Mem Cogn (2015) 43:1007–1020
DOI 10.3758/s13421-015-0522-6

Exploring the knowledge behind predictions in everyday cognition: an iterated learning study

Rachel G. Stephens[1] **· John C. Dunn**[1] **· Li-Lin Rao**[2] **· Shu Li**[2]

Published online: 3 April 2015

Abstract Making accurate predictions about events is an important but difficult task. Recent work suggests that people are adept at this task, making predictions that reflect surprisingly accurate knowledge of the distributions of real quantities. Across three experiments, we used an iterated learning procedure to explore the basis of this knowledge: to what extent is domain experience critical to accurate predictions and how accurate are people when faced with unfamiliar domains? In Experiment 1, two groups of participants, one resident in Australia, the other in China, predicted the values of quantities familiar to both (movie run-times), unfamiliar to both (the lengths of Pharaoh reigns), and familiar to one but unfamiliar to the other (cake baking durations and the lengths of Beijing bus routes). While predictions from both groups were reasonably accurate overall, predictions were inaccurate in the selectively unfamiliar domains and, surprisingly, predictions by the China-resident group were also inaccurate for a highly familiar domain: local bus route lengths. Focusing on bus routes, two follow-up experiments with Australia-resident groups clarified the knowledge and strategies that people draw upon, plus important determinants of accurate predictions. For unfamiliar domains, people appear to rely on extrapolating from (not simply directly applying) related knowledge. However, we show that people's predictions are subject to two sources of error: in the estimation of quantities in a familiar domain and extension to plausible values in an unfamiliar domain. We propose that the key to successful predictions is not simply domain experience itself, but explicit experience of relevant quantities.

Keywords Everyday reasoning · Iterated learning · Bayesian inference · Cross-cultural comparison

Survival depends upon making successful predictions about the future. This is hard enough in domains with which we are familiar but often predictions are also required in unfamiliar domains. A prosaic example is offered by the office worker who has a good understanding of how long to wait for the elevator in her building; how long should she wait in another, unfamiliar, building? If three futile minutes have already passed, should she head for the stairs? Given that these kinds of predictions are commonplace, two main questions arise: on what are the predictions based and how accurate are they?

Surprisingly, there is evidence that people can make accurate predictions even in domains for which they might have limited or even no direct experience. Griffiths and Tenenbaum (2006) asked participants to estimate quantities in several different domains that varied in familiarity (though they were collectively referred to as "everyday phenomena"), such as male life-spans, the baking time of cakes, movie grosses, and the lengths of pharaohs' reigns. Participants were asked to make a single prediction for each domain based on an observed (probe) value of that quantity. For example, they could be asked: given that a man is 39 years old, what is the best estimate of his total life span? Strikingly, responses generally reflected the actual distribution of the relevant quantity (as calculated from publicly available data), and were consistent with an optimal Bayesian updating rule (see also, Griffiths & Tenenbaum, 2011). Therefore, people behave as if they have accurate knowledge of the distributions of both familiar

✉ John C. Dunn
john.c.dunn@adelaide.edu.au

[1] School of Psychology, University of Adelaide, North Terrace, Adelaide, SA 5005, Australia

[2] Institute of Psychology, Chinese Academy of Sciences, Beijing, People's Republic of China

and unfamiliar quantities in the world and can use this knowledge to make sensible predictions.

The results found by Griffiths and Tenenbaum (2006) were extended by Lewandowsky, Griffiths, and Kalish (2009) using an iterated learning procedure (instead of the single-prediction approach). In this procedure, participants make multiple predictions in response to uniformly distributed probe values based on their previous responses. Griffiths and Kalish (2007) had earlier shown that this procedure converges to participants' subjective distribution of a quantity as long as their responses are independent and consistent with a Bayesian updating rule. Supporting Griffiths and Tenenbaum, Lewandowsky et al. (2009) found that estimates of the subjective distributions of quantities broadly matched their true distributions. Furthermore, because the estimates were based on multiple responses from individual participants, they could show that this result was not simply an artefact of aggregating across individuals, each of whom was employing a simple rule or heuristic – the so-called "wisdom of crowds" effect, in which an aggregated group is found to be accurate despite the fact that separate individuals make errors (this heuristics explanation had been proposed by Mozer, Pashler, & Homaei, 2008).

The results found by Griffiths and Tenenbaum (2006) and Lewandowsky et al. (2009) also contrast with a large body of earlier research suggesting that people tend not to be sensitive to prior probabilities but rely on simple-to-implement heuristics which, while effective, may also lead to systematic errors such as base-rate neglect (Tversky & Kahneman, 1974). Instead, Griffiths and colleagues have shown that when people predict the total extent or duration of a real phenomenon, they show an impressive sensitivity to the true distribution of that particular event.

If people's predictions are based on an internal representation of the actual distribution of quantities, a new question is posed: how is this knowledge acquired? One obvious answer is from direct experience, which may well account for domains such as movie run-times or cake baking times, or from indirect experience (such as reading), which might be the case for domains such as movie grosses. However, Lewandowsky et al. (2009) found that people's predictions of the lengths of pharaoh's reigns also matched the actual distribution. It is difficult to see how this knowledge could have been gained either directly or indirectly as it is unlikely that many of their undergraduate participants were experts in Egyptology. Interestingly, in the Griffiths and Tenenbaum study, participants were not as accurate in their judgments of this quantity and, although the distribution of responses tended to follow the correct (Erlang) distribution, the estimates were slightly too high. To account for this, Griffiths and Tenenbaum suggested that responses may have been based on an analogy to the reigns of modern monarchs, with which people would be more familiar, followed by a downward adjustment to account for the presumed shorter life spans of ancient Egyptians. In their study, this downward correction was not quite sufficient, whereas in Lewandowsky et al. this strategy appears to have been more successful.

Because we are often faced with unfamiliar events (cf. elevator example), what is the role of direct experience in generating accurate predictions? This raises two questions. First, if people vary in their familiarity with a domain, how are their predictions affected? Second, how do they extrapolate their knowledge of a familiar domain to make accurate (or perhaps inaccurate) predictions in a similar but less familiar domain? Though it is likely that some of the domains investigated by Griffiths and colleagues were more commonplace than others, the effect of domain-familiarity has not yet been directly studied.

In order to investigate more directly the role of familiarity in predicting quantities, we compared two groups of people who live in different cities, one in Australia and the other in China. We used the same iterated learning procedure employed by Lewandowsky et al. (2009) and elicited predictions across four different domains; pharaoh reign lengths, movie run-times, cake baking times, and the lengths of Beijing bus routes. The last of these had not been previously investigated and was selected to be familiar to the group resident in China (Beijing) but unfamiliar to the group resident in Australia (Adelaide). Thus, as well as providing points of comparison with the results found by Lewandowsky et al., the two groups allowed us to examine the effect of familiarity for different groups but in the same domain. We expected that pharaoh reign lengths would be equally unfamiliar to both groups, that movie run-times would be equally familiar, that cake baking times may be more familiar to Adelaide-resident than Beijing-resident participants (because of the relative rarity of traditional Western cake baking in Chinese households), and that the lengths of Beijing bus routes would be more familiar to Beijing residents.

Following the results of Lewandowsky et al., we expected that both groups of participants would be equally well (or poorly) calibrated in their judgments of pharaoh's reign lengths and movie run-times. Of greater interest would be their performance in domains with which one group is more familiar than the other: Beijing bus routes and cake baking times. Following Griffiths and Tenenbaum, we examined whether predictions for the relatively unfamiliar domains, (a) conform to the shapes of the true distributions, and (b) show evidence of some kind of adjustment.

