Individual Performance in Optimal Bayesian Inference

Rakesh Patel (rakesh_patel_1@brown.edu)
Department of Cognitive, Linguistic, and Psychological Sciences, Brown University
Providence, RI 02912 USA

Abstract

Applications of Bayesian inference to human decision-making have met with mixed success, but new theoretical developments and experimental paradigms are helping to form a clearer picture of the role that inference plays in human cognition. We combine the latest ideas to provide evidence that at a computational level, the mind's ability to make predictions may be grounded in Bayesian theory. Our results support the idea that the mind's capacity for statistical reasoning is more sophisticated than previously hypothesized.

Keywords: Bayesian inference; iterated learning; prediction; prior distribution.

Introduction

Imagine trying to predict your final grade in a math course, given only your grade on one homework assignment of many. Or consider being stuck in a traffic jam for 30 minutes, and deciding whether to wait until the traffic subsides or to exit the freeway and take slower surface streets. On a daily basis, we are bombarded by these types of situations in which we must make a decision on the basis of a very sparse set of data, often just one relevant observation. A powerful paradigm in cognitive science is the heuristics and biases approach (Gigerenzer, 1991, p. 84), which suggests that humans have difficulty making probability assessments. In particular, Tversky (1974) suggests that people cannot produce correct posterior probabilities on a simple Bayesian inference task with just 2 hypotheses (p. 1125). Instead, he claims, they produce estimates by transforming given data with simple linear or constant functions, a phenomenon known as the anchoring bias. He cites an experiment in which subjects were given either the number 10 or the number 65, and were asked if the number under- or over-estimated the percentage of African countries in the United Nations. Those who were given 10 as a starting number gave an average of 25 as the true answer, and those who started with the number 65 gave an average of 45 as the true answer. The anchoring effect explains this bias, Tversky claims, because people's responses depended heavily upon initial values.

Not content with Tversky's classical view, some researchers now believe that people use given data and prior knowledge to make intelligent estimates in accordance with statistical theory. After all, if somebody asks you a question such as, "Do you think the population of Russia is greater or fewer than 180 million?" you assume that the person's given estimate, 180 million, is not purely random and serves as a reasonable initial guess at the true answer. The estimate of 180 million is the only information you have, so why not make use of it by treating it as given data?

When results such as these are analyzed from the perspective of Bayesian inference, many of the biases disappear. To perform a Bayesian inference on a set of data, one first hazards a guess on the distribution of the data. After observing some real-world data, the guessed distribution is then revised – its mean, density, and overall shape may differ considerably from the initial guess. When the real-world data consists of a single datapoint, known as a *probe value*, the *median* of the revised, or posterior, distribution given the data is the optimal prediction. Tenenbaum and Griffiths (2006) tested 350 college students' ability to predict the outcome of everyday events (p. 767) by asking them questions such as such as, "Suppose that in 2000 B.C. a certain pharaoh has been ruling Egypt for 11 years. How long do you predict the total length of his reign to be?" After aggregating the results, they found that people's responses closely matched the optimal predictions given the single datapoint contained in the question. The finding is remarkable because different phenomena have wildly different distributions and thus very dissimilar optimal prediction functions. Life spans are normally distributed, movie run times follow a power-law distribution, and durations of pharaoh reigns follow an Erlang distribution, for example (see Figure 1). Tenenbaum and Griffiths concluded that people implicitly store knowledge of the distributions of these everyday phenomena (p. 771), and use this knowledge to make optimal predictions.

Mozer, Pashler, and Homaei (2008) presented the first theoretical challenge to these findings. They claimed that people's responses followed real-world distributions only when aggregated, as in Tenenbaum and Griffiths' analysis (p. 1134). As a demonstration of this effect, Mozer et al. recalled a country fair in which hundreds of people were asked to give the weight of an ox, and the average of their responses differed from the true value by just one pound (p. 1134). Their minimum-of-k-samples, or Min(k), algorithm produced the same results as in Tenenbaum and Griffiths' experiment, even though it provided responses by simply choosing the minimum value from a very sparse (just 2 or 3) number of samples. Because the collection of everybody's limited number of real-world samples must match the actual distribution of real-world data, their responses should also match the appropriate distributions when aggregated. When tested hundreds of times on Tenenbaum and Griffith's questions, the Min(k) algorithm's responses gave near-optimal predictions (p. 1145), even when assuming that each person might have access to just 2 samples.

