Eliciting Human Beliefs using Random Generation

Pablo León-Villagrá[1] **Lucas Castillo**[2] **Nick Chater**[3] **Adam Sanborn**[2]
[1]Brown University, USA [2]University of Warwick, United Kingdom
[3]Warwick Business School, United Kingdom

Abstract

Elicitation methods, such as asking people to produce the deciles of a distribution, are standard practices in policy or applied statistics. However, these approaches often only capture a rough outline of what people know. We investigated whether tasks in which participants generate random sequences of items can be used to elicit people's implicit beliefs about the distribution of these items. Because it remains unclear if, and at what level of detail, people represent distributions, we applied both decile elicitation and random generation tasks to uncover the kinds of environmental statistics investigated by Griffiths and Tenenbaum (2006). We found that random generation is competitive with decile elicitation in predicting participants' expectations. Both random generation and decile elicitation revealed that people know the rough shapes of environmental distributions. Random generation, however, goes beyond decile elicitation in establishing the novel finding that people are aware of fine details of environmental distributions.

Keywords: belief elicitation; sampling for inference

Introduction

"How long does this cake have to go in the oven?"
"I don't know... Maybe 45 minutes?"

People routinely make inferences about everyday events or the magnitude of common quantities and often give reasonable guesses even without explicitly knowing the answer. Furthermore, when these beliefs are not simply gut instinct, but the result of acquired expertise, opinions and suggestions play an essential role in guiding all areas of society, from political consultants to sports pundits or weather forecasters.

What is the basis of this ability? Theoretical accounts range from proposals suggesting that people only acquire knowledge of a small set of representative instances (Mozer, Pashler, & Homaei, 2008), to peoples' beliefs being based on rough summaries (Tran, Vul, & Pashler, 2017), or complex probability distributions (Griffiths & Tenenbaum, 2006).

The extent of theoretical dispute is partly due to the difficulty of directly accessing peoples' beliefs. While experts with statistical training might be able to express their beliefs in statistical or computational formalisms, these ways of sharing knowledge might be biased and coarse. Also, the beliefs of statistical non-experts might not be readily communicated or accessible.

Belief Elicitation Methods

Given the importance of accessing peoples' beliefs and the challenges faced in eliciting them, fields ranging from applied statistics to psychology have devised methods to maximize efficiency and standardize the results of belief elicitation methods. Most prominent elicitation methods, such as SHELF (Gosling, 2018), query experts for statistics of the distributions. For example, an expert might be asked to give the limits, mean, quartiles, or a combination of these measures (O'Hagan et al., 2006). These procedures can sometimes also be administered visually, for example, by asking participants to assign chips proportionally to the frequency of histogram bins or by displaying the best-fitting distribution corresponding to an elicited statistic to the participant (Jones & Johnson, 2014; for a recent review of elicitation methods, see Mikkola et al., 2021).

While these techniques have been studied extensively in applied statistics, it is less clear how useful they are when used with non-experts. In addition, these methods can usually only produce rough, low-dimensional outlines, and the elicitation can be biased to particular distributions. Furthermore, in setting up prior elicitation tasks, experimenters face many degrees of freedom, with potentially very different results (Stefan, Evans, & Wagenmakers, 2020), and there is no satisfactory theoretical framework for assessing how these degrees of freedom influence the task (Mikkola et al., 2021).

From a psychological perspective, there is strong evidence that people's estimates can be biased. Most relevant to belief elicitation, participant judgments can be biased towards previous values (the anchoring bias, Tversky & Kahneman, 1974). Furthermore, even experts are prone to exhibit an overconfidence bias, producing overly narrow estimates (McKenzie, Liersch, & Yaniv, 2008). Finally, it is unclear if people can accurately learn distributions at all. In experiments by Tran et al. (2017), participants were not able to reproduce the bimodal distribution over previously encountered objects (but also consider Griffiths & Tenenbaum, 2006; Sanborn & Beierholm, 2016, arguing that people can reproduce distributional knowledge).

Random Generation as a Belief Elicitation Method?

