What does the crowd believe? A hierarchical approach to estimating subjective beliefs from empirical data

Michael Franke, Fabian Dablander, Anthea Schöller
{michael.franke, fabian.dablander, anthea.schoeller}@uni-tuebingen.de
Department of Linguistics, Wilhelmstraße 19
72074 Tübingen, Germany

Erin Bennett, Judith Degen, Michael Henry Tessler, Justine Kao, Noah D. Goodman
{ebennett, jdegen, mtessler, justinek, ngoodman}@stanford.edu
Department of Psychology, Stanford University
450 Serra Mall, Stanford, CA 94305 USA

Abstract

People's beliefs about everyday events are both of theoretical interest in their own right and an important ingredient in model building—especially in Bayesian cognitive models of phenomena such as logical reasoning, future predictions, and language use. Here, we explore several recently used methods for measuring subjective beliefs about unidimensional contiguous properties, such as the likely price of a new watch. As a first step towards a way of assessing and comparing belief elicitation methods, we use hierarchical Bayesian modeling for inferring likely population-level beliefs as the central tendency of participants' individual-level beliefs. Three different dependent measures are considered: (i) slider ratings of (relative) likelihood of intervals of values, (ii) a give-a-number task, and (iii) choice of the more likely of two intervals of values. Our results suggest that using averaged normalized slider ratings for binned quantities is a practical and fairly good approximator of inferred population-level beliefs.

Keywords: subjective beliefs, hierarchical modeling, Bayesian data analysis, Bayesian cognitive models

Motivation

When trying to understand observed behavior we readily ascribe beliefs and desires to fellow agents. This happens intuitively, in folk psychology, but also in science. Scientific ascription of latent mental states plays an important role in many explanations of higher-order cognition: decision making, reasoning, language use, etc. It is therefore vital to have methods for validating explanatory mental state ascriptions.

A family of models where this is particularly pressing are Bayesian models of cognition which seek to explain task behavior in a variety of domains as partially informed by what participants believe about mundane events—their *prior beliefs*. Take interpretation of language. "That watch cost a million dollars," tends to be interpreted as hyperbole: conveying affect rather than literal truth. Empirical data on whether similar statements are understood as hyperbole can be explained well by a Bayesian model of utterance interpretation (Kao, Wu, Bergen, & Goodman, 2014). This model assigns a crucial role to an empirical measure of participants' expectations about the likely or normal price of a watch—only when the uttered price is sufficiently unlikely *a priori* will hyperbolic interpretation be possible. Other examples of domains in which empirically successful models have included a measure of participants' prior beliefs include making predictions about everyday events (Griffiths & Tenenbaum, 2006), referential reasoning (Frank & Goodman, 2012), strength of pragmatic enrichments (Degen, Tessler, & Goodman, 2015), or quantifier interpretation (Schöller & Franke, 2015).

Many methods have been used to build prior distributions used in Bayesian cognitive models. One method of getting at subjective beliefs is to take actual frequencies as an approximation to subjective beliefs (e.g. Griffiths & Tenenbaum, 2006). Unfortunately, frequency data may be unavailable (e.g., one-shot events) or deviate from participants' subjective beliefs in crucial respects.

Another common approach is to empirically measure subjective beliefs by *give-a-number* tasks. The simplest version would be this: being told that John just bought a new watch, participants are asked for a single numerical estimate of its price. This task is easy to comprehend and implement, but a single number does not provide much information about the subjective belief that it is a manifestation of. One solution is to infer which parameterized distribution best explains the observed number choices, either at the individual level (Manski, 2004) or at the population level (Tauber & Steyvers, 2013). More sophisticated *give-a-number* tasks give more information, but can be difficult to implement and analyze.

More complex elicitation methods include *scoring rules* (Savage, 1971; Andersen, Fountain, Harrison, & Rutström, 2014; Schlag, Tremewan, & van der Weele, online first), prominent in economics, and the *iterated learning paradigm* (Lewandowsky, Griffiths, & Kalish, 2009). These methods are powerful but difficult to implement, and often assume very specific behavior from participants—such as optimal decision making under perfect knowledge of the payoff scheme.

