Bayesian sampling in the interpretation of quantifiers? Evidence from Markov chain Monte Carlo with People

Fabian Dablander[1]

[1] University of Tübingen

Author Note

This research was conducted during an internship supervised by Dr. phil Michael Franke at the University of Tübingen. Most of the ideas presented here are his. I want to thank him for his creative input and patience.

Correspondence concerning this article should be addressed to Fabian Dablander. E-mail: dablander.fabian@gmail.com

Abstract

How do people use quantifiers such as *some* and *many*? Within the growing discipline of probabilistic pragmatics, we take first steps in answering this question by comparing three Bayesian models that estimate participants' subjective probability distribution of ten different quantifiers. We use an experimental design in which participants' responses are viewed as states in a Markov chain Monte Carlo algorithm—"Markov chain Monte Carlo with People" (Sanborn & Griffiths, 2007). Fourty-five participants were presented with fifty forced-choice trials for four different, randomly chosen quantifiers in which two images were presented which displayed different numbers of red dots. Participants had to indicate which image was a better description for the specific quantifier in use. Three models of the data generating process were developed. Interestingly, the model which closely mirrors the experimental design and assumes participants are Bayesian samplers had higher prediction error than a model which assumes participants soft-max prefer the number which is closer to the mode of the subjective probability distribution. Limitations of our preliminary finding as well as implications for further research are discussed.

Keywords: Bayesian modeling, experimental pragmatics, quantifier interpretation, Markov chain Monte Carlo with People

Word count: 2554

Bayesian sampling in the interpretation of quantifiers? Evidence from Markov chain Monte Carlo with People

How do people interpret quantifiers? Say I arrive hungry at a party where I gleefully discover twelve cookies on the kitchen table. The host, recognising my desire, says that I can have *some*. Can I eat two? Sure. What about ten? Probably not.

Why does two seem more likely than ten? Following the data-oriented approach of probabilistic pragmatics (Franke & Jäger, 2016), we tackle this question by estimating the subjective probability distribution over possible values (in the above example: cookies) for different quantifiers. To arrive at this distribution, we utilize an experimental design which uses participants' answers as states in a Markov chain Monte Carlo (MCMC) algorithm—"Markov chain Monte Carlo with People" (Sanborn & Griffiths, 2007). Additionally, we develope three Bayesian models of the data generating process, and directly test the hypothesis of a "Bayesian brain" (Sanborn & Chater, 2016) in quantifier use. Let's unpack the underlying ideas in turn.

Bayesian inference. Bayesian approaches are becoming increasingly popular as (a) an alternative to classical statistical data analysis (Wagenmakers et al., 2015), (b) a tool to estimate (hierarchical) cognitive models (Lee, 2011), and (c) a theory about how the brain works (Sanborn & Chater, 2016). The latter idea is known as the "Bayesian brain", a theme that offers to unify different aspects of cognition (Chater, Oaksford, Hahn, & Heit, 2010); the brain is viewed as a Bayesian inference machine which approximates complex probability distributions.

The most common objection to this idea is the observation that human cognition seems far from *optimal* or *rational* (Marcus, 2009; Marcus & Davis, 2013); instead, it seems more plausible to assume that our actions are guided by evolved satisficing strategies that lead to common reasoning errors such as the conjunction fallacy (e.g., Tversky & Kahneman, 1983). In a recent paper, however, Sanborn & Chater (2016) explain these satisficing strategies and the resulting reasoning errors by viewing the brain not as an ideal Bayesian reasoner, but as

Bayesian sampler. Only asymptotically do we follow the rules of probability exactly and act according to rational choice; with finite time and, thus, finite samples, reasoning errors result.

In this paper, we seek to directly test the notion that participants are Bayesian samplers in the domain of quantifier interpretation. We utilize an experimental design—"Markov chain Monte Carlo with People" (Sanborn & Griffiths, 2007)—which is inspired by computational challenges of Bayesian inference, and which allows us to sample from the subjective probability distribution of quantifiers.

Bayesian inference starts with a prior belief $p(\theta)$ over some parameter vector θ. The likelihood function $\mathcal{L}(\theta|\mathbf{y})$ specifies how the parameter vector relates to the observed data $\mathbf{y}$. Combining the two using Bayes' rule two yields the posterior distribution over the parameter vector, $p(\theta|\mathbf{y})$:

$$p(\theta|\mathbf{y}) = \frac{p(\theta)\mathcal{L}(\theta|\mathbf{y})}{\int p(\theta)\mathcal{L}(\theta|\mathbf{y})\mathrm{d}\theta}$$

In most cases, the denominator is a high dimensional integral which cannot be calculated analytically. To this end, sampling-based methods such as Markov chain Monte Carlo techniques are used which avoid computing the integral altogether (for an introduction, see e.g., Ravenzwaaij, Cassey, & Brown, 2016). Concretely, one sets up a Markov chain which has as its stationary distribution the normalized posterior distribution $p(\theta|\mathbf{y})$. Various algorithms have been proposed, but the canonical one is the Metropolis-Hastings (MH) method (Hastings, 1970; Metropolis, Rosenbluth, Rosenbluth, Teller, & Teller, 1953).