Experiment 1

In this experiment, we used iterated learning to determine people's subjective prior distribution of quantities in four domains: pharaoh reign lengths, movie run-times, cake baking times, and Beijing bus route lengths. Two groups of participants were compared: a group resident in Adelaide, Australia (a city of approximately one million people) and a group resident in Beijing, China (with a population of approximately 20 million). The aim of the experiment was to determine the accuracy of predictions as a function of domain familiarity.

Method

Participants

The required sample size for each group was guided by Lewandowsky et al. (2009; they recruited 35 participants). Our data-collection stopping rule was 50 participants, though for the Beijing-resident group we were constrained by a 2-week recruitment period. The resulting Adelaide-resident group consisted of 50 psychology undergraduates from the University of Adelaide, who received course credit for participation. Ages ranged from 18 to 32 years, with 14 subjects being males. The Beijing-resident group consisted of 40 undergraduate students recruited from three Beijing-based universities: China Agricultural University, Beijing Forestry University, and the Chinese Academy of Sciences. They were each paid 30 RMB for participating in the experiment. Ages ranged from 19 to 34 years, with 24 subjects being males. Although we did not formally test this knowledge, we assumed that the Beijing-resident group would be, on average, more familiar with the Beijing bus transport system than the Adelaide-resident group.[1] The data from four Adelaide participants and four Beijing participants were removed because they did not follow instructions or failed to give plausible responses.

Materials

Predictions were elicited for each of four domains: pharaoh reign lengths, movie run-times, cake baking times, and Beijing bus route lengths. According to publicly available data collected by Griffiths and Tenenbaum (2006), actual pharaoh reign lengths follow an Erlang distribution, movie run-times are approximately normally distributed, while cake baking times follow an "irregular" distribution. We obtained a list of the actual lengths (i.e., total number of stops) of 822 bus routes in the Beijing metropolitan area from the website: http://wenku.baidu.com/ (accessed November 2010). Figure 1 shows the distribution of these lengths. Based on visual inspection, it roughly follows a log normal distribution.

The experiment was a computer-based task using Matlab with the Psychophysics Toolbox extensions (Brainard, 1997). All materials were presented in English for the Adelaide-resident group and in Mandarin for the Beijing-resident group.

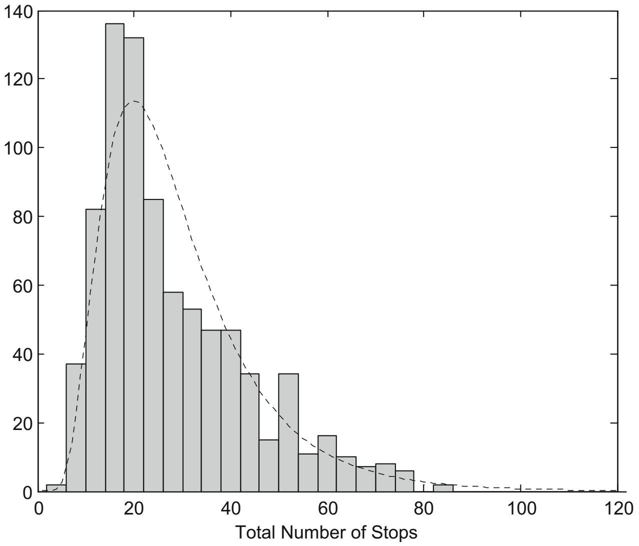

Fig. 1 Actual distribution of lengths (number of stops) of Beijing bus routes obtained from http://wenku.baidu.com/. Superimposed is the best-fitting log normal distribution

The stimuli consisted of eight chains of 20 prediction trials; one chain for each domain. The four domains of interest were intermixed with four other domains (movie grosses, poem lengths, male life-spans, and phone waiting times) also examined by Griffiths and Tenenbaum (2006) and Lewandowsky et al. (2009). These other domains served as filler trials designed to maximize the independence of judgments by minimizing the effects of memory across trials. Responses to the filler trials were not analysed.

Table 1 shows summary statistics for the four domains of interest. The set of seed values followed those used by Lewandowsky et al. (2009) and are explained below. Each participant was asked the following questions in relation to each domain.

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

Movie lengths If you made a surprise visit to a friend's place and found that they had been watching a movie for *t* 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 *t* minutes. How long do you expect the total amount of time to be that the cake needs to bake (in minutes)?

Lengths of Beijing bus routes If you caught a bus in Beijing, China and noticed that it had already passed *t* stops, what do

[1] Anecdotally, the lack of car ownership in the undergraduate population necessitates extensive use of the bus system for travel around Beijing. Bus tickets are also heavily discounted for students, facilitating the use of buses.

Table 1 The initial seed values used on the first trial of each domain/chain and the number of trials to reach convergence on each chain for the Adelaide-resident and Beijing-resident groups in Experiment 1

Domain/chain	Seeds					Trials to convergence	
						Australia-resident	China-resident
Movie lengths	30	60	80	95	110	1	1
Reign of Pharaohs	1	2	7	11	23	2	1
Cake-baking times	10	20	35	50	70	2	2
Beijing bus routes	12	20	37	53	68	2	3

you think would be the total number of stops on this bus route?

Prior to the commencement of the experiment, participants were given a set of practice trials. Following Lewandowsky et al. (2009), these consisted of equivalent questions about chrysalis ages, puddle durations, house painting durations, and shark-spotting durations.

Procedure

We used the same iterated learning procedure described by Lewandowsky et al. (2009). Participants answered a series of 20 questions (called a *chain*) for each of eight domains. The eight chains were intermixed from trial-to-trial to maximize independence of responses within each chain. There were always at least two trials separating trials that referred to the same domain. Each participant answered a total of 160 questions excluding practice trials. Following Lewandowsky et al. (2009), responses to the last ten trials on each chain were defined as being drawn from the final "stationary" distribution.

The response on each trial defined the participant's estimate of the total quantity, t_{total}, conditional on the current value, t. On each trial, the value of t was a random sample from the uniform distribution defined on the range, [1, t_{total}(previous)], where t_{total}(previous) was the participant's estimate of t_{total} from the previous trial in the current chain (which may have been several actual trials before because of the intermixing of chains). On the first trial of a chain, t was one of the corresponding seed values for that chain (randomly selected; see Table 1). Participants entered their responses using the computer keyboard. Only integer values were accepted and responses could not be less than the current probe value, t. Participants were permitted to edit their response prior to pressing the enter key on the keyboard. Participants were then asked to rate their confidence in their estimate (these ratings were not analysed). The next trial commenced after 1 s.

Participants were asked to take their time and to consider each question carefully, paying attention to the probe value. They were told that we were interested in their intuitions and formal calculation was not required. Before beginning the experimental trials, participants responded once to each of the four practice queries, presented in a random order. Overall, the experiment took around 25–40 min.

Results

Convergence analysis

In the iterated learning procedure, chain convergence occurs when responses are no longer dependent upon the initial seed values. Following Lewandowsky et al. (2009), we compared participants' responses across the five seed conditions for each chain using a Kruskal-Wallis test with a Bonferroni-adjusted $\alpha = 0.0025$. Convergence was defined as having occurred on the first trial to yield a non-significant χ^2 value across the five seed groups. For both the Adelaide-resident and Beijing-resident groups, the four critical chains converged quite quickly, in only one to three trials (see Table 1). We were therefore confident that responses from the final ten trials of each chain could be regarded as samples from the prior distribution. These final responses were aggregated across participants, constituting the stationary distribution for each domain.

Analysis of stationary distributions

Figure 2 presents histograms of the stationary distributions for the two groups (middle and lower rows). Histograms of the corresponding actual distributions are shown in the top row. Visual inspection suggests that for both groups, the stationary distributions are broadly similar in shape to the corresponding actual distributions.