In sum, there is a tradeoff between simplicity of paradigms and their information content. Ideally, we would like to have an experimental method for measuring subjective beliefs that (i) provides sufficient and reliable-enough information about subjective beliefs to derive testable predictions from cognitive models that rely on such information, (ii) is easy to understand by participants, (iii) is easy to implement, and that (iv) does not require sophisticated means of data analysis. Moreover, the ideal method would be (v) flexible enough to allow inferred subjective belief distributions beyond standard pa-

Item	Bins ({min, max} step; units)	Context sentence	GAN question	BH frame	PC frame
coffee	{<44, >200} 2; degrees	X has just fetched himself a cup of coffee from the office vending machine.	What do you think the temperature of his coffee is?	His coffee was the following temperatures	The temperature of his coffee is N degrees.
commute	{0, >98} 7; minutes	X commuted to work yesterday.	How many minutes do you think she spent commuting yesterday?	She commuted for the following numbers of minutes yesterday	She spent N minutes commuting.
joke	{0, 14} 1; children	X told a joke to 14 kids.	How many of the kids do you think laughed?	The following number of kids laughed	N of the children laughed.
laptop	{0, >7500} 500; dollars	X bought a laptop.	How much do you think it cost?	The laptop cost the following numbers of dollars	The laptop cost $N.
marbles	{0, 14} 1; marbles	X threw 14 marbles into a pool.	How many of the marbles do you think sank?	The following number of marbles sank	N of the marbles sank.
movies	{0, > 210} 16; minutes	X just went to the movies to see a blockbuster.	How many minutes long do you think the movie was?	The movie was the following numbers of minutes long	The movie was N minutes long.
TV	{0, >43} 3; hours	X watched TV last week.	How many hours do you think he spent watching TV last week?	He watched TV for the following numbers of hours last week	He spent N hours watching TV.
watch	{0, >750} 50; dollars	X bought a watch.	How much do you think it cost?	It cost the following numbers of dollars	The watch cost $N.

Table 1: Experimental items. *X* was a randomly generated name (different on each trial). *N* was one of the bins.

rameterized distributions and (vi) would provide a good approximation of the central tendency or average of subjective beliefs in a given population. The latter would allow using one set of participants for measuring prior beliefs and another for whatever other task is of interest, so as to avoid potential cross-over effects.

A simple technique that seems to fit this bill is the *binned histogram* task that has recently been used with apparent success (e.g. Kao, Wu, et al., 2014; Kao, Bergen, & Goodman, 2014; Tessler, 2015; Schöller & Franke, 2015). Participants adjust sliders to express (relative) subjective beliefs about how likely it is that the value of an uncertain contiguous quantity lies in some interval of possible values (a bin). E.g., to report beliefs about a watch's price, participants adjust sliders whose endpoints are labelled "extremely likely" and "impossible." Each slider corresponds to one of 15 bins that partition the range of plausible values (established by a pre-test), such as "$0-$50", "$50-$100", etc. up to "$700-$750" and "more than $750." Each participant's slider ratings are normalized and the results averaged across participants. The resulting *mean slider ratings* look like plausible population-level beliefs and give good results when fed into cognitive models that predict task behavior based on prior expectations.

Explanatory success aside, the question remains whether mean slider ratings measure what we would like them to. Are these good approximations of a central tendency of participants' individual beliefs? Are these measures consistent with participants' behavior in other tasks, such as *give-a-number*? To scrutinize the binned histogram task we address these issues by collecting data in a within-subject paradigm from three task types: (i) *binned histograms*, (ii) *give-a-number* and (iii) *paired comparisons*. The latter asks participants for a direct comparison of two bins from the *binned histograms* task, and was included as a further consistency check. We used a Bayesian hierarchical model to analyze this data jointly. The model infers latent subjective beliefs, where each subjective belief is, intuitively, a noisy perturbation of a population-level belief. We find that mean *a posteriori* population-level beliefs are approximated well by mean slider ratings from the binned-histograms task, suggesting that this may indeed be a sound and easy measure of what the crowd believes. Moreover, the model is able to capture participants' behavior in all three task types reasonably well, suggesting that what mean slider ratings measure is what we think it is: a population-level central tendency of latent subjective beliefs.