Metropolis-Hastings. The key insight here is that we do not need to have access to the posterior distribution but need only be able to compute its density, $\pi(x)$, for values of x. The MH algorithm works as follows (see also the pseudocode below). Start at random initial state x. In each step, generate a new sample $x^\star$ based on the current value using a proposal function. Choose $x^\star$ over x using a specified acceptance function. Required some mathematical conditions (see e.g., Jackman, 2009, p. 201), this constitutes a Markov chain of first order which has $p(\theta|\mathbf{y})$ as its stationary distribution. Once the stationary distribution

is reached, the generated samples are equivalent to samples from $p(\theta|\mathbf{y})$. The samples before that time in point are discarded as "burn-in".

If we further assume that the probability of proposing a new state $x^\star$ based on the current state x is the same as the probability of proposing x based on $x^\star$, i.e., the proposal distribution is symmetric, we can use the Barker acceptance function (Barker, 1965)

$$A(x; x^\star) = \frac{\pi(x^\star)}{\pi(x^\star) + \pi(x)}$$

where $\pi(x)$ indicates the density of the value x under the posterior distribution.

Algorithm 1 Barker Random-Walk Metropolis-Hasting

```
procedure METROPOLIS(σ, N)
    samples[1] ← RandomState()
    samples[2:N] ← 0
    for i = 2 to N do
        x ← samples[i - 1]
        x* ← x + Normal(μ = 0, σ = σ)
        A ← π(x*) / (π(x*)+π(x))
        if A ≥ Uniform(0, 1) then
            samples[i] ← x*
        else
            samples[i] ← x
        end if
    end for
    return samples
end procedure
```

In our experimental design we use participants' responses as states in a Markov chain, with the stationary distribution being the subjective probability distribution over values provided a specific quantifier. We then estimate three models of the data generating process. One of them, referred to as the Barker model, describes the experimental design. If participants are Bayesian samplers, this model should describe the data best.

The document was written in a reproducible manner using Rmarkdown and the papaja R package (Aust & Barth, 2016). All materials, including data, code for the experiment and data analysis, and manuscript, can be found at https://github.com/fdabl/MCMCP.

Methods

We report how we determined our sample size, all data exclusions, all manipulations, and all measures in the study. In a pilot study, five participants were recruited in order to test the technical setup of the experiment. Those participants are not included in the final analysis.

Participants

Fifty participants were recruited from Amazon Mechanical Turk (Buhrmester, Kwang, & Gosling, 2011). After inspection of the data, five participants were removed[1], leaving a total of fourty-five participants. On average, the experiment took 9.38 minutes to complete. The experiment seemed to be engaging and not too difficult (Mean engagement = 7.35, Mean difficulty = 3.73; scale: 1-10). Fourty-two Participants were self-reported native English speakers, three were of different mother tongue (French, Romanian, Russian).

Material

Images of random dot patterns were used. Each image showed 432 dots, of which any amount could be coloured red. The other dots were coloured black. For each possible pattern (0 - 432 red dots), ten images were generated. On each trial, images were randomly chosen out of the generated pool of images.

Procedure

After reading the instruction, each participant went through four blocks of 50 trials each. Each block consisted of a quantifier randomly chosen out of *Half*, *About half*, *Less than half*, *Few*, *Very few*, *Many*, *Most*, *The majority*, *Some*, *Almost all*. No participant saw a quantifier twice, and *About half* was never followed by *Half* and *Very few* never by *Few* (and vice versa). On each trial, the participant had to choose which of the two images fitted the

[1]See appendix for the rationale and the excluded participants' data patterns.

description best (see Figure 1 for an example trial). On the following trial, new images were generated.

