human life spans, G&T claim that individuals have available a distribution that specifies the probability of living to age t_{total} for any t_{total}.

3. **Likelihood function**: Prediction within the Bayesian framework requires an assumption about how the query was generated, i.e., how the experimenter selects a value of t_{cur}. In the prediction equation, this assumption is cast as $p(t_{cur}|t_{total})$. G&T hypothesize that individuals assume that the experimenter first has in mind a value of t_{total}, and then chooses a t_{cur} from a uniform distribution over the interval $[0, t_{total}]$. In Tenenbaum and Griffiths (2001), this assumption is referred to as the *size principle*.

4. **Prediction function**: Formulating a scalar prediction for t_{total} requires summarizing the posterior distribution, $p(t_{total}|t_{cur})$, in some manner. G&T assume that individuals compute the median of the distribution.

An Alternative Approach: Reasoning From Samples

Suppose that individuals do not have available veridical prior distributions over each domain, but can recall merely a sample of instances of size k that they have encountered or heard about. Let's refer to this conjecture as the *k-sample* assumption. If k is small, each individual has sparse knowledge. For example, knowing about $k = 2$ poems that have 5 and 12 lines total is hardly what one would consider to be a "close correspondence" to the veridical prior distribution, $p(t_{total})$, over poem lengths, which requires knowledge of the proportion of all poems that have length t_{total} for all t_{total}.

Even though individuals have sparse knowledge by the k-sample assumption, the collective mind of the crowd may have a complete picture of the veridical prior distribution. Further, the k-sample assumption does not preclude the possibility that individuals reason according to the G&T Bayesian model, with noisy sample-based prior distribution replacing the veridical prior distribution.

Our investigation of the k-sample assumption asks two distinct but related questions. First, how small can k be and still obtain predictions of comparable accuracy to the G&T Bayesian model? Second, can the computation of the G&T Bayesian model—even with the veridical prior distribution replaced by a small-sample prior distribution—be simplified by some heuristic algorithm? To anticipate our results, we find that a heuristic algorithm with $k = 2$ obtains fits as good as if not better than G&T. This result suggests a different perspective on everyday reasoning than the G&T Bayesian model implies.

The Minimum-of-k-Samples Model

We now elaborate the k-sample assumption into a simple heuristic model, which we refer to as the $minimum-of-k-samples$ model, or Min$k$Samples. Like the G&T Bayesian model, Min$k$Samples predicts a quantity t_{total} given a value of the query point, t_{cur}, for some domain. The model may not have the theoretical elegance of the G&T Bayesian model, but it is intuitive and directly maps to cognitive mechanisms.

Given a query, Min$k$Samples posits that an individual first retrieves a sample of k instances from memory. The model is neutral as to whether memory retrieval is implicit or explicit. Of the retrieved samples, only those with values at least as large as t_{cur} are relevant to the query. (If the query specifies a movie has already grossed \$20M, then any movie known to gross less than \$20M is irrelevant because it fails to satisfy the presupposition of the query.) Discarding the irrelevant samples, the individual reports the minimum value of the remaining samples. When all available samples are irrelevant to the query, the individual ventures a guess that is proportional to the query point, t_{cur}. (For example, if the query concerns the total baking duration of a cake that has been in the oven for 60 minutes, the individual might simply guess 25% above the current baking time, or 75 minutes.)

Formally, Min$k$Samples operates as follows:

1. A set of k samples, $S = s_1, s_2, ...s_k$, is drawn from the prior distribution of the domain.

2. Irrelevant samples are discarded, forming a new set $S' = \{s_i | s_i \geq t_{cur}\}$.

3. If $|S'| > 0$, the model's prediction is $\min_i s'_i$.

4. If $|S'| = 0$, the model's prediction is a proportion g larger than the query, i.e., $(1+g)t_{cur}$.

Methodology

Griffiths and Tenenbaum (2006) reported results from eight domains: cake baking times (in minutes), terms of U.S. representatives (in years), life spans (in years), movie grosses (in hundreds of million dollars), pharaoh reigns (in years), poem lengths (in lines), movie run times (in minutes), and waiting times (in minutes).

For each domain, G&T collected data from over 125 participants: 126 participants for cakes, 130 for U.S. representatives, 197 for life spans, 174 for movie grosses, 191 for pharaoh reigns, 197 for poems, 136 for movie run times, and

158 for waiting times. Each participant was queried with five values of t_{cur} for a domain; for example, the query values for cake baking times were 10, 20, 35, 50, and 70 minutes.