Experimental elicitation of prior beliefs

Participants, procedure and materials We recruited 20 self-reported English native speakers over Mechanical Turk and collected responses for eight different items (listed in Table 1) using the three different dependent measures mentioned above. Each participant rated each item using each dependent measure. Trial order was randomized.

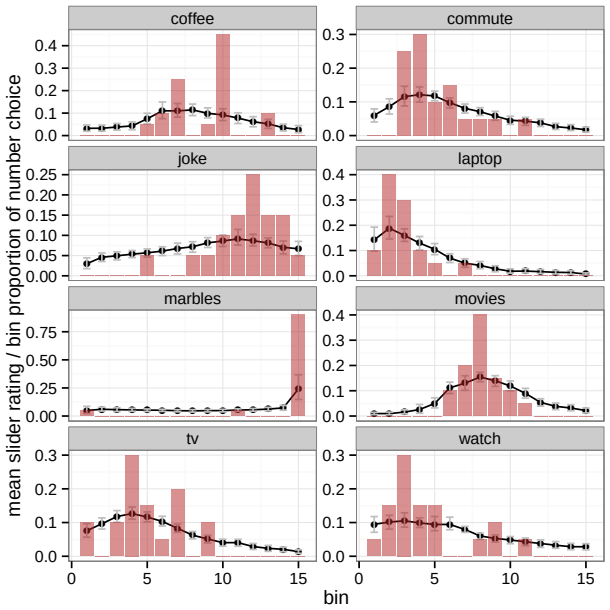
On *give-a-number* (GAN) trials, participants saw the context sentence and GAN question for each item (see Table 1) and provided one number by adjusting a slider with endpoints labeled as the lowest and highest number for that item. Min and max numbers are shown in Table 1 and were taken from previous studies that had used these items (Degen et al., 2015;

Schöller & Franke, 2015; Kao, Wu, et al., 2014). The selected number appeared above the slider so participants knew exactly what the value of the slider would be.

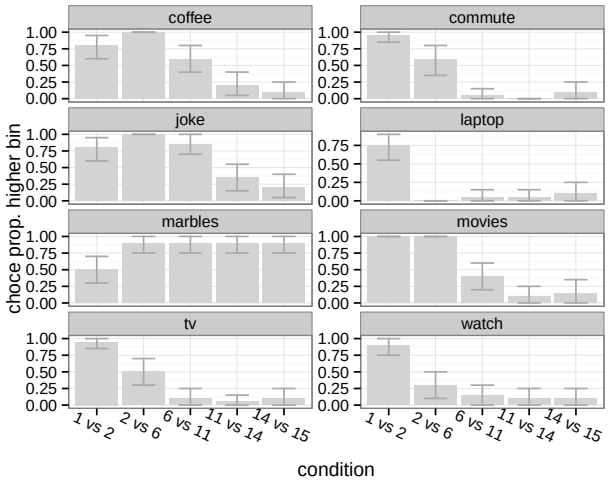
On *binned histogram* (BH) trials, participants saw an item's context sentence and were asked to *Please rate how likely it is that Y* (where *Y* came from the corresponding BH frame in Table 1). They adjusted 15 continuous sliders, one per bin, with five points labeled *impossible*, *not very likely*, *neutral*, *very likely*, and *extremely likely*. Bins were determined by dividing the interval spanned by the item's minimum and maximum value into equally sized bins (Table 1).

