Cognitive Science 33 (2009) 969–998
Copyright © 2009 Cognitive Science Society, Inc. All rights reserved.
ISSN: 0364-0213 print / 1551-6709 online
DOI: 10.1111/j.1551-6709.2009.01045.x

The Wisdom of Individuals: Exploring People's Knowledge About Everyday Events Using Iterated Learning

Stephan Lewandowsky,[a] Thomas L. Griffiths,[b] Michael L. Kalish[c]

[a]*School of Psychology, University of Western Australia*
[b]*University of California, Berkeley*
[c]*University of Louisiana, Lafayette*

Received 4 November 2008; received in revised form 6 March 2009; accepted 9 March 2009

Abstract

Determining the knowledge that guides human judgments is fundamental to understanding how people reason, make decisions, and form predictions. We use an experimental procedure called ''iterated learning,'' in which the responses that people give on one trial are used to generate the data they see on the next, to pinpoint the knowledge that informs people's predictions about everyday events (e.g., predicting the total box office gross of a movie from its current take). In particular, we use this method to discriminate between two models of human judgments: a simple Bayesian model (Griffiths & Tenenbaum, 2006) and a recently proposed alternative model that assumes people store only a few instances of each type of event in memory (Min*K*; Mozer, Pashler, & Homaei, 2008). Although testing these models using standard experimental procedures is difficult due to differences in the number of free parameters and the need to make assumptions about the knowledge of individual learners, we show that the two models make very different predictions about the outcome of iterated learning. The results of an experiment using this methodology provide a rich picture of how much people know about the distributions of everyday quantities, and they are inconsistent with the predictions of the Min*K* model. The results suggest that accurate predictions about everyday events reflect relatively sophisticated knowledge on the part of individuals.

Keywords: Iterated learning; Optimal predictions; Bayesian models of cognition

Correspondence should be sent to Stephan Lewandowsky, School of Psychology, University of Western Australia, Crawley, WA 6009, Australia, E-mail: lewan@psy.uwa.edu.au

1. Introduction

Suppose you call your phone company to resolve a perplexing charge for a call to Burkina Faso on your bill. After 2 min on hold listening to insipid music, punctuated by assurances that your call is important to the company, you wonder how much longer you will have to hold. Or suppose you have been waiting for the person ahead of you to finish their transaction with the bank teller for 10 min; will your total waiting time chew up your entire lunch break? Daily life abounds with situations like these, in which we must predict the future on the basis of a single observation. Notwithstanding the impoverished nature of the data—the time already spent waiting—most people would hazard a guess about the ultimate outcome. How accurate are those guesses and what processes determine them? Predictions about the future necessarily involve prior knowledge, but what is the nature of that knowledge? Is that knowledge being used optimally? How could we ever know unless we find a good assay of prior knowledge?

In an initial exploration of these questions, Griffiths and Tenenbaum (2006) asked people to predict outcomes (e.g., a person's life span) from a single sample (e.g., a person's age). Griffiths and Tenenbaum suggested that people's responses were consistent with Bayesian inference using the appropriate prior distribution. For example, when predicting a person's likely total lifetime, responses were consistent with use of the appropriate Gaussian prior (ignoring infant mortality, human life span is approximately normally distributed). Likewise, when predicting the total gross box office intake of a movie from its performance to date, people's responses were consistent with a power-law prior (movie grosses in fact follow a power-law distribution, with most movies making a small amount of money but some movies making a fortune). This concordance between people's responses and known statistical properties of the environment was observed for several other variables, including the rather esoteric task of predicting the total length of the reign of an Egyptian Pharaoh. Griffiths and Tenenbaum concluded that their data suggested ''… a far closer correspondence between optimal statistical inference and everyday cognition than suggested by previous research'' (p. 771). This conclusion must be considered quite provocative and counter-intuitive in light of the plethora of instances in which people's predictions have been shown to be at variance with optimal statistics (e.g., Tversky & Kahneman, 1983).

Indeed, Griffiths and Tenenbaum's (2006) conclusions have not gone unchallenged. In a recent paper, Mozer, Pashler, and Homaei (2008) revisited the method of Griffiths and Tenenbaum and suggested that their data did not permit a strong conclusion because each participant only provided a single guess about each variable under consideration. Hence, the correspondence between Bayesian inferences using the correct prior probabilities and people's responses was only observed in the aggregate, after averaging across a large number of participants, raising the possibility that the result may have been an artifact of aggregation (cf. Estes & Maddox, 2005). Mozer et al. (2008) explored this possibility and showed that the results of Griffiths and Tenenbaum (2006) could be reproduced by a simple non-Bayesian model, called Min*K*, that merely assumed that each person had access to no more than a few relevant instances in memory; when aggregated across individuals, Mozer et al. showed that this impoverished individual knowledge not only captured the aggregate responses observed by Griffiths and Tenenbaum but did so with greater precision than the Bayesian model, suggesting that the results of Griffiths and Tenenbaum arose from the ‘‘the wisdom of crowds’’ (Surowiecki, 2004). However, Mozer et al. (2008) provided no direct empirical test of the proposition implied by their model, namely that individuals *lack* ‘‘wisdom.’’

This article seeks to pinpoint the knowledge underlying people’s performance in the aggregate as well as at the individual level through an experiment that adjudicates between the predictions of Griffiths and Tenenbaum’s (2006) Bayesian model and Mozer et al.’s (2008) Min*K* model. In order to discriminate between the wisdom of crowds and that of individuals, we use a within-subject version of the prediction task in which people make repeated judgments about a single quantity. In particular, we use a novel methodology known as *iterated learning* (e.g., Griffiths & Kalish, 2007; Kalish, Griffiths, & Lewandowsky, 2007) for which the two models make starkly contrasting predictions. In iterated learning, the *input* for one learning episode is based on the *response* from a prior learning event. A prime example is the transmission and evolution of language, which is characterized by speakers learning their language from the data provided by other speakers who were once learners themselves (e.g., Griffiths & Kalish, 2007; Kirby, 2001; Kirby, Dowman, & Griffiths, 2007; Smith, Kirby, & Brighton, 2003; for a recent review, see Griffiths, Kalish, & Lewandowsky, 2008).

To foreshadow our principal conclusions, the experiment provides strong support for the idea that people’s predictions are best described as a Bayesian inference based on the appropriate prior distribution. The data also refute the Min*K* model by challenging its key prediction about how people should behave in an iterated version of the future-prediction task. Our results also demonstrate how the iterated-learning method can be useful for testing psychological models, magnifying what might be a small difference in predictions in a single iteration of learning into a large difference after several iterations (see, e.g., Reali & Griffiths, 2008).

The plan of the paper is as follows. We begin by presenting the prediction task used by Griffiths and Tenenbaum (2006) and its critique by Mozer et al. (2008) in more detail. We then summarize the formal and experimental work that identifies iterated learning as a valuable tool to reveal the knowledge underlying human inductive inference and present the contrasting predictions of the two models. These analyses are followed by the experiment that used an iterated version of the future-prediction task and that supported the predictions of the Bayesian model and rejects those of Min*K*. We conclude that performance on the future-prediction task more closely resembles what we would expect from knowledge of the appropriate prior distribution than from reliance on a small, fixed set of samples from that distribution. Our results thus highlight the wisdom of individuals, rather than just the wisdom of crowds.

2. The future-prediction task: Two competing models

In the prediction task, the optimal Bayesian model for predicting the total duration or extent of a quantity t_{total} when probed with the observation t is:

$$p(t_{\text{total}}|t) \propto p(t|t_{\text{total}})p(t_{\text{total}}), \tag{1}$$

where $p(t_{\text{total}})$ is the prior distribution, and $p(t|t_{\text{total}})$ the likelihood of observing a particular datum (Eq. 1 from Griffiths & Tenenbaum, 2006). This is equivalent to the more general form of Bayes' rule (see Eq. 2 below), with t_{total} playing the role of the hypothesis, and t the role of the data. Thus, the probability assigned to any possible value of t_{total} (the posterior distribution) depends on two factors: the prior distribution of possible values of t_{total} and the likelihood of encountering any particular time t under a given hypothesis about t_{total}. In this article, as in Griffiths and Tenenbaum (2006), the likelihood $p(t|t_{\text{total}})$ is always assumed to be uniformly distributed. The nature of the prior distribution, by contrast, differs with the variables under consideration. For example, whereas movie grosses follow a power-law distribution in the real world, life span is roughly normally distributed, and so on.

Griffiths and Tenenbaum (2006) showed that the *median* predictions across a large number of participants were in accord with this optimal Bayesian model—that is, the observed median matched the median of the predicted posterior distribution. However, the data of Griffiths and Tenenbaum were subject to a number of constraints. First, their study only provided limited snapshots of people's posterior distribution, by recording $p(t_{\text{total}}|t)$ for a highly sparse set of t values. Thus, although the results of Griffiths and Tenenbaum were consistent with the assumption that people had knowledge of the appropriate actual prior, that ''psychological'' prior remained unobserved.

Second, because each participant provided only a single estimate of t_{total} for a given t, Griffiths and Tenenbaum's results did not address the process by which people generated their responses. For example, the data were equally consistent with sampling from the posterior distribution—people generating any possible t_{total} according to its posterior probability—and deterministic responding with, say, the median of the posterior—people providing the same ''best guess'' in response to repeated polling with a constant t.

Expanding on this point, Mozer et al. (2008) provided an alternative account of the results of Griffiths and Tenenbaum that sidestepped the Bayesian framework altogether. Specifically, Mozer et al. rejected the idea that individuals have access to veridical prior distributions and instead assumed that people can merely recall a small number (k) of relevant instances (e.g., the age of one's grandfather at his death) and respond with the smallest recalled value—hence the name Min*K*. If the probe value, t, exceeds the value of all recalled instances, people are assumed to respond with a proportional guess; that is, $t_{\text{total}} = (1 + g) \times t$. Intriguingly, Mozer et al. could quantitatively account for the results of Griffiths and Tenenbaum when k was assumed to be as small as two, leading Mozer et al. to reject the need for a complex Bayesian model and suggesting instead that ''individual minds may reason from only a small number of instances…and the mechanisms…may be simple heuristic algorithms.'' (Mozer et al., p. 1145). It is only when people's responses are considered in the aggregate, by averaging across many samples of size $k = 2$, that the predictions approximate the prior distribution. In a nutshell, whereas Griffiths and Tenenbaum (2006) suggested that people reason in an optimal manner when confronted with impoverished data, Mozer et al. (2008) took a radically opposing view by assuming that people rely on the rather modest ability to recall two instances.

These conflicting viewpoints can be resolved empirically by a within-subject version of the future-prediction task. At first glance, this might appear quite straightforward because all that seems to be required is to present each participant with a suitably large number of trials, whereupon the two competing models could be fit to each individual's data and the ''better'' one is identified by standard model-selection techniques (e.g., Hélie, 2006; Myung, 2000; Myung, Navarro, & Pitt, 2006). However, upon closer inspection, the matter turns out to be considerably more complex—not because standard model-selection techniques are limited but because of some idiosyncratic properties of Min*K*.

