Cognitive Science 32 (2008) 1133–1147

ISSN: 0364-0213 print / 1551-6709 online
DOI: 10.1080/03640210802353016

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

Michael C. Mozer[a], Harold Pashler[b], Hadjar Homaei[a]

[a]*Department of Computer Science and Institute of Cognitive Science, University of Colorado*
[b]*Department of Psychology, University of California at San Diego*

Received 7 March 2008; received in revised form 24 June 2008; accepted 26 June 2008

Abstract

Griffiths and Tenenbaum (2006) asked individuals to make predictions about the duration or extent of everyday events (e.g., cake baking times), 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. On the conjecture that the accuracy of the group response is chiefly a consequence of aggregating across individuals, we constructed simple, heuristic approximations to the Bayesian model premised on the hypothesis that individuals have access merely to a sample of k instances drawn from the relevant distribution. The accuracy of the group response reported by Griffiths and Tenenbaum could be accounted for by supposing that individuals each utilize only two instances. Moreover, the variability of the group data is more consistent with this small-sample hypothesis than with the hypothesis that people utilize veridical or nearly veridical representations of the underlying prior distributions. Our analyses lead to a qualitatively different view of how individuals reason from past experience than the view espoused by Griffiths and Tenenbaum.

Keywords: Bayesian models; Everyday reasoning; Normative models; Wisdom of crowds; Instance-based reasoning

1. 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 (1,197 pounds) was strikingly close to the measured weight of the ox (1,198 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

Correspondence should be sent to Michael Mozer, Department of Computer Science, University of Colorado, Boulder, CO 80309-0430. E-mail: mozer@colorado.edu

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 movie 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

1. If you were assessing an insurance case for an 18-year-old man, what would you predict for his lifespan?
2. 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?
3. 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?
4. 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" (p. 767). 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.

2. 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:

$$p(t_{\text{total}}|t_{\text{cur}}) = \frac{p(t_{\text{cur}}|t_{\text{total}})p(t_{\text{total}})}{\int_\tau p(t_{\text{cur}}|\tau)p(\tau)}. \quad (1)$$

2. Prior distribution: Past real-world experience provides individuals with a veridical prior distribution over the domain in question, $p(t_{\text{total}})$. For example, in the case of predicting

human lifespans, 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 $p(t_{\text{cur}}|t_{\text{total}})$. 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 $[0, t_{\text{total}}]$. 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, $p(t_{\text{total}}|t_{\text{cur}})$, in some manner. G&T assume that individuals compute the median of the distribution.

3. 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* hypothesis. 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, $p(t_{\text{total}})$, over poem lengths, which requires knowledge of the proportion of all poems that have length t_{total} for all t_{total}.

Although individuals have sparse knowledge according to the k-sample hypothesis, the collective mind of the crowd may have a far more complete picture of the veridical prior distribution. One should thus be wary of assuming that each individual possesses whatever may be implied by the aggregate data. The history of experimental psychology contains many such examples of misleading aggregate data (e.g., see Estes, 1956; Maddox, 1999; Siegler, 1987).

Our investigation of the k-sample hypothesis 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 noisy sample-based 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, the G&T-Bayesian model. This result suggests a different perspective on everyday reasoning than the G&T-Bayesian model implies.

3.1. *The minimum-of-*k*-samples model*

We now elaborate the k-sample hypothesis into a simple heuristic model, which we refer to as the *minimum-of-k-samples* model, or Min$k$. Like the G&T-Bayesian model, Min$k$ 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$k$ 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 $20 million, then any movie known to gross less than $20 million 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} (e.g., 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 min).

Formally, Min*k* operates as follows:

1. A set of k samples, $S = s_1, s_2, \ldots, s_k$, is drawn from the prior distribution of the domain.
2. Irrelevant samples are discarded, forming a new set $S' = \{s_i | s_i \geq t_{\text{cur}}\}$.
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., $(1 + g)t_{\text{cur}}$.

We wish to emphasize that Min*k* does not require any individual to internalize the veridical prior. Only nature needs to know the prior in order to serve up k samples to any individual.

