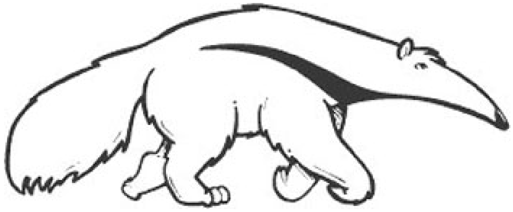
Bayesian methods in cognitive modeling

Michael Lee
mdlee@uci.edu
Department of Cognitive Sciences,
University of California, Irvine



Bayesian methods

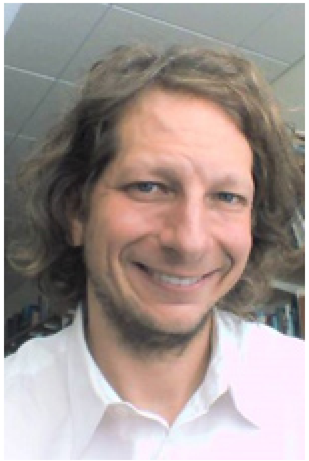
- Bayesian inference
  - represents uncertainty about parameters by probability distributions
  - uses probability theory in a complete and coherent way to update this uncertainty based on relevant information
  - works in the same simple way for any probabilistic model that relates parameters to data
- This allows principled, complete, and coherent
  - **Inferences** about parameters
  - **Evaluation** and comparison of models
  - **Predictions** about data

Bayesian methods

- Bayesian methods let you infer parameters, evaluate models, and understand and make predictions about data
- Three types of application in psychology

Bayesian methods

- Bayesian methods let you infer parameters, evaluate models, and understand and make predictions about data
- Three types of application in psychology
  - **Bayes in the head:** Use Bayes as a theoretical metaphor, assuming that when people make inferences they apply Bayesian methods (at some level)



Josh Tenenbaum

Tom Griffiths

Nick Chater

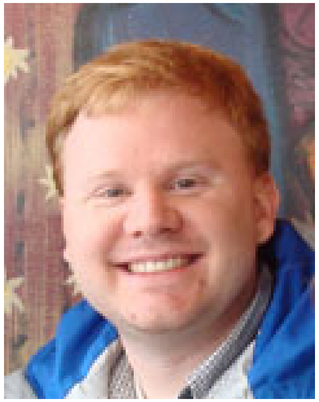
Charles Kemp

Bayesian methods

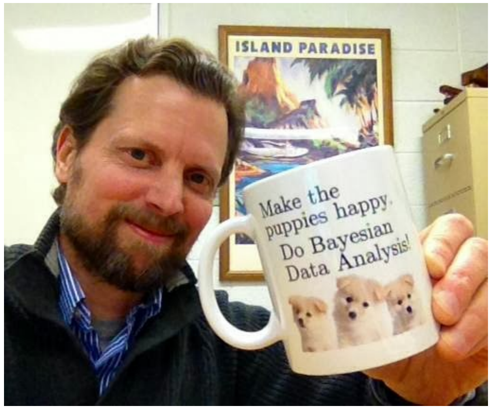
- Bayesian methods let you infer parameters, evaluate models, and understand and make predictions about data
- Three types of application in psychology
  - **Bayes in the head**
  - **Bayes for data analysis:** Instead of using frequentist estimation, confidence intervals, null hypothesis testing, and so on, use Bayesian inference to analyze data

EJ Wagenmakers

Jeff Rouder



Richard Morey

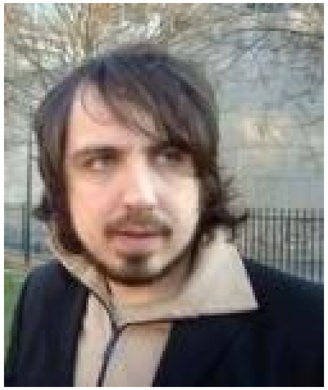


John Kruschke

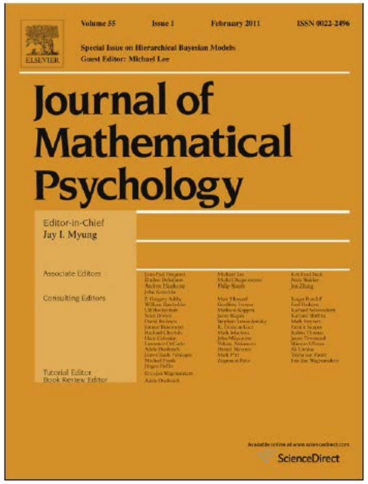
Bayesian methods

- Bayesian methods let you infer parameters, evaluate models, and understand and make predictions about data
- Three types of application in psychology
  - **Bayes in the head**
  - **Bayes for data analysis**
  - **Bayes for cognitive modeling**: Use Bayesian inference to relate models of psychological processes to behavioral data

EJ Wagenmakers



Wolf Vanpaemel



Volume 55 Issue 1 February 2011 ISSN 0022-2496

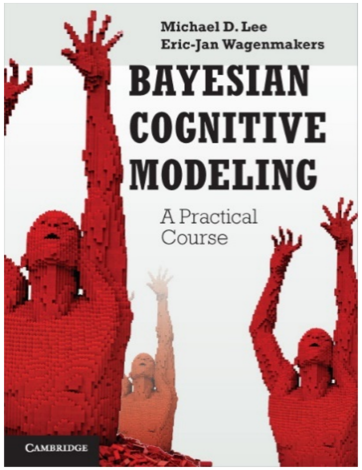
Special Issue on Hierarchical Bayesian Models
Guest Editor: Michael Lee