But because this algorithm assumes that prior knowledge is contained in just 2 or 3 samples, it should perform poorly when one set of samples (i.e. one person) is used to answer a large number of prediction questions – it would select identical answers to almost all questions. To test this idea, Griffiths and Kalish (2005) introduced iterated learning (p. 1), a paradigm in which a subject is repeatedly asked prediction questions about one distribution. They argued that in the perpetual process of gathering data, forming a hypothesis, gathering more data, and refining the hypothesis, the generation of each successive hypothesis depends only upon the current hypothesis and the current data. Thus, the hypotheses form a Markov chain whose transitional probabilities eventually converge to the true prior probability distributions, meaning that a subject's responses to successive prediction questions will approach optimal predictions. In an iterated learning experiment, Lewandowsky, Griffiths, and Kalish (2009) asked each of 35 subjects 160 prediction questions about 8 possible distributions, randomly generating each question's probe value from between 1 and the previous question's response (p. 976). They claimed that only sampling from the true prior distributions could have produced their finding that each individual made optimal predictions. They also considered the performance of the Min(k) algorithm, whose responses ultimately failed to replicate the subjects' performance on these prediction questions. As predicted, Min(k) was only capable of producing identical answers for most questions, whereas individual subjects' predictions demonstrated their knowledge of the true prior distributions.

However, Lewandowsky et al.'s analysis of individual, and not aggregate, performance on prediction tasks was limited by the fact that only 35 students participated, which was enough to demonstrate that their paradigm could discredit Min(k), but still not enough to thoroughly understand the nature of individual performance on these prediction tasks. Encouraged by their findings, we will perform an experiment with a method very similar to theirs, but with more participants and fewer experimental trials. More participants will allow us to more clearly visualize the distributions from which subjects selected their responses. We use fewer experimental trials because Lewandowsky et al. demonstrated that subjects were able to make optimal predictions after only 4 or 5 trials for each distribution – this means that the transition probabilities in the Markov chain of hypotheses converges rapidly for these prediction tasks. A web-based experiment will allow us both to reach more participants and to consider responses from those who only partially completed the experiment.

Experiment

Method

Participants 72 anonymous subjects participated in the Web-based experiment, which was advertised on a popular social networking site. It is likely that many of the participants were college students. Further, analytics revealed that the majority of participants were accessing the internet through Brown University servers. 37 subjects completed the experiment in full; the remaining 35 subjects, who answered at least one question, responded to an average of 9.8 questions each.

Apparatus and Design The experiment was implemented using a PHP web application which stored responses in a MySQL database. Subjects were presented with 4 chains of 10 prediction questions each. The 4 chains corresponded to the 4 distributions: life spans of males, total movie box office grosses in US dollars, length of pharaohs' reigns in ancient Egypt, and U.S. Representative term lengths. Each question contained a statement and a probe value, *t*. An example question for the movie grosses distribution with *t*=6 would be, "Imagine you hear about a movie that has taken in 6 million dollars at the box office, but don't know how long it has been running. What would you predict for the total amount of box office intake for that movie, in millions of dollars?" The first question in each chain was seeded with a value randomly chosen from the set of 5 possible seed values for the chain, given in Table 1. The value of *t* for the *nth* question in each chain was an integer randomly selected from the interval [1, t_{n-1}], where t_{n-1} was the value of *t* for the *(n-1)th* question in the chain. Finally, questions from all chains were mixed and presented in random order.

Table 1: The 4 phenomena, their distributions in nature, and the seed values used to initiate the chains.

Chain	Distribution	Seed Values
Life Spans	Gaussian	18, 39, 61, 83, 96
Movie Grosses	Power Law	1, 6, 10, 40, 100
Pharaohs' Reigns	Erlang	1, 3, 7, 11, 23
U.S. Rep. Terms	Erlang	1, 3, 7, 15, 31

Procedure Upon visiting the website where the web application was hosted, subjects were greeted with a welcome page indicating that the experiment consisted of 40 questions that should be answered in no more than a few seconds each, and that all responses would be kept anonymous. If they agreed to these terms, they clicked on a link that took them to the first question. Subjects entered their answers in a text box and were then taken to the next question's page. After completing all 40 questions, the subject was taken to a page containing a brief thank-you message.