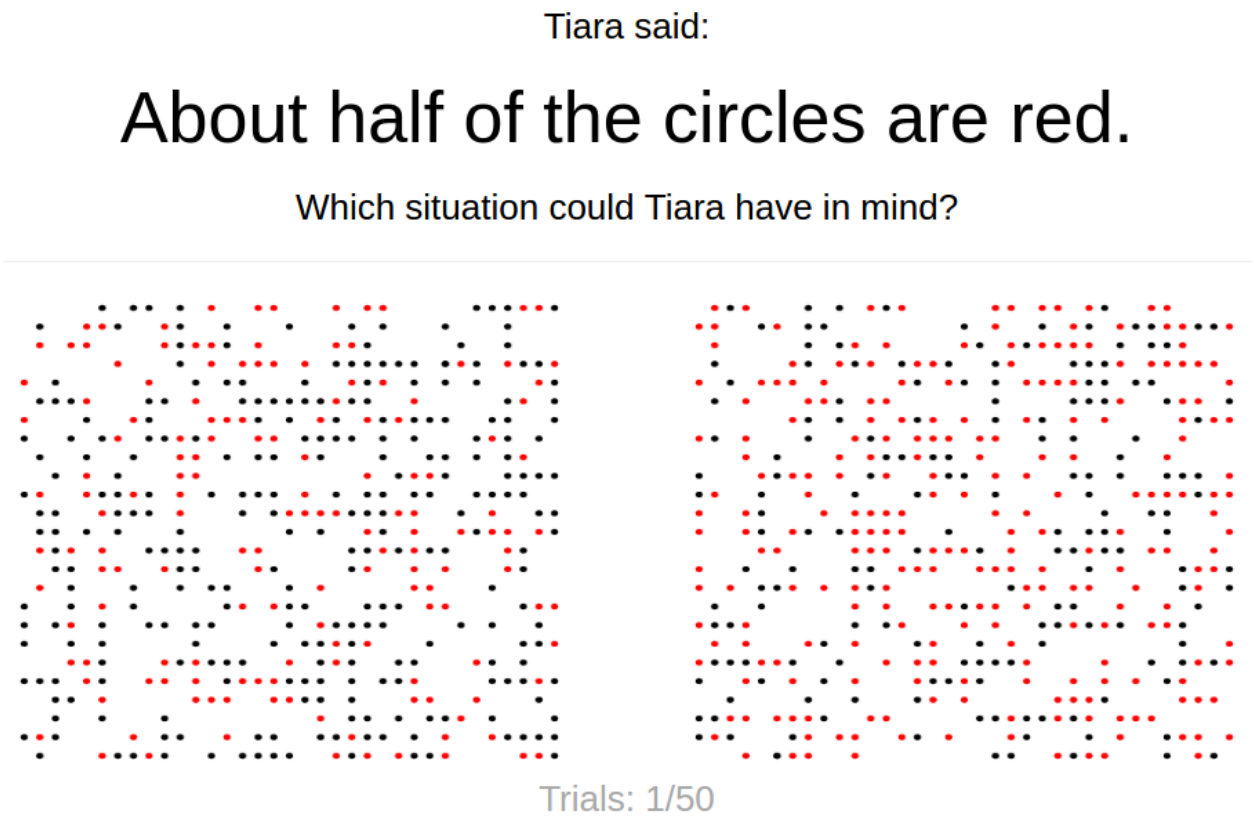

Figure 1. Example Trial. Participants had to choose between the left or right image. Names in the description were chosen randomly out of a pool of fifty names.

Analogously to the steps of a MH algorithm, the number of red dots in the images on the first trial of each block were generated randomly. On subsequent trials, samples based on the current number of red dots were proposed. Note that the support of the distribution we want to elicit is bounded by the interval $[0, 432]$. Therefore, to ensure a symmetric proposal function, our MH algorithm proceeded as follows. Given the chosen number of red dots x, uniformly generate points for the second image within the interval $[x - \delta, x + \delta]$ with probability $1 - \epsilon$. With probability ϵ, points were generated uniformly outside the interval. Again, symmetry is crucial to enable the use of the Barker acceptance function. We set $\delta = 20$ and $\epsilon = .4$.

As an example, assume the participant chose the image with 420 red dots. Out of a

hundred cases, in sixty cases the newly selected image will have points from the set $\{400, \ldots, 432, 0, \ldots, 8\}$. In fourty cases, it will have points selected from the set $\{8, \ldots, 400\}$.