Here we suggest a novel method to capture people's beliefs — asking participants to produce a sequence of random instances of the domain. This approach is deceptively simple: instead of carefully constructing choice situations to debias participants, our method asks people to quickly come up with random instances. This simplicity makes random generation

In J. Culbertson, A. Perfors, H. Rabagliati & V. Ramenzoni (Eds.), *Proceedings of the 44th Annual Conference of the Cognitive Science Society.* ©2022 The Author(s). This work is licensed under a Creative Commons Attribution 4.0 International License (CC BY).

an attractive experimental method. Furthermore, people can produce many random values in a short amount of time, allowing the uncovering of distributional properties of people's beliefs. Finally, since no numerical information is given to participants, the task may be less susceptible to anchoring and other biases.

People's systematic deviations from randomness when explicitly instructed to produce random sequences have been studied since pioneering work by Baddeley (1966). While the focus of this line of research has been on characterizing deviations from pure randomness, mostly studying random sequences of integers or letters (for a review, see Nickerson, 2002), some early results suggested that the random sequences people produce reflect features of the domain. For example, when studying how people produce random sequences for a small set of alternatives, people tend to produce uniform distributions (Teraoka, 1963), but not when the alternative set is large (Rath, 1966; Bakan, 1960). Interestingly, these results also showed that people produced the number one disproportionately often, which matches environmental frequencies of small digits (Benford's law, Rath, 1966). The idea that people use distributional knowledge of the domain to access random samples when tasked to produce random sequences was only recently more directly studied. Motivated by the theoretical idea of human inference as sample-based (for a recent review, see Chater et al., 2020), results by Castillo, León-Villagrá, Chater, and Sanborn (2021) showed that manipulating distributional features of the domain in which participants were tasked to generate sequences resulted in manipulation-dependent signatures in their sequences. Furthermore, recent work suggests that random generation can be performed in varied domains, and sequences obtained in these experiments exhibit domain-specific signatures (Castillo, León-Villagrá, Chater, & Sanborn, 2022).

Here, we evaluated if random generation tasks could predict participants' beliefs about environmental statistics, and how these predictions compared to predictions based on decile elicitation. We hypothesized that random generation would generally be on-par with decile elicitation, but be preferable when the target domain is not a simple parametric distribution. We also evaluated potential carry-over between the tasks. For example, it is plausible that performing the random generation task would highlight outliers, subsequently resulting in broader decile distributions.

Experiment

Participants

The study was certified following departmental ethics guidelines. We recruited 60 participants, ($M_{age} = 41.08$, $SD_{age} = 12.39$, 28 female, 30 male, 2 non-binary) on Amazon Mechanical Turk, following pre-registered criteria[1]. Participants had to have an MTurk Masters qualification, be located in the United States, and have more than 100 approved HITs with an approval rate of 95% or higher. Twenty-four additional participants were excluded following one or several pre-registered criteria[2]. We also excluded an additional participant who only provided one unique value in the conditional estimation task and predicted all baking times to be 90 minutes long. Given that this prediction pattern represented an extreme outlier (the median number of unique values was 5, $M = 4.27$, $SD = 0.94$) and the normalized root mean squared deviation is not defined for zero-variance responses, we decided to exclude this participant. Participants were paid a flat fee of $3. The experiment took about 10 minutes to complete ($M = 9.31$, $SD = 2.80$).

Materials

As more than 15 years have passed since the publication of Griffiths and Tenenbaum (2006), we collected up-to-date data from similar sources, see Table 1.

Table 1: Data sources used to establish environmental statistics. For cake baking times, we parsed all cake recipes in the dessert category. For movie lengths, we excluded series and other non-movies and excluded movies with runtimes longer than 10 hours. All sources were accessed in October 2021.

Data set	Source	Points
Lifespans	2018 Life Tables (Arias & Xu, 2020)	
Movie Gross	www.worldwideboxoffice.com	5460
Movie Length	www.imdb.com	372655
NFL	www.pro-football-reference.com	33836
Pharaohs	www.metmuseum.org	182
Cakes	www.allrecipes.com	3849