Finding the minimum of the samples seems arbitrary; why not the mean or the max? One argument is that individuals may treat the task as a search for a sample that is similar to but larger than the query, t_{cur}. The minimum of the samples is the nearest neighbor to t_{cur}. In essence, the Min*k* model suggests that individuals treat the task as a memory retrieval task, and base their response on similarity of the query to the stored instance.

3.2. *The G&T* k-*samples model*

In addition to the heuristic Min*k* model, we also explored two variants of the G&T-Bayesian model that utilize only k samples. Rather than utilizing the veridical prior distribution, these variants base the prior solely on the k samples. In the first variant, which we refer to as GT*k*Guess, the prior is defined to have probability mass

$$\hat{P}(t_{\text{total}}) = \frac{1}{k} \sum_{i=1}^{k} \delta_{s_i, t_{\text{total}}}, \tag{2}$$

where $\delta_{\cdot,\cdot}$ is the Kronecker delta. Just as Min*k* needed to handle the situation in which the available samples are irrelevant to the query, so must GT*k*Guess because the posterior is undefined in this situation and cannot serve as the basis for a response. One solution—as the "guess" in GT*k*Guess implies—is to make the same assumption that Min*k* does: Individuals formulate a guess that is a fixed proportion g above the query point. The formal definition of GT*k*Guess is, therefore identical to Min*k*, with the "min" function in Step 3 of the algorithm replaced by the median of the posterior (Equation 1), computed using the sample prior (Equation 2).

The second variant of the G&T-Bayesian model we propose, which we refer to as GT*k*Smooth, solves the undefined-posterior problem in perhaps a more principled manner

than GT*k*Guess, and in a manner more in the spirit of probabilistic approaches. GT*k*Smooth assumes that individuals treat the samples not as single observations, but as representative of a distribution of cases (e.g., a Gaussian distribution centered on the sample, truncated at zero)—that is, the prior distribution is a smoothed version of that used in GT*k*Guess:

$$\hat{P}(t_{\text{total}}) \propto \begin{cases} \sum_{i=1}^{k} \exp\left(\frac{(t_{\text{total}} - s_i)^2}{2\sigma^2}\right) & \text{for } t_{\text{total}} \in \mathbb{N} \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

Because the prior has nonzero probability mass for all $t_{\text{total}} \in \mathbb{N}$, the posterior can be determined regardless of whether the sample values are larger than t_{cur}. Although this blurring of the samples is a more elegant assumption than the g guessing factor, it nonetheless also involves one free parameter, σ.

4. Methodology

Griffiths and Tenenbaum (2006) reported results from eight domains: cake baking times (in minutes), terms of U.S. representatives (in years), lifespans (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 lifespans, 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 one of five values of t_{cur} for a domain; for example, the query values for cake baking times were 10, 20, 35, 50, or 70 minutes.

To obtain predictions from our three models—Min*k*, GT*k*Guess, and GT*k*Smooth—we simulated the same number of participants for each query as G&T tested. 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. Each of our models had one free parameter: g for Min*k* and GT*k*Guess, and σ for GT*k*Smooth. We coarsely tuned the free parameters by hand to obtain the best performance, which yielded the multiplicative guessing factor $g = 0.30$ and the prior-smoothing parameter $\sigma = 10$.