Journal of Mathematical Psychology

Editor-in-Chief
Jay I. Myung

Associate Editors

Consulting Editors

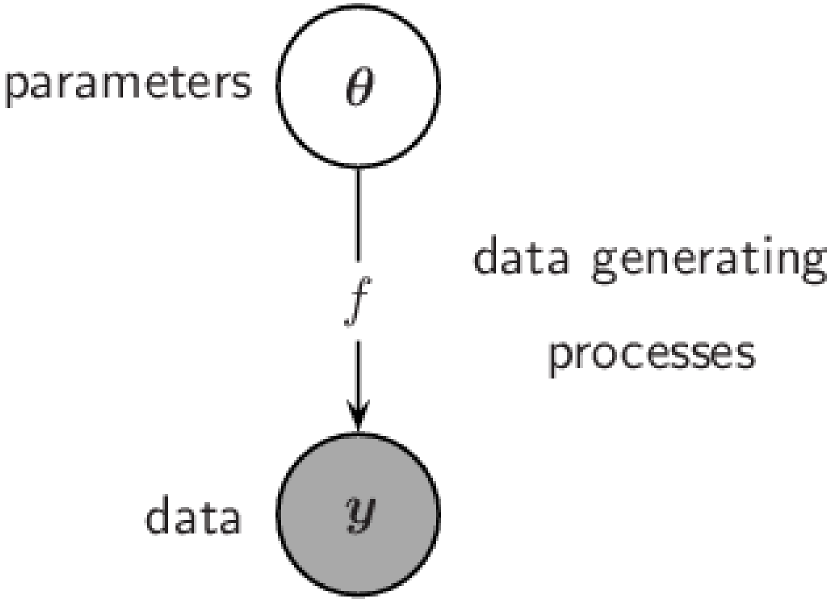


Michael D. Lee
Eric-Jan Wagenmakers

BAYESIAN COGNITIVE MODELING
A Practical Course

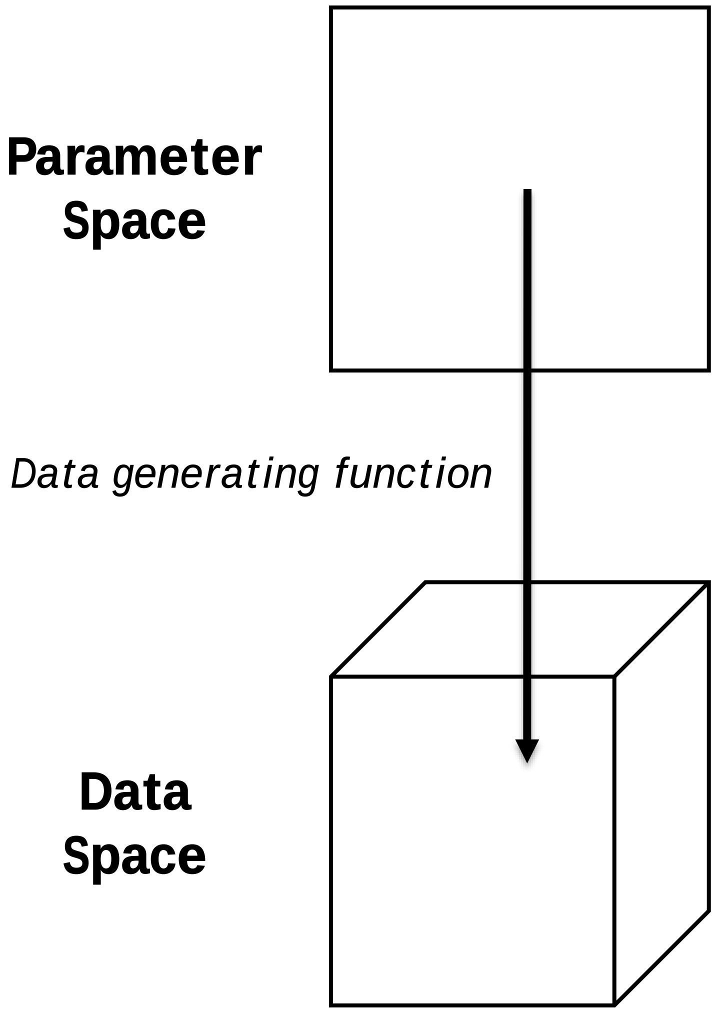
Bayesian methods for cognitive modeling

- Bayesian methods are a way of relating
  - **parameters**, representing psychological variables
  - and **models**, assumptions about how parameters generate behavior
  - to **data** that can be observed and measured



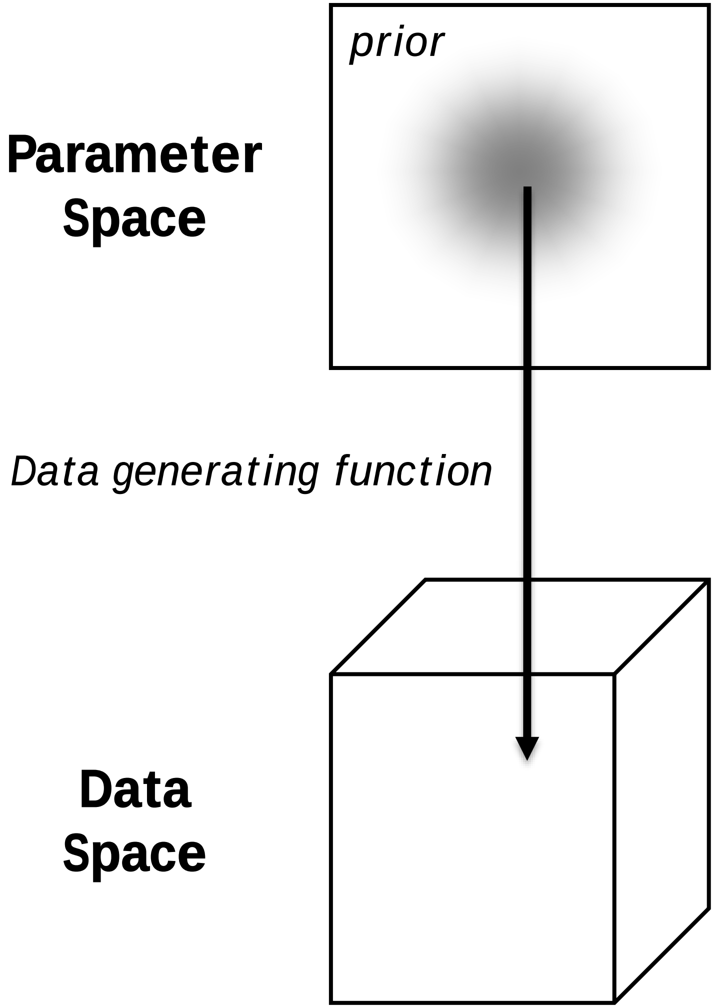
Bayesian methods for cognitive modeling

- Psychological models can be thought of as cognitive processes, controlled by psychological variables, that generate data



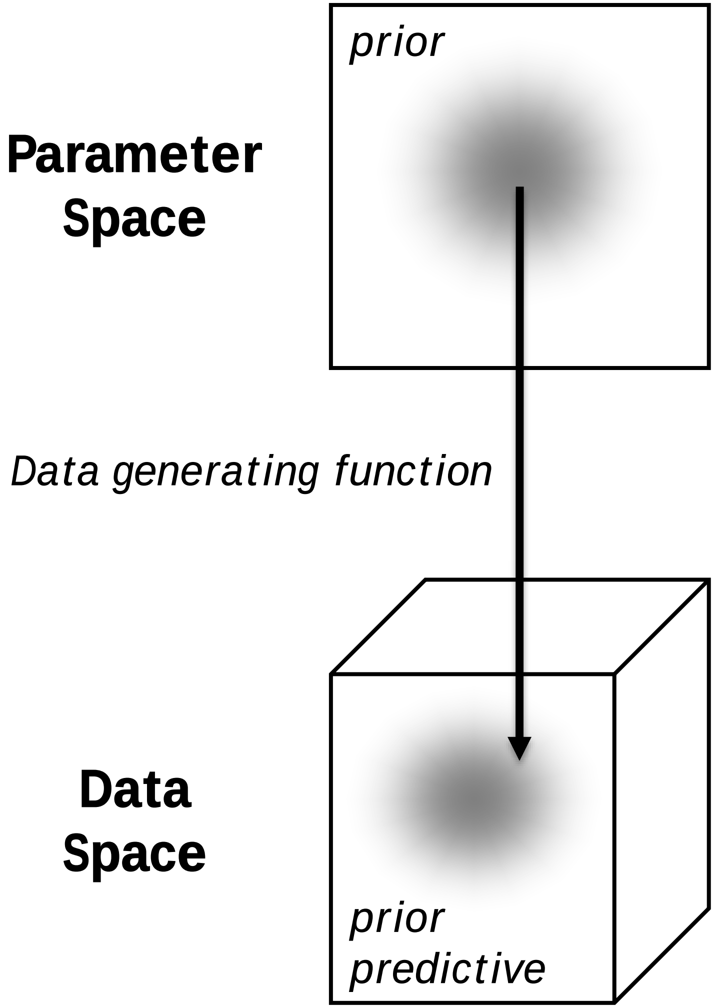
Bayesian methods for cognitive modeling

- The data generating function and the prior distribution on parameters formalize the model