To obtain data from Min*k*Samples we performed a simulation experiment with the same number of simulated participants for each query as G&T studied. The procedure for obtaining a prediction from each simulated participant is presented above. Tom Griffiths provided us with the empirical prior distributions from six of the domains, obtained from sources on the world-wide web (see Griffiths & Tenenbaum, 2006). For the other two domains—wait times and pharaohs—G&T did not use an empirical prior, but instead used hypothetical priors—a power-law distribution for wait times and an Erlang distribution for pharaohs. Each of these distributions had one free parameter that G&T fit to the human data. (The Erlang has two free parameters, but one was constrained such that the mean of the distribution matched participants' estimate of the average reign of pharaohs.) Although we could legitimately have set these parameters to obtain the best fit to our model, we instead used the same parameters as G&T. The one free parameter of Min*k*Samples is the multiplicative guessing factor, g.

G&T summarized the outcome of each experiment by determining the median response of the participants to each query. We did the same with Min*k*Samples, yielding a single prediction from the model for each simulation experiment. We performed 100 replications of the simulation experiment, and obtained the mean and standard deviation over replications of the simulation experiment.

Results

Figure 1 presents results for the eight domains studied by Griffiths and Tenenbaum (2006). Each graph includes the median responses of human participants in the G&T experiments (blue squares), predictions from the G&T-Bayesian model (red dashed lines), and predictions from Min*k*Samples with $k = 2$ (i.e., Min2Samples) and $g = 0.3$ (solid green lines). The error bars on the human data and on Min2Samples will be discussed shortly.

To quantify the goodness of fit of each model to the data, we computed the normalized root mean squared error (NRMSE) between the models and the data at the query points, defined as

$$NRMSE = \left(\frac{\sum_i (h_i - m_i)^2}{\sum_i (h_i - \bar{h})^2} \right)^{1/2}, \qquad (1)$$

where h_i and m_i are the human data and model prediction for query i. Examining Table 1, Min2Samples achieves a lower NRMSE than G&T-Bayesian for six of the eight domains, performing worse only on pharaoh reigns and life spans. Min3Samples also achieves a better fit than G&T-Bayesian for six of the eight domains, performing worse only on run times and life spans. Min1Samples (not shown) does not perform as well as either Min2Samples or Min3Samples.

Where does Min2Samples fail? Although the NRMSE is higher for Min2Samples than G&T-Bayesian on pharaoh reigns, it is impossible to see a qualitative difference in performance between the models when examining the graph (third row, first column) in Figure 1. G&T-Bayesian does come closer than Min2Samples to human data for query points $t_{cur} = 1, 7, 11$, but as the error bars suggest, these are the least reliable data. (More on the error bars shortly.) Moreover, the predictions of G&T-Bayesian for this particular data set were based not on a veridical prior distribution, but on a hypothetical prior distribution constructed by G&T. G&T found that their model produced a poor fit to the data using the veridical prior. Consequently, G&T assumed that participants did not have much knowledge of pharaoh reigns beyond the general shape and mean of the distribution. G&T therefore elected to use an Erlang distribution with one free parameter to fit the data. (The Erlang has two free parameters, but one was constrained by the mean reign.) We did not tune the parameter for fits with Min*k*Samples. Therefore, G&T-Bayesian had an additional degree of freedom that Min*k*Samples did not.

The second domain for which Min2Samples underperformed G&T-Bayesian was life spans. Examining the graph (third row, second column of Figure 1), it is evident that the poor fit of Min2Samples stems from the rightmost query point, $t_{cur} = 96$. For $t_{cur} = 96$, Min*k*Samples is unlikely to sample an individual who lived beyond this age; consequently, the model will guess using the g factor, which will produce a prediction of 124.8 years for the life span. Certainly participants in the G&T experiment are aware that people rarely live to this age, and as a result might lower their guess. Min*k*Samples lacks this world knowledge. Because g has a significant effect on only the final query point, we might lower g for this domain to reflect general knowledge about life spans. Reducing g by a factor of ten, Min2Samples outperforms G&T-Bayesian, shown in Figure 2 and quantified in terms of NRMSE in column 4 of Table 1.