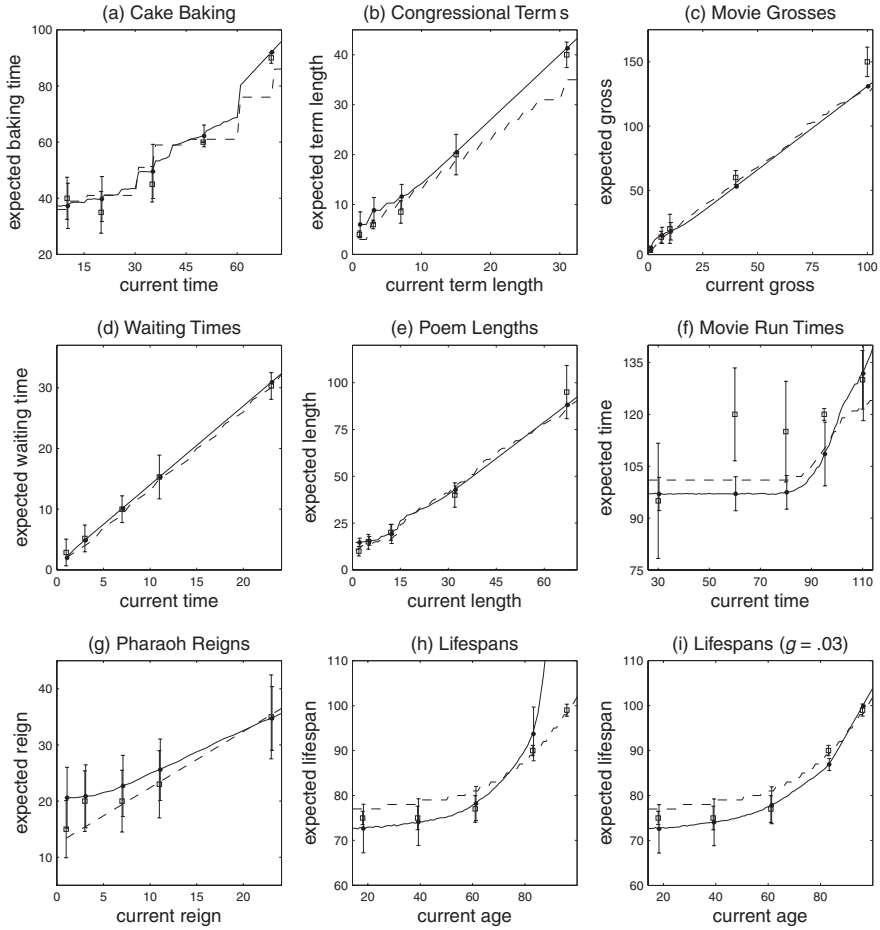

Fig. 1. (a)–(h) Human and simulation results on eight everyday prediction tasks. *Note*: The squares indicate median human responses from the experiments of Griffiths and Tenenbaum (G&T; 2006). The G&T-Bayesian model prediction is indicated by the dashed line. The Min2 prediction, with $g = 0.3$, is indicated by the solid line. The error bars surrounding the human data and the Min2 predictions at the query points denote ± 2 SDs in experimental outcome. The error bars for the human data were obtained by G&T via bootstrap sampling; the error bars for Min2 were obtained via 100 replications of the simulation experiment. (i) The lifespan simulation with $g = 0.03$.

5. Results

5.1. Min*k* versus G&T-Bayesian

G&T summarized the outcome of each experiment via the median response of the participants to each query. We did the same for Min$k$ with $k = 2$ (i.e., Min2), 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.

Fig. 1a through h shows predictions for the eight domains studied by Griffiths and Tenenbaum (2006). Each graph includes the median predictions of human participants in the G&T experiments (blue squares), predictions from the G&T-Bayesian model (dashed lines), and predictions from Min2 $g = 0.3$ (solid lines). The error bars on the human data and on Min2 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

$$\text{NRMSE} = \left(\frac{\sum_{q=1}^{Q} (h_q - m_q)^2}{\sum_{q=1}^{Q} (h_q - \bar{h})^2} \right)^{1/2}, \tag{4}$$

where h_q and m_q are the human data and model prediction for query q, and $\bar{h}$ is the mean human response across queries. The second and third columns of Table 1 show the NRMSE for G&T-Bayesian and Min2 for each of the eight domains. All simulations use $g = 0.3$, except for the score in parentheses for the lifespan simulation, which we discuss shortly. The Min2 scores that outperform G&T-Bayesian are highlighted in boldface type. By the NRMSE measure, Min2 achieves a better fit on five of the eight domains.

The fourth and fifth columns of Table 1 show a different measure of fit for the two models: the mean normalized deviation (MND), which indicates the deviation of the model from the data relative to the uncertainty in the data:

$$\text{MND} = \frac{1}{Q} \sum_{q=1}^{Q} \frac{|h_q - m_q|}{s_q}, \tag{5}$$

where s_q is a bootstrap estimate of the standard deviation in the outcome of the human experiment (details follow). The MND places less weight on the less reliable data points.