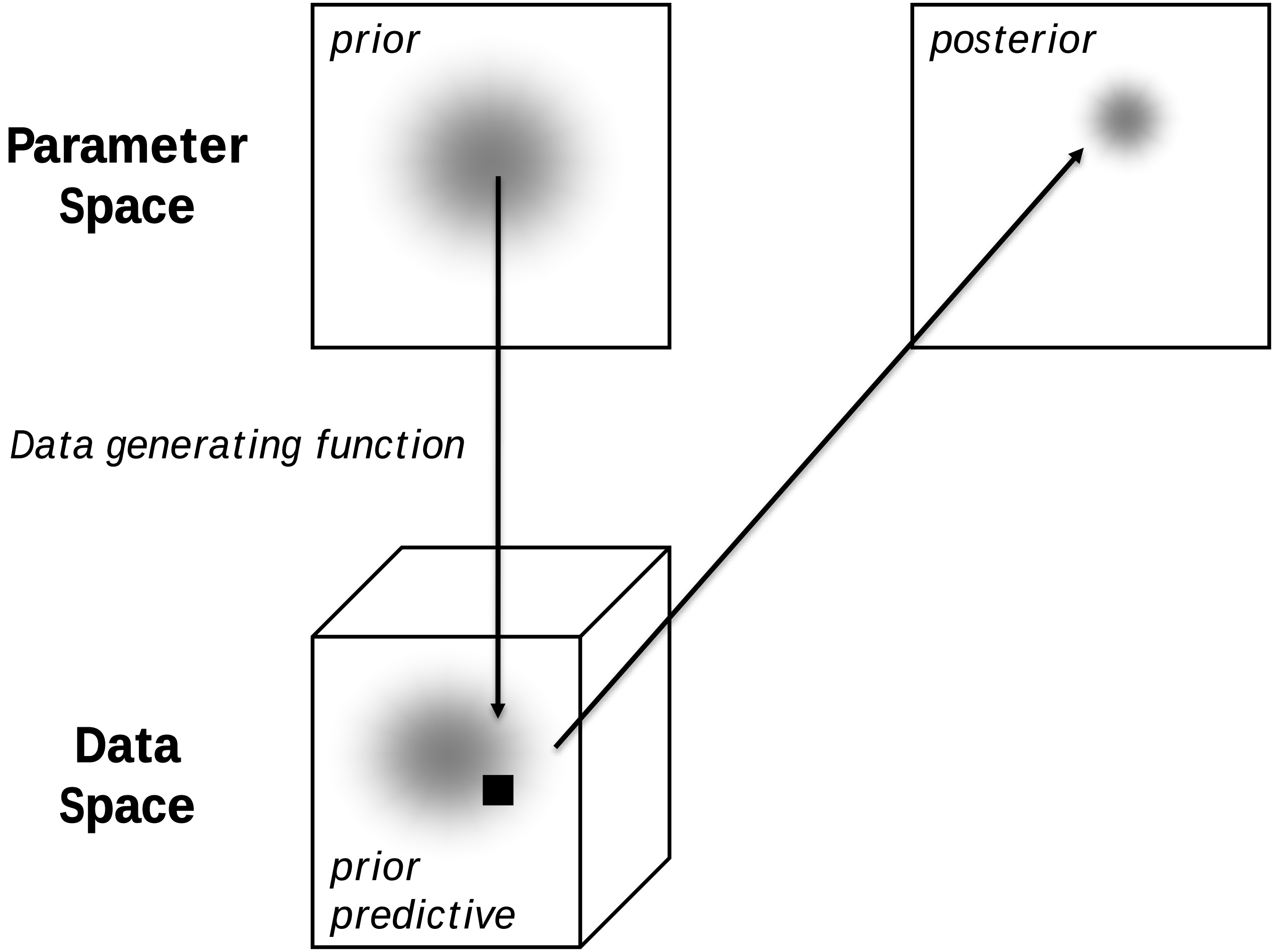
Bayesian methods for cognitive modeling

- This model, which combines the prior and data generating function (aka likelihood function), predict observed data



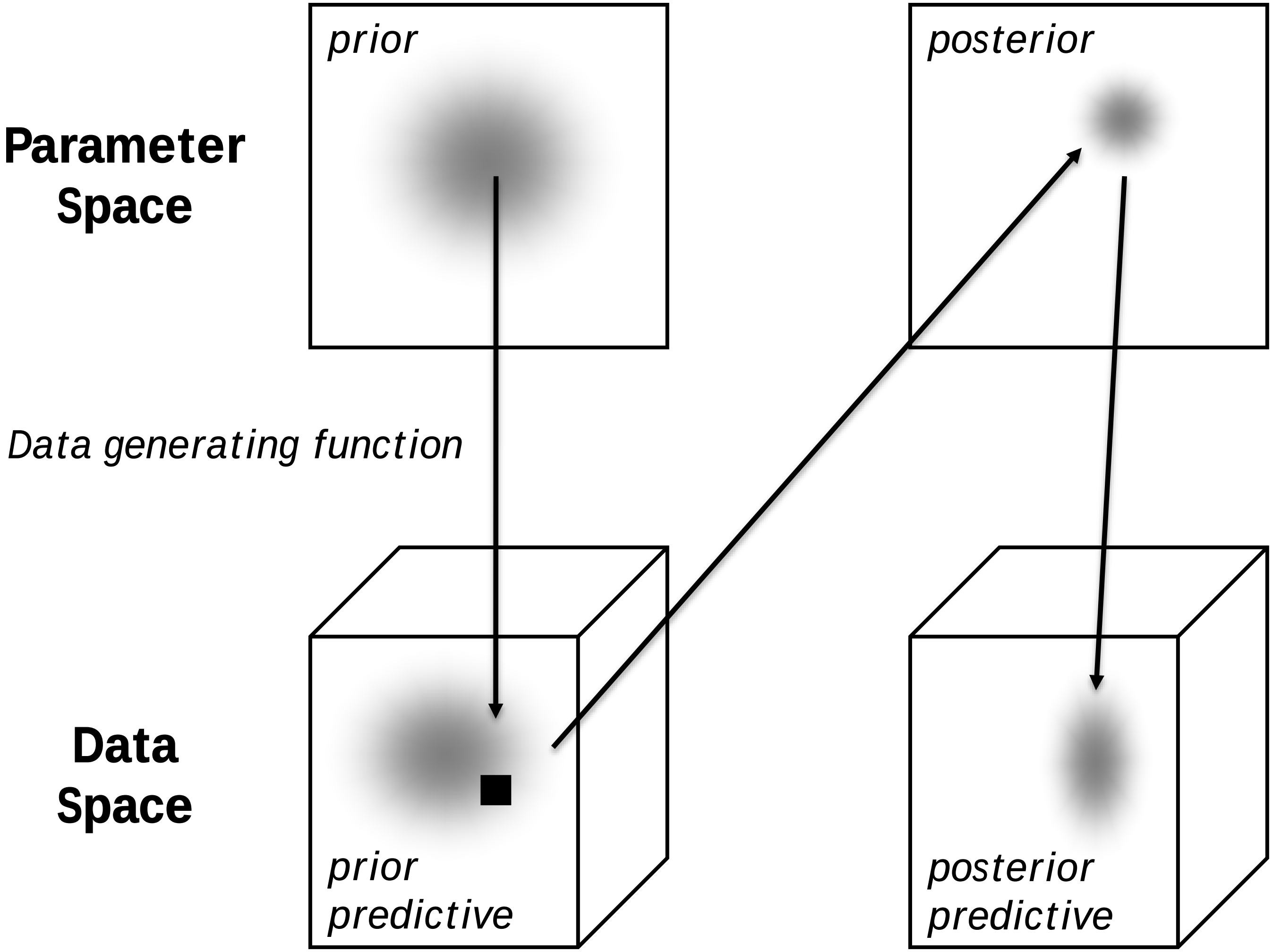
Bayesian methods for cognitive modeling

- Once data are observed, probability theory (via Bayes theorem) allows the prior over parameters to be updated to a posterior