Results

Responses from every subject who answered at least one question were included in the analyses of aggregate performance, but only those subjects who fully completed all 40 questions were considered in the individual analyses. Website analytics revealed that, excluding the first five questions and the last question, subjects spent an average of 9 seconds on each question (the figure includes page loading time, which may have been 1 second or more). The web-based experiment included basic controls to ensure

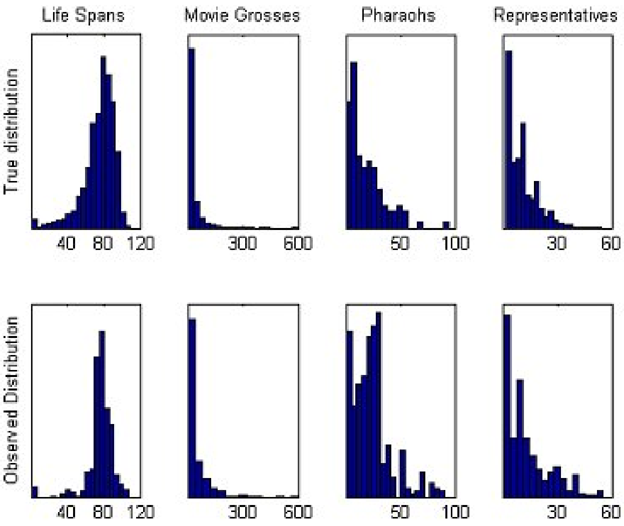

Figure 1: On the top row are the actual distributions of the phenomena. On the bottom row are the distributions of the 1,823 datapoints entered by participants.

high-quality data. For example, subjects could not leave a question blank, or enter a response that was less than the probe value. In total, 1,828 responses to questions were recorded. Of these, 2 responses to pharaoh reign questions and 3 responses to U.S. representative questions were greater than 100 and were removed. All of the remaining 1,823 responses were considered in the analyses of aggregated data.

Though it is difficult to enforce the rule that each person should participate in the experiment only once, the nature of this experiment allows the same person to repeatedly participate and still generate prediction data that can be analyzed. Participants were never given the correct (optimal) predictions after answering each question, so the act of taking the experiment does not improve their ability to make predictions. Only 2 IP addresses were duplicated among those who answered at least one question, and these could still each be 2 different participants using the same computer.

We wanted to test both the hypotheses that the wisdom of crowds effect would still hold across a variety of distributions, and that each individual's responses represented optimal predictions. Figure 1 shows the distributions of all responses entered by all participants underneath the true distributions of the phenomena, which appear to moderately correspond. To verify this correspondence, we created quantile-quantile, or Q-Q, plots of the 4 phenomena, to help us compare the observed and true distributions. In a Q-Q plot, the quantiles of the first dataset are plotted against the same quantiles of the second dataset – if the plotted points lie on the line $y = x$, then the data very likely have the same distribution. These Q-Q plots are shown in Figure 2.

That people's responses and real-world datapoints come from similar distributions is confirmed by the correlation coefficients between 25-quantiles of the observed and true distributions for each type of phenomena: $R^2 = .95$ for lifespans, 0.99 for movie grosses, 0.95 for pharaoh reigns, and 0.97 for representative term lengths.

An optimal prediction curve for a given distribution is a function whose input is an actual instance from the population, and whose output is the predicted "total life" of that instance, based on the population's distribution. We define the total life to be the median of the posterior distribution when Bayesian inference is performed using the input value as the sole observed datapoint.