Table 1
Comparison of G&T-Bayesian and Min2

	NRMSE		MND	
Domain	G&T-Bayesian	Min2	G&T-Bayesian	Min2
Cake baking	0.38	**0.17**	1.79	**1.13**
Congressional terms	0.20	**0.16**	1.64	**1.10**
Movie grosses	0.20	**0.17**	2.12	**1.83**
Waiting times	0.08	**0.05**	0.52	**0.25**
Poem lengths	0.15	**0.13**	1.20	**0.94**
Movie run times	1.06	1.21	1.73	**1.69**
Pharaoh reigns	0.33	0.45	0.47	0.74
Lifespans	0.26	1.26 (**0.20**)	2.39	4.71 (**1.26**)

Note. G&T = Griffiths and Tenenbaum (2006); NRMSE = normalized root mean squared error; MND- mean normalized deviation.

The Min2 scores that outperform G&T-Bayesian are highlighted in boldface type.

The MND measure yields essentially the same result as the NRMSE: Min2 outperforms G&T-Bayesian on six of the eight domains.

Let's examine where Min2 fails. Although the pharaoh-reigns NRMSE is higher for Min2 than G&T-Bayesian, it is difficult to see a qualitative difference in performance between the models (Fig. 1g). G&T-Bayesian does come closer than Min2 to human data for query points $t_{\text{cur}} = 1, 7, 11$; 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 Min2. Therefore, G&T-Bayesian had an additional degree of freedom that Min2 did not.

For the movie run times (Fig. 1f), G&T-Bayesian is a bit closer on query points $t_{\text{cur}} = 60, 80$—although both models significantly underpredict the data—and Min2 is a bit closer on query points $t_{\text{cur}} = 30, 110$. As the MND measure suggests, when the unreliability of the data is taken into account, performance of the two models is comparable.

The third domain for which Min2 underperformed G&T-Bayesian was lifespans. As Fig. 1h makes evident, the poor fit of Min2 stems from the rightmost query point, $t_{\text{cur}} = 96$. For $t_{\text{cur}} = 96$, Min2 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 lifespan. Certainly, participants in the G&T experiment are aware that people rarely live to this age, and as a result might lower their guess. Because g has a significant effect on only the final query point, we might lower g for this domain to reflect general knowledge about lifespans. Reducing g by a factor of 10, Min2 outperforms G&T-Bayesian, shown in Fig. 1i and quantified by NRMSE and MND (numbers in parentheses in Table 1).

We again emphasize that Min2 outperforms G&T-Bayesian on five of the eight domains. We discussed the three remaining domains in detail to convince the reader that Min2 shows no pathological deficiency. Our aim is not to argue than Min2 is superior to G&T-Bayesian, but only that a preference for one model or the other cannot be justified on the grounds of data fit. In fairness, we note that Min2 has a 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. We also had freedom in choosing how g influences the prediction. Although g might have been an additive constant, a multiplicative constant makes the most sense given that we wanted to allow only one free parameter for all eight domains, the currency of the domains vary, and a multiplicative constant is scale invariant, whereas an additive constant is not. Although the free parameter does make a comparison between G&T-Bayesian and Min2 challenging, we will show in the next two sections that (a) the need for the free parameter is due to the fact that the Min2 is sample based, and not the fact that Min2 lacks the theoretical foundation of G&T-Bayesian; and (b) Min2, as well as other sample-based models, is able to explain aspects of the data that G&T-Bayesian cannot address.

5.2. Comparing sample-based models

In the previous section, we illustrated that Min2, a model based on a very small sample size, provides as good an account of individuals' everyday predictions as the G&T-Bayesian model. Min2 serves as a proof of the sufficiency of two samples to explain the data.

In this section, we argue for the robustness of the result obtained with Min2 by exploring three small-sample models—Min$k$, GTkGuess, and GTkSmooth—for sample sizes k ranging from 1 to 5, and showing that many models based on small sample sizes are adequate to explain the data. The simulation results we report here are for a single simulation experiment with 10,000 simulated participants for each model, a large enough population to ensure the reliability of the median response, and therefore the estimate of NRMSE.