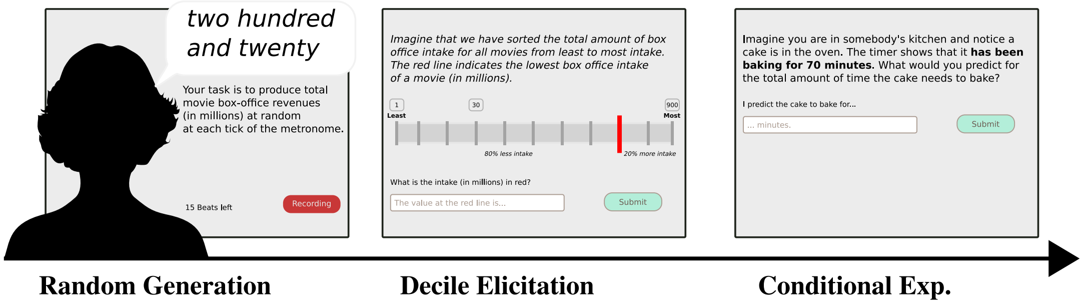
Procedure

To ensure that participants had a functioning audio setup and were comfortable saying numbers aloud at the pace of the metronome (30 ticks per minute), immediately after the first set of instructions, participants completed a 15-second practice trial. The practice trial followed the same structure as the random generation block, but participants were prompted to produce numbers between 0 and 10 for 15 seconds. Once participants completed the practice, they could listen to their recording, and if they judged the audio quality to be insufficient, re-record the block. Then, participants were shown the experiment instructions and had to pass a comprehension check. Finally, participants proceeded to either the random generation (RG) or decile elicitation (DE) task. For an overview of the procedure, see Figure 1.

Random Generation Block (RG): Participants first received brief instructions about the domain for which they

[1] For the full preregistration, including exclusion criteria and planned analysis, see https://osf.io/a6g3r

[2] 11 did not produce any numbers in the RG block, 4 produced invalid CE values (≥ 2 initial submissions lower than the observed value), 7 produced invalid DE submissions (the submitted value was not greater than or equal to the previous decile), and 9 did not record the familiarization task.



Your task is to produce total movie box-office revenues (in millions) at random at each tick of the metronome.

Imagine that we have sorted the total amount of box office intake for all movies from least to most intake. The red line indicates the lowest box office intake of a movie (in millions).

What is the intake (in millions) in red?

Imagine you are in somebody's kitchen and notice a cake is in the oven. The timer shows that it **has been baking for 70 minutes.** What would you predict for the total amount of time the cake needs to bake?

I predict the cake to bake for...

Figure 1: The experiment consisted of three main blocks. In the random generation block, participants had to say values aloud at random. In the decile elicitation block, participants had to judge the value corresponding to a proportion of the domain. In the conditional expectation task, participants were tasked to predict the total final value given an observation.

had to sample values, e.g., "Your task is to produce the total amount of box office revenues (in millions) at random at each tick of the metronome". Next, they were instructed to imagine that they had written all items on sheets of paper, put them in a hat, shuffled, and retrieved them at random, one at a time. Then they proceeded to the two-minute random generation task, where audio was recorded, and the domain instructions were shown again. A metronome played a ticking sound at 30 ticks per minute to ensure that participants kept pace throughout the task.

Decile Elicitation Block (DE): Participants had to provide values for the deciles at 11 points (0, 10%, ..., 100%), following a prompt, e.g., "Imagine that we have sorted the total amount of box office intake for all movies from least to most intake. The red line indicates the box office intake of a movie (in millions). $\{p\}$% of all box office intakes were lower than the movie intake in red. $\{100 - p\}$% were higher than the movie intake in red.", similar to Achtypi, Ashby, Brown, Walasek, and Yechiam (2021).

The range of sorted values was presented as a horizontal bar, with marks at the 11 points and the current query highlighted. Participants typed their responses into a textbox, and once they submitted their choice, the value appeared above the corresponding mark. Since the two extreme points (0, 100%) could influence subsequent submissions, we presented these first (starting with 0). The remaining 9 points were presented in random order. Participants could correct previous submissions at any point of the task by clicking on them and entering a new number.

Conditional Expectation (CE): After the RG and DE blocks, participants proceeded to the conditional expectation (CE) task. They were presented with 15 questions in which they had to predict a value, given an observation. These questions corresponded to three domains with five observed values each (shown in random order). For domains used by Griffiths and Tenenbaum (2006), we used the same questions.

Following Tauber, Navarro, Perfors, and Steyvers (2017), we did not allow participants to submit values smaller than the observed value and instead prompted participants to resubmit a valid value. Participants received the same domain as in the RG and DE tasks and two additional domains. Domains were shown interlaced, with the RG and DE domain always being shown third in the sequence.

Survey: Finally, participants completed a short survey about their expertise for the domains in the CE block (on a scale from 0 [no knowledge] to 6 [expert]), as well as their age and gender, which they could decline to answer.