Bayesian methods for cognitive modeling

- The posterior distribution over parameters quantifies uncertainty, and makes new predictions



prior

posterior

Parameter Space

Data generating function

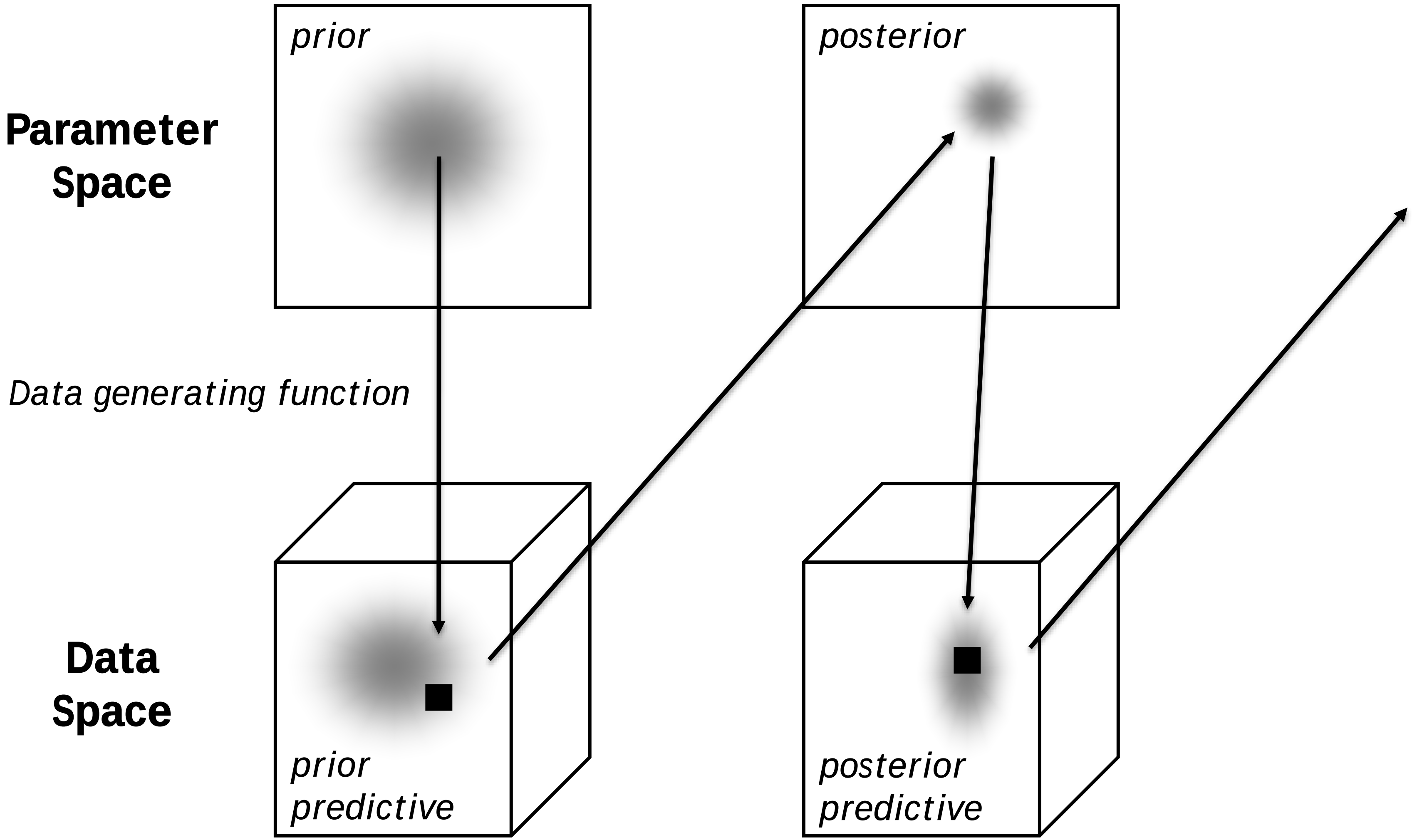
Data Space

prior predictive

posterior predictive

Bayesian methods for cognitive modeling

- Bayesian inference is a complete framework for representing and incorporating information