In order to determine the level of correspondence between actual and stationary distributions, we examined the quantile-quantile (Q-Q) plots for each domain and group. The results are shown in Fig. 3 for the set of quantiles from 5 % to 95 % in steps of 5 %. If the stationary and actual distributions are identical then all the points on the Q-Q plot should fall on the main diagonal (the solid line in each plot) corresponding to an angular slope of 45°. We measured departures from identity in two ways. First, we calculated the angular slope, θ, of the best fitting straight line through the origin of each

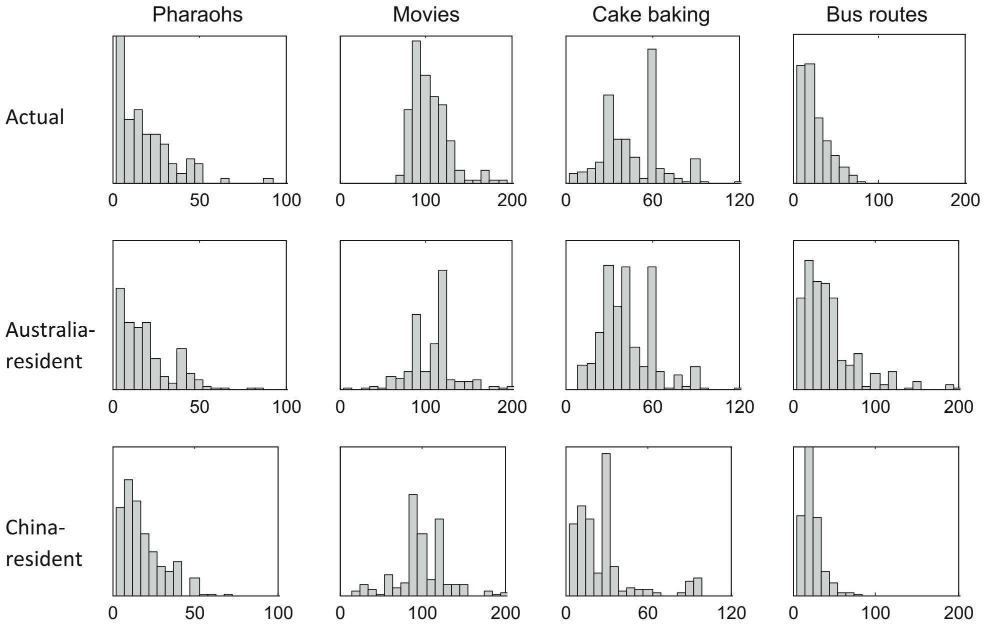

Fig. 2 Histograms of the actual and stationary distributions from Experiment 1

plot. Values of θ greater than 45 indicate systematic *under*-estimation of the actual quantity while values of θ less than 45 indicate systematic *over*-estimation. Second, we calculated the normalized root-mean squared deviation or coefficient of variation, V, of the points from the line of best fit (with slope θ) through the origin. If V is substantially greater than zero then it

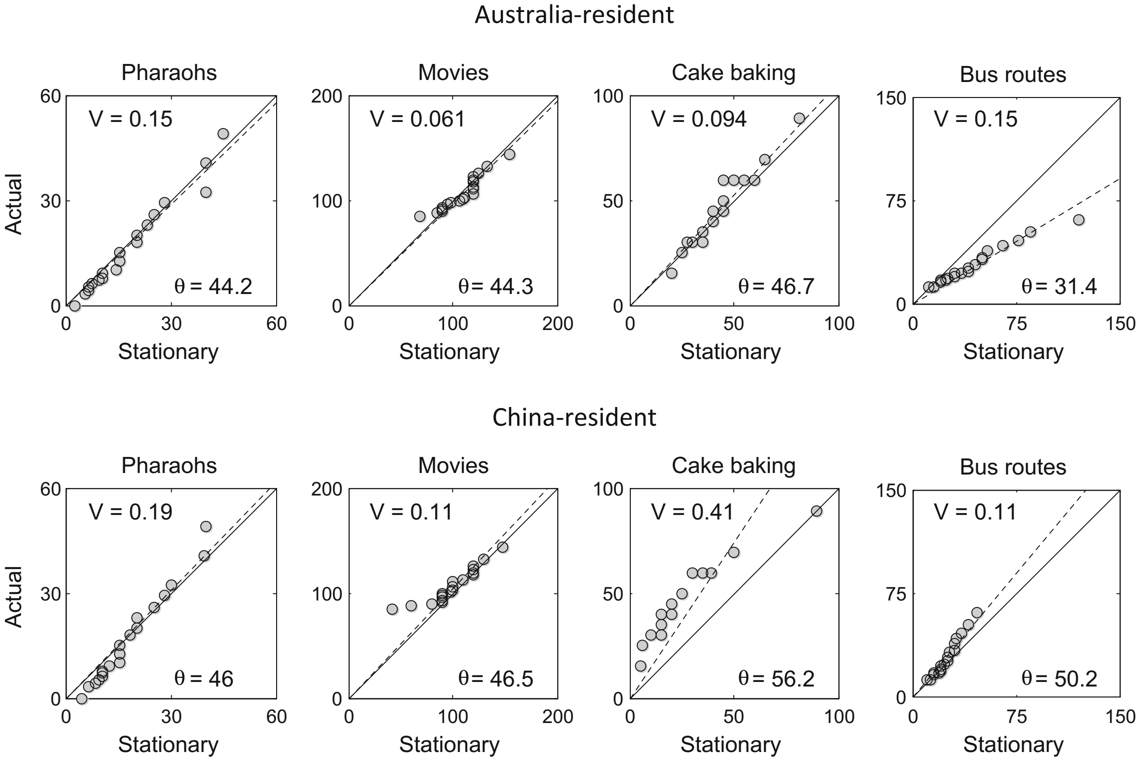

Fig. 3 Quantile-quantile (Q-Q) plots showing the correspondence of the stationary distributions from the Australian-resident (top row) and Chinese-resident (bottom row) groups with the actual distributions in Experiment 1. The data points mark the quantiles in 5 % increments from the 5th to the 95th percentiles. If the stationary and actual distributions are identical, all the data points should fall on the main diagonal (dark line). The dotted line presents the line of best fit (with the intercept fixed at zero). θ is the mean slope of this line in degrees where deviation from 45 indicates departure from identity. V is the coefficient of variation based on deviation of the points from the line of best fit based on θ

indicates that despite systematic over- or under-estimation, the stationary distribution has a different shape to the corresponding actual distribution. We tested the statistical significance of the observed values of θ and V through bootstrap re-sampling of the actual distributions. For each bootstrap sample we calculated the corresponding values of θ and V and compared them to the obtained values. We did this 10,000 times for each chain and group and estimated the proportion of sample estimates that exceeded the obtained values. The mean values of θ and V for each group and domain are shown in Fig. 3. We report the results of our bootstrap analysis in terms of the standardized difference between the observed and actual parameter values (corresponding to an effect size, *d*), the standard error of the bootstrap sample, and the observed *p*-value.

Visual inspection of Fig. 3 suggests that participants in the Adelaide-resident group were reasonably well calibrated in their predictions of the lengths of pharaohs' reigns, movie run-times, and cake baking times. The subjective distribution for each domain generally approximated the correct shape (Erlang, normal or irregular), scale, and location. These results replicate those found by Lewandowsky et al. (2009) although, in the current study, some short movie run-times were predicted despite short films not having been included in the actual data. In contrast to the first three domains, the Adelaide-resident group systematically over-estimated the lengths of the Beijing bus routes although they did capture the appropriate shape – an approximate log normal distribution. Nevertheless, using the bootstrap significance test, the hypothesis that the observed value of θ was equal to 45 could be rejected for movie run-times (d = -2.38, SE = 0.285, p = .037), cake baking (d = 3.82, SE = 0.515, p = .002), and Beijing bus routes (d = -17.86, SE = 0.764, p < .0001). Similarly, the observed values of V were significantly greater than zero for all four domains. In other words, although the stationary distributions approximated to varying degrees the actual distributions for pharaohs' reigns, movie run-times, and cake baking times, the observed deviations were sufficiently great to formally reject the hypothesis that the two sets of distributions were identical.