We emphasize once again that Min2Samples and Min3Samples outperform G&T-Bayesian on *six of eight* domains. We discussed the two remaining domains in detail to discount the concern that Min*k*Samples shows any pathological deficiency.

Free Parameters

Min*k*Samples has one free parameter, g. Although this free parameter was chosen to fit the human data, it has a relatively weak effect on the model's predictions, and its effect is primarily seen for the rightmost query point of each graph, where the set of samples drawn beyond the query point is most likely to be empty.

Individual Variability

The key claim of Min*k*Samples is that each participant reasons from a very small number of examples. Consequently, response variance among participants should be quite high. Min*k*Samples could be ruled out as a candidate explanation

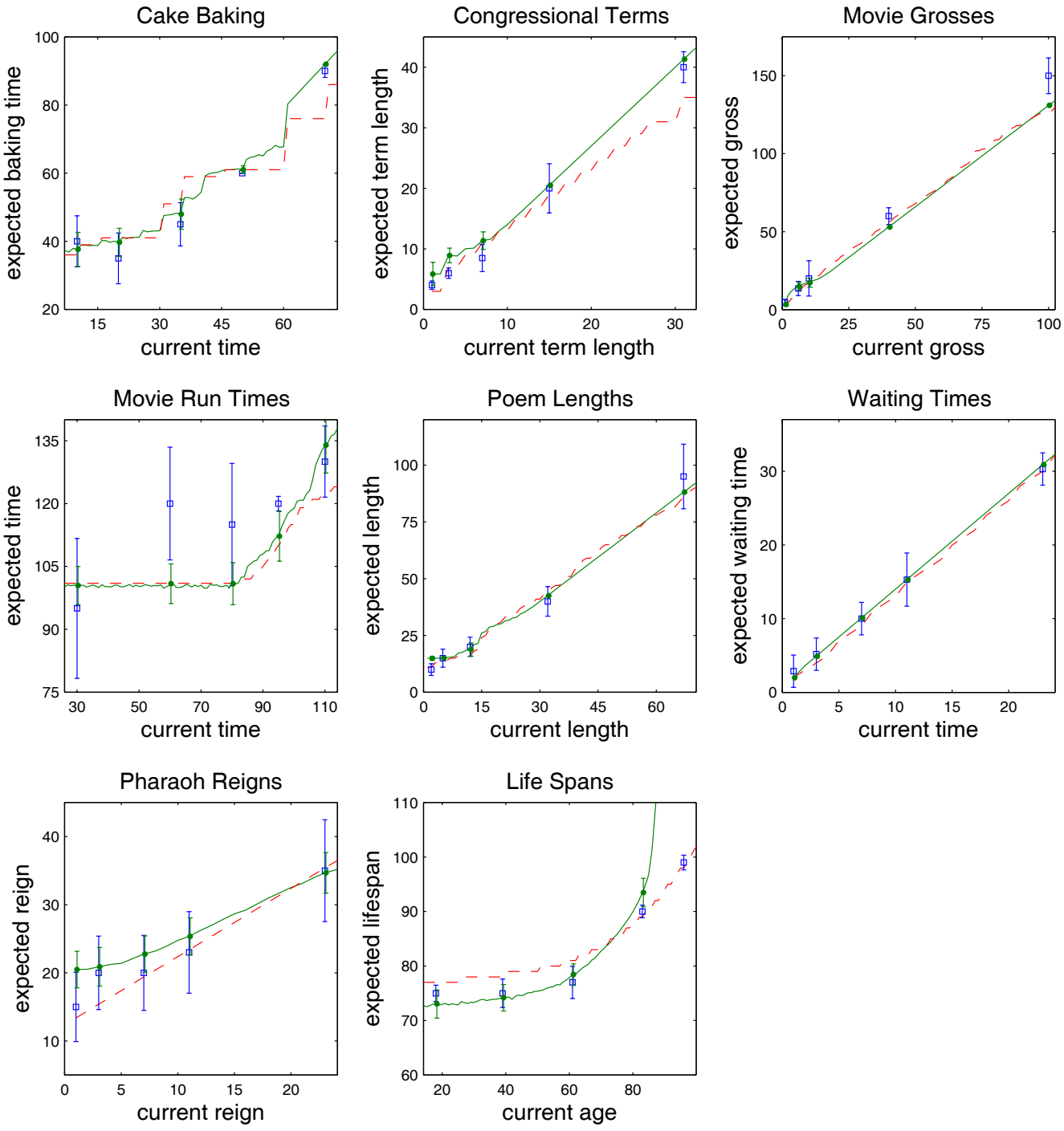

Figure 1: Human and simulation results on eight everyday prediction tasks. The blue squares indicate median human responses from the experiments of Griffiths and Tenenbaum (2006). The G&T Bayesian-model prediction is indicated by the dashed red line. The Min2Samples prediction, with $g = 0.3$, is indicated by the solid green line. The error bars surrounding the human data and the Min2Samples predictions at the query points denote ± 2 standard deviations in experimental outcome. The error bars for the human data were obtained by G&T via bootstrap sampling; the error bars for Min2Samples were obtained via 100 replications of the simulation experiment.

Table 1: NRMSE Comparison of G&T-Bayesian and Min$k$Samples

Domain	G&T-Bayesian	Min2Samples	Min2Samples with $g = .03$	Min3Samples	Min3Samples with $g = .03$
Cake Baking	0.3804	0.1455		0.1996	
Congressional Terms	0.1960	0.1573		0.1109	
Movie Grosses	0.1980	0.1712		0.1706	
Movie Run Times	1.0629	1.0004		1.3156	
Poem Lengths	0.1489	0.1284		0.1262	
Waiting Times	0.0835	0.0503		0.0503	
Pharaoh Reigns	0.3301	0.4418		0.3102	
Life Spans	0.2572	1.2537	0.1898	1.3067	0.4543

for the data if it produces greater variability than G&T's participants.