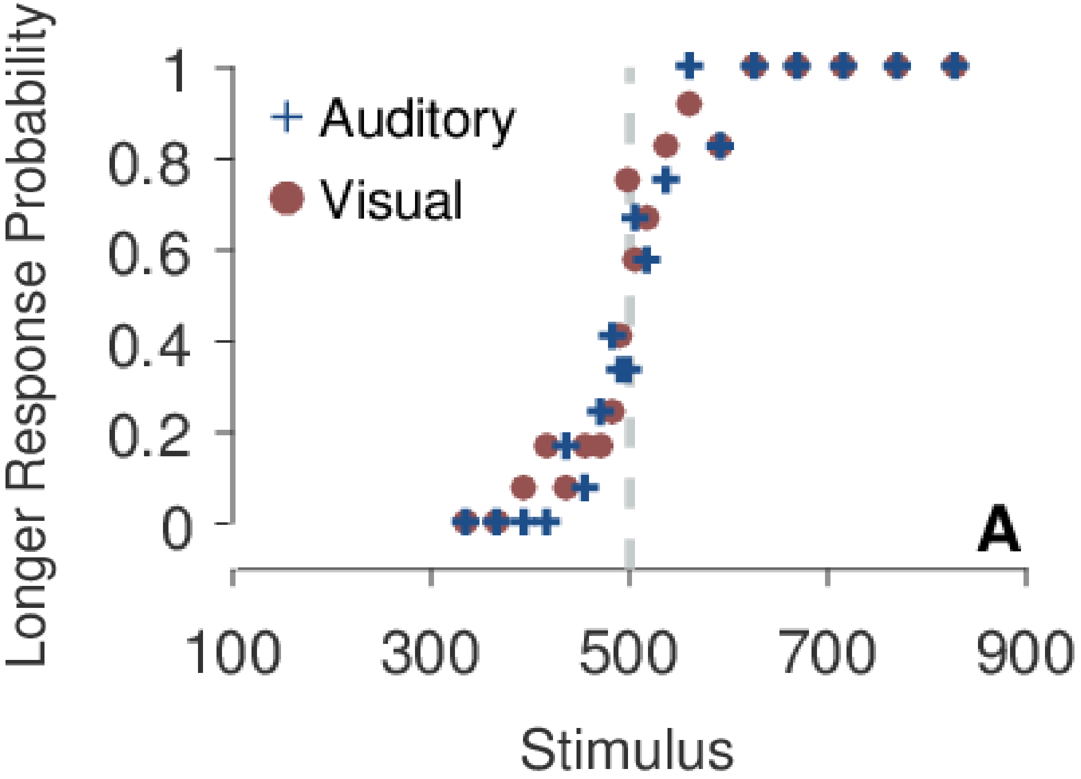


A case study in psychophysical modeling

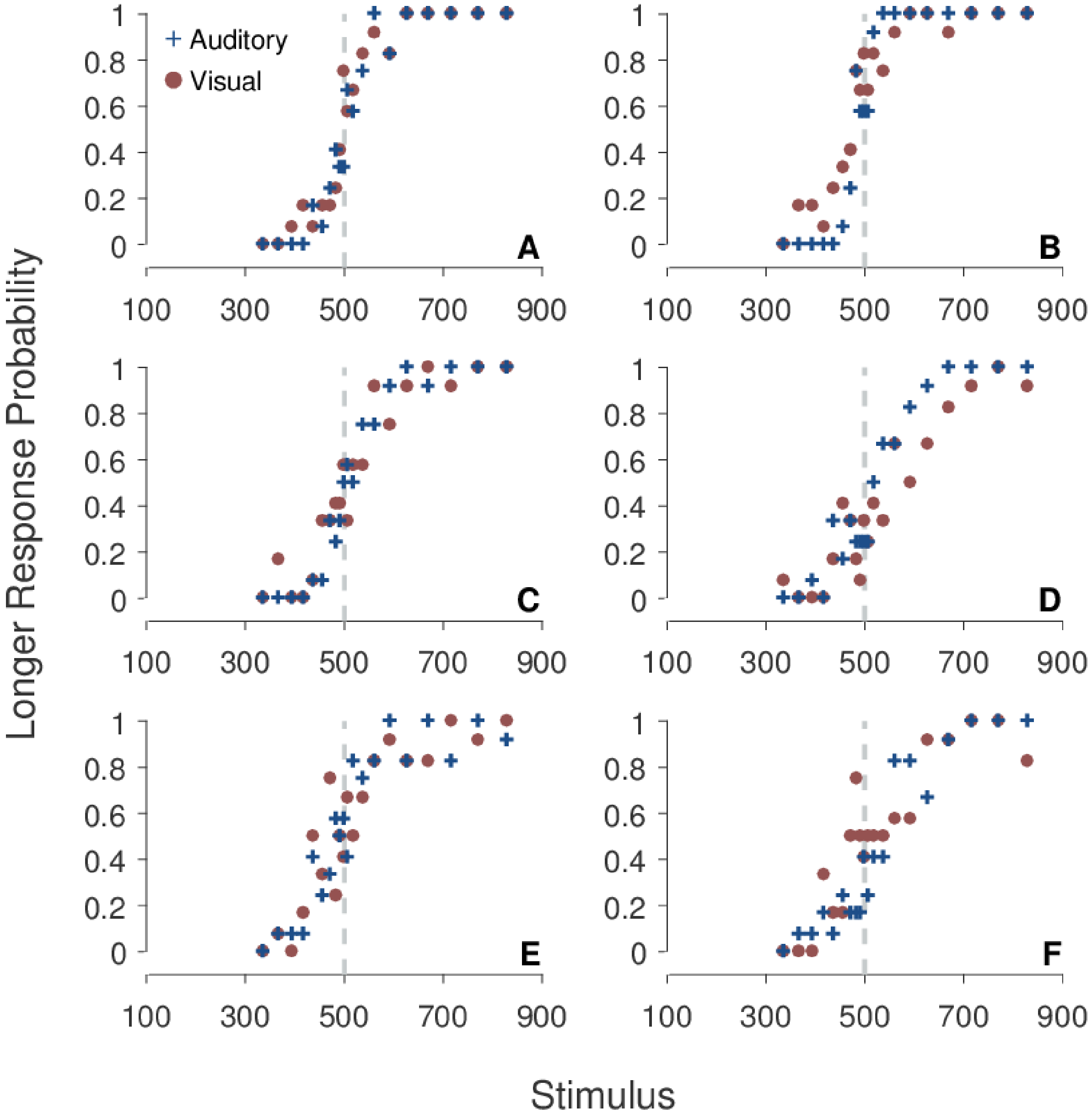
Joram
van Driel

Experiment and data

- 19 subjects did visual and auditory psychophysical discrimination tasks with 240 trials each
  - Whether an LED light or a beep was shorter or longer than a 500ms standard



Behavioral data for six subjects

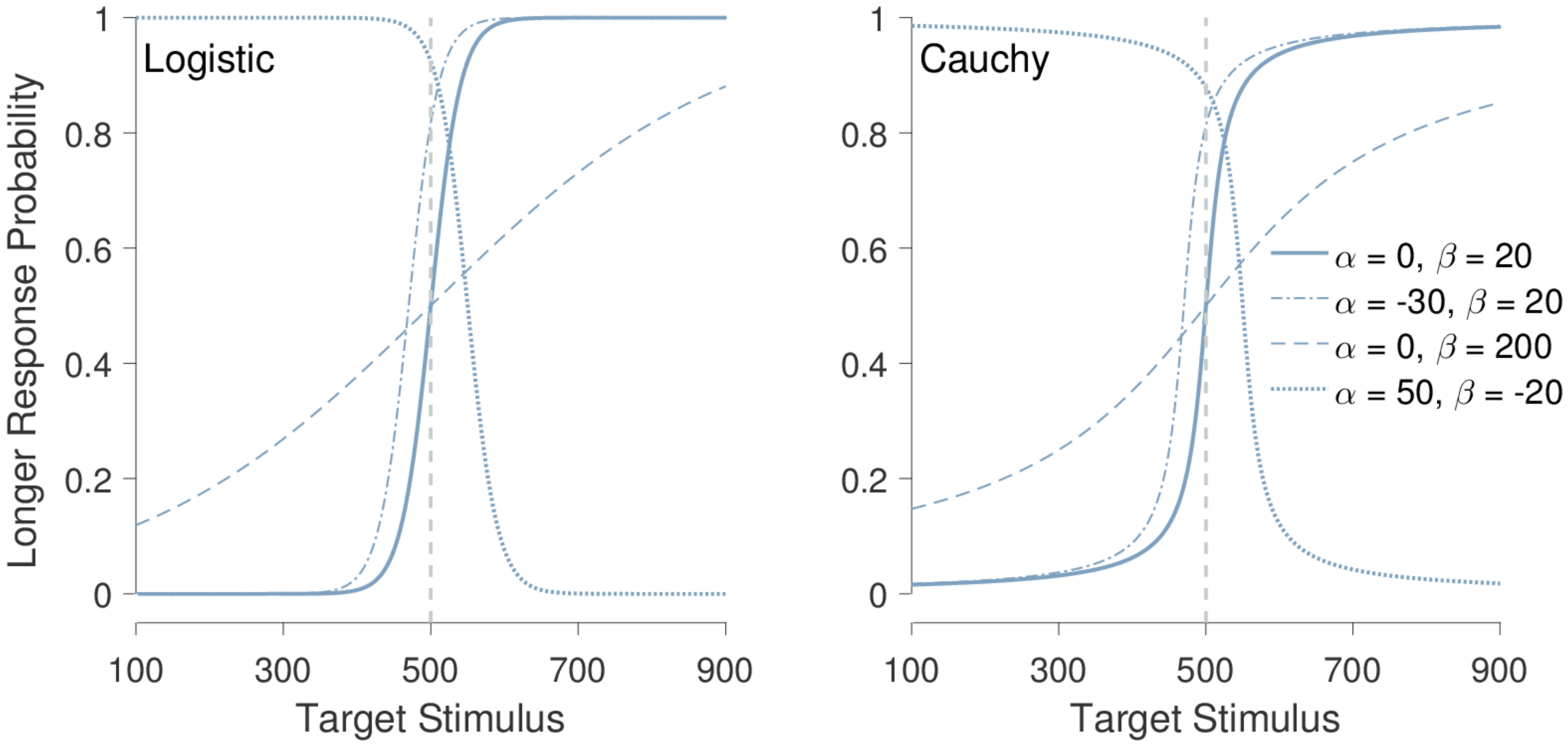


Research questions

- Some possible research questions include
  - What’s the form of the psychophysical function?
  - What parameters of that function describe each subject?
  - Are there individual differences in the function or parameters?
  - Are there differences depending on the modality?
  - Are there sequential dependencies in responding, or any sort of adaptation or learning?
  - Are there lapses that lead to contaminant trials?
  - ….