Fig. 2 contains one bar graph for each of the eight domains, and a final bar graph for lifespans with $g = .03$ (as in Fig. 1). The ordinate of each graph is the model NRMSE. The horizontal dashed line shows the NRMSE for G&T-Bayesian. The bars show NRMSE for the three sample-based models as a function of k. Let's first consider the domains in Fig. 2a through e. In all of these domains, GTkSmooth performs poorly relative to the other models. However, Min$k$ and GTkGuess all match or beat the performance of G&T-Bayesian for all values of k, except perhaps for the poem domain and $k = 1$. Now consider the final three domains in Fig. 2f through h. In these domains, GTkSmooth performs well for all $k > 1$. Min$k$ and GTkGuess perform well for certain values of k in Fig. 2f and g, although GTkGuess seems more robust than Min$k$ as a function of k. For Fig. 2h, the lifespan simulation, neither Min$k$ nor GTkGuess perform well, unless the guessing constant g is lowered (Fig. 2i).

Overall, we see that GTkSmooth, the most principled Bayesian sample-based approach, is least convincing in its performance. Min$k$ does well for a small k, but in some domains, the fit worsens as k increases. This result is due to the fact that the model's predictions drop to the query point as k increases. Nonetheless, Min$k$ was proposed for situations where k is likely to be small.

GTkGuess can be viewed as a hybrid of Min$k$ and G&T-Bayesian. For $k = 1, 2$, GTkGuess yields exactly the same predictions as Min$k$. As $k \to \infty$, GTkGuess approaches G&T-Bayesian because the sampling and guessing assumptions of GTkGuess become irrelevant. Thus, it is not surprising that GTkGuess is quite robust over values of k. We therefore view GTkGuess not as a competitor to Min$k$, but rather as a general description of a class of models that is equivalent to Min$k$ for small k—the focus of our discussion.

5.3. Individual variability

We have argued for the plausibility of Min$k$ but the results thus far do not allow us to distinguish between: (a) the view espoused by Min$k$ that each individual reasons based on a few samples, and accuracy arises by combining predictions over individuals; and (b) the view espoused by G&T-Bayesian that each individual holds a veridical prior distribution in his or her mind; or (c) some intermediate state of affairs.

The key empirical distinction that arises between these perspectives is in the variability of individual responses. If each individual reasons from a small number of samples, and the samples available to one individual are independent of the samples available to another