Results

Decile Elicitation

Participants only rarely corrected their estimates ($M = 11.63$, $SD = 1.58$ submissions for 10 deciles). To allow us to derive conditional expectations corresponding to the DE task, we first fitted three target distributions (Normal, Gamma, and Pareto) to participants' data by minimizing mean-squared error (MSE) on the deciles, similar to the fitting procedure in Gosling (2018).

Participants were best fit by the Normal distribution (73%) or the Gamma distribution (27%), and no participant was best-fit by the Pareto distribution. This split across Normal and Gamma distributions was fairly consistent across domains, with NFL (100%), lifespans (90%), movie gross (70%), and baking times (70%) all strongly favoring Normality. Only for movie lengths (60% Normal) and pharaohs (50% Normal), participants were more frequently fit by a Gamma distribution. For individual DE data and the corresponding environmental data, see Figure 2.

Random Generation

Overall, participants produced numbers close to the targeted 60 numbers ($M = 59.08$, $SD = 1.15$). We again fitted the three candidate distributions for direct comparison to the DE task. In contrast to the DE data, we now fitted distributions directly to the samples via maximum-likelihood estimation. In contrast to the DE task, we find most responses were fit better by the more flexible Gamma distribution (82%) over the Normal distribution (18%).

This preference was consistent across conditions, with baking times, movie gross, pharaohs (100%), and movie lengths (80%) being better fit by Gamma distributions, and NFL

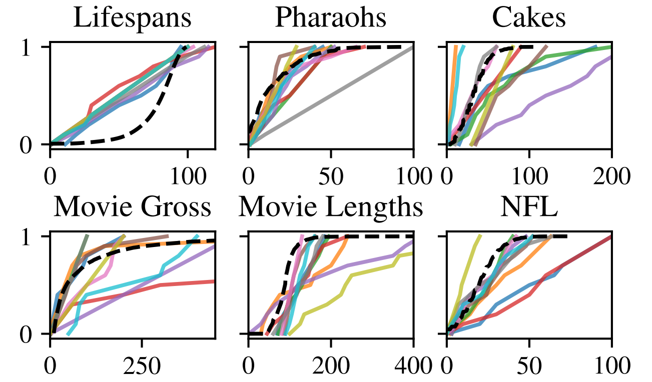


Figure 2: Per-participant cumulative density functions (colored lines) corresponding to the DE task along with environmental data (dashed line). Lifespans and Pharaoh reigns are in years, cake baking times and movie lengths in minutes, movie gross earnings in millions of USD.

(60% Normal) and lifespans (50% Normal). For the empirical cumulative distribution function (ECDF) corresponding to participants' RG data, see Figure 3.

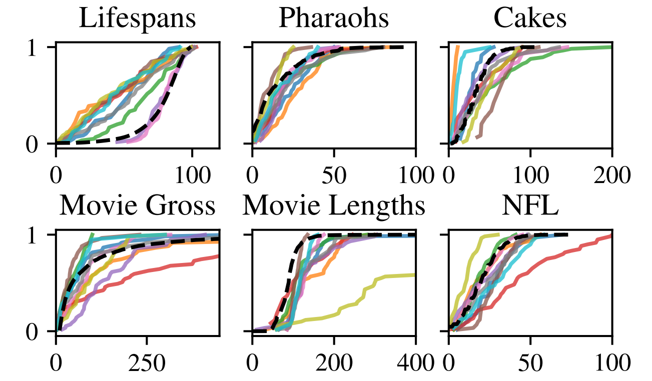


Figure 3: Per-participant cumulative density functions (colored lines) corresponding to the RG task along with environmental data (dashed line).

Matching Environmental Statistics

Next, we assessed how closely both tasks could reproduce the environmental data. One deviation from environmental statistics was immediately apparent – participants' DE and RG data for lifespans. These results were surprising since we had expected participants to have well-established beliefs about lifespans. One possible explanation for this consistent mismatch could be that participants misinterpreted the prompt, producing random sequences and deciles matching the age distribution of living people instead.