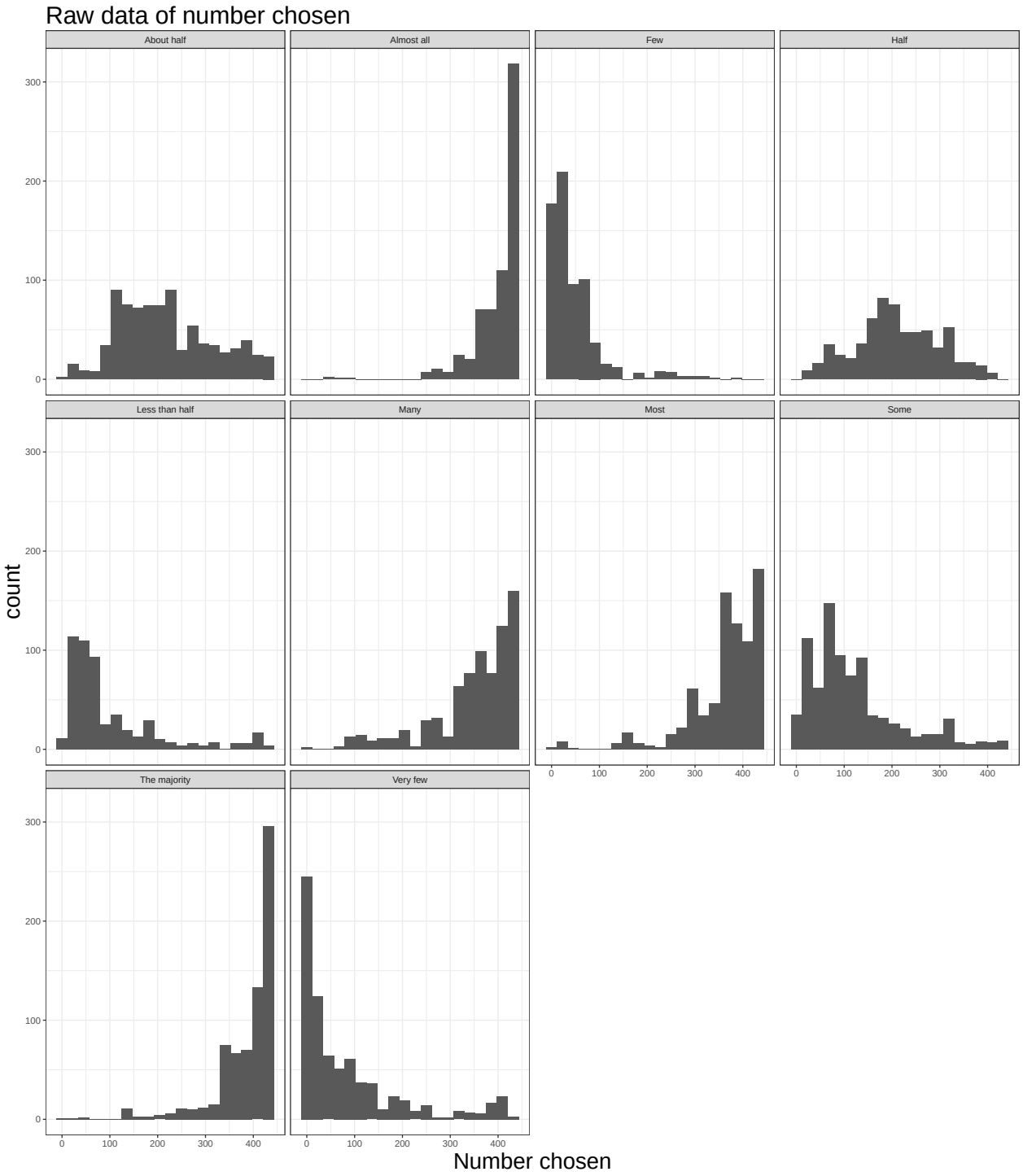

Figure 2. Histograms of the raw data (number chosen) for all quantifiers pooled over participants.

Data analysis

All analyses were completed using the open-source statistical programming language R (R Core Team, 2016). We removed five participants whose response pattern was highly unusual (see appendix). The data we want to explain is the choice the participants make in each trial: do they pick the image with the higher number of red dots? Figure 2 shows a visualization of pooled participants' data for all quantifiers. See Figure 3 for a visualization of participant-wise choice data for the quantifier *Some*. The full dataset can be interactively explored at https://fdabl.shinyapps.io/MCMCP/.

Model specification

We developed three models of the data generating process. Common to all is the parameterization of the distribution over number of red dots for each quantifier as a beta-binomial distribution, and the Bernoulli likelihood function.

Two of the models, thereafter *Closer model* and *Distance model*, assume the participant is soft-max preferring the image in which the number of red dots is closer to the mode of the subjective distribution for that quantifier. The Distance model uses information about the distance of the choices to the mode, while the Closer model uses a categorical measure of distance (closer or not closer). The models are similar to the one discussed by Franke et al. (2016) for the *bin comparison* task.

The third model, which we call the *Barker model*, does not use the mode but instead compares the likelihood of the respective number of red dots—just like the Barker acceptance function in the MCMC algorithm. See Figure 4 for the graphical model specifications using the notation of Lee & Wagenmakers (2014).