Visual inspection also reveals that participants in the Beijing-resident group were similarly well calibrated in their predictions of the lengths of pharaohs' reigns and movie run-times (again with some short films predicted). However, they were poorly calibrated in the predictions of cake baking times and, surprisingly, in their predictions of the lengths of Beijing bus routes. This group tended to under-estimate the duration of cake baking times and, as Fig. 2 also shows, their aggregate stationary distribution did not capture the shape of the actual distribution. They also systematically under-estimated the lengths of Beijing bus routes, capturing the shape but not the scale of the actual distribution. As for the Adelaide-resident group, bootstrap testing revealed that the best-fitting value of θ deviated significantly from 45 for movie run-times, cake baking times, and lengths of Beijing bus routes (p < .0001). As found for the Adelaide-resident group, the coefficient of variation, V, was also significantly greater than zero in all four domains (p < .0001).

Discussion

The aim of Experiment 1 was to examine how people's predictions of quantities depended upon their level of experience or familiarity with the domain in question. There were two main results. First, we had expected that pharaoh reign lengths would be equally unfamiliar to both groups and that movie run-times would be equally familiar. In each domain, both groups were reasonably well calibrated – more so for pharaoh reign lengths than movie run times. This demonstrates that even without direct experience, people are able to produce estimates of quantities that largely reflect the actual distributions.

Second, we compared domains of varying familiarity between the two groups. For cake baking times, as expected, predictions by the Adelaide-resident group (mean θ = 46.7) tended to be better calibrated than those of the Beijing-resident group (mean θ = 56.2), who consistently under-estimated the actual times and did not accurately reflect the shape of the distribution. Following the suggestion by Griffiths and Tenenbaum (2006), we speculate that the estimates from this group may be based on a distribution of baking (or cooking) times for more familiar food and insufficiently adjusted (upwards), although it is impossible to know the base-distribution considered by this group.

Although we expected that Beijing bus routes would be more familiar to the Beijing-resident group than to the Adelaide-resident group, both groups were found to be poorly calibrated in their estimates but in different ways – they were over-estimated by the Adelaide-resident group and under-estimated by the Beijing-resident group. It is possible that the responses of the Adelaide-resident group are consistent with the extrapolation strategy suggested by Griffiths and Tenenbaum (2006), in which the more familiar distribution of bus routes lengths in their home city were adjusted according to the expectation that routes in Beijing should be longer due to the much greater size and population of that city. However, in order to confirm this strategy, it is necessary to know the subjective distribution of Adelaide bus route lengths for this group. It is possible that this distribution matches exactly the responses made by the Australia-resident group, suggesting that no adjustments were made when predicting Beijing routes. Furthermore, it is noteworthy that participants in the Beijing-resident group tended to under-estimate the actual lengths of the Beijing routes. This result was unexpected as we had thought that this group would be familiar with and well calibrated to the distribution of bus routes in their home city. It potentially draws into question the idea that people are consistently able to gain accurate knowledge of everyday events with which they are familiar.

Experiment 2

In order to shed light on the strategies used by the Adelaide-resident group and, indirectly, those used by the Beijing-resident group, we conducted a second experiment. This was identical to Experiment 1 except that all participants were residents of Adelaide and the question concerning Beijing bus routes was replaced by an equivalent question concerning Adelaide bus routes. We wanted to determine the subjective distribution of local route lengths of an Adelaide-resident group in order to evaluate whether and how the similar group in Experiment 1 extrapolated this knowledge to estimate the lengths of Beijing bus routes.

Method

Participants

Because of the confirmatory nature of Experiment 2, our data-collection stopping rule was 25 participants, though due to non-attendance our final sample size was 24 people. The participants were residents of Adelaide, South Australia (mostly students of the University of Adelaide), who had lived in Adelaide for M = 14.7 years (range 1–40 years). They received AU$15 for their contribution. Ages ranged from 19 to 40 years, with eight subjects being males.

Materials and procedure

The stimuli were the same as those used in Experiment 1 with one difference. The question concerning the lengths of bus routes referred to Adelaide instead of Beijing. The same seed values were used for this question as were used in the corresponding question in Experiment 1. The procedure was the same as in Experiment 1.

Figure 4 shows the actual distribution of bus route lengths in the Adelaide metropolitan area based on data from http://adelaidemetro.com.au (accessed 25 January to 7 February, 2013). It is apparent that this distribution is different to the distribution shown in Fig. 1. First, it is not well described by a log normal distribution and has a more irregular shape. We have superimposed a best-fitting gamma distribution which provided a better fit (than a log normal). Contrary to intuition, bus routes in the Adelaide metropolitan area tend to be longer rather than shorter than those in Beijing. This may be attributable to the relative population densities of the two cities. Like all Australian cities, Adelaide has a relatively low population density, estimated at approximately 600 persons per square kilometre,[2] while the population density of Beijing is estimated at approximately 5000 persons per square kilometre.[3] With greater density, shorter bus routes can be sustained with comparable carrying capacity.

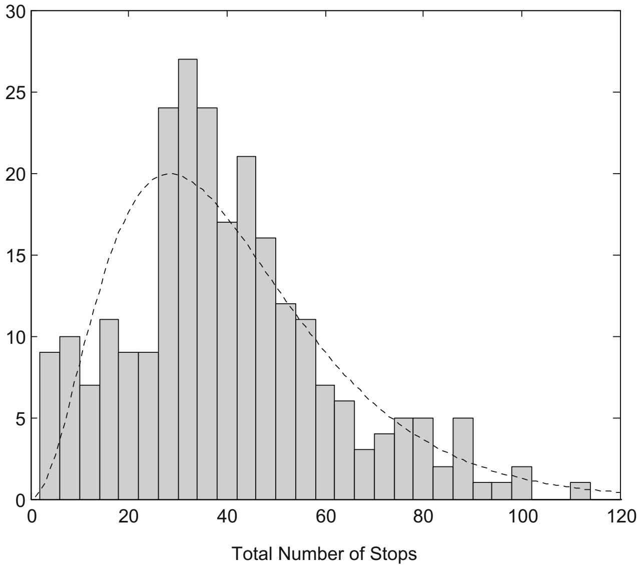

Fig. 4 Actual distribution of lengths (number of stops) of Adelaide bus routes obtained from http://adelaidemetro.com.au. Superimposed is the best-fitting gamma distribution

Results

Convergence analysis

Using the same procedure as Experiment 1, we again confirmed that the chains from the five different seeds converged after a maximum of two trials to a stationary distribution for each question. The final ten responses in each chain were aggregated across participants to form the stationary distributions.

Analysis of stationary distributions

Figure 5 shows the histograms of the actual and stationary distributions and corresponding Q-Q plots for the four questions of interest. Note that "bus routes" now refers to the lengths of Adelaide bus routes. These results generally replicated those from Experiment 1 for pharaoh reign lengths, movie lengths, and cake baking times. Participants tended to be well calibrated although, as found by Griffiths and Tenenbaum (2006), the participants in Experiment 2 slightly over-estimated pharaoh reign lengths. In this case, bootstrap

[2] From http://www.epa.sa.gov.au/soe_resources/education/population_and_urban.pdf.

[3] From http://www.newgeography.com/content/002808-world-urban-areas-population-and-density-a-2012-update.