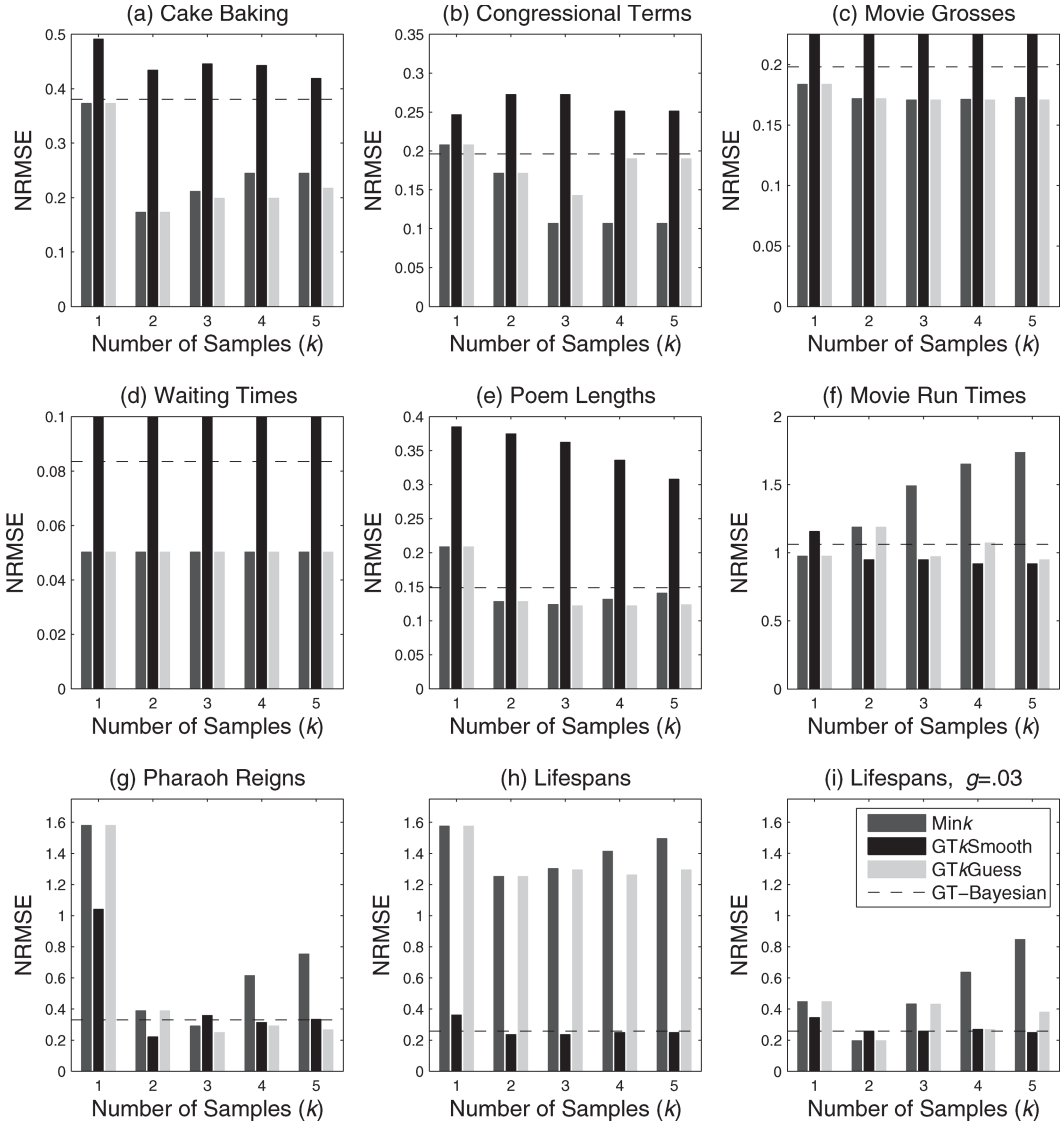

Fig. 2. (a)–(h) Normalized root mean squared error (NRMSE) evaluation of Min$k$, GTkGuess, and GTkSmooth for $k = 1 \ldots 5$ on eight domains. *Note*: Min$k$ and GTkGuess scores are based on $g = 0.30$, and GTkSmooth scores are based on $\sigma = 10$. (i) NRMSE evaluation of the Lifespans data set with $g = 0.03$ for Min$k$ and GTkGuess.

individual, responses will be highly variable. As sample size grows, the sample statistics of two individuals will become more similar, and response variability will drop. Thus, the response variability of human participants can be used to infer the sample size on which reasoning is based. (If we are incorrect in our assumption concerning the independence of samples available to two individuals, variability will be low and not diagnostic of k. Fortunately, it turns out that observed variability is high, as we will show.)

G&T report one measure of variability: 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 predictions. The human data in Fig. 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 variance with Min2, and the Min2 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 1,000 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$k$. Thus, although Min$k$ 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.

A more direct measure of variability than inter-experiment variance is the inter-participant variance. Because G&T report as a summary statistic the median participant response, not the mean, the inter-participant variance is not equivalent to the inter-experiment variance. Nonetheless, they should be strongly related. Inter-participant variance was not discussed by Griffiths and Tenenbaum (2006), but Tom Griffiths kindly provided us with the raw data to compute inter-participant variance. Fig. 3 presents a measure of the inter-participant variability in human responses relative to the inter-participant variability of model responses for domain d,

$$V(d) = \frac{1}{Q}\sum_{q=1}^{Q} \log_{10} \frac{v_{\mathrm{h}}(d,q)}{v_{\mathrm{m}}(d,q)}, \tag{6}$$

where $v_{\mathrm{h}}(d,q)$ is the inter-participant variance for domain d and query point q from the human experiment, $v_{\mathrm{m}}(d,q)$ is the same measure from the model, and Q is the total number of query points (5 in these experiments). If the human variability matches the simulation variability, $V(d)$ will be zero; if individuals are more variable than the simulation predicts for a given k, the value will be positive.