Different distributions produce wildly different optimal prediction curves. A simple application of Bayes' Law with a power-law prior reveals that the optimal prediction curve is a straight line passing through the origin, with slope dependent on the parameter of the power function (Tenenbaum and Griffiths, 2006, p. 773). Similarly, the optimal prediction curve for Erlang-distributed data is a straight line with slope 1 and y-intercept dependent on the Erlang parameter. The optimal prediction curve for normally-distributed data has no simple analytical form. We should expect optimal prediction curves with shapes similar to these for our real-world data. Figure 3 shows all of the ordered pairs of datapoints, with the probe value in the question as the independent variable, and the subjects' response as the dependent variable. We fit a cubic polynomial to the lifespan data as an approximation for its optimal prediction curve. That it reasonably matches the true optimal prediction curve for actual lifespan data, despite the curve being fit to over 450 datapoints, supports the idea that people are capable of making optimal predictions about life spans. We performed a linear regression on the movie grosses, pharaoh reigns, and representative term length data, since we expect their optimal prediction curves to be linear. For movie grosses, we obtained a line with y-intercept 2.1379. Considering the range of data values, this line very nearly passes through the origin. However, it is much steeper than the true optimal prediction curve. This is representative of a power-law distribution with a longer tail – one in which more movies earned large amounts of money. The real-world movie data were gathered in 2003, so it is not unreasonable to expect that people's prior distributions for movies have been adjusted recently to account for the growing number of billion-dollar blockbusters. For pharaohs and representative term lengths, we obtained lines with slopes 1.0223 and 1.2398, respectively, both of which do not pass through the origin. The observed prediction curve for pharaohs is consistently above the prediction curve for the actual data – Tenenbaum and Griffiths (2006) also found that people consistently but reliably overestimated the length of pharaoh reigns (p. 771), explaining that they most likely did not realize how low the average life span was in ancient Egypt, which produced subjects' overestimated predictions for the length of pharaoh reigns.

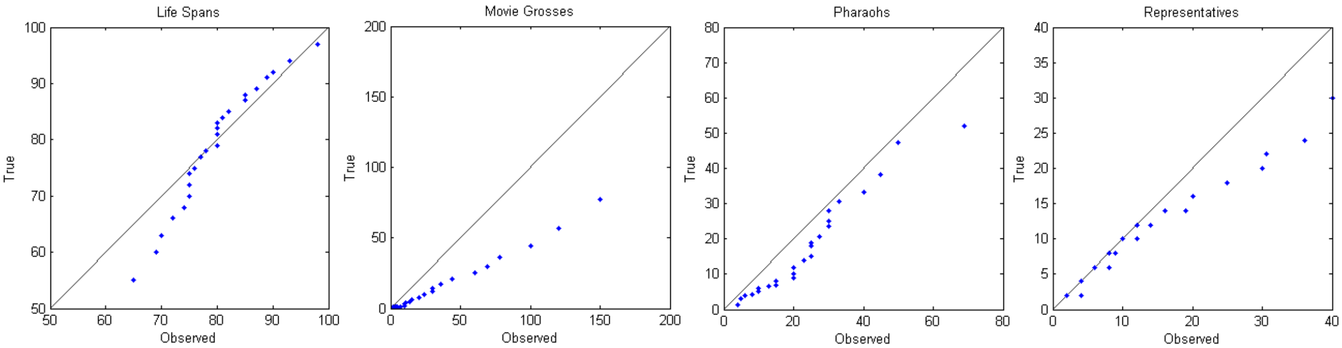

Figure 2: Q-Q plots of the distributions of responses entered by participants (observed) against the actual distributions (true) for each phenomenon. We removed the top and bottom 3% of each distribution and used 4% increments to plot the quantiles. If the points lie on the line *y* = *x*, then the two distributions are virtually identical. Any sort of linear correspondence, however, indicates that the data may still come from similarly-shaped distributions with different parameters.

The nature of the Min(k) hypothesis suggests that the few values in the long tail for movie distributions might just be outliers that subjects enter for lower probe values, and not optimal predictions. We counter by noting that for all subjects whose seed value for movie questions was 1, the highest probe value was 6 and the highest response was 10. The responses for modest-grossing movies resemble a power-law shape, with no outliers, suggesting that people do sample from a power-law distribution even when the probe values come from the densest part of the power law graph.

We must now test whether individual subjects were able to make optimal predictions. For all 37 subjects who completed the experiment, we should expect that even though their prior distributions may not have the same means, the line of best fit through their prediction data should have positive slope. We would expect this even for life span data, despite the fact that the optimal prediction curve for normally-distributed data is not linear. We performed linear regression on all 37 subjects for each of the 4 chains, and found that the mean slope of the regression line was above zero for all 4 chains. The results and associated statistics are shown in Table 2. Moreover, the slope for representative term lengths is very close to 1, in accordance with the fact that the optimal prediction curve for an Erlang distribution has slope 1. We were not able to observe such remarkable results for pharaoh reigns, but we note that the mean slope for subjects' prediction curves for pharaohs is significantly above 0. The mean slope for lifespan prediction curves is slightly positive (0.1765) with y-intercept approximately 69, which is expected if the true optimal prediction curve is horizontal but tends slightly upward for ages close to and greater than the mean of 76. We observed dramatically higher levels of significance (as indicated by the respective p-values) than Lewandowsky et al. (2009) when they performed the same analysis for their data (p. 988).