The core assumption of Min*K* is that each individual has access to only a few (usually $k = 2$) memorized instances that represent samples from the prior and are assumed to be pre-experimentally acquired. It follows that Min*K* is committed to accounting for the results of a single individual by relying on a single sample of k memorized instances. This commitment throws up a number of difficulties because the predictions of Min*K* differ considerably between different idiosyncratic samples of instances even when k is constant. For example, a person who has memorized life spans of 14 and 36 years will make very different predictions in response to $t = 50$ from someone who recalls 70 and 89 years. Thus, notwithstanding the fact that Min*K* was expressly designed to mimic the performance of individuals, it is actually very difficult to fit to the data of individuals.

One way in which one might apply Min*K* to within-subject data involves repeated sampling, such that the data produced by an individual are modeled by aggregating across a large number of different samples, each of size k. This approach, however, defeats the principal reason for the development of Min*K* because it opens the door to the aggregation fallacy. Another possibility is to fit each individual on the basis of a single sample, but with k additional free parameters that determine the value of each person's memorized instances. This approach, however, is at odds with Min*K*'s core postulate that the memorized instances are samples from the appropriate prior. Treating each memorized instance as a free parameter would also increase the complexity of Min*K* even further, which is discomforting in light of the fact that the competing Bayesian predictions are parameter free.

The challenges of testing Min*K* using data derived from individuals in a conventional within-subject experiment suggest that we need to seek an alterative means of discriminating between Min*K* and the Bayesian model. In the remainder of the paper, we show that this can be done by using an iterated-learning methodology, in which the predictions that people produce in one trial affect the stimuli they see on future trials. In a within-subject iterated-learning experiment, the predictions of Min*K* deviate starkly and inevitably from those of the Bayesian model, allowing the models to be discriminated without requiring any parameter fitting or use of sophisticated model selection techniques.

3. Iterated learning reveals prior knowledge

In iterated learning, the input during each learning episode is based on the response emitted at a previous episode. For example, during language evolution, a given learner is producing utterances—which in turn constitute the stimuli for a subsequent generation of learners—using a grammar that was imputed from the input received from an earlier generation of learners. What are the long-term consequences of such iterated learning? Do learners ultimately converge onto a predictable outcome?

There is no a priori reason to suspect that iterated learning would necessarily converge to an equilibrium; instead, iteration might degenerate from structure into noise, it might descend into random or unpredictable alternation from one solution to another, or people might blend their prior knowledge with the to-be-learned data to form consistent ‘‘compromise’’ solutions. Indeed, if the ‘‘Chinese Whispers’’ party game (also known as ‘‘Telephone,’’ depending on the location of the party) is anything to go by, then one might expect iterated learning to yield entirely unpredictable results.

It may thus come as a surprise that Griffiths and Kalish (2007) were capable of showing by mathematical analysis that iterated learning will—under certain plausible circumstances—necessarily converge to an equilibrium reflecting the learners' prior knowledge or expectations, irrespective of the data provided to each generation. Griffiths and Kalish (2007) analyzed iterated learning for a chain of Bayesian agents that process and transmit information. Each agent receives data from the previous agent in the chain and, using Bayes' rule, seeks to infer from that data the hypothesis entertained by the previous agent. This is done by computing the posterior probability of each hypothesis, $p(h|d)$, by combining its prior probability before seeing any data, $p(h)$, with the likelihood of the observed data under that hypothesis, $p(d|h)$, to give:

$$p(h|d) = \frac{p(d|h)p(h)}{\sum_{h\in\mathcal{H}} p(d|h)p(h)}, \tag{2}$$

where the denominator simply ensures that the resulting probabilities sum to one. The posterior distribution can then be used to select the hypothesis entertained by the agent. For example, each agent could sample a hypothesis from this distribution according to its probability, and then use the sampled hypothesis to generate the data provided to the next learner.

If we assume that the data for each generation are produced by sampling from the likelihood function $p(d|h)$, then the probability that the nth agent chooses hypothesis i given that the previous agent chose hypothesis j is:

$$p(h^{(n)} = i|h^{(n-1)} = j) = \sum_{d} p(h^{(n)} = i|d)p(d|h^{(n-1)} = j), \tag{3}$$

where $p(h^{(n)}=i|d)$ is the posterior probability obtained from Eq. 2. Eq. 3 describes the transition matrix of a Markov chain, with the hypothesis chosen by each agent depending only on the choice of the previous agent. Griffiths and Kalish (2005, 2007) showed that the stationary (i.e., ultimately attained) distribution of this Markov chain is $p(h)$; that is, the chain converges onto the prior distribution assumed by the agents. The chain will converge to this distribution irrespective of the nature of the data and hypotheses involved, provided some simple conditions on the properties of the transition matrix are satisfied (e.g., Norris, 1997). By implication, the probability that the last in a long line of learners chooses a particular hypothesis is simply the prior probability of that hypothesis, regardless of the initial data provided to learners.

In support of the idea that iterated learning should converge to the prior, Kalish et al. (2007) showed that when people learned a continuous function concept (e.g., the relationship between driving speed and stopping distance), the responses of the last in a chain of generations of learners conformed to people's known preferences for linearly increasing functional relationships. Crucially, the stimuli provided for learning at the outset proved irrelevant in the long run: Chains of learners initialized with a range of functions, including negative linear, nonlinear (quadratic), and random relationships between two variables, all reliably converged to positive linear functions after just eight generations of iterated learning.

Iterated learning is not confined to inter-generational transmission—that is, situations in which one person provides data to another—but can also be observed *within* individuals across learning episodes. For example, Griffiths, Christian, and Kalish (2008) used each participant's response on one categorization trial to create stimuli for the same person on later trials. The experiment used category structures defined on three binary dimensions originally introduced by Shepard, Hovland, and Jenkins (1961). These structures reduce to six basic types, with a robust known order of their relative difficulty of learning. Participants were shown a subset of the members of a category (the data) and were asked to identify the most likely structure from which these members had been drawn (the hypothesis). The hypotheses selected in one block of testing were used to generate the data seen in the next block. Notwithstanding the shift from inter-generational to intra-individual transmission, convergence to people's prior expectations was observed, and the structures known to be the easiest to learn rapidly came to dominate people's choice of hypotheses.

4. Iterated learning and the prediction task

The key point of contention between the original Bayesian analysis of Griffiths and Tenenbaum (2006) and the Min*K* model proposed by Mozer et al. (2008) is the nature of the knowledge that people use to make predictions. The fact that iterated learning can reveal the prior knowledge that guides people's inferences suggests that this methodology may be useful in discriminating between these models. In particular, a within-subjects design, in which convergence occurs across trials produced by each participant, would seem to be an ideal tool to examine how people perform the future-prediction task used by Griffiths and Tenenbaum (2006). In this section, we consider the predictions that the two models make for how people should behave in such an experiment.

4.1. Iterated learning and the Bayesian model

The Bayesian model of the prediction task described in Eq. 1 quite naturally entails an instantiation of the transition matrix defined in Eq. 3, which describes the iterated learning of Bayesian agents. Specifically, in our study, people repeatedly provided a prediction for t_{total} in response to a value of t for a number of different quantities. The probe value of t was sampled uniformly from the interval $(0, t_{\text{total}}^{(n-1)}]$, where $t_{\text{total}}^{(n-1)}$ was the person's response on the previous trial involving that chain. The iterations across trials within an individual are therefore summarized by:

$$Q_{ij} = p(t_{\text{total}}^{(n)} = i | t_{\text{total}}^{(n-1)} = j) = \sum_t p(t_{\text{total}}^{(n)} = i | t) p(t | t_{\text{total}}^{(n-1)} = j), \tag{4}$$

which defines the transition matrix, $\mathbf{Q} = (Q_{ij})$, of the Markov chain that converges to the prior distribution $p(t_{\text{total}})$.

It follows that the Bayesian model of the prediction task proposed by Griffiths and Tenenbaum (2006) makes clear predictions about the distributions onto which each individual should converge across multiple iterated-learning trials. These predictions are illustrated in the top row of panels in Fig. 1, which show the actual prior distributions of some of the variables considered by Griffiths and Tenenbaum (and in our experiment below). Owing to the properties just discussed, those priors are also the distributions onto which people are predicted to converge during iterated learning.

Although the Bayesian model expects the distributions in Fig. 1 to hold for each individual as well as in the aggregate, with the modest number of responses from each participant

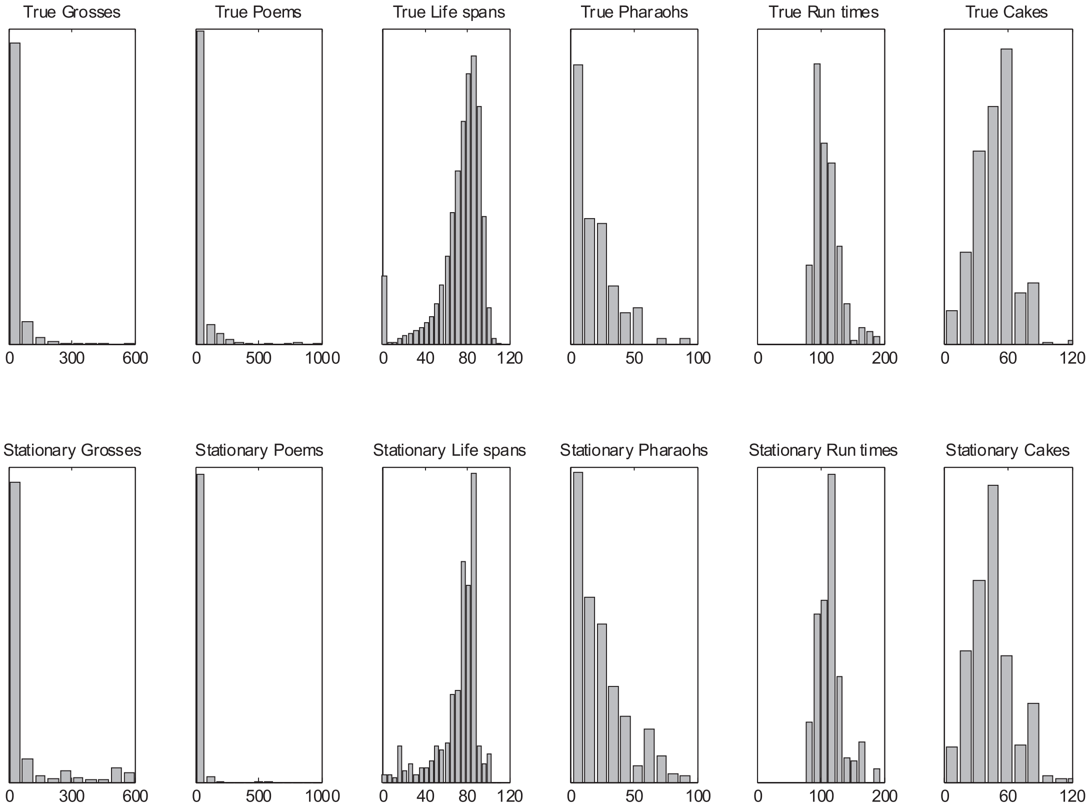

Fig. 1. Actual prior distributions (top row) and stationary distributions observed in the experiment (bottom row) for six of the variables considered in the experiment and by Griffiths and Tenenbaum (2006).

that are typical of most experiments, we can gather more information from individual data by examining the prediction functions each person produces. Prediction functions show responses (t_{total}) as a function of the probe values (t). The relationship between t and t_{total} that is expected under the Bayesian model depends on the nature of the prior distributions (see Griffiths & Tenenbaum, 2006, for a derivation, and see Appendix A for a summary). The top row of panels in Fig. 2 illustrate the prediction functions we should expect to see for different prior distributions. According to the Bayesian model, the responses of each individual should be characterized by the appropriate prediction function for the variable under consideration.