On *paired comparison* (PC) trials, participants saw an item's context sentence and had to click on one of two options shown side by side – whichever one they thought was more likely. Each option used the appropriate PC frame in Table 1 and numbers were filled in by comparing the following bins: 1 vs 2, 2 vs 6, 6 vs 11, 11 vs 14, and 14 vs 15. Paired comparison trials always occurred as a block of five trials, with order of comparisons randomized within block.

Results Fig. 1 shows mean slider ratings alongside the frequencies of number choices in the *give-a-number* task. It looks like the latter could be samples from the former, with a tendency to modal choices. But some *give-a-number* results seem influenced by a tendency towards round or salient numbers. E.g., in item *coffee* more than 40% of participants guessed that a coffee from the vending machine was ca. 150°F (bin 10). The results from the *paired comparison* seem consistent with the mean slider ratings as well, albeit not as straightforwardly as one might think. This is clearest from the *marbles* case. While the mean slider ratings suggest that bins 2, 6 and 11 were considered roughly equiprobable on average, almost every participant chose the higher bin in direct bin comparisons 2 vs 6 and 6 vs 11. A potential explanation for this, which we will implement formally in the data-generating model, is that the *paired comparison* condition invited participants to think about what they thought was mostly likely the case (usually: bin 15 in the *marbles* case) and to choose the bin that is closest to that.



(a) Red: proportion of bins corresponding to number choices on *give-a-number* trials. Black: mean slider ratings on *binned histogram* trials.



(b) Proportion of higher bin choices on *paired comparison* trials.

Figure 1: Average data. Bars are bootstrapped 95% CIs.

Model

The data we would like to explain are: (i) the normalized slider ratings $s_{ijk} \in [0;1]$ of participant $i \in \{1,\ldots,20\}$ for item $j \in \{1,\ldots,8\}$ and bin $k \in \{1,\ldots,15\}$ from the binned histogram task; (ii) the bins $n_{ij} \in \{1,\ldots,15\}$ in which participant i's number choice for item j was in the give-a-number task; and (iii) the binary choices $c_{ijl} \in \{0,1\}$ of whether participant i selected the higher bin for item j in the paired comparison task for each comparison l. There are two simplifications in need of commenting. In (i), we focus on slider ratings after normalizing for each participant, because we assume that slider adjustments reflect relative, not absolute estimates of subjective beliefs. In (ii), we focus on bin choices, not actual number choices, in order to avoid, as much as possible, considerations of salience of particular numbers, and also because otherwise data from items with smaller domains of plausible numbers would get more weight than data from items with a wider range of number choices. (Future work should investigate the relation to an explicit *give-a-bin* task.)

All three pieces of data are to be explained as functions of subjective beliefs P_{ij}, with P_{ijk} being participant i's belief about the relative likelihood of bin k for item j. Each P_{ij} defines a likelihood for our data, via appropriate link functions. Variance in subjective beliefs is harnessed by a population-level hyper-prior with central tendency Q_j. The structure of this model is pictured in Fig. 2.

To fill the structure in Fig. 2 with life, we need to spell out three parameterized link functions, one for each task type,

and the relation between population-level belief Q_j and individual beliefs P_{ij}. Let's start with the latter. The idea is that P_{ij} are noise-perturbed variants scattered around Q_j, with some parameter w to determine how much perturbation we should expect. To realize this, the model assumes that P_{ij} are distributed according to a Dirichlet distribution with weights given by wQ_j:

$$Q_j \sim \text{Dirichlet}(1,\ldots,1) \qquad w \sim \text{Gamma}(2,0.1)$$
$$P_{ij} \sim \text{Dirichlet}(wQ_j)$$

The higher w, the more likely it is that P_{ij} is "close" to Q_j.