Discussion

Our experiment is but the most recent in a line of studies that demonstrate people's ability to make optimal predictions when presented with a single datapoint. This finding holds across data with different distributions, and suggests that people might be able to perform quite sophisticated Bayesian inference even without conscious awareness. The first important conclusion from this experiment is that because the 1,823 responses entered by subjects closely matched the true distributions of the respective datasets, we cannot attribute their responses to the Wisdom of Crowds effect – subjects must have had at least some familiarity with the true prior distributions of data, beyond just a few relevant examples, in order to consistently enter data that matched the real-world distributions.

Both the large number of subjects and the number of questions each subject answered allows us to more accurately determine the methods by which humans make predictions. Individual subjects' prediction curves closely resembled true optimal prediction curves, even with a variety of seed values and probe values from each end of the respective distributions, and the iterated learning paradigm produced rapid convergence to prior distributions – even for subjects who partially completed the experiment. This lends some support in favor of the idea that humans store implicit statistical knowledge about real-world distributions of data, and then recall this knowledge to make predictions when asked. More experimentation with this paradigm should help form a clearer picture of the powers and limits of human statistical inference.

Most subjects had at least some familiarity with the real-world distributions used in these and previous experiments. An important extension of these results would be the use of the iterated learning paradigm to capture people's statistical knowledge for unfamiliar distributions, including data for which it is difficult or impossible to determine a real-world distribution. The proliferation of online prediction markets,

Table 2: Summary of individual linear regression statistics for the 37 subjects who fully completed the experiment. We give the mean slope and y-intercept of the 37 regression lines for each chain; the associated t-statistics with 36df; p-values for the hypothesis that the mean slope is 0 against the hypothesis that it is different from 0; 95% confidence intervals for the slope; and in the last column, the number of individuals for whom the line of best fit had non-positive slope.

Chain	**Mean Slope**	**Mean Intercept**	**1-sample *t***	**p-value**	**95% CI**	**Slope ≤ 0**
Life Spans	**0.1785**	**69.3968**	**4.3850**	**$< 10^{-4}$**	**(0.0959, 0.2610)**	**6**
Movie Grosses	**1.1277**	**22.0560**	**10.4560**	**$< 10^{-11}$**	**(0.9090, 1.3465)**	**2**
Pharaohs' Reigns	**0.6514**	**17.6097**	**7.3599**	**$< 10^{-7}$**	**(0.4719, 0.8309)**	**5**
U.S. Rep. Terms	**0.9524**	**6.6026**	**15.0075**	**$< 10^{-16}$**	**(0.8237, 1.0811)**	**0**

and their relatively high degree of success in making predictions about unknown parameters, might serve as the ideal place to test the power of iterated learning beyond tightly-controlled experiments. Making use of our individual capacity to perform optimal statistical inference could dramatically improve collective prediction making, and consequently our ability to make decisions in uncertain conditions. After all, the intelligence of crowds depends solely upon the intelligent individuals that constitute them.

Acknowledgements

Many thanks to Professor Philip Fernbach, who provided valuable feedback for this experiment, and to Professor Tom Griffiths, who provided the datasets from his original research.

References

Gigerenzer, G. (1991). How to Make Cognitive Illusions Disappear: Beyond Heuristics and Biases. *European Review of Social Psychology*, *2*, 83-115.

Griffiths, T. L. & Kalish, M. L. (2005). A Bayesian view of language evolution by iterated learning. *Proceedings of the 27th Annual Conference of the Cognitive Science Society*.

Griffiths, T. L. & Tenenbaum, J. B. (2006). Optimal Predictions in Everyday Cognition. *Psychological Science*, *17*, 767-773.

Lewandowsky, T., Griffiths, T. L., & Kalish, M. L. (2009). The Wisdom of Individuals: Exploring People's Knowledge About Everyday Events Using Iterated Learning. *Cognitive Science*, *33*, 969-998.

Mozer, M. C., Pashler, H., & Homaei, H. (2008). Optimal Predictions in Everyday Cognition: The Wisdom of Individuals or Crowds?. *Cognitive Science*, *32*, 1133-1147.

Tversky, A. & Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases. *Science*, *185*(4157), 1124-1131.