Min$k$, GTkGuess, and GTkSmooth all show the same pattern: The models produce less variability as k increases, just as one would expect. What is surprising is that for $k > 1$, $V(d) > 0$ (i.e., all models show *less* variability than the human participants). This result provides strong evidence that human participants do not rely on a large number of cases in making everyday predictions. If participants relied on, say, $k = 20$ cases, then according to the three models, the inter-participant variance would be far smaller than is actually observed. What might be problematic for the sample-based models is the fact that human participants show more variability than the models for k as small as 1 or 2. We have two explanations for this finding. First, when the models formulate a guess based on the g factor, the resulting variability will be small, and it might be more realistic to treat g not as a constant but a random variable, which will increase the variability of the model's predictions. Second, eyeballing the human data, participants clearly seem to be rounding their predictions in a domain-appropriate manner. For example, for cake baking times, waiting times, and movie run times, most reported

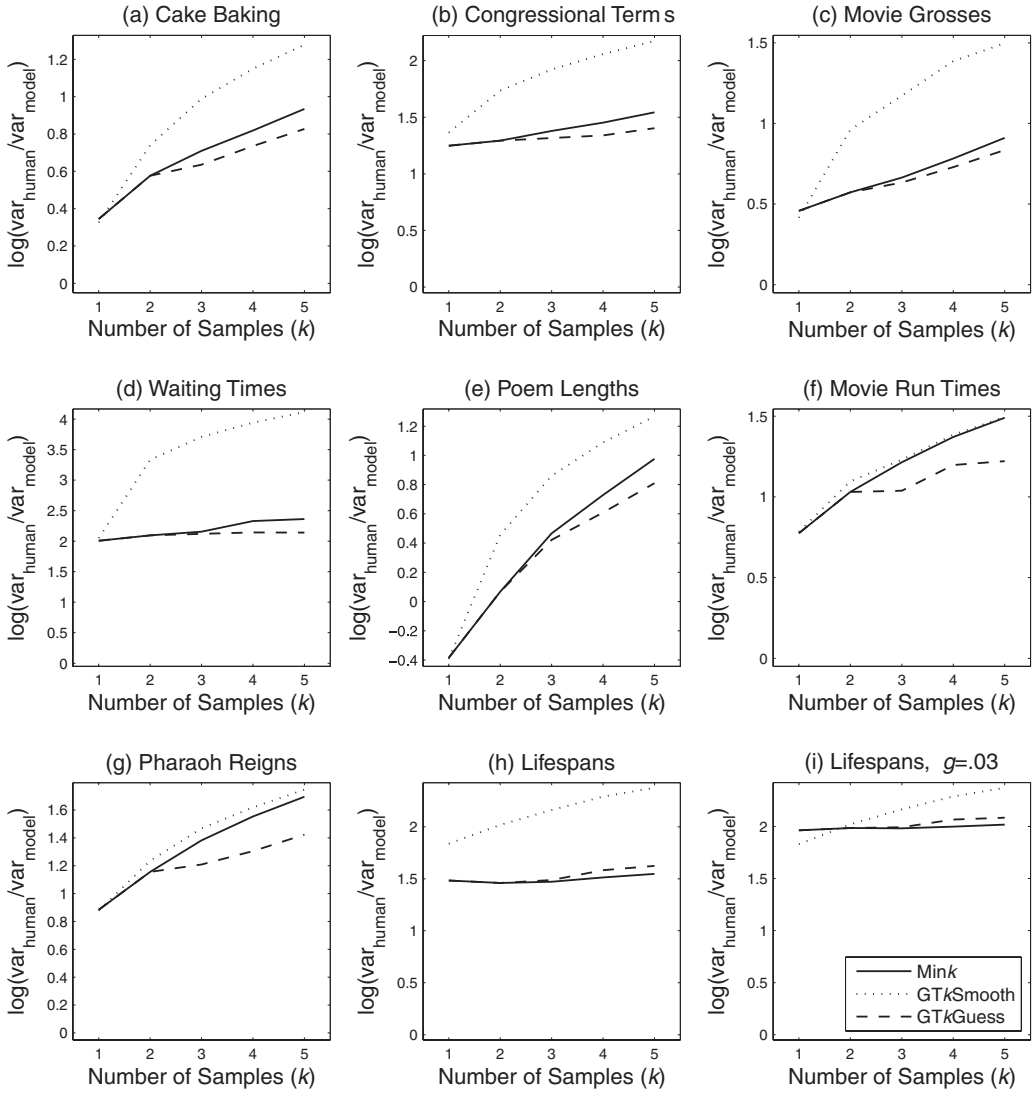

Fig. 3. Inter-participant variance of humans relative to inter-participant variance of simulation models (see Equation 6), for eight domains and three models—-Min$k$ (solid lines), GTkGuess (dashed lines), and GTkSmooth (dotted lines)—as a function of k. *Note*: An ordinate value of zero corresponds to equal variance of humans and simulation; the larger the value, the greater the variance of humans relative to the simulation.

times on the 15- or 30-min marks. It is easy enough to justify the introduction of additional noise sources to the models to raise their response variability to the level of human participants. However, had we obtained the result that the models with small k yielded far *more* variability than the participants, we would have been forced to admit the failure of the models and, consequently, reject the small-sample hypothesis.

6. Discussion

When the Griffiths and Tenenbaum (2006) article 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*:

> [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. (pp. 70–71)

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—one, 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* models—both be consistent with the data? One answer lies in the wisdom of crowds. Even if any one individual has very limited knowledge and inference capabilities, combining estimates over a population allows the population to be well characterized from a Bayesian perspective.