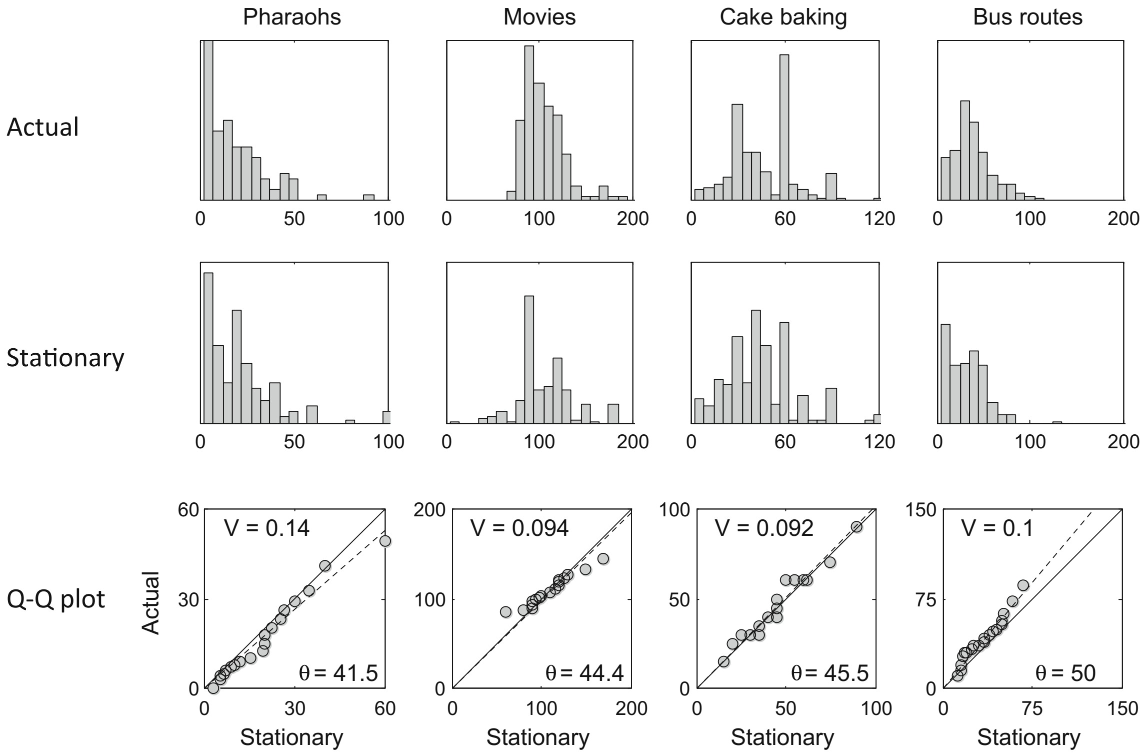

Fig. 5 Results for Experiment 2. Top row: Histograms of actual distributions, including Adelaide bus routes. Middle row: Histograms of stationary distributions. Bottom row: Q-Q plots showing correspondence of stationary distributions with actual distributions. The data points mark the quantiles in 5 % increments from the 5th to the 95th percentiles. If the stationary and actual distributions are identical, all the data points should fall on the main diagonal (dark line). The dotted line presents the line of best fit (with the intercept fixed at zero). θ is the mean slope of this line in degrees where deviation from 45 indicates departure from identity. V is the coefficient of variation based on deviation of the points from the line of best fit based on θ

tests of θ were significant for this domain (d = -2.14, SE = 1.62, p = .024) but not for movie lengths or cake baking times. Despite this apparent correspondence, the value of V in each domain was significantly different from zero (all p's < .05).

The Adelaide-resident group, like the Beijing-resident group in Experiment 1, systematically under-estimated the lengths of their local bus routes. Before examining this surprising result in more detail, we wanted to know whether Adelaide-participants' predictions of the lengths of Beijing bus routes are directly based on their (under-estimation) of the lengths of Adelaide bus routes or whether they make a further upward adjustment based on their (erroneous) beliefs about Beijing. This question is answered in the Q-Q plot shown in Fig. 6 which presents percentiles of the stationary distribution generated by the Adelaide-resident group in Experiment 2 (estimating the lengths of Adelaide bus routes) against those produced by the Adelaide-resident group in Experiment 1 (estimating the lengths of Beijing bus routes). It shows that the subjective lengths of Beijing bus routes are over-estimated in comparison to the subjective lengths of Adelaide bus routes (d = -13.07, SE = 0.699, p < .0001). This suggests that if the Adelaide-resident group in Experiment 1 based their estimates on the lengths of Adelaide bus routes, they adjusted these lengths upwards to compensate for the

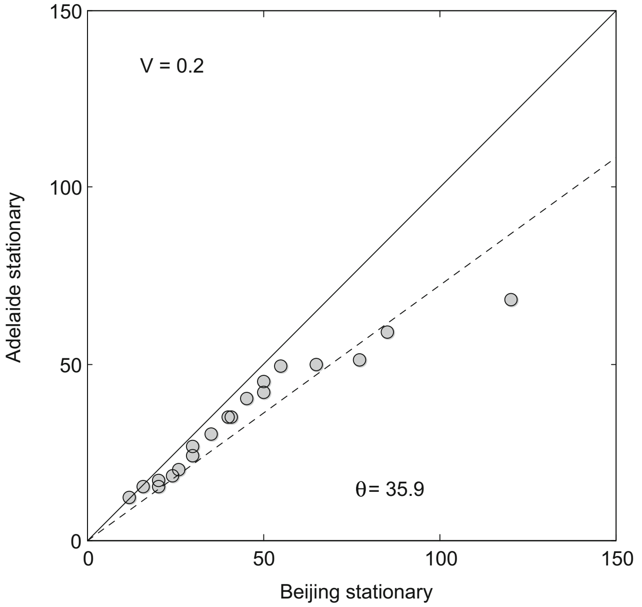

Fig. 6 Q-Q plot showing the correspondence of the stationary distributions for estimates of the lengths of Adelaide bus routes from the Australia-resident group in Experiment 2 against estimates of the lengths of Beijing bus routes from the Australia-resident group in Experiment 1. The data points mark the quantiles in 5 % increments from the 5th to the 95th percentiles

(false) perception that the bus routes in Beijing should be longer.

Discussion

The results of Experiment 2 support the view that the Adelaide-resident group in Experiment 1 used an analogy-with-adjustment strategy for predicting the unfamiliar Beijing bus routes. This Adelaide group did not simply apply their knowledge of local Adelaide routes (as elicited in Experiment 2), but, if they did consider this knowledge, they adjusted it upwards, perhaps based on the incorrect belief that Beijing bus routes are longer.

However, the results also showed that predictions for local bus route lengths can be systematically biased. The participants in Experiment 2, consistent with the Beijing-resident group in Experiment 1, consistently under-estimated the lengths of their local bus routes (d = 4.83, SE = 1.01, p < .0001). We had expected that since participants would be familiar with their local bus system, at least as much as they may be familiar with movie run or cake baking times, they would have developed an accurate representation of the lengths of the routes. However, both Beijing-resident and Adelaide-resident participants systematically under-estimated this quantity. One possible conclusion is that the hypothesis that people are generally able to acquire accurate knowledge of the distributions of quantities in familiar domains, is not correct. However, it may also be the case that we asked the wrong question. It is possible that people have accurate knowledge of the lengths of bus *journeys* they take but have difficulty extrapolating this knowledge to estimate the lengths of entire bus routes (of which journeys are only a part).

If bus journeys are familiar events, people may well have acquired accurate knowledge of their distribution, as with cake baking times and movie lengths. Thus, if we ask people to make predictions about the length of local bus journeys, they should be better calibrated than for total route lengths. On this scenario, people have no direct knowledge of the total length of their bus route but must estimate it from the length of their journey, analogous to the proposal that pharaoh reign lengths are based on the reign lengths of modern monarchs.