4.2. Iterated learning and MinK

The Min*K* model posits that upon presentation of the probe, t, the set of k recalled instances is first pruned by discarding all instances whose values fall below t (because they

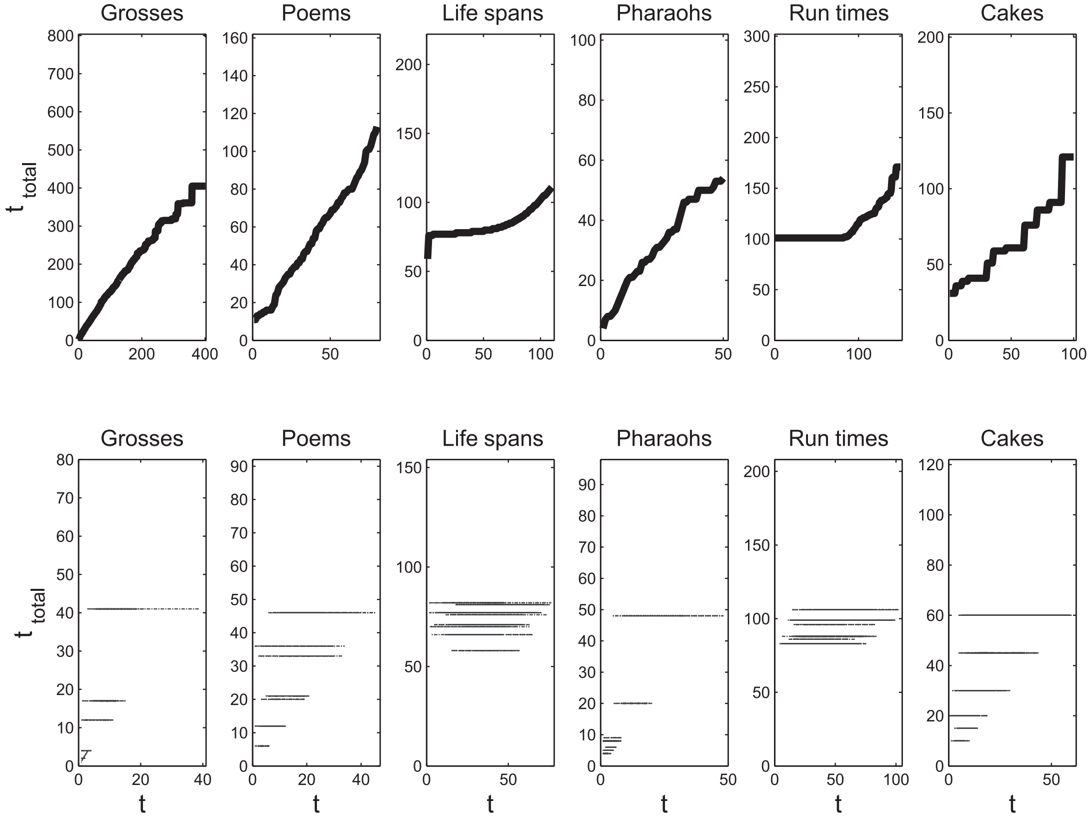

Fig. 2. Prediction functions for Bayesian model (top row) and Min*K* (bottom row) for different prior distributions. For the Bayesian model, each prediction function shows the median of the posterior distribution on t_{total} when probed with t. For Min*K*, these are the prediction functions for 10 randomly chosen replications (each with an idiosyncratic set of $k = 2$ instances) obtained after iterated learning has converged, which always conform to the smallest of k samples of t_{total} drawn from the prior and span a range of values of t less than this smallest sample.

are irrelevant to the decision). The minimum value of the remaining instances is reported as t_{total}. For example, if a person is asked to predict the age at death of a 50-year-old, then supposing that person remembers the instances 40, 80, and 84, the response would be 80. When t exceeds all retrieved instances (e.g., if $t = 85$ for this example), people are presumed to respond with a guess that is proportional to the probe value t. Specifically, the model's prediction then is a proportion g larger than t; that is, $t_{\text{total}} = (1 + g) \times t$.

For the iterated-learning paradigm, Min*K*'s predictions diverge drastically from those of the Bayesian model. In particular, Min*K* predicts that people converge onto a single constant response that, once reached, is emitted ad infinitum. This is because in the iterated-learning paradigm just presented, each t (bar the first one) is randomly chosen from the interval $(0, t_{\text{total}}^{(n-1)}]$, where $t_{\text{total}}^{(n-1)}$ was the person's previous response. It follows that if $t_{\text{total}}^{(n-1)}$ was equal to or below the overall minimum of a person's k instances (call that $t_{\min}$), the next value of t will necessarily never exceed that minimum and hence all subsequent responses will be equal to $t_{\min}$. Alternatively, if $t_{\text{total}}^{(n-1)}$ was greater than $t_{\min}$ (either because it was a proportional guess or one of the other instances $>t_{\min}$), then the next value of t may still exceed $t_{\min}$, but because of the random sampling process, it will necessarily be closer to $t_{\min}$ than $t_{\text{total}}^{(n-1)}$. This downward migration of each successive t will continue until a $t_{\text{total}}^{(n-1)} \leq t_{\min}$ occurs, at which point the chain will have converged onto the constant response $t_{\text{total}}^{(n)} = t_{\min}$. Notably, these convergence predictions are independent of g because guessing is absent once a t falls below $t_{\max}$ (the person's largest memorized instance). The predictions of Min*K* thus depend only on k.

Illustrative aggregate convergence predictions of Min*K* are shown in Fig. 3 for various values of k. Each distribution represents the simulated distribution of the minima of 1,000 samples of k instances from the appropriate prior. The figure reveals several noteworthy features. First, in all instances, the predicted distributions shift to the left and become more peaked as k increases. Second, in some instances (movie grosses, poems, and reigns of Pharaohs), the predicted distributions converge onto single points near zero as k increases. Third, even with $k = 2$, the value preferred by Mozer et al. (2008), there is considerable deviation between these predictions and those of the Bayesian model shown in Fig. 1, attesting to the empirical differentiability of the two models with reasonable (i.e., published) parameter settings.

The differences between the two models are particularly striking when shown at the level of individual prediction functions. Fig. 2 shows the prediction function for 10 randomly chosen samples, each of size $k = 2$, for movie grosses. In all cases, the first stimulus was $t = 100$, and all subsequent stimuli were sampled from a uniform distribution between zero and the model's immediately preceding response. The model was run for 20 iterated-learning trials, the last 10 of which are shown in the figure. As expected from the preceding analysis, Min*K* had converged onto a constant response (the minimum of the k instances) in each instance after only 10 iterated-learning trials, and thus produces a constant prediction function for all subsequent trials.

To confirm the generality of this convergence behavior, the model was run on 1,000 samples for each of the chains shown in Figs. 1 and 3 and for each of five different seed values (used in the present experiment; see Table 2). The median slope of the prediction function (across the 1,000 replications) for the last 10 of 20 iterated-learning trials was zero in all

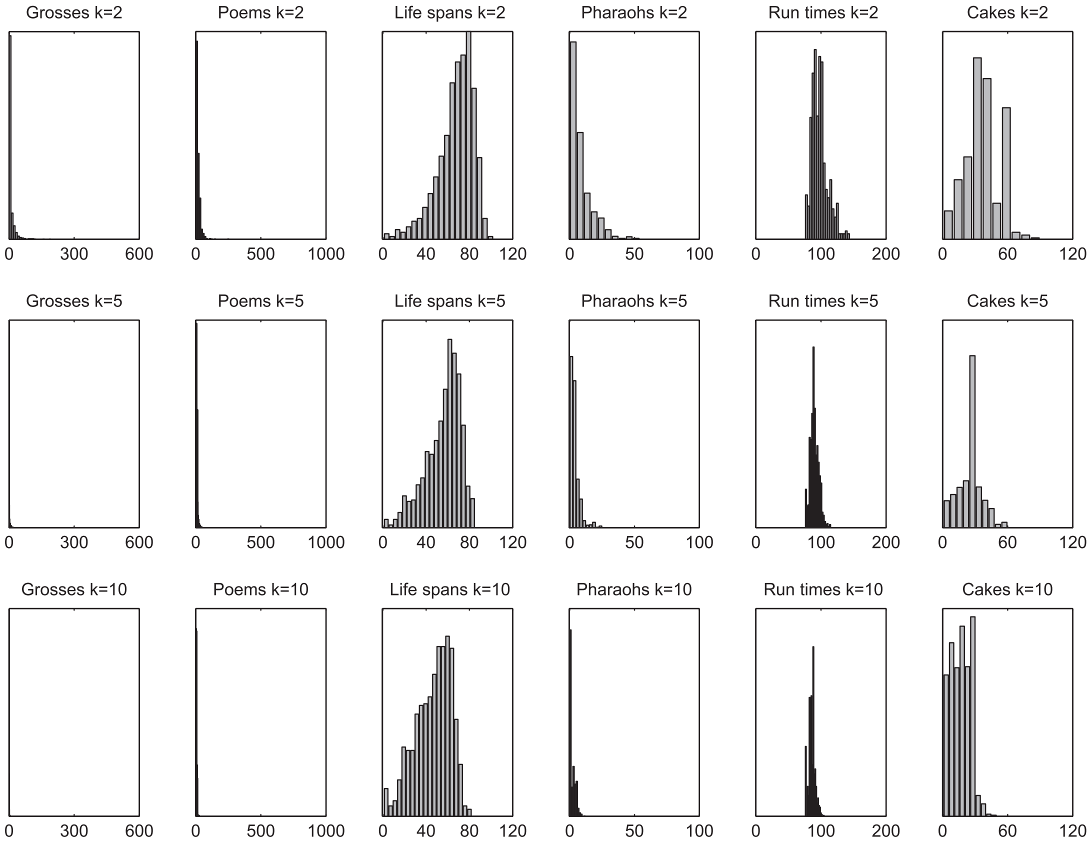

Fig. 3. Predicted stationary (converged) distributions of responses obtained from Min*K* for three different values of k ($k = 2$, 5, and 10, in the top, center, and bottom row of panels, respectively) and for six of the chains considered in the experiment and by Griffiths and Tenenbaum (2006).

cases. The mean values were slightly greater than zero, reflecting the fact that very occasionally the model had not converged after 10 iterated-learning trials. Table 1 shows those mean values together with the proportion of total replications that were equal to zero (within a narrow tolerance window). It is clear from the table that with one exception (for movie grosses), Min*K* predicts that between 94% and 100% of participants should converge onto a constant response after only 10 iterated-learning trials.