We calculated the maximum absolute difference between the environmental data and either the DE task or the ECDF of the RG task (Kolmogorov–Smirnov or KS distances). To

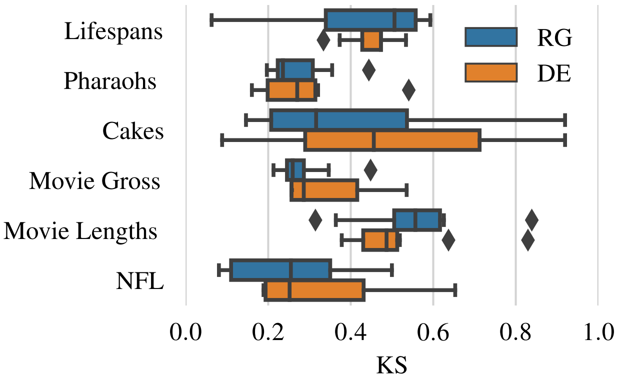


Figure 4: KS distance between environmental distributions and elicitation data.

ensure that all cumulative density functions (CDF) were defined on the same domains, we used a linear interpolation of the ground-truth ECDF and limited the evaluation points for RG to the 11 deciles, as presented in the DE task. While KS distances were slightly lower for the RG task compared to DE ($M_{DE} = 0.40$, $SD_{DE} = 0.18$; $M_{RG} = 0.36$, $SD_{RG} = 0.19$), KS distances varied considerably across domains, see Figure 4. We did not find a significant effect of task on KS distances when fitting a linear mixed-effects model, $KS \sim task$, with varying intercepts for domain and participants ($\beta = 0.02$, $SE = 0.01$, $t(59) = 1.75$, $p = .085$).

As an exploratory analysis, we compared the variance of the environmental data to that of participants' RG and DE data. If the RG or DE variance is smaller than the environmental variance, this suggests that people recall a small set of environmental values to perform the tasks, similar to the model suggested by Mozer et al. (2008). If participants instead reproduce environmental statistics, as suggested by Griffiths and Tenenbaum (2006), we would expect RG and DE variance to match the environmental variance. Finally, if RG or DE variance is larger than the environmental variance, this suggests that participants use a more complex and potentially adaptive (Feldman, 2013) statistical inference. To reproduce variances from DE data, we sampled a large (10000) random sample via inverse sampling. We then subtracted the task variance (RG or DE) from the environmental variance. Since domains considerably differed in their variance, we standardized the variances by dividing by the domain-specific SD. A linear mixed-effects model with sum-contrast codes for task type (RG or DE) and random intercepts per participant and domain found a small negative effect of task type ($\beta = -0.21$, $SE = 0.07$, $t(5) = -3.00$, $p = .039$), with marginal means for DE suggesting a larger difference between task and environmental variance ($M = -0.85$, $SE = 0.28$) than for RG ($M = -0.42$, $SE = 0.28$). Overall, there was a negative, albeit insignificant, intercept, suggesting that task variance was at least not smaller than empirical variance ($\beta = -0.63$, $SE = 0.27$, $t(5) = -2.35$, $p = .066$).

Testing Signatures of NFL Knowledge in RG: One advantage of RG over alternative prior elicitation methods is that RG can potentially investigate idiosyncratic distributions. This is especially important when distributions exhibit "gaps" of low density or even impossible values. Alternative elicitation methods would require the experimenter to focus on these regions a-priori and ensure that methods exhibit appropriate resolution in the low-density areas. In practice, this can be problematic since prior elicitation is usually adopted when detailed knowledge of the distribution is not available. In contrast, eliciting knowledge via RG can, in principle, reveal these characteristic gaps, as knowledgeable participants would likely deem them salient and memorable features of their acquired knowledge. To test this hypothesis, we evaluated if participants in the NFL condition reproduced characteristics of the distribution over NFL scores and how their self-reported expertise affected this relationship. Due to the scoring rules in the NFL, some scores are improbable, or have not occurred in professional practice at all, see Figure 5.

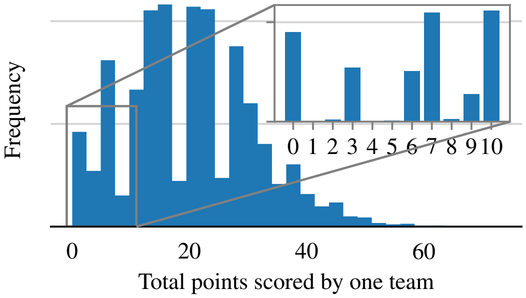


Figure 5: Points scored by one team in the history of the NFL. Due to its complex scoring rules, some final scores are extremely infrequent, or have not occurred at all.