Model inference

We used JAGS (Plummer, 2003) to estimate the model parameters. 100.000 samples were obtained from two chains with a thinning rate of 2 after burn-in of 5000 that ensured

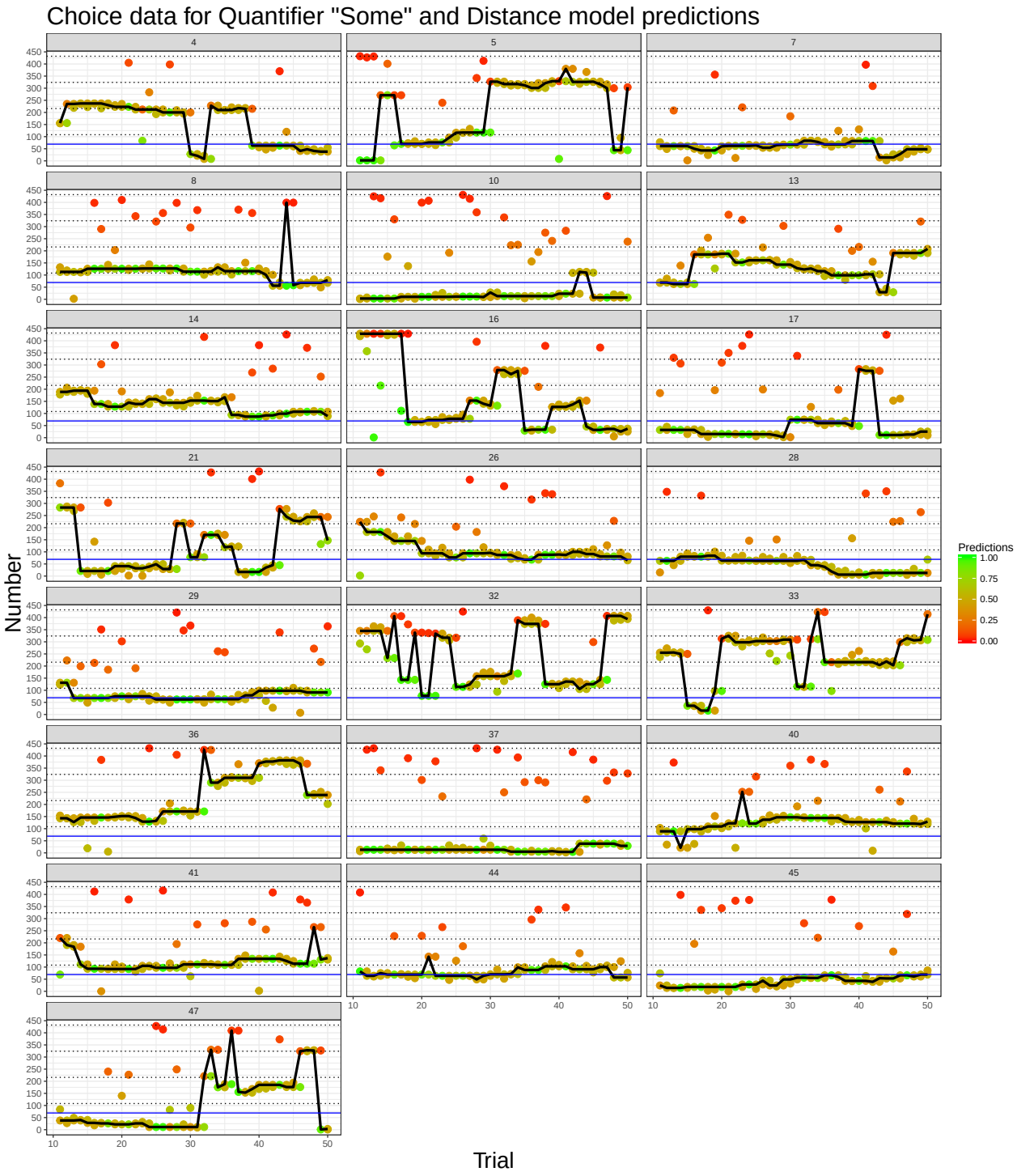

Figure 3. Shows the raw data of all participants who completed a block with the quantifier *Some*. Colour indicates the predictions of the Distance model for each respective data point. Blue line indicates the mode of the subjective distribution. Note that the first ten trials were discarded as burn-in.