Two psychophysical functions

- Both have shift and scale parameters

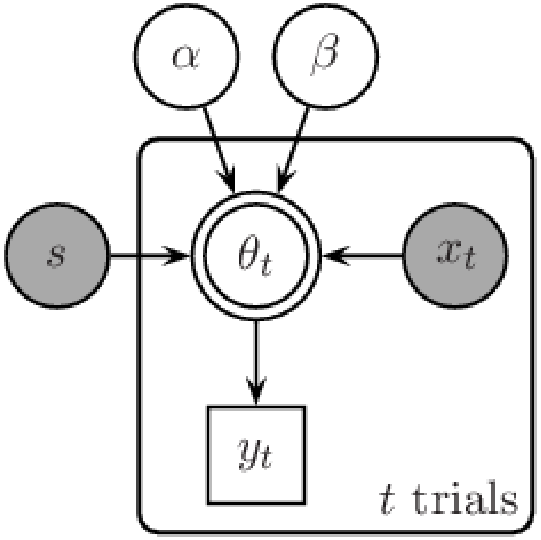


$$\theta_i = 1/\left(1 + \exp\left(-\frac{x_i - s - \alpha}{\beta}\right)\right)$$

$$\theta_i = \arctan\left(\frac{x_i - s - \alpha}{\beta}\right)/\pi + \frac{1}{2}.$$

Graphical model representation of a logistic model

- Graphical models are a useful and fairly general language for implementing probabilistic models of cognition
  - A good language for using BUGS, JAGS, Stan to automate computational Bayesian inference



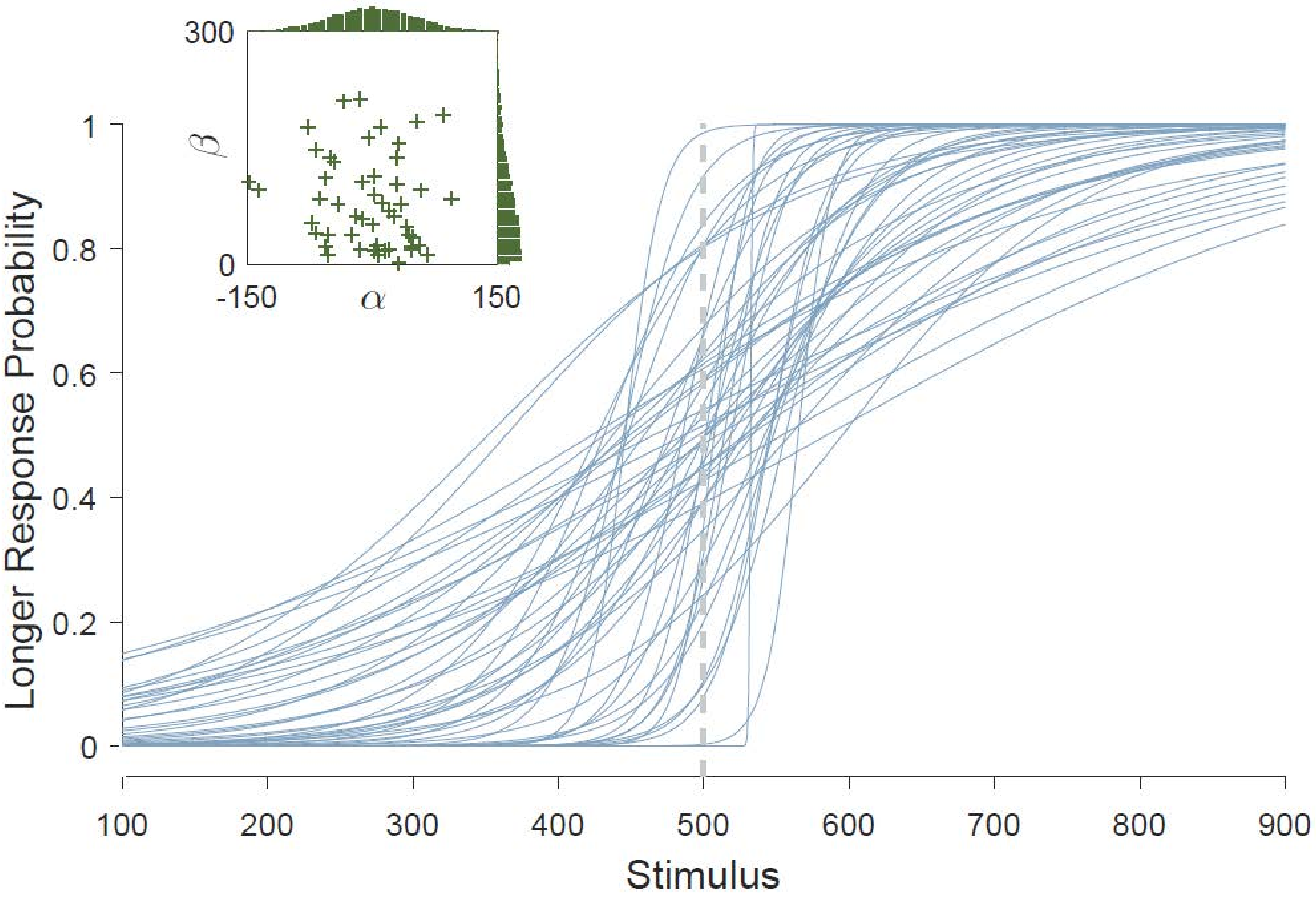
$$\alpha \sim \text{Gaussian}(0, 1/50^2)$$

$$\beta \sim \text{TruncatedGaussian}_+(0, 1/100^2)$$

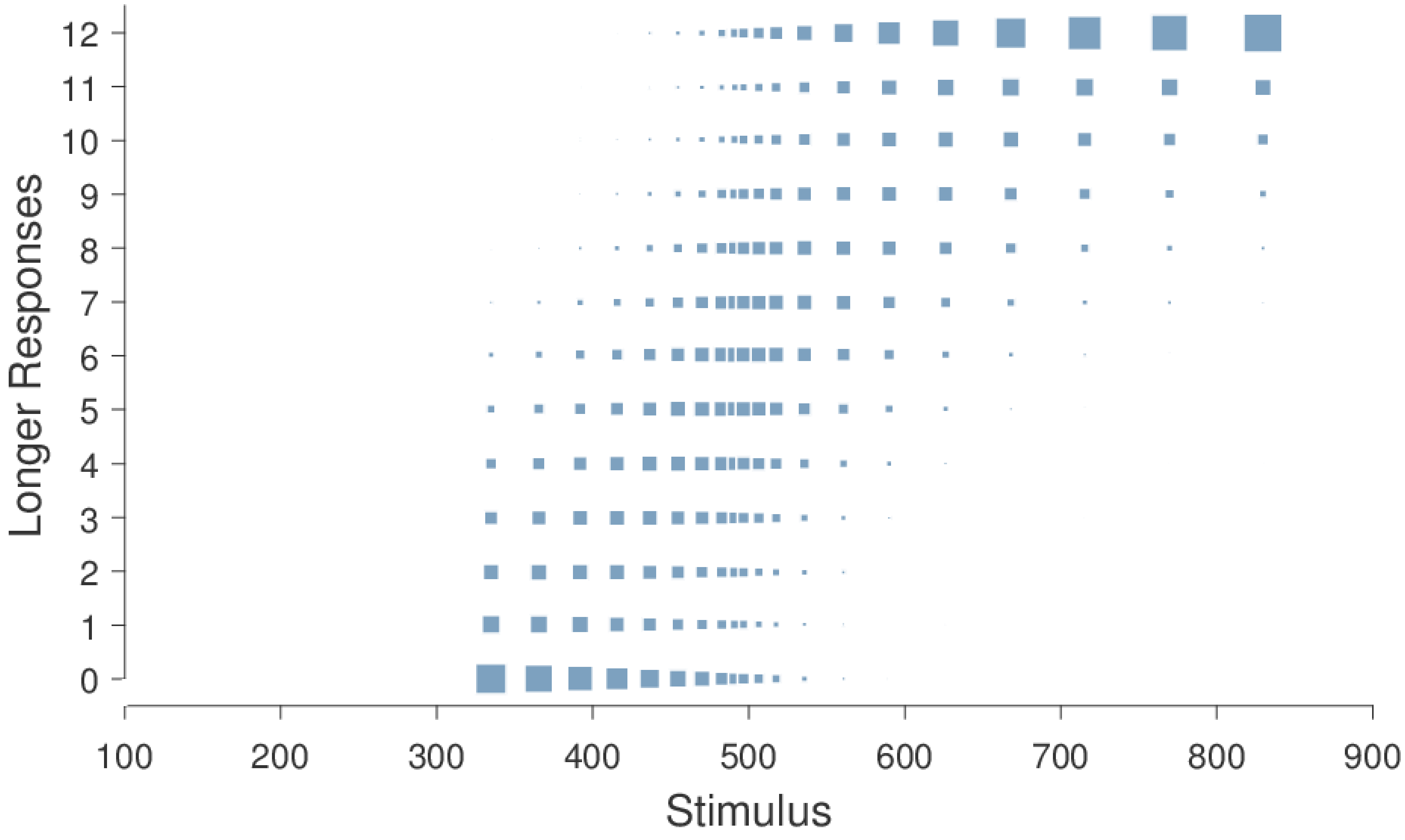
$$\theta_t = 1/\left(1 + \exp\left(-\frac{x_t - s - \alpha}{\beta}\right)\right)$$