Experiment 3

The aim of Experiment 3 was to determine whether Adelaide-resident participants are calibrated in their estimates of local bus *journeys*. This experiment was identical to Experiment 2 with the exception that the question concerning Adelaide bus routes was replaced by an equivalent question about Adelaide bus journeys.

Method

Participants

Again our data-collection stopping rule was 25 participants, which was met. The participants were residents of Adelaide, South Australia (mostly students of the University of Adelaide), who had lived in Adelaide for M = 11.3 years (range 0.5–32 years). They received AU$15 for participation. Ages ranged from 18 to 47 years, with subjects including 14 males. One participant was removed for failing to follow instructions.

In order to evaluate the extent of their direct experience, after the experiment we asked participants about whether they travelled by bus "multiple times per week," "a few times per month," "a few times per year," "once per year or less," or "never" (or "other," which was never selected). The modal response was "multiple times per week," with all participants catching the bus at least "a few times per year." We also asked for the details of their most frequent bus trip and used this to calculate their average journey and route lengths (for this trip). The average journey length was 20.4 stops (SD = 13.8) and the average route length was 56.8 stops (SD = 24.2).

Materials and procedure

The stimuli and procedure were the same as in Experiments 1 and 2, with one difference. The bus question was changed to the following:

> Imagine you are on a bus in Adelaide and a stranger tells you she has already passed t stops since she got on the bus. What do you think would be the total number of stops along her journey, including the one where she gets off? (Do not count the stop at which she got on.)

In order to acquire the actual distribution for bus journeys in Adelaide, we surveyed people waiting at ten bus stops in the Adelaide central business district during the evening rush hour, from 4.45 pm to 6 pm, on 6 August 2014. The ten bus stops were selected to capture a wide range of bus routes, heading out from the city centre in a variety of different directions. Two researchers were stationed at each bus stop, and they approached as many waiting commuters as they could, ultimately contacting 855 valid responders. No personal details were collected.

The survey asked individuals to report the number of the bus they were waiting to catch and the name (or number) of the bus stop at which they would get off the bus. Based on this information, we calculated the actual length of each person's journey and the associated actual route length (based on based on data from http://adelaidemetro.com.au; accessed 14 August to 9 September 2014). In addition, we also asked the survey participants to estimate the length of their journey and the associated total route length.

Figure 7 shows the obtained distributions of actual bus route lengths and journey lengths based on the commuter survey. The distribution of route lengths is irregular and appears to selectively omit shorter routes that are present in the complete set of Adelaide bus routes (cf. Fig. 4). On the other hand, the distribution of journey lengths appears more systematic and is well described by a Weibull distribution.

As they necessarily are, journey lengths are shorter on average (M = 20.3 stops; SD = 11.4) than total route lengths (M = 54.9 stops; SD = 22.5). We note that these values are close to those reported by the participants in the iterated learning part of this experiment, suggesting that they are representative of the larger surveyed sample.

Results and discussion

Convergence analysis

Using the same procedure as Experiments 1 and 2, for each iterated learning question we again confirmed that the chains from the five different seeds converged after a maximum of two trials to a stationary distribution. The final ten responses in each chain were aggregated across participants to form the stationary distributions.

Analysis of stationary distributions and survey estimates

Figure 8 shows the histograms of the actual and stationary distributions and corresponding Q-Q plots for the four critical questions. Note that "bus journeys" refers to the lengths of Adelaide bus journeys (as opposed to the lengths of entire routes) compared to the distribution of bus journey lengths obtained from the commuter survey. These results generally replicated those from Experiments 1 and 2 for pharaoh reign lengths, movie lengths, and cake baking times. Participants tended to be well calibrated, although again the participants in Experiment 3 slightly over-estimated pharaoh reign lengths, and the irregular distribution for cake baking times was not captured quite so well. However, bootstrap tests of θ were all significant except for cake baking times while the tests of V were all significant (p's < .05).

Of most interest were the results for local bus journeys. In this case, although we had expected them to have well calibrated knowledge of a familiar domain, the participants in Experiment 3 consistently under-estimated the lengths of local bus journeys (d = 2.36, SE = 1.61, p < .02).

It is possible that this under-estimation may have been an artefact of the iterated learning procedure. However, the responses made by participants in this task are remarkably similar to those found in the commuter survey. As part of this survey, we asked respondents to estimate the number of stops they expected to pass from the current bus stop to when they got off the bus. Figure 9 (left panel) shows the Q-Q plot of these estimates plotted against the stationary distribution for bus journeys (based on the iterated learning task in Experiment 3). The two sets of estimates are quite closely calibrated (d = -1.21, SE = 1.18, p = 0.17). Interestingly, the same correspondence is found for estimates of the total lengths of the bus routes (i.e., comparing responses from the survey against those from Experiment 2; Fig. 9, right panel), d = 0.52, SE = 1.07, p = 0.40.

In summary, despite the apparent familiarity of the domain, people are not well calibrated in their estimates of the lengths of bus journeys they typically take. Estimates obtained from the commuter survey (assessing the length of the forthcoming journey) and estimates obtained from the iterated learning task (assessing the lengths of typical journeys), while equivalent, *both* under-estimated the actual lengths of journeys.

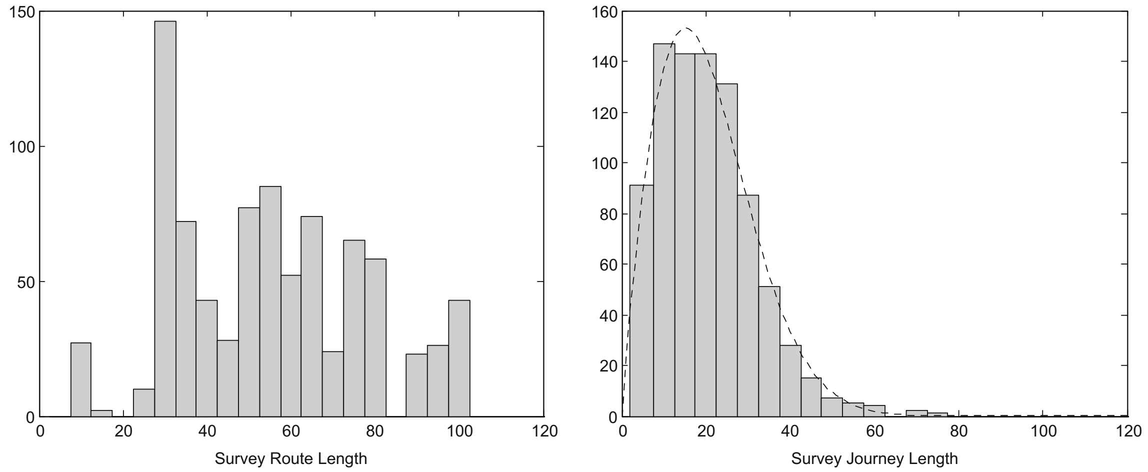

Fig. 7 Actual distribution of route lengths and journey lengths obtained from the commuter survey. The dotted line in the right hand panel shows the best-fitting Weibull distribution