4.3. *Differentiating the models*

The preceding discussion has identified two testable predictions to differentiate between the two models with an iterated-learning methodology. First, at an aggregate level, the predicted stationary distributions of responses differ between the Bayesian model (top panels of Fig. 1) and Min*K* (Fig. 3). Second, at the level of individual subjects, the Bayesian model predicts that people’s prediction function should have a slope greater than zero and should

Table 1
Summary statistics for slopes of prediction functions predicted by Min*K* after 10 iterated-learning trials. Each table entry summarizes 5,000 replications (1,000 for each of five different seed values)

Chain	Median	Mean	Prop 0[a]
Movie grosses	.00	.15	.77
Length of poems	.00	.04	.97
Lifespan	.00	.00	1.00
Reign of Pharaohs	.00	.04	.94
Movie runtimes	.00	.00	1.00
Cake baking time	.00	.02	.99

Note: [a]Proportion of 5,000 replications that were within .00001 of zero.

follow the appropriate forms shown in the top panels of Fig. 2, regardless of the number of iterations of learning. Min*K*, by contrast, predicts that people's prediction functions should have zero slope after just 10 iterations of learning, and that across subsequent repeated trials people should respond with a constant value, as shown in the bottom panels of Fig. 2. These predictions are independent of Min*K*'s free parameter *g* and depend only on the number of samples on which judgments are based, *k*.

These two distinguishing predictions allow us to test whether people's judgments are more consistent with the Bayesian model or Min*K* for different values of *k*. In particular, we can compare Min*K* with $k = 2$ (the value advocated by Mozer et al., 2008) against the Bayesian model, allowing us to determine whether the wisdom of crowds is sufficient to explain the results of Griffiths and Tenenbaum (2006).

5. Method

The experiment sought to determine: (a) if people converge onto a stationary distribution during iterated learning, (b) whether the observed stationary distributions matched the actual prior distributions in the environment, and (c) whether each person's prediction functions were best characterized by the Bayesian sampling model or by Min*K*. People were presented with eight chains of trials, each of which involved 20 prediction trials that were linked in the manner described by Eq. 4. Each chain was ''seeded'' with one of five probe values (t) for the first trial. Our analyses focused on the convergence properties of the various chains and on examination of the resulting stationary distributions.

5.1. Participants

The participants were 35 members of the campus communities at the University of Western Australia ($N = 17$) and the University of Louisiana ($N = 18$) who participated voluntarily in exchange for course credit (some Australian participants also received AU$5 reimbursement).

5.2. Apparatus and stimuli

The experiment was controlled by a Windows-based PC that presented stimuli and recorded responses using a MatLab program written with the aid of the Psychophysics Toolbox (Brainard, 1997; Pelli, 1997).

The stimuli involved eight chains of 20 prediction trials. The chains involved the variables used by Griffiths and Tenenbaum (2006) whose actual prior distributions are known. (The number of terms served by members of the U.S. House of Representatives was omitted because this was unlikely to be common knowledge among the Australian participants.) Table 2 shows the eight variables, the shapes of their actual prior distributions, and the five possible seed values (histograms of these actual prior distributions are provided in the top panel of Fig. 1). The seed values were the values of t used by Griffiths and Tenenbaum and were selected to cover a wide range of the actual distributions (11th–99th percentile).

For each participant and each variable, the chain was seeded with one of the five values at random. For all subsequent trials of that chain, the next value of t was drawn at random from the interval $(0, t_{\text{total}}^{(n-1)}]$, where $t_{\text{total}}^{(n-1)}$ was the person's previous response. This process defined an iterated-learning chain of the form specified earlier, with the uniform sampling of t corresponding to the assumption in the model that the likelihood $p(t|t_{\text{total}})$ is uniform. Trials from all chains were randomly interleaved without constraints to form the total sequence of 160 trials.

5.3. Procedure

Each trial was initiated with the presentation of a statement and the probe value t. For example, for the life span variable, a trial with $t = 39$ would involve the statement: ''Insurance agencies seek to predict people's life spans—their age at death—based upon demographic information. If you were assessing an insurance case for a 39-year-old man, how old would you expect him to be at death?'' The complete list of questions appears in Appendix B.

Table 2
Chains and their actual prior distributions and seeds used in the experiment, and the observed number of trials to convergence

Chain	Prior	Seeds					Trials[a]
Movie gross	Power[b]	1	6	10	40	100	2
Length of poems	Power	2	5	12	32	67	2
Life span	Gaussian	18	39	61	83	96	2
Reign of Pharaohs	Erlang[c]	1	2	7	11	23	1
Duration of marriages	?	1	3	7	11	23	1
Movie run times	Gaussian	30	60	80	95	110	1
Cake baking time	Irregular	10	20	35	50	70	3
Waiting times	Power	1	3	7	11	23	4

Notes: [a]Trials to convergence.
[b]Power-law priors have the form $p(t_{\text{total}}) \propto t_{\text{total}}^{-\gamma}$ for some $\gamma > 0$.
[c]Erlang priors have the form $p(t_{\text{total}}) \propto t_{\text{total}} \exp(-t_{\text{total}}/\beta)$ for some $\beta > 0$.

People entered their response (t_{total}) using the keyboard and the next trial commenced 1 s later. The experimental trials were preceded by four practice trials involving another unique set of variables (e.g., the time required for a puddle to dry if it stopped raining *t* hours ago).

6. Results

6.1. Convergence analysis

We first examined the convergence of the chains from the various seeds. Convergence was defined as the point along the sequence of 20 trials within each chain at which responses no longer differed between sequences originating with different seed values. Fig. 4 shows an illustrative sequence of responses for all seed values for two chains (each data point is based on the median across participants who received a given seed), with movie grosses in the left panel and the length of poems on the right. For both chains, convergence appears to occur fairly rapidly, within approximately five trials.

Statistical confirmation for the rapid convergence was provided by comparing responses across seed values at each trial for each chain. Using a Kruskal–Wallis test with a Bonferroni-adjusted $\alpha = .0025$ (to maintain a level of significance of .05 across trials for

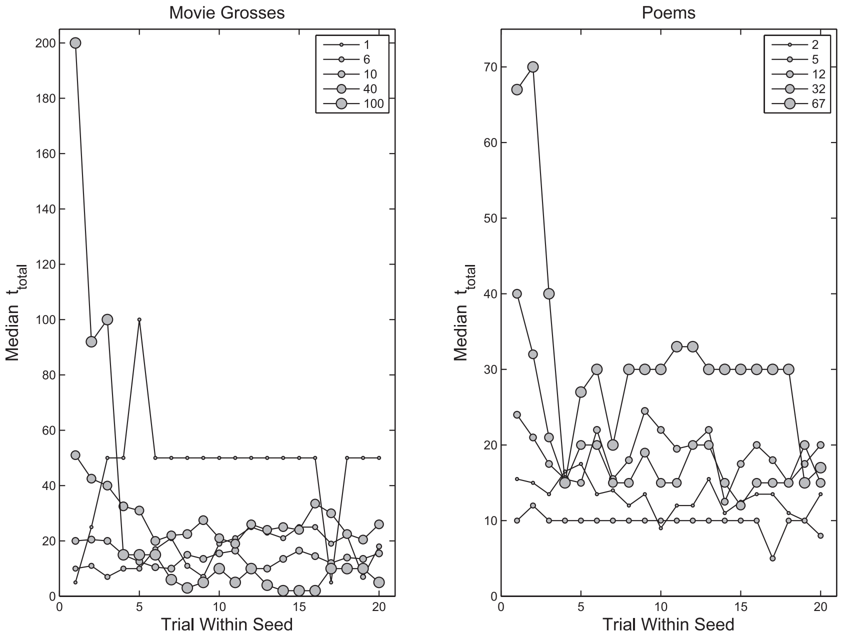

Fig. 4. Illustrative sequence of responses across trials for two chains. The left panel shows responses (median t_{total}) for movie grosses, and the right panel for length of poems. Within each panel, the five parameters represent the different seed values, with the size of plotting symbols proportional to the magnitude of the seed. Exact seed values for each chain are shown in the legend.

each chain), the number of trials to convergence (i.e., the first trial to yield a nonsignificant value of χ^2) was found to range from 1 to 4 (see Table 2).

A general result in iterated learning is that the second eigenvalue of the transition matrix in Eq. 4, λ_2, reveals the rate at which the chain converges, with higher values of λ_2 associated with slower convergence (see Griffiths & Kalish, 2005, 2007, for details). In confirmation, the observed correlation between the number of trials to convergence and the second eigenvalue (λ_2) of the transition matrix $\mathbf{Q}$ computed using the actual prior distributions of t_{total} was in line with the expectations of the Bayesian model.

6.2. Stationary distributions

The aggregate stationary distributions, obtained by considering responses from all participants during the last 10 trials of each chain, are shown in the bottom row of panels in Fig. 1. Note that waiting time is omitted from this and all further analyses because the actual distribution of that variable is unknown in the context of the particular question asked of participants. Likewise, we did not consider the duration of marriages because their actual attributes differ widely between Australia and the United States, thus preventing a meaningful analysis of our overall sample of participants.

Visual comparison of the data (bottom panels in Fig. 1) and the actual priors (top panels) suggests that people's responses: (a) differed widely between the different chains and (b) often mirrored the actual distributions within each chain, as predicted by the Bayesian model, but (c) deviated considerably from the predictions of Min*K* in at least some instances. We now present a quantitative evaluation of the performance of the two models.

6.2.1. Stationary distributions and predictions of the Bayesian model

The match between people's responses and the actual distributions is further illustrated in Fig. 5, which shows quantile–quantile (Q–Q) plots of the actual and stationary distributions. When two distributions are identical, all points in a Q–Q plot will fall along the diagonal. If the points fall on a straight line but differ in intercept (or slope) from the diagonal, the distributions are identical in shape but differ in location (or spread, respectively). It follows that the observed mismatch between stationary and actual distributions in some cases reflects miscalibration of the spread of people's distributions (e.g., for Pharaohs and movie grosses), whereas other cases hint at systematic deviations between the distributions' shape (e.g., movie run times and cake baking times).