$$y_t \sim \text{Bernoulli}(\theta_t)$$

Prior on psychophysical function



Prior predictive



Models need a likelihood and prior

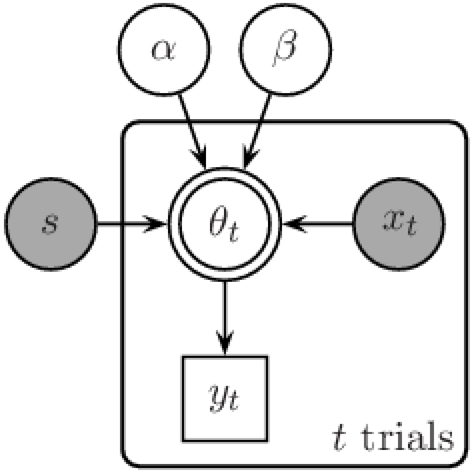
- A defining property of a scientific model is that it makes predictions (Feynman, 1994)
  - this requires both a likelihood (a cognitive process) and a prior (information about the variables that control that process)
- Even proponents of Bayesian inference are sometimes apologetic about priors, viewing them as a sort of necessary evil (e.g., Myung & Pitt, 1997)
- Our view is that it is a key feature of the Bayesian approach that the prior distribution over parameters has the same status as the likelihood as a vehicle to formalize theory and assumptions (Vanpaemel & Lee, 2012; Lee & Vanpaemel, 2015)

Vague vs informative priors

- It is common practice to use “non-informative” (“weakly-informative”, “vague”, “flat”, …) priors in cognitive modeling
- Conceptually, the idea seems to be something like “letting the data speak for themselves” or “not letting assumptions influence the results”
- In practice the approach is something like “[t]ypically, a non-informative prior would be represented by a distribution with a relatively flat density, where the different values the parameter can take on have approximately equal likelihood under the distribution” (Depaoli & van de Schoot, 2015)
- Our view is that these are conceptual and practical mistakes, and priors should be our best attempt to formalize what we know and assume, as we do for likelihoods

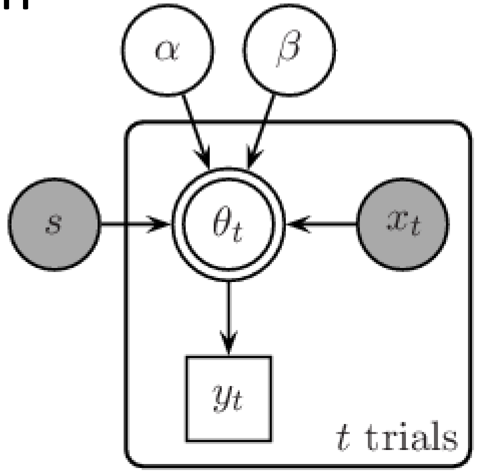
Alternative models

- "vague" priors on the shift and scale parameters



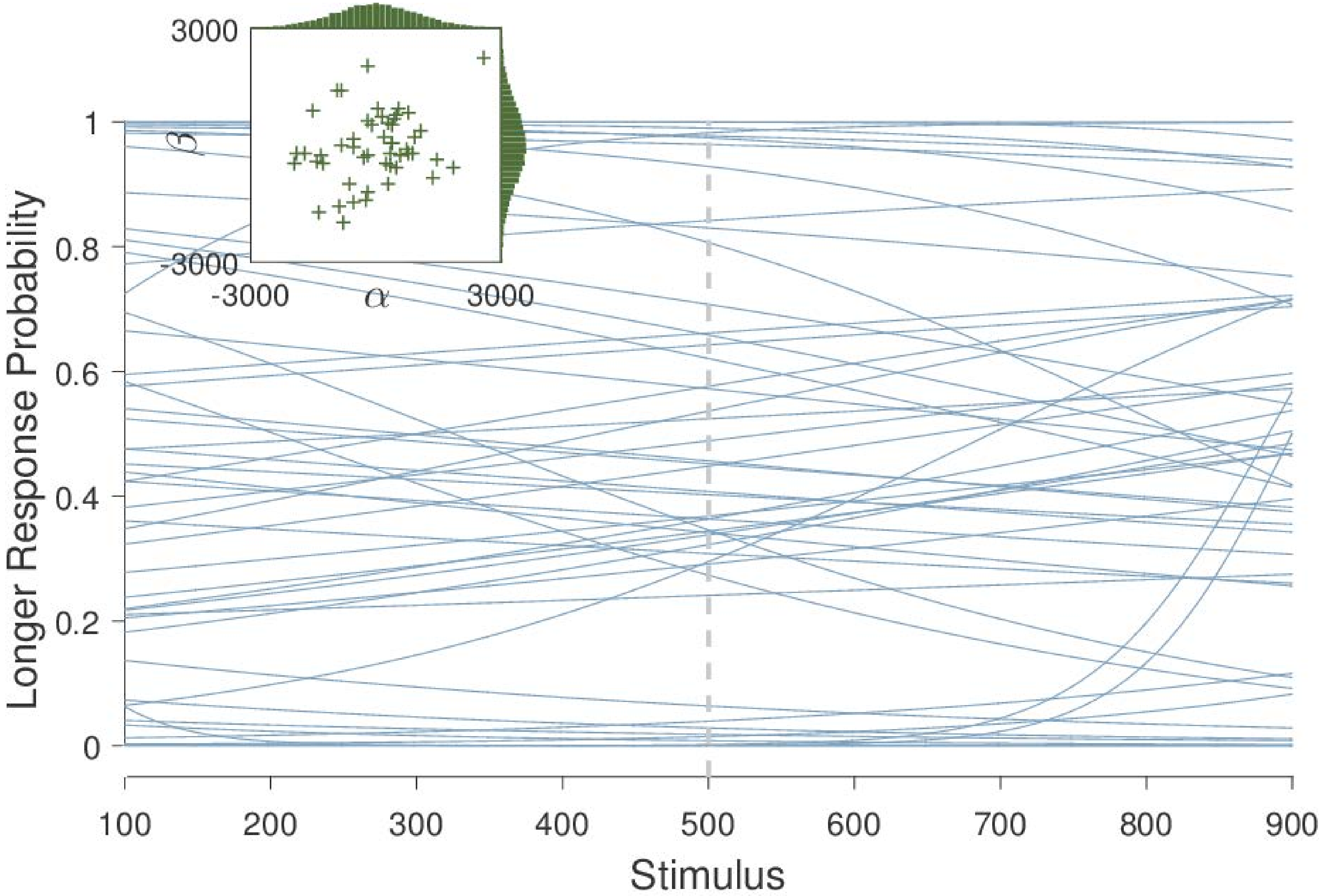
$$
\begin{aligned}
\alpha &\sim \text{Gaussian}(0, 0.000001)\\
\beta &\sim \text{Gaussian}(0, 0.000001)\\
\theta_t &= 1/\left(1+\exp\left(-\tfrac{x_t - s - \alpha}{\beta}\right)\right)\\
y_t &\sim \text{Bernoulli}(\theta_t)
\end{aligned}
$$

- "vague" priors on the shift and scale parameters, in a different parameterization