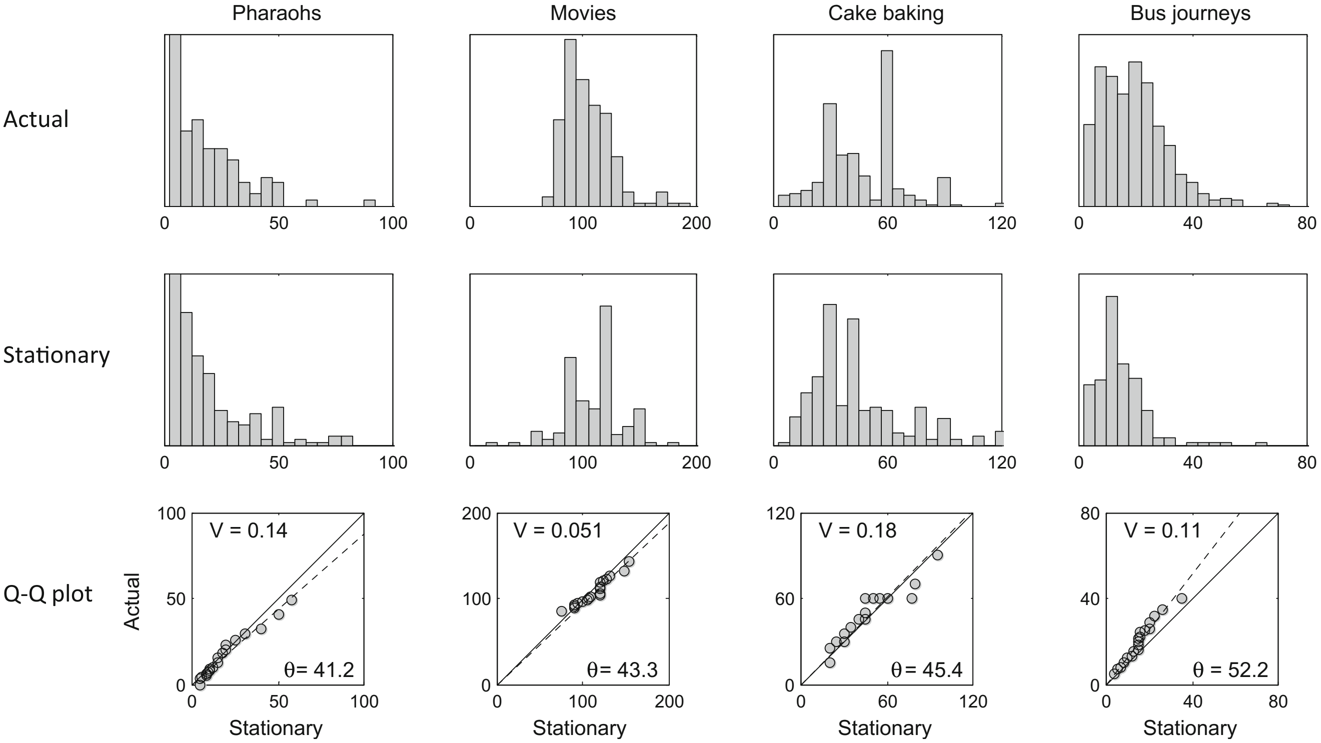

Fig. 8 Results for Experiment 3. Top row: Histograms of actual distributions, now including Adelaide bus journeys. Middle row: Histograms of stationary distributions. Bottom row: Q-Q plots showing correspondence of stationary distributions with actual distributions. The data points mark the quantiles in 5 % increments from the 5th to the 95th percentiles. If the stationary and actual distributions are identical, all the data points should fall on the main diagonal (dark line). The dotted line presents the line of best fit (with the intercept fixed at zero). *θ* is the mean slope of this line in degrees where deviation from 45 indicates departure from identity. *V* is the coefficient of variation based on deviation of the points from the line of best fit based on *θ*

Both Adelaide residents and Beijing residents also under-estimated the total lengths of their local bus routes. It is possible that this under-estimation is a direct consequence of the under-estimation of journey length. That is, although people under-estimate their journey length they correctly estimate the proportion of the total route length that the journey forms. To check this, we calculated the mean actual journey and route lengths from the commuter survey and compared these to the mean estimated journey and route lengths (from the same survey). The mean *actual* journey length was 20.3 stops and

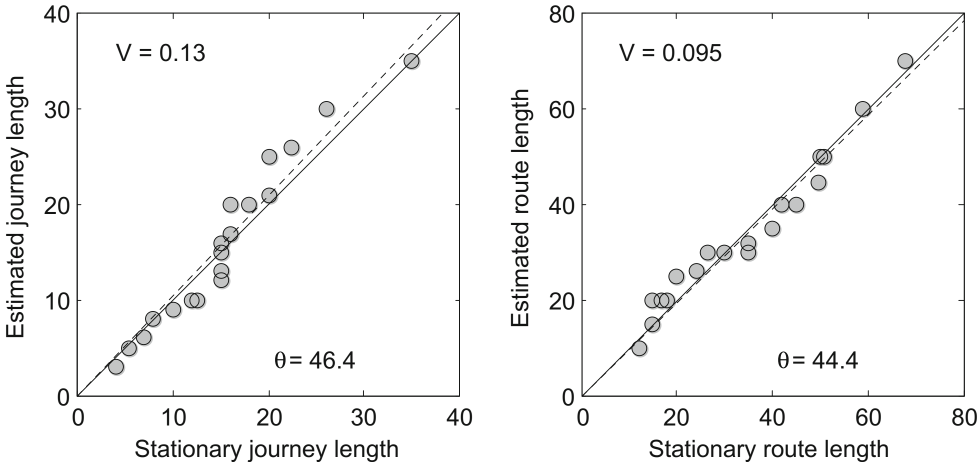

Fig. 9 Left panel: Q-Q plot showing the correspondence of the stationary distribution for estimates of Adelaide bus journeys in Experiment 3 against estimates of bus journeys from our survey. Right panel: corresponding Q-Q plot showing the stationary distribution for estimates of Adelaide bus routes in Experiment 2 against estimates of bus routes from our survey. The data points mark the quantiles in 5 % increments from the 5th to the 95th percentiles

the mean actual route length was 54.9 stops. The former is 0.37 of the latter. The mean *estimated* journey length was 16.4 stops (an under-estimate) and the mean estimated route length was 34.7 stops (also an under-estimate). The mean estimated journey length is therefore 0.47 of the mean estimated route length. We tested the hypothesis that these two proportions are equal using bootstrap resampling[4] which led to its easy rejection ($p < .0001$). As well as under-estimating journey length, participants also over-estimated the extent to which this journey forms part of the total route.

General discussion

Across three experiments, we explored how people use their existing knowledge and experience to predict quantities in familiar and unfamiliar domains. Our results revealed three main findings. First, consistent with similar results found by Griffiths and Tenenbaum (2006) and Lewandowsky et al. (2009), people are reasonably well calibrated in their understanding of the distribution of quantities in several different domains. That is, their estimates did not markedly over- or under-estimate the quantity of interest although, in many cases, even relatively small deviations are statistically significant. Thus, we found that four different groups, three resident in Australia and one in China, were able to generate well calibrated predictions for a domain with which all might be expected to have direct experience – movie run-times. In addition, the same groups (although less so for the groups in Experiments 2 and 3) were able to generate well calibrated predictions in a domain with which they were likely to have little direct experience – the lengths of the reigns of pharaohs of ancient Egypt. This suggests that at least in some situations, people are able to extrapolate knowledge from one or more familiar domains (presumably including knowledge of life spans) to an unfamiliar domain. Yet this is not without its limits. The Beijing-resident group in Experiment 1 were unable to extrapolate to the unfamiliar domain of cake-baking times, generating predictions that deviated from both the shape and scale of the actual distribution.

Second, in contrast to previous research, we found that people may be surprisingly poorly calibrated even in highly familiar domains. We had initially supposed that people would be familiar with the lengths of bus routes in their city but found that these were systematically under-estimated (Experiments 1 and 2). On reflection, it was apparent that people would rarely have direct experience of a journey along an entire bus route and so we supposed that they might be well calibrated on the more familiar journey length. Instead, people also systematically under-estimated this quantity (Experiment 3).