First, we tested how strongly participants' RG distributions correlated with the environmental distribution. We correlated individual participants' frequencies with environmental frequencies for low team scores (0-10 points scored), and found that on average, participant frequencies correlated significantly with environmental frequencies ($M = 0.43$, $SD = 0.29$, $t(9) = 4.7$, $p < .001$). In addition to establishing that participants' frequencies were positively related to environmental statistics, we were also interested in the relationship between participants' self-reported expertise and how closely the shape of their RG distribution over low NFL scores resembled the true distribution. We calculated the Wasserstein, or earth mover's distance, for frequencies of scores in the range of 0-10 for each participant. The subset of participants in our study who performed the RG and DE tasks for the NFL domain did not consider themselves knowledgeable for NFL scores ($M = 1.7$, $SD = 1.57$, $min = 0$, $max = 4$, out of a maximum score of 6). Nevertheless, we found a small negative relationship between self-reported expertise and similarity of the RG distributions to the low-range environmental scores when regressing Wasserstein distances with centered expertise ratings ($\beta = -2.08$, $SE = 0.90$, $t(8) = -2.325$, $p = .049$).

Conditional Expectation

On average, participants' CE data matched ground-truth distributions well, see Figure 6, though we are mainly interested in how well RG and DE predict these judgments as a measure of their internal validity.

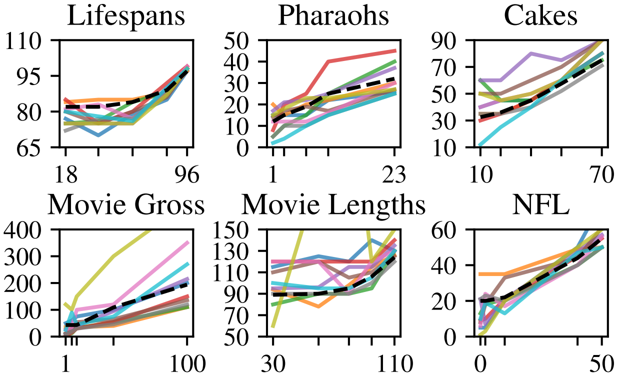


Figure 6: Participants' conditional expectations (y-axis) given an observation (x-axis) in the CE task and conditional expectations derived from environmental data (dashed lines).

To assess how well the distributions elicited in RG and DE tasks matched participants' CE judgments, we followed previous work by sampling from the conditional distribution elicited in the DE or RG task and predicting the median of the obtained distribution (Griffiths & Tenenbaum, 2006; Mozer et al., 2008; Tauber et al., 2017). We then used the normalized root mean-squared deviation (NRMSE; Mozer et al., 2008) to assess how well this prediction fit the CE task.

To obtain samples from the RG task, we sampled with replacement from the empirical distribution and dropped values smaller than the observed value in the CE task. To sample from the DE task, we used two approximations of the distribution implied by participants' DE data. First, we used the best-fitting distributions and sampled from the conditional distribution. Alternatively, we sampled using inverse transform sampling, which allowed us to map uniform samples to the max entropy distribution implied by DE via the CDF.

Overall, both RG and DE data were similar in predicting participants' CE behavior. Thirty-two participants were best fit by RG (52%), whereas 28 were best fit by DE (15 via best-fit distributions, 13 via max entropy, 25%, and 22%, respectively).

We then assessed which of the two tasks produced lower errors by fitting a linear mixed-effects model with a main effect of task and random intercepts for participants and domains. We found a significant effect of task on NRMSE ($\beta = 0.18$, $t(59) = 2.06$, $SE = 0.09$, $p = .044$), with marginal means for RG ($M = 1.17$, $SE = 0.2$) lower than DE ($M = 1.53$, $SE = 0.2$), see Figure 7. Finally, we also assessed how the aggregation of RG data performed in predicting participants'