To put these discrepancies into a proper context, people's stationary distributions were next compared to the predictions of the Bayesian model when applied to the particular sequence of trials shown to participants. In this model, as in the analysis of iterated learning presented above, we assumed that participants sampled a value of t_{total} from the posterior distribution. That is, for each value of t presented to participants in the experiment, we obtained a predicted distribution of $p(t_{\text{total}}|t)$ based on the actual distributions, $p(t_{\text{total}})$, as described by Eq. 1. Those predictions, in turn, were summed across trials and participants to yield aggregate posterior distributions. The result is a distribution that takes into account the

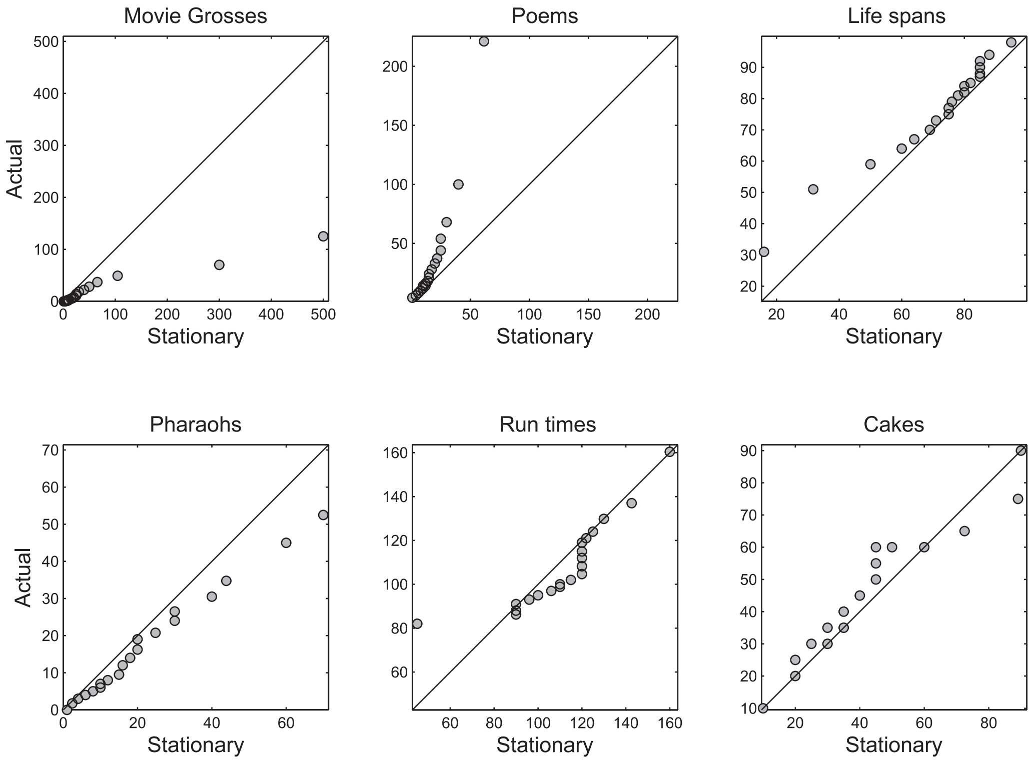

Fig. 5. Quantile–quantile (Q–Q) plots relating the observed stationary distributions to the actual prior distributions. Quantiles are from the 2nd to 98th percentile in 5% increments. A perfect overlap between the two distributions would correspond to all data points falling along the diagonal in each panel. Conversely, if the two distributions differed in location (mean) or spread (variance), the Q–Q plot would deviate from the diagonal in intercept or slope, respectively. Nonlinear deviations from the diagonal reveal more subtle differences in shape between the two distributions.

fact that we are only observing a relatively modest sample of responses. The Q–Q plots of these predicted and stationary distributions are shown in Fig. 6.

The Bayesian model clearly provides a good account of people's stationary distributions, as confirmed by the r^2 values (.99, .98, .95, .99, .79, and .89, for movie grosses, poems, life span, Pharaohs, movie run times and cakes, respectively, $M = .93$; see Table 3 for further summary statistics). Notably, the discrepancy between distributions for the length of poems observed in Fig. 5 is nearly absent in Fig. 6, suggesting that the original deviation was primarily due to the particular selection of stimuli shown to participants. Note that the two sets of Q–Q plots in the figures are conceptually identical; the only difference is that Fig. 6 converts the actual prior distributions into optimal Bayesian predictions based on the particular sequence of stimuli shown to participants and corrects for distortions resulting from approximating the stationary distribution by a small sample—the fact that this model fits well without any free parameters suggests that people: (a) have knowledge of the appropriate actual

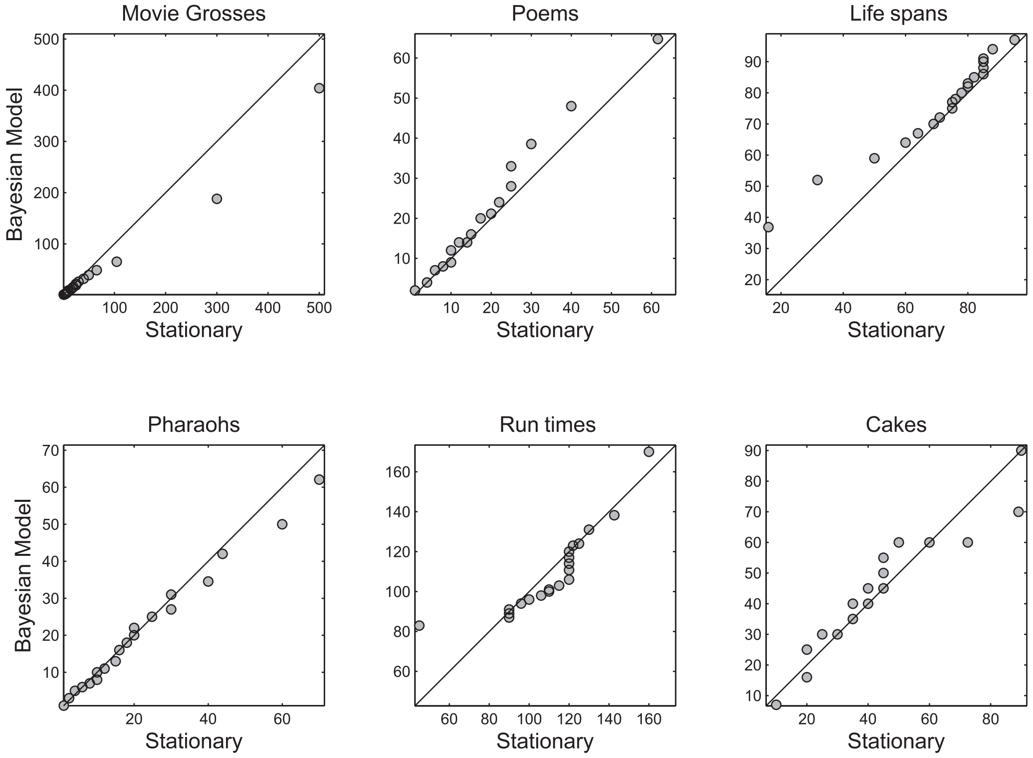

Fig. 6. Quantile–quantile plots relating the observed stationary distributions to the distributions predicted by the Bayesian model (see text for details). Quantiles are from the second to 98th percentile in 5% increments. A perfect overlap between the two distributions would correspond to all data points falling along the diagonal in each panel.

distributions, (b) generate predictions by combining this prior knowledge or expectations with the information provided by the probe stimuli in an optimal manner, and (c) sample their responses from the resulting posterior distribution.

Table 3
Comparison of the observed stationary distributions and model predictions

		MinK		
Chain	Bayes	$k = 2$	$k = 5$	$k = 10$
Movie gross	34.59	125.15	132.76	134.16
Length of poems	3.56	2.86	12.76	15.56
Life span	7.42	6.43	16.81	26.11
Reign of Pharaohs	3.33	16.85	24.01	26.21
Movie runtimes	10.74	18.55	28.07	31.73
Cake baking time	6.97	10.92	22.48	28.60
Mean	11.10	30.13	39.48	43.73

Note: All table entries are RMSDs that summarize the deviation of the predicted quantiles from the diagonal in Figs. 6 and 7.

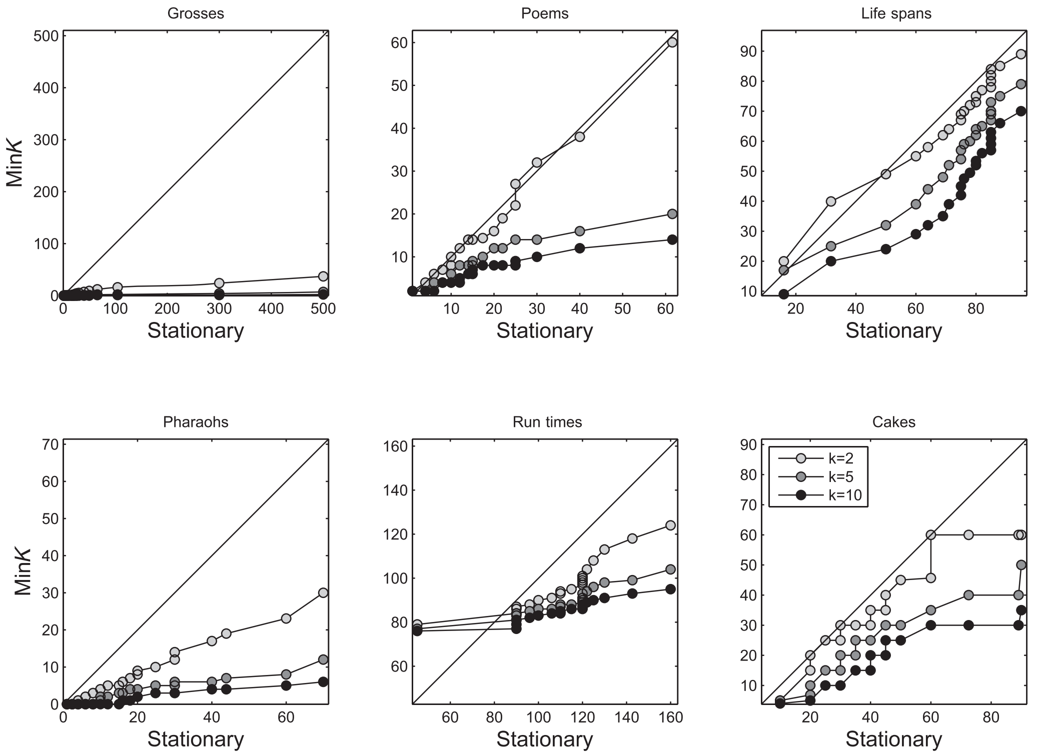

Fig. 7. Quantile–quantile plots relating the observed stationary distributions to the distributions predicted by Min*K* for three values of *k* (see text for details). Quantiles are from the 2nd to 98th percentile in 5% increments and the values of *k* are shown in the legend.

6.2.2. Stationary distributions and predictions of MinK

Fig. 7 compares the observed stationary distributions to the predictions of Min*K* in a further set of Q–Q plots. Each panel in the figure contains three parameters that correspond to separate Q–Q plots for three values of *k* (viz., 2, 5, and 10). The figure thus plots the quantiles of the predicted distributions shown at the outset (Fig. 3) against the data.

The figure reveals that in the vast majority of cases, Min*K* was incapable of capturing people's convergence behavior. Only when $k = 2$ did the model succeed in predicting people's performance, and then only for poems and life spans. In all other cases, the model's predictions deviated considerably from the data. Table 3 provides statistical confirmation of the obvious pattern in the Q–Q plots and additionally provides a comparison with the performance of the Bayesian model. The entries in the table are the root mean-squared deviations (RMSDs) between the diagonal and the points in each Q–Q plot, which provide an indication of the extent of deviation from identity of the two distributions. The table confirms our principal conclusion: Whereas the Bayesian model captured people's behavior in the iterated-learning variant of the prediction task without any free parameters, Min*K* was incapable of doing so in all but a limited number of cases despite the aid of a parameter.