6.1. 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* is a performance theory. That is, Min*k* is a mechanistic approximation of the G&T-Bayesian theory in which only a small number of instances are accessible to an individual for reasoning. Min*k* does not preclude the possibility that different instances are available to an individual at different times.

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* is an algorithmic

level theory. Min*k* and G&T-Bayesian are similar, in some sense: The predictions of the two models for a large population average are similar (Fig. 1).

Moreover, there is some non-accidental correspondence between Min*k* and G&T-Bayesian. Min*k* 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. The Min*k* approximation is best for small k, and small k seems to obtain the best fits to human data (Figs. 2 and 3).

If our investigations had found that Min*k* 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* and G&T-Bayesian accounts have to be viewed as qualitatively distinct. Certainly, the sort of interpretation described in the *Economist* article would not be consistent with Min*k*.

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. Although there is no doubt that it is often very illuminating to view human reasoning from a Bayesian perspective, an overemphasis on the ways in which reasoning conforms to Bayesian principles may draw attention away from important psychological distinctions, and may obscure important memory and processing limitations of human reasoning.

Acknowledgments

This research was supported by National Science Foundation Grants BCS–0339103, BCS–0720375, and CSE–SMA 0509521; Institute of Education Sciences Grant SBE–0542013 (to G. Cottrell, PI); and U.S. Department of Education Grants R305H020061 and R305H040108 (to H. Pashler, PI). Many thanks to Tom Griffiths for providing us with the prior distributions used in G&T, as well as the Bayesian model predictions and human data. This work also benefited from the thoughtful and constructive reviews of Tom Griffiths and two anonymous reviewers, and from comments on an earlier draft of the manuscript by Victor Ferreira, David Huber, Don MacLeod, and John Wixted.

References

Bayes rules. *The Economist, 378* (8459), 70–71.

Dalton, F. (1907). Vox populi. *Nature, 75*, 450–451.

Estes, W. K. (1956). The problem of inference from curves based on group data. *Psychological Bulletin, 53*, 134–140.

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

Maddox, W. T. (1999). On the dangers of averaging across observers when comparing decision bound models and generalized context models of categorization. *Perception & Psychophysics, 61*, 354–374.

Siegler, R. S. (1987). The perils of averaging data over strategies: An example from children's addition. *Journal of Experimental Psychology: General, 116*, 250–264.

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: Random House.

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