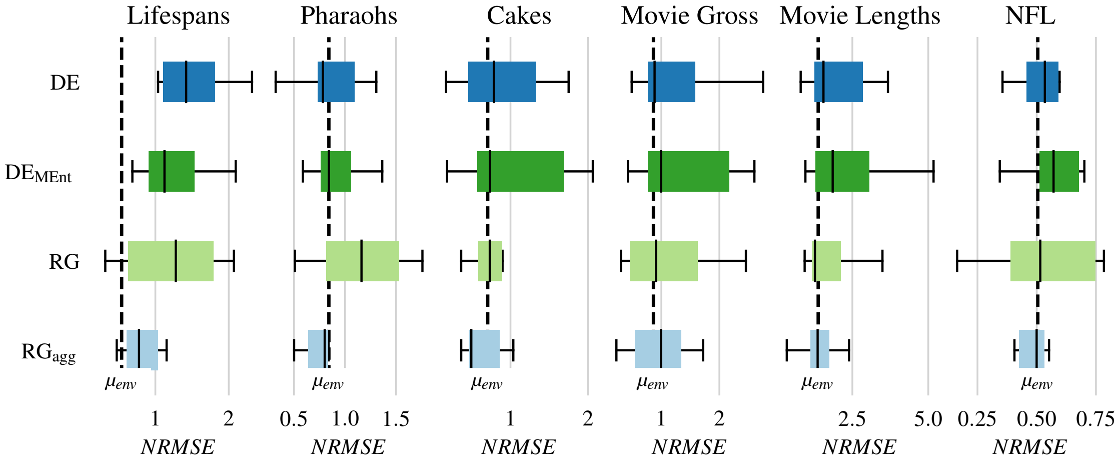


Figure 7: Predicting participants' CE from their RG data, the distribution corresponding to DE (or max-entropy interpolation, DE$_{\text{MEnt}}$). Both DE and RG were competitive with predictions based on the full set of environmental data (dashed line). Overall, RG produced lower NRMSE than DE and this fit could be further improved by aggregating across participants (RG$_{\text{agg}}$).

CEs. A linear mixed-effects model contrasting RG and aggregations of RG on NRMSE, with random intercepts for domain and participant, found a significant effect ($\beta = 0.14$, $t(59) = 3.01$, $p = .004$), with marginal means for aggregation significantly lower ($M = 0.89$, $SE = 0.14$) than non-aggregated RG ($M = 1.17$, $SE = 0.14$). These results show that pooling small groups of RG data can produce consistent elicitation of participants' implicit knowledge and suggests that longer generation tasks may produce even better results.

Order Effects and Distribution Spread

To assess if the order of tasks affected the standardized range of the DE distributions, we fitted a linear mixed-effects model with a random intercept per domain. In contrast to our hypothesis that there would be carryover effects, we found no significant effect of task order on DE ranges ($\beta = 0.67$, $t(53) = 0.03$, $SE = 24.76$, $p = .98$) and ranges for orders in which DE was performed first ($M = 154$, $SD = 63.1$) were almost identical to RG first ($M = 153$, $SD = 63.1$).

Discussion

Our work investigated whether random generation tasks can be used to elicit people's implicit beliefs about the distribution of everyday quantities. When we assessed the internal validity of random generation by predicting participants' judgments in a conditional expectation task, we found that overall, random generation could capture participants' judgments as well as decile elicitation. Furthermore, these predictions could be significantly improved by aggregating small groups of participants. Next, we found that random generation could reproduce environmental data as accurately as decile elicitation. We then explored the ability of random generation tasks to reproduce fine details of the environmental distributions — information that is challenging to uncover with alternative elicitation methods. Our results suggest that people are generally aware of the shapes of environmental distributions, even those fine details embodied in the distribution of NFL final scores, and people's ability to reproduce these details increased with expertise. In addition, we found that the elicited variances were as large or larger than those of the environmental distributions, suggesting that people infer distributions of values, rather than perfectly remembering a small set of experienced values (Spicer, Sanborn, & Beierholm, 2020).

While these results suggest that random generation is on-par with decile elicitation in assessing peoples' beliefs, our results also highlight the complementary strengths of the approaches. For example, our results suggest that decile elicitation can bias participants towards linear interpolation of deciles which could explain the better fit of Gaussian distributions across domains. In contrast, the random generation task exhibited more skewed distributions, suggesting that participants produced more flexible, or possibly noisy, distributions. In addition to the advantages highlighted here, random generation has an additional advantage as an elicitation method — it is not limited to numerical domains. Thus, it could be used more widely to elicit people's beliefs for domains that can be easily verbalized, and we are currently exploring this possibility.

However, we do not believe that a method will be universally preferable or that random generation does not exhibit potential task-specific biases. Instead, we take these results to suggest that prior elicitation should ideally use several complementary methodologies, and random generation offers attractive features such as producing data at a high pace and not requiring assumptions about the underlying distribution (Mikkola et al., 2021). Future work should establish the performance of combinations of elicitation methods and assess random generation tasks in the context of modern alternatives (Sanborn, Griffiths, & Shiffrin, 2010).