6.3. Individual performance

Fig. 8 shows representative prediction functions for two participants (one in each row of panels). In each panel, the data are shown by large plotting symbols, with the last 10 trials identified by black squares. The predictions of the Bayesian model are represented by thick solid lines (these are the posterior medians shown in Fig. 2). For the reasons noted at the outset, Min*K* cannot be fit to the results from individual subjects and hence its predictions cannot be shown in the same figures, although we note that Min*K* expects all prediction functions to be completely flat after convergence to the minimum of the set of *k* samples, as illustrated in the bottom panel of Fig. 2.

It is clear from the figure that Min*K*'s predictions were at odds with the behavior of these particular subjects. In all instances, subjects' predictions—even for the last 10 (postconvergence) trials—deviate considerably from the constant responses expected by Min*K*.

Lest one think that those subjects might not be representative of the sample as a whole, we fit a linear regression to responses from the last 10 trials (with t and t_{total} as independent

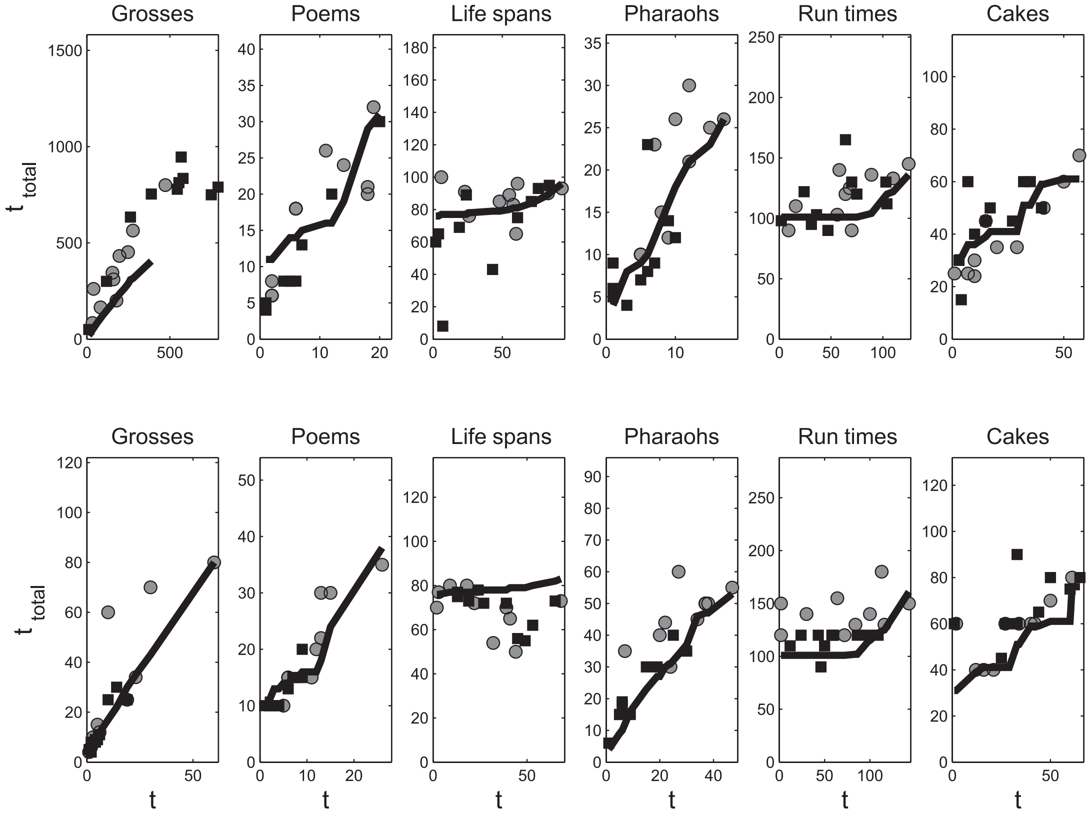

Fig. 8. Representative prediction functions for two participants (each row of panels corresponds to one participant). Data are shown by large plotting symbols and the last 10 trials are identified by black squares. Each panel also shows the predictions of the Bayesian model (thick solid line).

and dependent variables, respectively) for each chain and for each of the 35 participants separately. The results of that analysis of each individual are shown in Table 4. For all chains, the mean slope across subjects was significantly greater than zero, attesting to the fact that people did not converge onto a constant response as predicted by Min*K*.

When interpreting the table, it must be borne in mind that the test is very conservative: Under a Bayesian model, for variables with a Gaussian prior, people's responses are expected to be constant up to a point that corresponds to the mean of the distribution; it is only for values of t beyond that point that an upturn is expected; see top panels in Fig. 2. The fact that the prediction functions for the two Gaussian variables (life span and movie runtimes) nonetheless exhibit mean slopes that are significantly greater than zero attests to the power of the analysis reported in Table 4 and to the consistency of behavior across subjects.

One might object that the mean predictions of Min*K* were close to, but not equal to zero, whereas the observed slopes were tested against the null hypothesis of zero. This objection can be evaluated by considering the two columns in Table 4 that show the 95% confidence bounds around the mean slope estimate. None of those confidence intervals spans the predicted mean slope of Min*K* in Table 1. Further confirmation of the consistency of the effect across subjects is provided in the final column of the table, which shows the number of subjects (out of 35) whose slopes were equal to or below zero. It is apparent that with the exception of the two Gaussian variables (for the reasons just noted), very few participants' slopes were ≤0.

Finally, Fig. 9 presents histograms of the distribution of individual postconvergence slope estimates for the six chains. In confirmation of the data in the table, the vast majority of subjects exhibited positive slopes of at least moderate magnitude. There are two exceptions to this pattern, both involving the Gaussian variables (life span and movie run times), for which slopes may differ only slightly from zero even under a Bayesian model for the reasons just noted.

We conclude that even when the data are considered at an individual level, there is clear evidence that people converged onto a stationary behavior very different from that predicted

Table 4
Summary statistics for linear regressions that predict the last 10 responses (t_{total}) on the basis of the last 10 stimuli (t) for each subject and chain separately

Chain	Intercept[a]	Slope[a]	One-sample t[b]	p	$\text{CI}^{c}_{\text{lower}}$	CI_{upper}	N (slope ≤0)[d]
Movie gross	43.71	.86	11.16	<.0001	.71	1.02	1
Length of poems	14.11	.65	8.07	<.0001	.49	.81	3
Life span	65.74	.22	3.73	<.001	.10	.35	9
Reign of Pharaohs	16.29	.82	6.39	<.0001	.56	1.08	3
Movie run times	104.90	.13	3.46	<.001	.05	.21	16
Cake baking time	41.05	.22	5.12	<.0001	.13	.31	5

Notes: [a]Average across subjects of individual estimates.
[b]Tested against the hypothesis that the mean slope is equal to 0, d.f. = 34.
[c]Lower and upper bounds of 95% confidence interval on mean slope.
[d]Number of participants whose slope was ≤0.

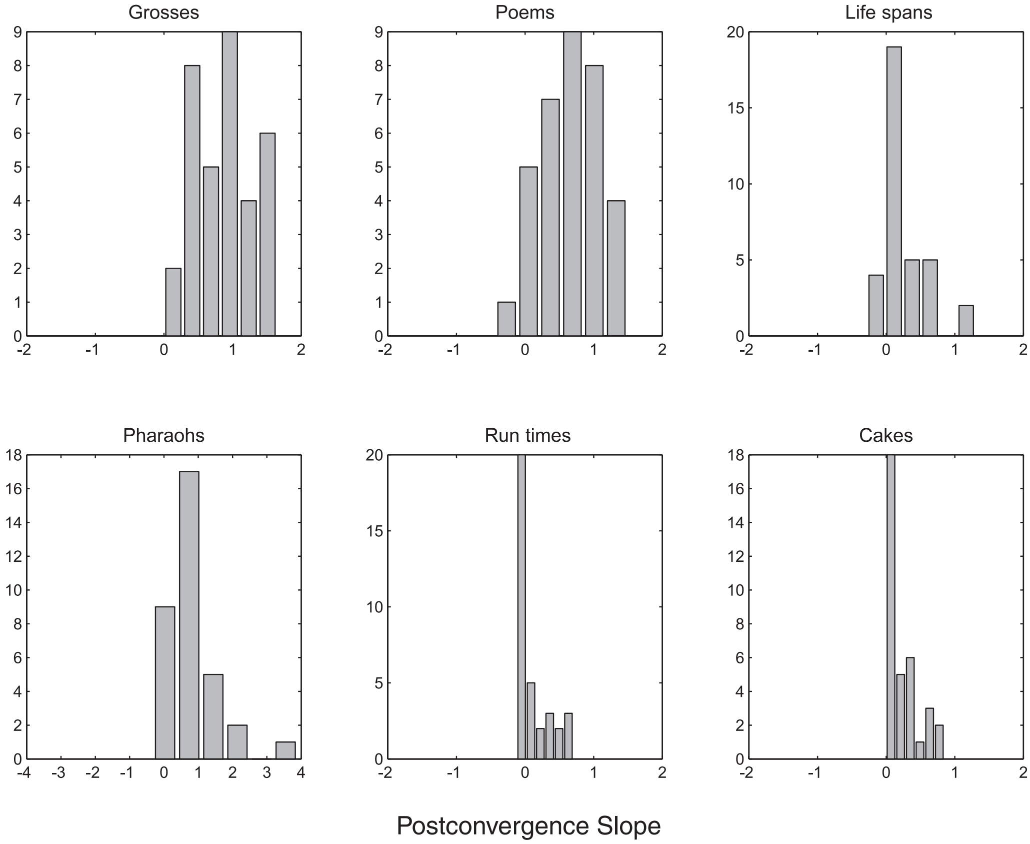

Fig. 9. Distribution of postconvergence slope estimates across participants for all six chains. Min*K* predicts slopes to be equal to zero. Note that variables with a Gaussian distribution (life span and run times) are characterized by small slopes even under a Bayesian model; see text for details.

by Min*K*. In confirmation of recent related reports (e.g., Vul & Pashler, 2008), people in our study individually exhibited considerable wisdom.

7. Discussion

The experiment and modeling permit a number of conclusions. First, for all chains, people's responses converged onto a stationary distribution as predicted by the iterated-learning model. Second, the individual prediction functions conformed more closely to the shape expected by the Bayesian model than to that predicted by Min*K*, suggesting that the conformity with the optimal Bayesian predictions reported by Griffiths and Tenenbaum (2006) was not merely a result of aggregating across participants as had been suggested by Mozer et al. (2008). Third, the predictions of Min*K* were clearly disconfirmed both at the aggregate and individual levels.

These conclusions have ramifications for a number of theoretical issues, which we explore further in the remainder of this article. We start by summarizing how these results illuminate people's behavior in the future-prediction task. We then turn to how they contribute to our understanding of iterated learning more broadly. Finally, we discuss how Min*K* falls short as a plausible alternative to a Bayesian model of how people form predictions, and we consider the implications of our results for other models of cognition that assume people rely only on a sparse set of instances stored in memory.