The link function for the **slider rating data** uses a logit transformation to project observed slider ratings s_{ijk} and latent probabilities P_{ijk}, which are bound to lie between 0 and 1, to the reals. The likelihood of logit-transformed observation s_{ijk} is given by a Gaussian with standard deviation σ around the logit-transformed predictor P_{ijk}. On top of that, there is a parameter κ, the steepness of the logit transform of P_{ijk}, that allows response likelihoods to capture end-point affinity for $\kappa > 1$ (values of P_{ijk} close to 0 or 1 are likely mapped to 0 or 1) or end-point aversion for $\kappa < 0$ (values of P_{ijk} are likely to be realized as more median), with a prior that expects $\kappa = 1$.

$$\text{logit}(s_{ijk}) \sim \text{Norm}(\text{logit}(P_{ijk},\kappa),\sigma)$$
$$\sigma \sim \text{Gamma}(0.0001,0.0001) \quad \kappa \sim \text{Gamma}(5,5)$$

The link function for **number choice data** treats each bin n_{ij} as a draw from a categorical distribution where the probability of bin k is proportional to $\exp(aP_{ij})$, i.e., a soft-max choice from P_{ij}. The higher parameter a, the more likely n_{ij} is the mode of P_{ij}. For $a \to 0$, all bins become equiprobable.

$$n_{ij} \sim \text{Categorical}(\exp(aP_{ij})) \qquad a \sim \text{Gamma}(2,1)$$

Finally, consider the link function for **bin comparisons**. We are interested in the likelihood with which participant i selects the higher bin for item j in bin comparison condition l. Suppose l is about comparing the lower bin b_l to the higher b_h. Perhaps the most natural approach would be to link the likelihood of choosing b_h over b_l to the difference between P_{ijb_h} and P_{ijb_l}. However, as discussed above, this does not appear to be what participants were doing. Indeed, a model that implements this idea blatantly fails to capture the relevant regularities in the data. Another plausible link function is to assume that what matters is the distance to the mode of P_{ij}: soft-max prefer the bin that is closer to the mode of P_{ij}; select randomly if both bins are equally far from the prototype.[1]

$$c_{ijl} \sim \text{Bern}((1+\exp(2b(1-p_{ijl}^{\text{high}})))^{-1}) \quad b \sim \text{Gamma}(2,1)$$

$$p_{ijl}^{\text{high}} = \begin{cases} 2 & \text{if } mode(P_{ij}) \text{ is closer to higher} \\ & \text{bin of } l \text{ than to lower bin} \\ 1 & \text{if equal distance} \\ 0 & \text{otherwise} \end{cases}$$

[1] $(1+\exp(2b(1-x)))^{-1} = \frac{\exp(bx)}{\exp(bx)+\exp(by)}$ if $x = 2-y$.

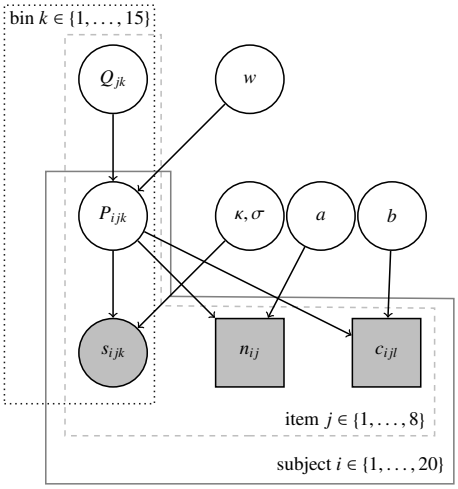


Figure 2: The data-generating model as a probabilistic graphical model, following conventions of Lee and Wagenmakers (2015). Shaded nodes are observed, white nodes are latent variables. Square nodes represent categorical, round nodes continuous variables. Boxes indicate scope of indices.

Inference

The model was implemented in JAGS (Plummer, 2003). 50,000 samples were obtained from two chains with a thinning rate of 2 after a burn-in of 100,000 that ensured convergence according to $\hat{R}$ (Gelman & Rubin, 1992).