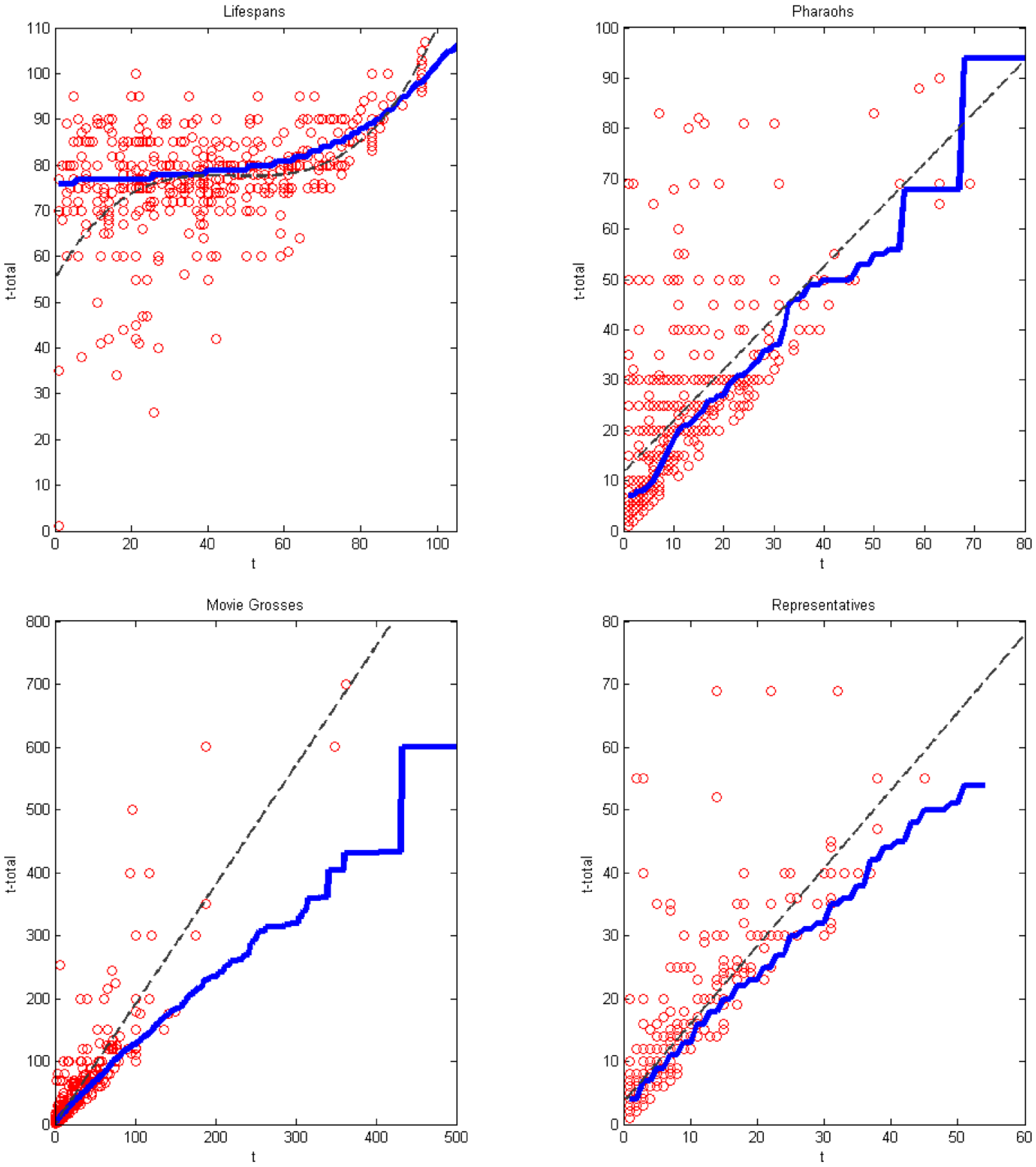

Figure 3. Red circles are question values (*t*) vs. subjects' responses (t_{total}). Solid blue lines represent true optimal prediction curves for the real-world data. Dashed grey lines represent best-fit curves for experimental data. For normally-distributed data, the optimal prediction curve has no simple analytic form, so we fit a cubic polynomial as an approximation. Optimal prediction curves for the other 3 distributions are linear, so we fit lines for these distributions.