Third, an additional source of error (beyond inaccurate knowledge of a familiar domain) can be introduced when people need to extrapolate from a familiar to an unfamiliar domain. Although Adelaide residents systematically under-estimated the lengths of their local bus routes, they systematically over-estimated the lengths of Beijing bus-routes. As suggested by Griffiths and Tenenbaum (2006), we hypothesize that these estimates were based on estimates of local bus routes adjusted by a factor to compensate for the larger size of Beijing. This factor was an (excessive) over-estimate. Indeed, if participants had offered their unadjusted under-estimates of the lengths of (the longer) Adelaide bus-routes, they would have been more accurate. Similarly, we also found systematic errors when people extrapolated from estimates of local journey lengths to total route lengths. In this case, people over-estimated the proportion of the total route that was their journey, which led to an additional under-estimation of the total route length.

Another unanticipated finding was that people assume that their own experiences are representative of typical quantities (or at least treat them as such). We found a close correspondence between judgments of *typical* local bus routes and journeys (as queried in our iterated learning experiments) and judgments of people's *own* forthcoming route and journey (from the commuter survey). If people apply their own experiences to events in general then their predictions can be accurate only to the extent that this is a sound assumption, and indeed to the extent that their own experiences are correctly understood or estimated.

An outstanding question is why people under-estimate even their *own* journey lengths. In a typical case, a journey would be experienced multiple times, yet when asked in our survey, people reported approximately 80 % of the true length. We suspect that if given sufficient time (or accuracy incentive), many people would be able to bring to mind the correct sequence of stops and calculate the correct answer (cf. the well known mnemonic of the "method of loci"). However, people often use a variety of heuristics to avoid this kind of mental effort. One possible candidate is the availability heuristic (Tversky & Kahneman, 1973) combined with anchoring-and-adjustment (Tversky & Kahneman, 1974). On this view, when participants are asked to estimate a bus journey length, they first generate a sample of stops that are readily accessible from memory. The number of such stops serves as a self-generated anchor that requires adjustment upwards (Epley & Gilovich, 2001), which tends to be insufficient because of the mental effort required (Epley & Gilovich, 2006). That is, people progressively adjust and test their estimate until it reaches a plausible range at which point the process terminates, resulting (in this case) in an under-estimate.

[4] This consisted of the following four steps: (1) fit a model to the observed means that minimized sum of squared error subject to equal ratios; (2) shifted the data by subtracting the observed means and adding the best fitting estimates from (1); (3) bootstrap resampled the shifted data and fit the model from (1) to the means of each sample; (4) calculated p as the proportion of bootstrap model fits that exceed the observed model fit.

The under-estimation of local route lengths can be similarly explained. In this case, people are anchored by their (under-estimate of) actual or typical journey lengths. This anchor is also adjusted until a plausible range is reached, leading to a further under-estimation of the route length and concomitant over-estimation of the relative journey length. A similar process may lead to over-estimation of the route lengths of Beijing buses by Adelaide residents. In this case, the self-generated anchor is the under-estimated route length of a typical Adelaide bus. This, in turn, is adjusted until a plausible value is reached. However, what counts as a plausible value is a generous over-estimate of the actual lengths of Beijing bus routes.

At the outset, we asked: what is the role of familiarity or direct experience in generating accurate predictions? It seems that domain familiarity itself is not crucial, because people make errors even for familiar bus routes and journeys. Instead, we suggest that it is the *type* of relevant experience that is important. It is possible that a key to success for familiar domains such as cake baking times (for residents of Australia) and movie lengths (for residents of Australia and China) is having explicit experience of t_{total}, the value to be predicted. We suggest that people have access to a discourse in which the total duration of movies and baking times are explicitly (although separately) discussed and that it is this knowledge that is applied to generate their predictions. Even for Pharaoh reign lengths, people can also draw upon relevant explicit examples, such as known life-spans, also frequently discussed. In contrast, people are unlikely to have access to a discourse on the lengths of bus routes and journeys (unless they happen to work in the public transport sector). Thus, while people are familiar with various buses, stops, and key destinations, they typically are not told that a given bus has a route or journey of 34 stops (for example). Perhaps people encode their own bus journeys in terms of typical *durations* rather than the number of stops, and so would be better calibrated if they were instead queried about average journey durations (though it would be difficult to obtain the actual distribution of bus journey durations for comparison). In short, for some predictions people have domain knowledge that is explicitly represented in a form that is useable, while for other (even familiar) domains, knowledge is embedded in an activity where it is rarely explicitly noticed, though possibly encoded in an alternative unit of measurement (e.g., duration rather than number of stops).

This explanation of our results is related to the “estimation modes” proposed by Brown (2002) to account for people’s estimates of specific real-world quantities, such as the population of Los Angeles. One mode is *numerical retrieval*, in which an estimate is based on the recollection of at least one relevant numerical fact. Within this mode, there are various processes that may occur: people may be able to *directly retrieve* the value of interest (3.9 million people), or retrieve a related value that must be *adjusted* or *transformed* (e.g., a few years ago the population was 3.5 million people). In the absence of such retrievable facts, Brown (2002) has suggested that people engage in an alternative mode, called *ordinal conversion*. This is a more complex process that involves establishing a plausible response range, then judging where the target item (e.g., the population of Los Angeles) is situated relative to other relevant items (e.g., the populations of other cities).

We suggest that in the present series of experiments, when people are asked about familiar quantities that they have experienced as explicit facts (such as movie run times), they engage in a form of direct numerical retrieval, the products of which tend to closely match the actual distribution. In contrast, when people are asked about familiar or unfamiliar quantities that they have not experienced as explicit facts (such as bus journey lengths), they cannot rely on direct numerical retrieval and instead engage in more complex processes, analogous to retrieval-with-adjustment or ordinal conversion. The products of this may be subject to anchoring and adjustment biases and not reflect the actual distribution, unless by chance. This hypothesis could be tested in future research that directly contrasts novel domains constructed to contain explicit experience of t_{total} (permitting *direct numerical retrieval*) with the same domains constructed to contain experience of t_{total} embedded in other activities (preventing any form of *numerical retrieval*), and with domains constructed to contain explicit experience of a related but different t_{total} (permitting *numerical retrieval* with *adjustment* or *transformation*).

Acknowledgements The research reported in this paper was supported by the National Basic Research Program of China (973 Program, No. 2011CB711000) awarded to Li, and the Chinese Academy of Sciences Visiting Professorship for Senior International Scientists awarded to Dunn. The authors wish to thank their team of zealous researchers who helped them to survey bus commuters in Adelaide, and Kathleen Wright for processing the survey data. The authors also thank the reviewers for their helpful suggestions.

References

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision, 10*(4), 433–436.

Brown, N. R. (2002). Real-world estimation: Estimation modes and seeding effects. *The Psychology of Learning and Motivation, 41,* 321–359.

Epley, N., & Gilovich, T. (2001). Putting adjustment back in the anchoring and adjustment heuristic: Differential processing of self-generated and experimenter-provided anchors. *Psychological Science, 12*(5), 391–396.

Epley, N., & Gilovich, T. (2006). The anchoring-and-adjustment heuristic: Why the adjustments are insufficient. *Psychological Science, 17*(4), 311–318.

Griffiths, T. L., & Kalish, M. L. (2007). Language evolution by iterated learning with Bayesian agents. *Cognitive Science, 31*(3), 441–480.

Griffiths, T. L., & Tenenbaum, J. B. (2006). Optimal predictions in everyday cognition. *Psychological Science, 17*(9), 767–773.

Griffiths, T. L., & Tenenbaum, J. B. (2011). Predicting the future as Bayesian inference: People combine prior knowledge with observations when estimating duration and extent. *Journal of Experimental Psychology: General, 140*(4), 725–743.

Lewandowsky, S., 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*(7), 1133–1147.

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology, 5*(2), 207–232.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*(4157), 1124–1131.