7.1. The future-prediction task

Our results go beyond related precedents (e.g., Griffiths & Tenenbaum, 2006) in several important ways. At the outset, we asked two questions about the specifics of how people make predictions: First, whether people really do use the appropriate actual prior in each domain and, second, whether individuals have extensive knowledge of this distribution, or whether the appearance that this is the case was merely a consequence of aggregating their responses.

Our study resolved these questions. The fact that iterated learning converged onto distributions that closely matched the actual distributions associated with everyday quantities supports the claim that these are the distributions underlying people's predictions. Moreover, using a within-subject variant of iterated learning allowed us to confirm that the prediction functions produced by individual participants conformed to those expected from an optimal Bayesian agent using the appropriate actual prior. Our conclusions thus mirror—and considerably extend—those recently reported by Vul and Pashler (2008), who investigated the ''crowd within'' (p. 645) and likewise concluded that ''…responses made by a subject are sampled from an internal probability distribution, rather than deterministically selected…'' (p. 647).

Our results thus bear out the basic conclusions of Griffiths and Tenenbaum (2006) regarding the optimality of human predictions for everyday events. In addition, our study provided a more complete picture of what people are doing when asked to form a prediction. The analysis of iterated learning presented by Griffiths and Kalish (2007) requires that Bayesian agents select hypotheses with a probability that is proportional to their posterior probability in order to guarantee convergence to the prior. The fact that we see convergence to the prior in our experiment suggests that people are performing the task in a way that is consistent with producing samples from the appropriate posterior distribution—that is, people produce responses that match their expected probabilities.

Probability matching is commonly observed in decision-making tasks (see Vulkan, 2000, for a review) and provides an explanation for how aggregating of responses nonetheless preserves key properties of people's predictions: If each prediction is a sample from a common posterior distribution, then a set of predictions from different people will approximate that distribution. Thus, the close correspondence between the posterior median and the median of people's predictions observed by Griffiths and Tenenbaum was not an artifact of averaging but the result of a deeper correspondence between the posterior distribution and the distribution of people's predictions.

We therefore conclude that people seem to: (a) use the correct prior and (b) sample from the posterior distribution. This conclusion must be qualified by the fact that neither (a) nor (b) can be independently observed: We can only infer the prior and the response process from people's judgments, and the interdependence of these two factors makes them unidentifiable. Nonetheless, the fact that people's stationary responses resemble the actual priors implies that whatever people do is equivalent to use of the correct prior and sampling from the posterior.

7.2. Iterated learning

The correspondence between the stationary distributions produced by iterated learning and the actual distributions provides some of the most compelling evidence to date that iterated learning converges to the prior expectations of human learners. Previous experiments (Griffiths, Christian, & Kalish, 2008; Kalish et al., 2007) have relied upon established but not immediately related results, such as the relative difficulty of learning functions and categories, in order to test the prediction that iterated learning should converge to the prior. In our study, by contrast, we have an objective indication of what the priors of human learners should be (that is, we have access to the actual distributions of the quantities in question), and we can compare those actual priors directly to the outcome of iterated learning.

The present experiment goes beyond previous research on iterated learning in several additional ways. Unlike Kalish et al. (2007), the experiment involved no supervised learning. Rather, people simply made predictions without any feedback and on the basis of fairly impoverished information, thus performing a generalization task more similar to that used by Griffiths, Christian, and Kalish (2008). However, unlike the method used by Griffiths, Christian, and Kalish, people were not restricted to a small discrete hypothesis space: They were allowed to enter integer values from an unbounded range as responses. By implication, our stationary distributions were defined over the set of all nonnegative integers rather than a small number of hypotheses. This combination of a naturalistic task and a rich hypothesis space resulted in a far more detailed picture of the knowledge that people were bringing to bear than previous studies.

The use of multiple interleaved chains also provided an important control for an alternative explanation for the outcome of iterated learning. In previous tests of the prediction that iterated learning should converge onto people's prior expectations, the prior probabilities of these hypotheses were inferred based on the ease with which they were learned. The hypothesis most readily learned (e.g., positive linear function concepts; Kalish et al., 2007) was taken to represent people's prior expectation. This analysis is open to the objection that the change in people's response patterns over time could reflect fatigue or other factors that lead people to favor the ''easier'' hypotheses. The present results indicate that this is not the case: The fact that each everyday quantity has a unique actual distribution and that people simultaneously converged to these quite disparate distributions in different chains indicates that people were not simply drifting toward some ''easiest'' hypothesis due to becoming fatigued. If fatigue were the explanation, we would expect to see similar stationary distributions across all chains—a result very different from that seen in our experiment.

Finally, our study represents the first instance in which iterated learning provided an opportunity to differentiate between two competing models that would otherwise have been difficult to test. The basic effect of iterated learning is to magnify the processes that influence people's judgments, with each iteration providing another opportunity for these processes to have an effect. At the behavioral level, this magnification property of iterated learning has been examined by Reali and Griffiths (2008). The same kind of magnification occurs with the predictions of models, implying that models that might be difficult to distinguish after only a single iteration can make quite different predictions after several iterations. We anticipate that this property of iterated learning will make it a valuable tool for testing other psychological theories, by exaggerating subtle differences in their predictions that could not otherwise be empirically differentiated.

7.3. Implications for MinK and other sparse instance models

Our data and modeling have fairly clear—albeit negative—implications for Min*K*. One of the model's basic predictions, namely the convergence onto a constant response during iterated learning, found no support in our data. We now briefly consider how the model performed well in the analysis of the original data from Griffiths and Tenenbaum (2006) and yet so poorly in the context of our experiment, and then discuss the implications of our results for other psychological models based on storing a sparse set of instances in memory.

One of the key factors in the success of Min*K* in modeling the results of Griffiths and Tenenbaum (2006) was aggregation. Because each participant in the original experiment contributed only one response in each prediction domain, any mechanism that produced a reasonable approximation to a single sample from the posterior distribution would be capable of mimicking the Bayesian model. This was part of the point raised by Mozer et al. (2008), who then used Min*K* to demonstrate that a relatively simple mechanism (requiring little knowledge on the part of each participant) could produce equivalent results. However, without aggregation, the prediction functions of Min*K* for any single set of samples look quite unlike the kind of prediction functions produced by the Bayesian model, being piecewise linear with sharp discontinuities at the location of each sample from the prior. In a within-subject experiment, where multiple predictions are obtained for each domain from each participant, the difference between these models becomes more apparent. Visual inspection of the prediction functions shown in Fig. 8 shows that they do not seem to display the kind of discontinuities that we would expect from Min*K*. The iterated-learning methodology exaggerates the way in which the models differ in their predictions for individual subjects even further, resulting in the dramatic differences documented at the outset.

Another factor that allowed Min*K* to perform well in fitting the aggregate data was the use of a linear ''guessing'' parameter, with predictions for values that are above the largest stored sample being $(1 + g) \times t$. For all three of the kinds of priors explored by Griffiths and Tenenbaum (2006), the predictions produced by the Bayesian model converge to a linear function as t becomes large. For power-law and Erlang priors, the posterior median is always a linear function of t (with slope between 1 and 2 for power-law and 1 for Erlang). For Gaussian priors, the posterior median approaches a linear function with slope 1 once

t becomes larger than the mean of t_{total}. Use of a linear guessing function with g as a free parameter thus provides an opportunity to produce predictions that match different classes of priors, with an extra degree of freedom that the Bayesian model lacked. Producing a good fit to predictions resulting from a power-law distribution should thus come as little surprise, and the quality of the approximation for Gaussian and Erlang distributions would depend on the range of values of t that are investigated, with worse performance for more extreme values. While Mozer et al. (2008) used a single value of g to fit all domains, they noted that this resulted in a very poor fit for the Gaussian-distributed life span data and advocated using a different guessing parameter in this case. Knowing what values of g are reasonable in a given domain is exactly the kind of knowledge that is necessary to make good Bayesian predictions and reflects the kind of sensitivity to environmental statistics that Griffiths and Tenenbaum (2006) considered notable.

One unusual property of the Min*K* model is that the more knowledge a person has, the less plausible his or her responses become. That is, as k increases, the minimum instance will necessarily get closer to the probed value (t). In the extreme case, if people have perfect knowledge of the appropriate prior distribution (so $k = \infty$), the model's predicted response will always be $t_{\text{total}} = t$, thus generating prediction functions that fall along the principal diagonal for all variables irrespective of their actual distribution. While many heuristics work well with moderate amounts of data and less well when more data are available (e.g., Goldstein & Gigerenzer, 2002), this property appears counter-intuitive, especially when we think of heuristics as approximate solutions to challenging problems. In the present case, the scaling of the model's behavior with sample size is particularly awkward because of Min*K*'s reliance on the *minimum* of a sample. Unlike other quantities that could be estimated from a sample, such as the mean and median, the expected minimum of a sample is always going to change (viz., decrease) as sample size increases (e.g., Gumbel, 1958). By implication, Min*K*'s behavior is necessarily very sensitive to sample size and, as we have just shown, its limiting behavior ($k = \infty$) is quite odd.

In support of this point, when Mozer et al. (2008) considered alternatives to Min*K* that embodied similar principles but asymptotically approached Bayesian inference with the correct prior as k became large, they found that the value of k had less effect on the fit of the models. Even for these models, we might expect better performance for smaller values of k because this is the range in which the free parameter g has the greatest effect. However, these results illustrate that the key property necessary for matching the aggregate data is producing predictions similar to a sample from the posterior distribution—something that Min*K* is more likely to do when k is small.

One might object to our critique of Mozer et al. (2008) by pointing out that Min*K* might best be considered an illustrative toy model whose sole purpose was to provide a quick alternative to the assumptions made by Griffiths and Tenenbaum (2006). In consequence, its refutation may have little bearing on the ongoing debate as to whether people are capable of optimal use of Bayesian priors (as strongly suggested by our data) or whether their performance is better described by some other heuristic model that shares with Min*K* the idea of a sparse set of instances but none of its other architectural commitments. In response, we note that our data challenge a whole class of ''sparse instance'' models that could be used to explain people's predictions. Any model in which a small, fixed set of instances is used to form predictions is going to produce stereotyped behavior in an iterated-learning setting. For example, a Rand*K* model, in which participants choose at random between all stored instances greater than t would produce prediction functions that assigned probabilities to only a constant set of values, corresponding to the different stored instances. Likewise, a Mean*K* model, averaging all exemplars greater than t, would converge to a constant in the same way as Min*K*. We believe that the match between the human stationary distributions and the priors, together with the nonconstant stationary distributions produced by our participants, severely limit the application of ''sparse instance'' explanations in this context.

Importantly, there is one class of ''sparse instance'' models that is not compromised by our results and that seems plausible as a heuristic account of how people might be solving the problem. Specifically, our results are consistent with any model in which responses rely on a new sample from the Bayesian posterior distribution on each trial. One such model would postulate that people have a relatively large reservoir of stored instances for each distribution, but they retrieve only a small number of these instances on each trial. If people then weight those instances of t_{total} that are greater than t by $1/t_{\text{total}}$ and sample an instance with probability proportional to these weights, the sample will approximate a draw from the posterior distribution. This kind of sparse instance model implements a Monte Carlo approximation to the posterior known as importance sampling (for details, see Shi, Feldman, & Griffiths, 2008) and would be indistinguishable from the full Bayesian model in our experiment, provided the reservoir of stored instances is reasonably large.