Ideally, we would like to know the inter-participant response variance, but this measure was not available in G&T's paper or in materials that they provided to us. Instead, G&T reported a bootstrap estimate of inter-experiment variance. This estimate indicates the variability one would expect if the entire experiment were replicated many times. Replicating the experiment involves obtaining data from 125+ participants, and then computing the median of their predictions. Because G&T use the median, not the mean, as a summary statistic, the inter-participant variance is not equivalent to the inter-experiment variance. Nonetheless, it is a close proxy, and the inter-experiment variance offers insight into the inter-participant variance. (When one is small, the other is small; when one is large, the other is large.)

The human data in Figure 1 includes error bars that denote ± 2 standard deviations on the inter-experiment distribution, as G&T estimated by a 1,000-sample bootstrap. We also estimated inter-experiment distribution with Min2Samples, and the Min2Samples predictions at the query points are shown with error bars that denote ± 2 standard deviations. Because simulation studies permit an unlimited supply of simulated participants, instead of bootstrap sampling a finite set of participants, we simply generated new participants for each of 100 replications of the experiment.

As the error bars clearly indicate, the variability of the human participants is at least as large as that obtained by Min$k$Samples. Thus, even though Min$k$Samples produces significant inter-participant variability because each response is based only on k samples, this variability is no larger than that observed in the G&T human studies.

Discussion

When the Griffiths and Tenenbaum (2006) paper first appeared, its conclusion that everyday reasoning can be cast as optimal (Bayesian) inference seemed astonishing and radical to many who learned of the work. Beyond surprise, many were swayed by the elegance of the work. The research also had an impact outside the academic community. Consider the following quote, from *The Economist*:

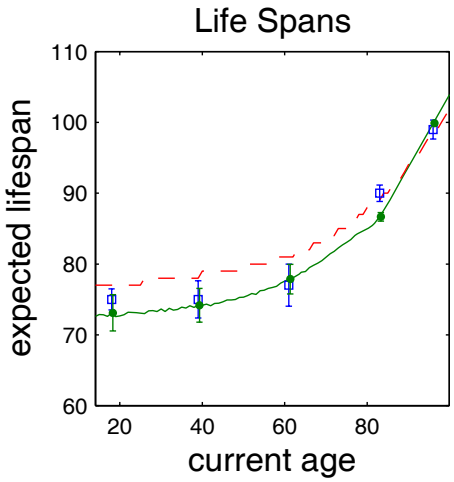

Figure 2: Life span simulation, with $g = .03$.

"[Griffiths and Tenenbaum]...put the idea of a Bayesian brain to a quotidian test. They found that it passes with flying colors.

The key to successful Bayesian reasoning is ... in having an appropriate *prior*, as it is known to the cognoscenti. This prior is an assumption about the way the world works—in essence, a hypothesis about reality—that can be expressed as a mathematical probability distribution of the frequency with which events of a particular magnitude happen...