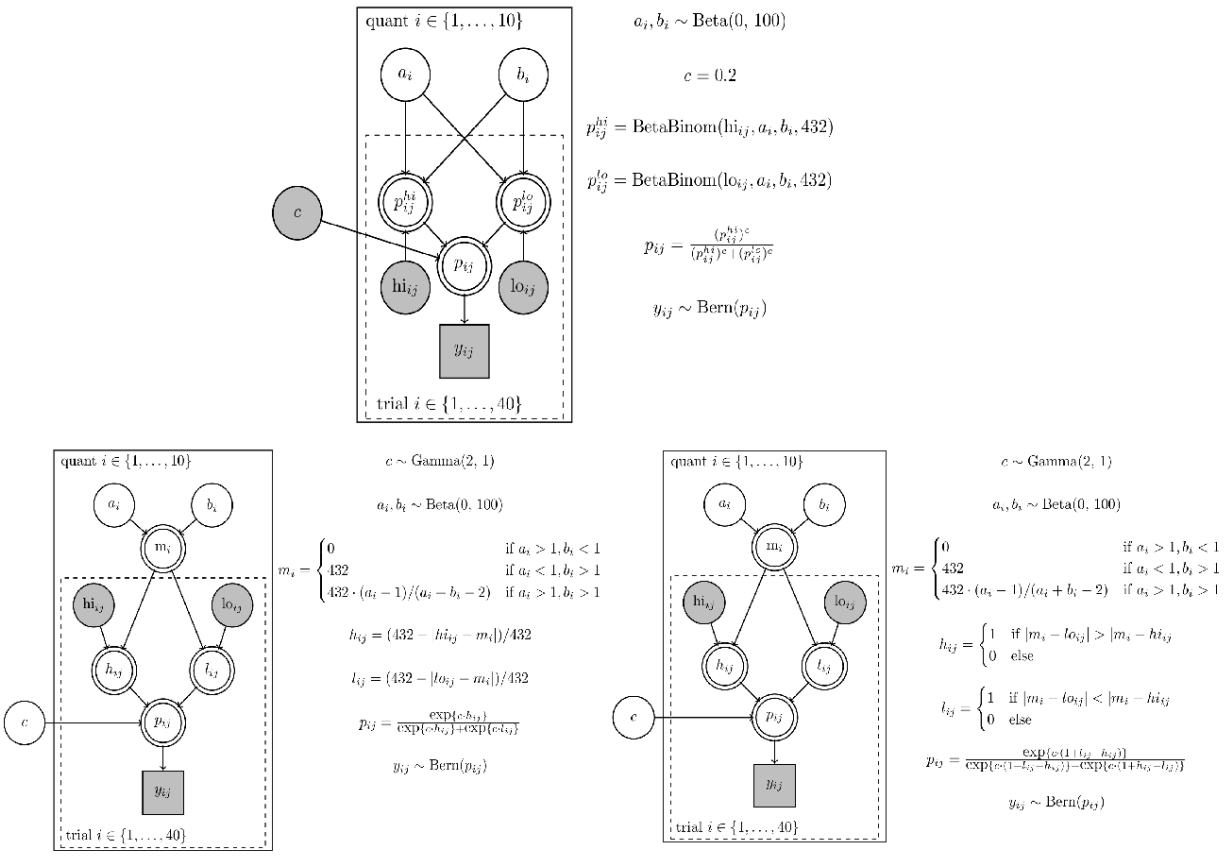

Figure 4. Graphical model specification for the Barker model (top), the Distance model (left), and the Closer model (right). Transparent nodes indicate parameters, shaded ones indicate observed values. Circles indicate continuous, rectangles categorical values. Nodes with double lines indicate deterministic nodes.

convergence according to $\hat{R}$ (Gelman & Rubin, 1992) for the Distance and Closer model. Even after increasing the samples to 200.000, the Barker model did not converge. This was because the parameters were underspecified; different values for a, b, and c resulted in the same likelihood. Therefore, we set the parameter $c = 1$ which resulted in convergence[2].

Results and Discussion

We compared the models using the likelihood as well as the Deviance Information Criterion (DIC; Spiegelhalter, Best, Carlin, & Van Der Linde, 2002), the latter being an

[2]Changing the parameter to c = .2 or c = .5 did not alter the main conclusion drawn. However, the model fitted best with $c = .2$; we report conclusions based on this fit.

Table 1
Results of model comparison.

	Distance	Barker	Closer
DIC	7848.373	8194.429	8654.730
Likelihood	4963.402	4690.150	4669.204

estimate for out of sample prediction error. The Distance model had lower prediction error than the Closer model (see Table 1). This is not surprising, because the Distance model uses information about the numerical distance of the choices to the mode, while the Closer model only cares about which choice is closer. Interestingly, the Distance model fares better than the Barker model. Note again that the Barker model closely mirrors the experimental design.

The underspecification of parameters with c as a free parameter in the Barker model required us to specify contraints. We did this by fixing c; c influences how strongly participants prefer higher values. For c approaching zero, participants have no preference, i.e. there choice whether to prefer the image with the higher number of red dots or the image with the lower number of red dots was random; for c greater than one, participants prefer higher values. The prediction error as measured by the DIC decreased with a decreasing c, indicating that the Barker model did not capture relevant regularities in the data.

Although we want to avoid drawing strong conclusions from these preliminary results, it seems that, in contrast to what the "Bayesian brain" hypothesis postulates, participants do not engage in sampling-based algorithmic behaviour in the domain of quantifier interpretation. Instead, it seems that participants infer likely values given the quantifier based on the distance to the mode of the subjective distribution over all values under that quantifier.