The idea that people represent the world using a small number of samples from their environment is one that is common in psychological process models, with exemplar models having been used to explain human category learning (Medin & Schaffer, 1978; Nosofsky, 1986), function learning (DeLosh, Busemeyer, & McDaniel, 1997), probabilistic reasoning (Juslin & Persson, 2002), and social judgment (Smith & Zarate, 1992). The success of iterated learning in distinguishing between Min*K* and a model assuming more abstract statistical knowledge in the context of predicting the future suggests that a similar approach might productively be used to investigate the adequacy of exemplar models in these other contexts. We view this as an important direction for future work, given the prominent role that exemplar models play throughout cognitive psychology.

8. Conclusions

Taken together with our previous experiments (Griffiths, Christian, & Kalish, 2008; Kalish et al., 2007), the present results provide strong support for the conclusion that iterated learning converges to an equilibrium that reflects people's knowledge or expectations about a task. In this particular instance, the data also show that people optimally predict future events on the basis of impoverished information, as suggested by Griffiths and Tenenbaum (2006). A similar method may prove valuable in evaluating other Bayesian models, providing an independent source of information about the prior distributions that people appear to use which can be compared with the assumptions embodied in the model. By the same token, use of an iterated-learning methodology allowed us to refute a competing model of the future-prediction task, namely the Min*K* model proposed by Mozer et al. (2008). The methodology is also likely to be valuable in exploring the predictions of other models based on the idea that people store a sparse sample of instances in memory.

Having established that iterated learning can reliably uncover people's prior expectations, we can now use this paradigm as a tool to reveal people's knowledge in situations in which it might not be readily observable by other means. In the context of judgment and decision making, one might use iterated learning to examine people's knowledge of distributions of economic indicators such as wages, wealth, or home ownership. More generally, we anticipate that the method will prove valuable in assessing the constraints that guide human learning in contexts where the nature of such constraints remains controversial, such as language acquisition, causal induction, and category learning.

Acknowledgments

Preparation of this paper was facilitated by various Discovery Grants from the Australian Research Council and an Australian Professorial Fellowship to Stephan Lewandowsky, and grants BCS-0704034 and BCS-0544705 from the National Science Foundation to Thomas L. Griffiths and Michael L. Kalish respectively. We thank Leo Roberts for assistance during data collection.

References

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision*, *10*, 433–436.

DeLosh, E. L., Busemeyer, J. R., & McDaniel, M. A. (1997). Extrapolation: The sine qua non of abstraction in function learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*, 968–986.

Estes, W. K., & Maddox, W. T. (2005). Risks of drawing inferences about cognitive processes from model fits to individual versus average performance. *Psychonomic Bulletin & Review*, *12*, 403–408.

Goldstein, D. G., & Gigerenzer, G. (2002). Models of ecological rationality: The recognition heuristic. *Psychological Review*, *109*, 75–90.

Griffiths, T. L., Christian, B. R., & Kalish, M. L. (2008). Using category structure to test iterated learning as a method for identifying inductive biases. *Cognitive Science*, *32*, 68–107.

Griffiths, T. L., & Kalish, M. L. (2005). A Bayesian view of language evolution by iterated learning. In B. G. Bara, L. Barsalou, & M. Bucciarelli (Eds.), *Proceedings of the twenty-seventh annual conference of the cognitive science society* (pp. 827–832). Mahwah, NJ: Erlbaum.

Griffiths, T. L., & Kalish, M. L. (2007). Language evolution by iterated learning with Bayesian agents. *Cognitive Science*, *31*, 441–480.

Griffiths, T. L., Kalish, M. L., & Lewandowsky, S. (2008). Theoretical and empirical evidence for the impact of inductive biases on cultural evolution. *Philosophical Transactions of the Royal Society B*, *363*, 3503–3514.

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

Gumbel, E. J. (1958). *Statistics of extremes*. New York: Columbia University Press.

Hélie, S. (2006). An introduction to model selection: Tools and algorithms. *Tutorials in Quantitative Methods for Psychology*, *2*, 1–10.

Juslin, P., & Persson, M. (2002). PROBabilities from EXemplars (PROBEX): A lazy algorithm for probabilistic inference from generic knowledge. *Cognitive Science*, *26*, 563–607.

Kalish, M. L., Griffiths, T. L., & Lewandowsky, S. (2007). Iterated learning: Intergenerational knowledge transmission reveals inductive biases. *Psychonomic Bulletin and Review*, *14*, 288–294.

Kirby, S. (2001). Spontaneous evolution of linguistic structure: An iterated learning model of the emergence of regularity and irregularity. *IEEE Journal of Evolutionary Computation*, *5*, 102–110.

Kirby, S., Dowman, M., & Griffiths, T. L. (2007). Innateness and culture in the evolution of language. *Proceedings of the National Academy of Sciences*, *104*, 5241–5245.

Medin, D., & Schaffer, M. (1978). Context theory of classification learning. *Psychological Review*, *85*, 207–238.

Mozer, M. C., Pashler, H., & Homaei, H. (2008). Optimal predictions in everyday cognition: The wisdom of individuals or crowds? *Cognitive Science*, *32*, 1132–1147.

Myung, I. (2000). The importance of complexity in model selection. *Journal of Mathematical Psychology*, *44*, 190–204.

Myung, I., Navarro, D., & Pitt, M. (2006). Model selection by normalized maximum likelihood. *Journal of Mathematical Psychology*, *50*, 167–179.

Norris, J. R. (1997). *Markov chains*. Cambridge, England: Cambridge University Press.

Nosofsky, R. (1986). Attention, similarity and the identification–categorization relationship. *Journal of Experimental Psychology: General*, *115*, 39–57.

Pelli, D. G. (1997). The video toolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision*, *10*, 437–442.

Reali, F., & Griffiths, T. L. (2008). The evolution of frequency distributions: Relating regularization to inductive biases through iterated learning. In B. C. Love, K. McRae, & V. M. Sloutsky (Eds.), *Proceedings of the 30th annual conference of the cognitive science society* (pp. 229–234). Austin, TX: Cognitive Science Society.

Shepard, R. N., Hovland, C. I., & Jenkins, H. M. (1961). Learning and memorization of classifications. *Psychological Monographs*, *75*, 1–42. (13, Whole No. 517)

Shi, L., Feldman, N. H., & Griffiths, T. L. (2008). Performing Bayesian inference with exemplar models. In B. C. Love, K. McRae, & V. M. Sloutsky (Eds.), *Proceedings of the 30th annual conference of the cognitive science society* (pp. 745–750). Austin, TX: Cognitive Science Society.

Smith, E. R., & Zarate, M. A. (1992). Exemplar-based model of social judgment. *Psychological Review*, *99*, 3–21.

Smith, K., Kirby, S., & Brighton, H. (2003). Iterated learning: A framework for the emergence of language. *Artificial Life*, *9*, 371–386.

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.

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, *90*, 293–315.

Vul, E., & Pashler, H. (2008). Measuring the crowd within: Probabilistic representations within individuals. *Psychological Science*, *19*, 645–647.

Vulkan, N. (2000). An economist's perspective on probability matching. *Journal of Economic Surveys*, *14*, 101–118.

Appendix A: Mathematical definitions of distributions and corresponding prediction functions

The Gaussian distribution is the familiar normal distribution, with:

$$p(t_{\text{total}}) \propto \exp(-(t_{\text{total}} - \mu)^2/2\sigma^2), \tag{A1}$$

where μ and σ are the mean and standard deviation, respectively. This distribution produces values of t_{total} centered around μ, with the rate at which they fall off on either side being exponential but determined by σ.

The power-law distribution has:

$$p(t_{\mathrm{total}}) \propto t_{\mathrm{total}}^{-\gamma}, \tag{A2}$$

where γ is a parameter determining the rate at which $p(t_{\mathrm{total}})$ decreases as t_{total} becomes large. The probability of observing a value of t_{total} decreases monotonically with t_{total}, but more slowly than other (specifically, exponential-tailed) distributions. As a consequence, it is possible to see extreme values of t_{total} relatively often. This property leads to the distribution being described as having ''heavy tails.''

The Erlang distribution is a one-parameter variant of the Gamma distribution, with

$$p(t_{\mathrm{total}}) \propto \exp(-t_{\mathrm{total}}/\beta), \tag{A3}$$

where β is a parameter of the distribution determining its mean and variance. The probability of t_{total} rises to a peak (with location determined by β) before falling off exponentially at a rate determined by β.

These three distributions result in different predictions when used as a prior. Griffiths and Tenenbaum (2006) computed the prediction functions that result from defining the optimal prediction to be the median of the posterior distribution, the value t^* such that $p(t_{\mathrm{total}} > t^*|t) = p(t_{\mathrm{total}} < t^*|t) = 0.5$, where $p(t_{\mathrm{total}}|t)$ is computed as described in the main text. The prediction function for the Gaussian does not have a simple analytic form but essentially indicates that one should guess the mean of the distribution $p(t_{\mathrm{total}})$ until the observed values of t approach this mean and then produce predictions that increase linearly with t. The power-law distribution produces the prediction function $t^* = 2^{1/\gamma}t$, indicating that one should predict a constant multiple of the observed value of t. The Erlang distribution produces the prediction function $t^* = t + \beta \log 2$, with predictions always being a little greater than the observed value of t. Derivations of these prediction functions are provided by Griffiths and Tenenbaum (2006).

Appendix B: Questions used in the experiment

The questions were presented to participants exactly as shown here, with the probe value t replacing the ''X'' in the text below.

Movie grosses

Imagine you hear about a movie that has taken in X million dollars at the box office, but you do not know how long it has been running. What would you predict the total box office intake for that movie to be?

Length of poems

If your friend read you her favorite line of poetry and told you it was line X of a poem, what would you predict the total number of lines to be?

(Male) Life span

Insurance agencies seek to predict people's life spans—their age at death—based upon demographic information. If you were assessing an insurance case for an X-year-old man, how old would you expect him to be at death?

Reign of Pharaohs

If you opened a book about the history of ancient Egypt and noticed that at 4000 BC a particular pharaoh had been ruling for X years, how many years total would you expect his reign to be?

Duration of marriages

A friend is telling you about an acquaintance whom you do not know. In passing, he happens to mention that this person has been married for X years. How many years do you think this person's marriage will last?

Movie run times

If you made a surprise visit to a friend's place and found that they had been watching a movie for X minutes, what is your prediction about the total length of the movie (in minutes)?

Cake baking times

Imagine you are in somebody's kitchen and notice that a cake is in the oven. The timer shows that it has been baking for X minutes. How long to you expect the total amount of time to be that the cake needs to bake?

Waiting time

If you were calling a telephone box office to book tickets and had been on hold for X minutes, how long would you expect to be on hold overall?