Means and bounds of 95% high-density intervals (HDIs) of the posteriors for model parameters are in Table 2. Posterior credible levels of w allow for a limited amount of slack around Q_j. Values for κ indicate that, on average, participants had no preference for or against extreme slider ratings. Relatively high values of a indicate that participants, on average, had a strong tendency to choose modal values in the *give-a-number* task.

	w	κ	σ	a	b
lower	14.65	0.98	0.26	22.04	1.11
mean	15.55	0.99	0.28	27.43	1.27
upper	16.48	0.99	0.31	32.95	1.42

Table 2: Summary statistics for posteriors on parameters

Posterior estimates of Q_j are the most relevant. Fig. 3 shows their means with their 95% HDIs, alongside the mean slider ratings. The latter provide a very good approximation of the inferred population-level beliefs. Inspection of posteriors of individual P_{ij} shows that there is ample variation between participants. Still, the way the model suggests we should think about harnessing the individual P_{ij}s under a population-level central tendency is closely approximated by mean slider ratings. Although the match is not perfect, it is good enough to say that the latter are a practical way of approximating what the crowd believes despite individual differences.

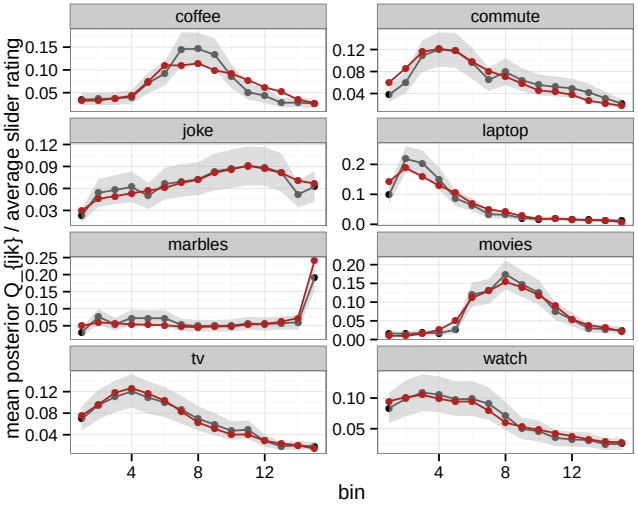
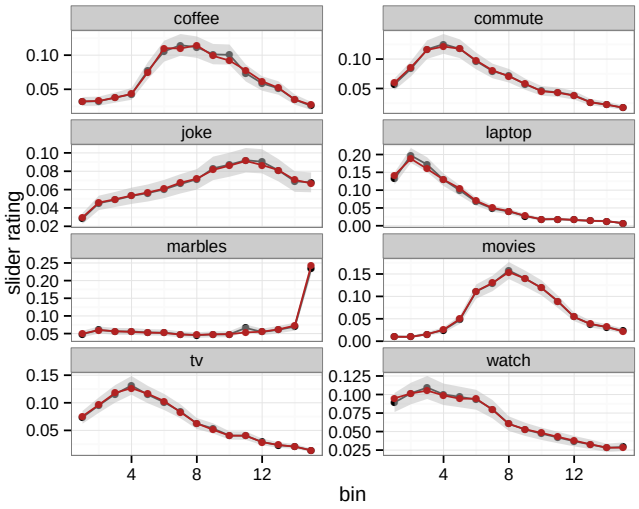


Figure 3: Means of posteriors over Q_j in black with gray area indicating 95% HDIs. Red: mean slider ratings.