There are a number of limitations that need to be addressed. With respect to the experimental design, it is unclear how many trials are needed for the Markov chain to converge, and for the resulting samples to be draws from the subjective probability distribution. In our analysis, we excluded the first ten trials as burn-in, only working with

the resulting fourty. Other choices might be equally, or more reasonable. This point is exaggerated by the observation that participant's choices are—by design—not independent; there is serial autocorrelation which violates our assumption of an independent Bernoulli likelihood. Along the same lines, it is unclear whether our parameter settings $\delta = 20$ and $\epsilon = .4$ for the proposal function which generated new images are adequate, i.e. whether this results in a good exploration of the state space.

However, these issues are analogous to the issues in Markov chain Monte Carlo based inference more broadly, an area where remedies have already been developed; future research should utilize approaches from this domain. For example, to assess convergence, one could repeatedly present the participants with blocks of the same quantifier, i.e. run more than one Markov chain, and compute statistics such as $\hat{R}$ (Gelman & Rubin, 1992).

In this paper, we assumed that participants share the same probability distribution over the number of red dots for each quantifier; this simplifying assumption need not be reasonable. Future research should utilize a hierarchical approach similar to Franke et al. (2016), estimating individual-level probability distribution as variations of a shared population-level belief.

Despite the limitations, we believe that this paper constitutes a novel contribution by extending the use of Markov chain Monte Carlo with People type experiments to the domain of language interpretation. Moreover, it casts initial doubt on the idea of the brain as a Bayesian sampler in quantifier interpretation. Avenues for future research abound.

References

Aust, F., & Barth, M. (2016). *papaja: Create APA manuscripts with RMarkdown*. Retrieved from https://github.com/crsh/papaja

Barker, A. (1965). Monte Carlo calculations of the radial distribution functions for a proton? electron plasma. *Australian Journal of Physics*, *18*(2), 119–134.

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk a new

source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, *6*(1), 3–5. doi:10.1177/1745691610393980

Chater, N., Oaksford, M., Hahn, U., & Heit, E. (2010). Bayesian models of cognition. *Wiley Interdisciplinary Reviews: Cognitive Science*, *1*(6), 811–823. doi:10.1002/wcs.79

Franke, M., & Jäger, G. (2016). Probabilistic pragmatics, or why Bayes' rule is probably important for pragmatics. *Zeitschrift Für Sprachwissenschaft*, *35*(1), 3–44. doi:10.1515/zfs-2016-0002

Franke, M., Dablander, F., Schöller, A., Bennett, E., Degen, J., Henry-Tessler, M., ... Goodman, N. D. (2016). What does the crowd believe? A hierarchical approach to estimating subjective beliefs from empirical data. In Papafragou A., Grodner D., Mirman D., & T. J. C. (Eds.), *Proceedings of the 38th Annual Conference of the Cognitive Science Society* (pp. 2669–2674).

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, 457–472.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, *57*(1), 97–109.

Jackman, S. (2009). *Bayesian analysis for the social sciences* (Vol. 846). John Wiley & Sons.

Lee, M. D. (2011). How cognitive modeling can benefit from hierarchical Bayesian models. *Journal of Mathematical Psychology*, *55*, 1–7. doi:10.1016/j.jmp.2010.08.013

Lee, M. D., & Wagenmakers, E. (2014). *Bayesian cognitive modeling: A practical course*. Cambridge University Press.

Marcus, G. (2009). *Kluge: The haphazard evolution of the human mind*. Houghton Mifflin Harcourt.

Marcus, G., & Davis, E. (2013). How robust are probabilistic models of higher-level cognition? *Psychological Science*, *24*(12), 2351–2360. doi:10.1177/0956797613495418

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical*

Physics, *21*(6), 1087–1092.

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In *Proceedings of the 3rd international workshop on distributed statistical computing* (Vol. 124, p. 125).

R Core Team. (2016). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Ravenzwaaij, D., Cassey, P., & Brown, S. D. (2016). A simple introduction to Markov Chain Monte–Carlo sampling. *Psychonomic Bulletin & Review*, 1–12. doi:10.3758/s13423-016-1015-8

Sanborn, A. N., & Chater, N. (2016). Bayesian brains without probabilities. *Trends in Cognitive Sciences*, *20*(12), 883–893. doi:10.1016/j.tics.2016.10.003

Sanborn, A. N., & Griffiths, T. L. (2007). Markov chain Monte Carlo with People. In *Advances in neural information processing systems* (pp. 1265–1272).

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *64*(4), 583–639.

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, *90*(4), 293–315.