With the correct prior, even a single piece of data can be used to make meaningful Bayesian predictions.

Indeed, one of the most impressive things Dr Griffiths and Dr Tenenbaum have shown is the range of distributions the mind can cope with. Besides Erlang, they tested people with examples of normal distributions, power-law distributions and, in the case of baking cakes, a complex and irregular distribution. They found that people could cope equally well with all of them, cakes included. Indeed, they are so confident of their method that they think it could be reversed in those cases where

the shape of a distribution in the real world is still a matter of debate." (*The Economist*, 1/5/2006)

The message transmitted by G&T's work is that individual minds encode complex prior distributions in domains casually encountered in daily life, and that individual minds are Bayesian and utilize these prior distributions to draw complex inferences. In contrast, the present article shows that the results are quite consistent with a far less dramatic possibility: individual minds may reason from only a small number of instances—two or three—and that the mechanisms of reasoning may be simple heuristic algorithms.

How can these two perspectives—embodied in the G&T Bayesian and Min*k*Samples models—both be consistent with the data? The answer lies in the wisdom of crowds. Even if any one individual has very limited knowledge and inference capabilities, combining estimates over a large population—greater than 125 in the G&T experiments—allows the population to be well characterized from a Bayesian perspective.

Levels of Analysis

A proponent of Bayesian approaches may argue that G&T-Bayesian is something like what linguists have referred to as a competence theory, whereas Min*k*Samples is a performance theory. That is, Min*k*Samples is a mechanistic approximation of the G&T-Bayesian theory. Alternatively, one might cast the two theories as being at different levels of analysis in the Marr sense: G&T-Bayesian is a computational level theory, whereas Min*k*Samples is an algorithmic level theory. Min*k*Samples and G&T-Bayesian are similar, in some sense: the predictions of the two models for a large-population average are similar (Figure 1).

Moreover, there is some non-accidental correspondence between Min*k*Samples and G&T-Bayesian. Min*k*Samples utilizes the heuristic of reporting the minimum value of the k samples recalled. This heuristic might be viewed as an approximation to the Bayesian size principle, which biases the posterior distribution to smaller hypotheses. However, the heuristic works well only for small k; for large k, the minimum of a sample will be smaller than the size-principle weighted median of samples.

If our investigations had found that Min*k*Samples or some other sample-based model required, say, $k = 20$ samples per individual to match the data, we would not have considered the sampling account to be a qualitatively different story than the G&T-Bayesian account. However, when $k = 2$ samples per individual accounts for the data, our sense is that the Min*k*Samples and G&T-Bayesian accounts have to be viewed as qualitatively distinct. Certainly, the sort of interpretation described in the *Economist* article quoted above would not be consistent with Min*k*Samples.

One point that a competence-performance or levels-of-analysis distinction makes is that the Bayesian formalism is sufficiently broad that nearly any heuristic or mechanistic account can be cast in Bayesian terms, given the right set of assumptions. While an increased awareness of Bayesian reasoning is obviously a healthy development in cognitive science, an almost religious devotion to this formalism may result in obscuring important psychological distinctions, and an obscuring of the important limitations that apply to human reasoning mechanisms.

Acknowledgments

We thank Tom Griffiths for providing us with the prior distributions used in G&T, as well as the Bayesian model predictions and human data. Thanks also to Victor Ferreira, David Huber, and John Wixted for comments on an earlier draft of the manuscript. This research was supported by National Science Foundation Grants BCS-0339103, BCS-0720375, CSE-SMA 0509521, Institute of Education Sciences Grant SBE-0542013 (G. Cottrell, PI), and US Department of Education Grants R305H020061 and R305H040108 (H. Pashler, PI).

References

Galton, F. (1907). Vox Populi. *Nature*, *75*, 450-451.

Griffiths, T. L., & Tenenbaum, J. B. (2006). Optimal predictions in everyday cognition. *Psychological Science*, *17*, 767-773.

Surowiecki, J. (2004). *The wisdom of crowds: Why the many are smarter than the few and how collective wisdom shapes business, economies, societies, and nations*. New York, NY: Random House.

Tenenbaum, J. B., & Griffith, T. L. (2001). Generalization, similarity, and Bayesian inference. *Behavioral and Brain Sciences*, *24*, 629-640.