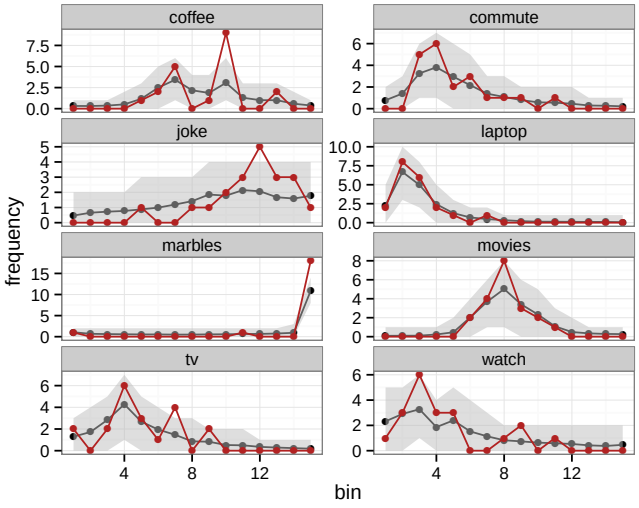
Model criticism

Inferences based on the model are only as reliable as the model itself is plausible. Model criticism is therefore important. Fig. 4 shows posterior predictive checks at the population-aggregate level for all of our three task types. For the binned histogram task, posterior predictions are spot-on. For the give-a-number task, some of the data is surprising despite the model being trained on this very data. This could have various reasons: (i) the give-a-number data does not have a huge influence on the posterior likelihood, (ii) number choices may be influenced by saliency and/or roundness of numbers after all (and thus, not accurately reflect the true beliefs Q_j). Finally, there is one condition in the paired comparison task that the model definitely got wrong. This is the choice of what is more likely: that one or that none of 14 marbles thrown into a pool would sink. The model predicts that almost everybody should answer that it is more likely that one marble sank. But that is not what we observe. It may be that participants revise beliefs about "normality" of the marbles, while holding on to an assumption that all marbles behave in the same way (Degen et al., 2015).

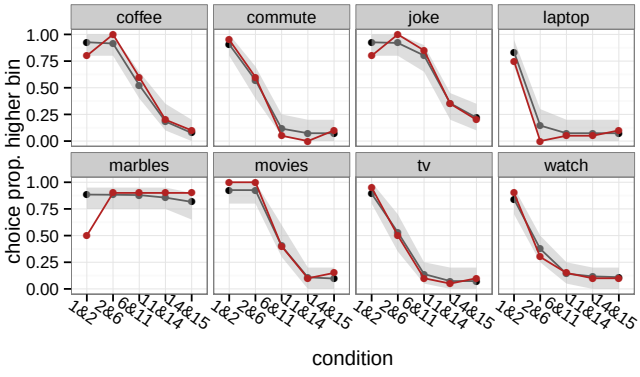
Posterior predictive checks indicate that the trained model captures patterns of answers at the aggregate population level well. To have a more fine-grained measure of model fit, we also looked at posterior predictive p-values (Gelman, Carlin, Stern, & Rubin, 2014) at the level of participants and items. We look at the binned histograms task, because this is our main focus here and population-level posterior predictive checks for BH revealed no systematic deviance. Fixing a participant and an item, observations and replicates are probability vectors of length 15. In a first analysis, we used the mean of these probability vectors as a test statistic. The minimum posterior predictive p-value over all 20 (participants) times 8 (items) cases was 0.13, suggesting that the means of observed s_{ij} are non-surprising to the trained model. In a second analysis, we used entropy as a test statistic. Two cases gave poste-



(a) Binned histogram task.



(b) Give-a-number task.



(c) Paired comparison task.

Figure 4: Posterior predictive checks for aggregate data. Red lines give empirical observations. Black lines are means of posterior predictive samples, gray areas are 95% HDIs.

rior predictive p-values lower than 0.05. These were from the two participants who gave a very extreme slider rating for the "marbles" item, basically assigning all "mass" to the last bin. What this suggests is that the model can cope reasonably well also with individual-level data, but, somewhat unsurprisingly, has problems accounting for "extreme" choices, given that the population-level hyper-prior on P_{ij} will lead to shrinkage.

Conclusion

The data and model presented here suggest that mean slider ratings are consistent with other measures of subjective belief, namely from give-a-number and paired comparison tasks. Future research should evaluate whether this holds for other possible measures of subjective beliefs as well, such as iterated-learning or scoring-rule tasks.