Wagenmakers, E., Verhagen, A., Ly, A., Matzke, D., Steingroever, H., Rouder, J., & Morey, R. (2015). The need for Bayesian hypothesis testing in psychological science. In Papafragou A., Grodner D., Mirman D., & T. J. C. (Eds.), *Psychological science under scrutiny: Recent challenges and proposed solutions*. John Wiley & Sons.

Appendix

The figure below shows the raw choice data for the participants that have been excluded. An online app where the full dataset can be interactively explored is hosted at https://fdabl.shinyapps.io/MCMCP/. All materials are available at https://github.com/fdabl/MCMCP.

The participant on the top left was excluded due to a bias towards small values, while three other participants were excluded due to a bias towards large values. The fifth participant who was excluded did not seem to understand the experiment, unsystematically alternating between images with a higher and lower number of red dots.

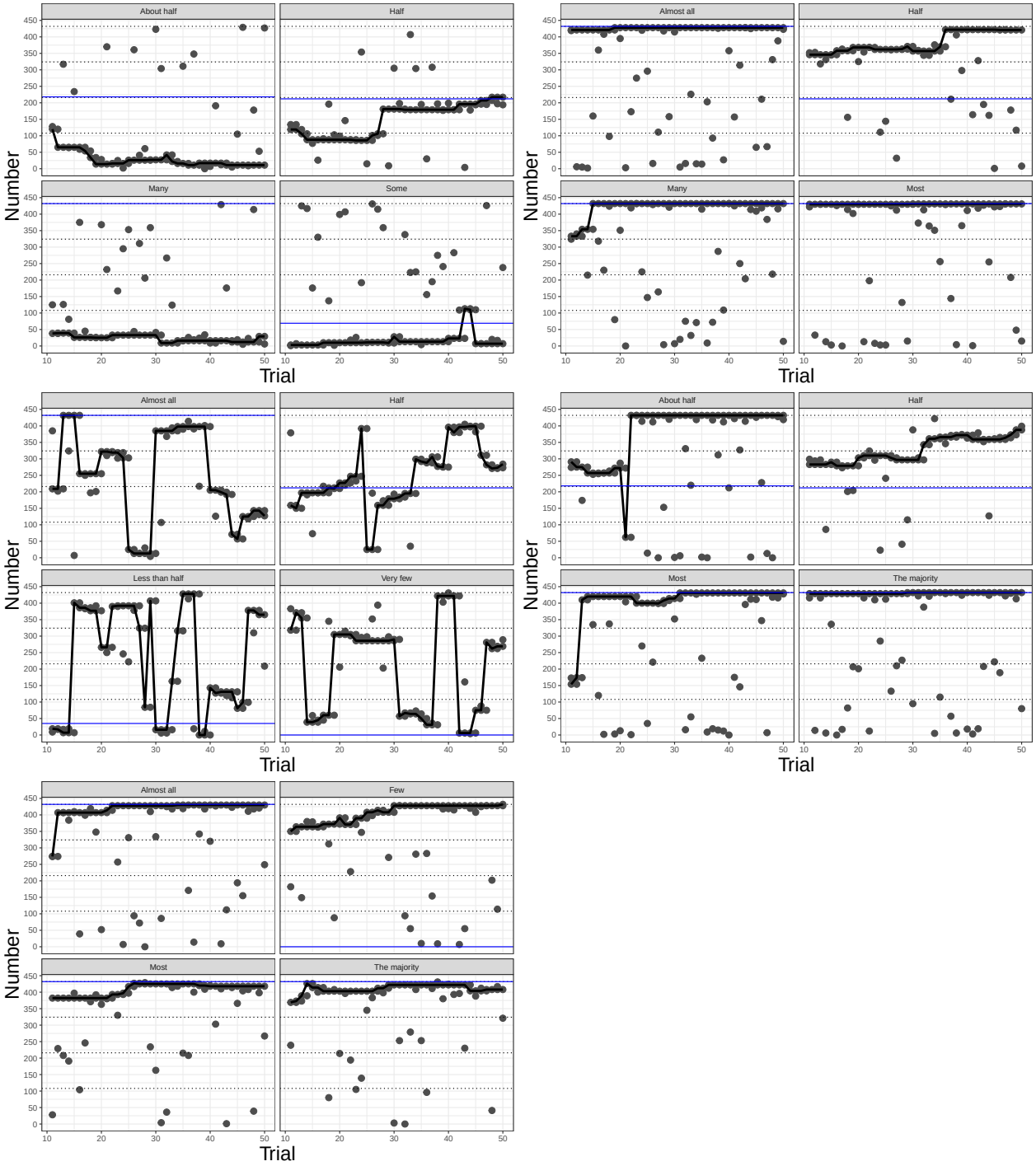

Figure A1. Raw data for the excluded participants. Blue line indicates the mode of the subjective probability distribution for that specific quantifier (estimated without excluded participants).