Acknowledgments

We thank Tom Griffiths and Sean Tauber for kindly sharing their experimental data and the three anonymous reviewers for their helpful comments. This work was supported by the European Research Council (ERC; 817492-SAMPLING).

References

Achtypi, E., Ashby, N. J., Brown, G. D., Walasek, L., & Yechiam, E. (2021). The endowment effect and beliefs about the market. *Decision*, *8*(1), 16.

Arias, E., & Xu, J. (2020). National vital statistics reports. *National Vital Statistics Reports*, *69*(12).

Baddeley, A. D. (1966). The Capacity for Generating Information by Randomization. *Quarterly Journal of Experimental Psychology*, *18*(2), 119–129.

Bakan, P. (1960). Response-tendencies in attempts to generate random binary series. *The American journal of Psychology*, *73*(1), 127–131.

Castillo, L., León-Villagrá, P., Chater, N., & Sanborn, A. (2021). Local Sampling with Momentum Accounts for Human Random Sequence Generation. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 43).

Castillo, L., León-Villagrá, P., Chater, N., & Sanborn, A. (2022). Random generation as local sampling with momentum. *Manuscript in preparation*.

Chater, N., Zhu, J.-Q., Spicer, J., Sundh, J., León Villagrá, P., & Sanborn, A. (2020). Probabilistic Biases Meet the Bayesian Brain. *Current Directions in Psychological Science*, *29*(5), 506–512.

Feldman, J. (2013). Tuning your priors to the world. *Topics in Cognitive Science*, *5*(1), 13–34.

Gosling, J. P. (2018). SHELF: the Sheffield elicitation framework. In *Elicitation* (pp. 61–93). Springer.

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

Jones, G., & Johnson, W. O. (2014). Prior elicitation: Interactive spreadsheet graphics with sliders can be fun, and informative. *The American Statistician*, *68*(1), 42–51.

McKenzie, C. R., Liersch, M. J., & Yaniv, I. (2008). Overconfidence in interval estimates: What does expertise buy you? *Organizational Behavior and Human Decision Processes*, *107*(2), 179–191.

Mikkola, P., Martin, O. A., Chandramouli, S., Hartmann, M., Pla, O. A., Thomas, O., … others (2021). Prior knowledge elicitation: The past, present, and future. *arXiv preprint arXiv:2112.01380*.

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.

Nickerson, R. S. (2002). The production and perception of randomness. *Psychological Review*, *109*(2), 330.

O'Hagan, A., Buck, C. E., Daneshkhah, A., Eiser, J. R., Garthwaite, P. H., Jenkinson, D. J., … Rakow, T. (2006). *Uncertain Judgements: Eliciting Experts' Probabilities*. John Wiley & Sons.

Rath, G. J. (1966). Randomization by humans. *The American Journal of Psychology*, *79*(1), 97–103.

Sanborn, A. N., & Beierholm, U. R. (2016). Fast and accurate learning when making discrete numerical estimates. *PLoS Computational Biology*, *12*(4), e1004859.

Sanborn, A. N., Griffiths, T. L., & Shiffrin, R. M. (2010). Uncovering mental representations with Markov chain Monte Carlo. *Cognitive Psychology*, *60*(2), 63–106.

Spicer, J., Sanborn, A. N., & Beierholm, U. R. (2020). Using Occam's razor and Bayesian modelling to compare discrete and continuous representations in numerosity judgements. *Cognitive Psychology*, *122*, 101309.

Stefan, A. M., Evans, N. J., & Wagenmakers, E.-J. (2020). Practical challenges and methodological flexibility in prior elicitation. *Psychological Methods*.

Tauber, S., Navarro, D. J., Perfors, A., & Steyvers, M. (2017). Bayesian models of cognition revisited: Setting optimality aside and letting data drive psychological theory. *Psychological Review*, *124*(4), 410.

Teraoka, T. (1963). Some serial properties of "subjective randomness". *Japanese Psychological Research*, *5*(3), 120–128.

Tran, R., Vul, E., & Pashler, H. (2017). How effective is incidental learning of the shape of probability distributions? *Royal Society Open Science*, *4*(8), 170270.

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