There are aspects of the data that the model does not capture well, but there are also natural explanations for these discrepancies. It therefore does not seem implausible that participants' latent beliefs could have generated the data from all three task types roughly in the way assumed by the model's link functions, with each subjective belief being an expression of a population-level central tendency. If that is so, then mean slider ratings from the binned histogram task are a practical and reliable approximation of what the crowd believes.

Future research should investigate whether and how our results can be extended to other types of uncertain variables. Here, guided by the needs of many previous Bayesian cognitive models, we focused on uncertainty about unidimensional contiguous variables. It remains to be seen whether the binned histogram task can be applied to higher-dimensional variables, like joint prior beliefs over, say, *height* and *weight* of an individual or, more abstractly, different dimensions of a multi-dimensional property like *intelligence*. Another challenge lies in devising means of eliciting prior expectations about properties that do not have clear and familiar measure terms that can be used to label bins—again like *intelligence*.

Finally, another aspect that our model has ignored so far are potential individual differences in the way subjective beliefs P_{ij} generate responses. It is not unlikely that there is individual variation in at least some link function parameters, such as κ, which expresses end-point attraction or end-point aversion in the binned histograms task. A detailed investigation of such individual-level differences must be left to future work. Still, we believe that the model presented here is an important first step towards finding reliable and practical means of measuring what the crowd believes.

Acknowledgments

MF's work was supported by the Institutional Strategy of the University of Tübingen (DFG, ZUK 63). MF, FD, AS and JD's work was supported by the Priority Program XPrag.de (DFG Schwerpunktprogramm 1727). This work was further supported by ONR grant N00014-13-1-0788 and a James S. McDonnell Foundation Scholar Award to NDG and an SNF Early Postdoc.Mobility Award to JD. Thanks to three anonymous reviewers for helpful suggestions and criticism.

References

Andersen, S., Fountain, J., Harrison, G. W., & Rutström, E. E. (2014). Estimating subjective probabilities. *Journal of Risk and Uncertainty*, *48*, 207–229.

Degen, J., Tessler, M. H., & Goodman, N. D. (2015). Wonky worlds: Listeners revise world knowledge when utterances are odd. In *Proceedings of CogSci 37*.

Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, *336*(6084), 998.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2014). *Bayesian data analysis* (3rd edition ed.). Boca Raton: Chapman and Hall.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences (with discussion). *Statistical Science*, *7*, 457–472.

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

Kao, J. T., Bergen, L., & Goodman, N. D. (2014). Formalizing the pragmatics of metaphor understanding. In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.), *Proceedings of CogSci 36*. Austin, TX: Cognitive Science Society.

Kao, J. T., Wu, J. Y., Bergen, L., & Goodman, N. D. (2014). Nonliteral understanding of number words. *PNAS*, *111*(33), 12002–12007.

Lee, M. D., & Wagenmakers, E.-J. (2015). *Bayesian Cognitive Modeling: A Practical Course*. Cambridge, MA: Cambridge University Press.

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.

Manski, C. F. (2004). Measuring expectations. *Econometrica*, *72*(5), 1329–1376.

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In K. Hornik, F. Leisch, & A. Zeileis (Eds.), *Proceedings of Distributed Statistical Computing 3*.

Savage, L. J. (1971). Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association*, *66*(336), 783–801.

Schlag, K. H., Tremewan, J., & van der Weele, J. J. (online first). A penny for your thoughts: a survey of methods for eliciting beliefs. *Experimental Economics*.

Schöller, A., & Franke, M. (2015). Semantic values as latent parameters: Surprising *few* & *many*. In S. D'Antonio, M. Moroney, & C. R. Little (Eds.), *Proceedings of SALT*.

Tauber, S., & Steyvers, M. (2013). Inferring subjective prior knowledge: An integrative Bayesian approach. In M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.), *Proceedings of the CogSci 35* (pp. 3510–3515).

Tessler, M. H. (2015). Understanding *Belief bias* by measuring prior beliefs for a Bayesian model of syllogistic reasoning. In *Proceedings of ESSLLI student session.*