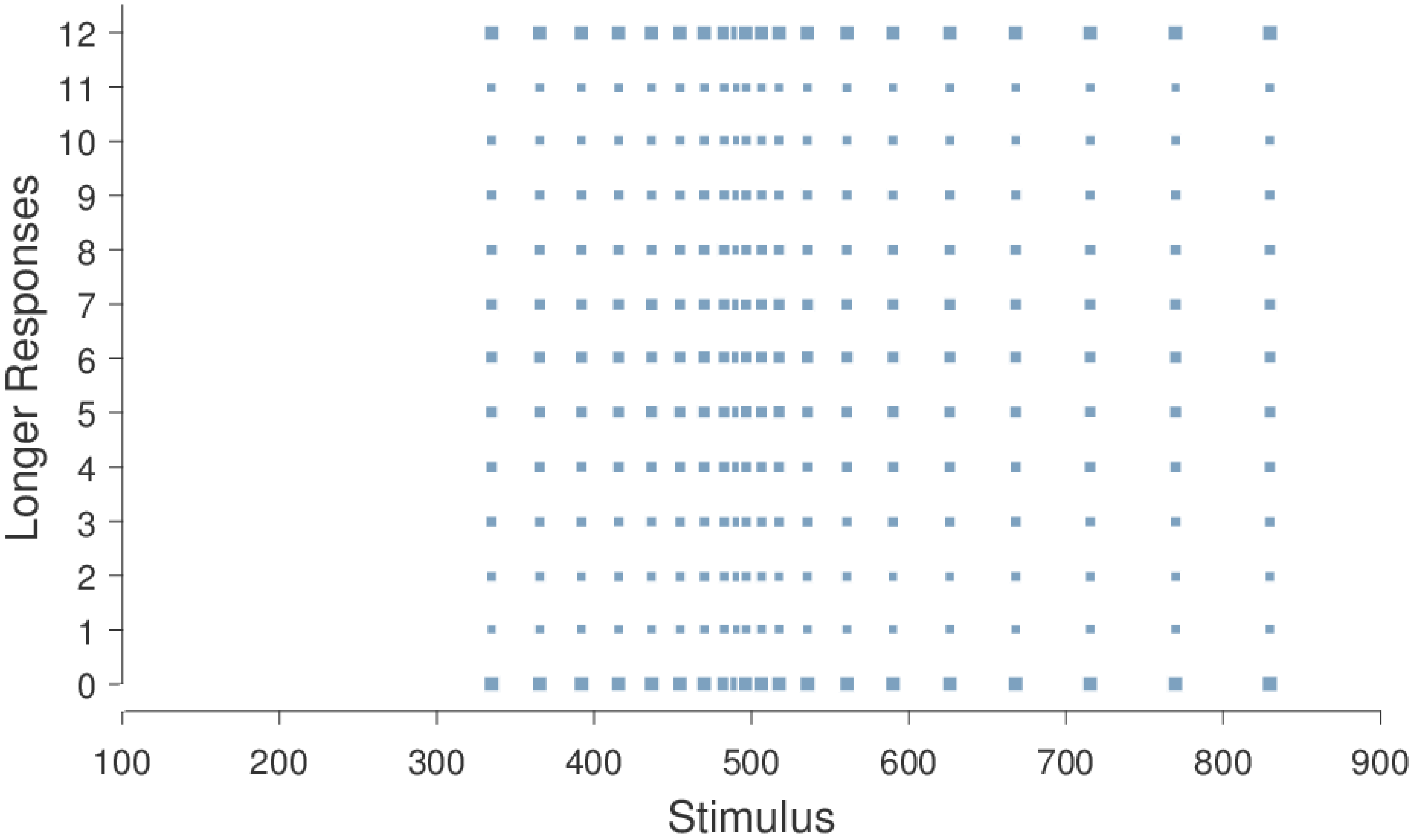


$$
\begin{aligned}
\alpha &\sim \text{Gaussian}(0, 0.000001)\\
\beta &\sim \text{Gaussian}(0, 0.000001)\\
\theta_t &= 1/\left(1+\exp\left(-\beta\left(x_t - s - \alpha\right)\right)\right)\\
y_t &\sim \text{Bernoulli}(\theta_t)
\end{aligned}
$$

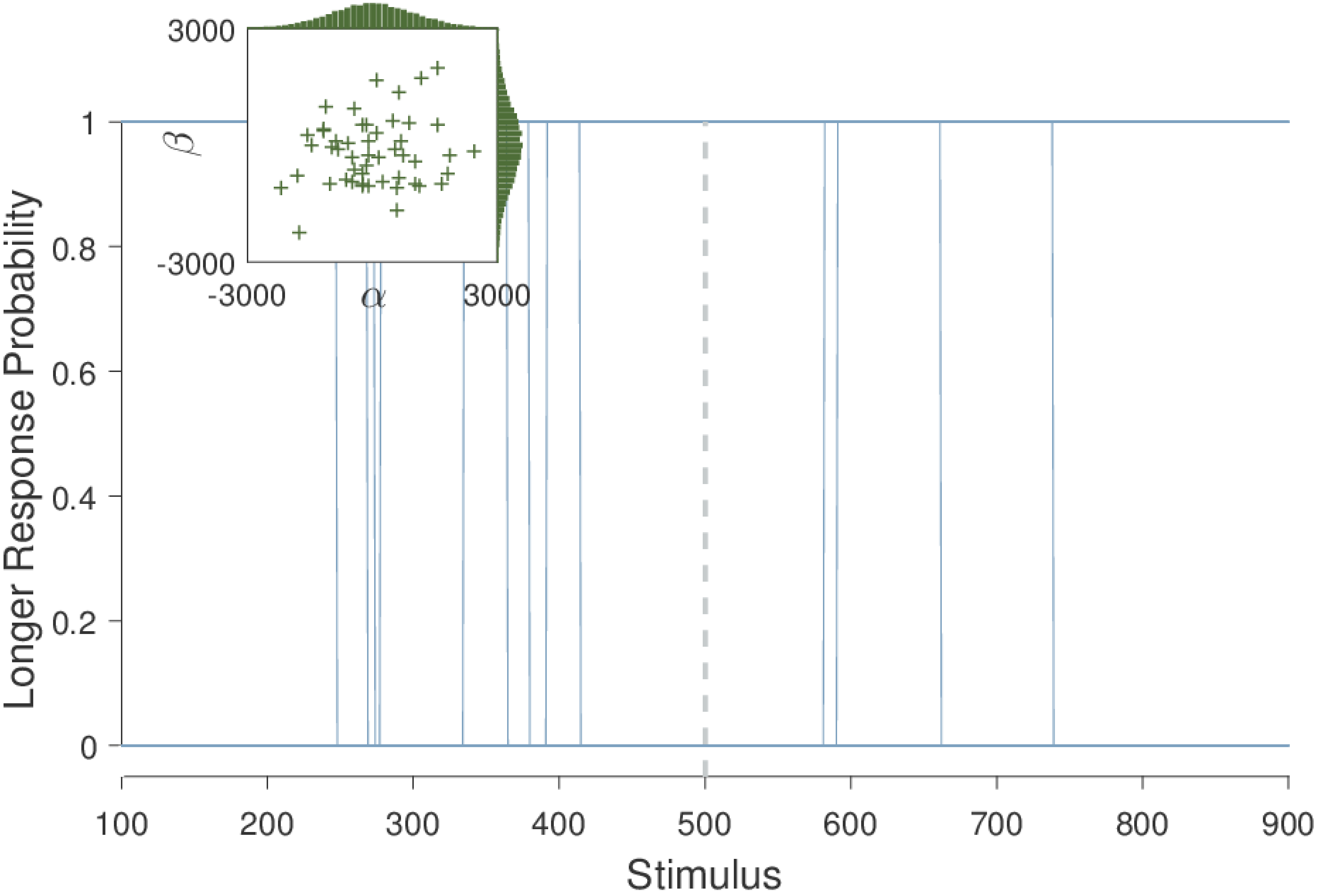
Prior on psychophysical function of model with flat prior



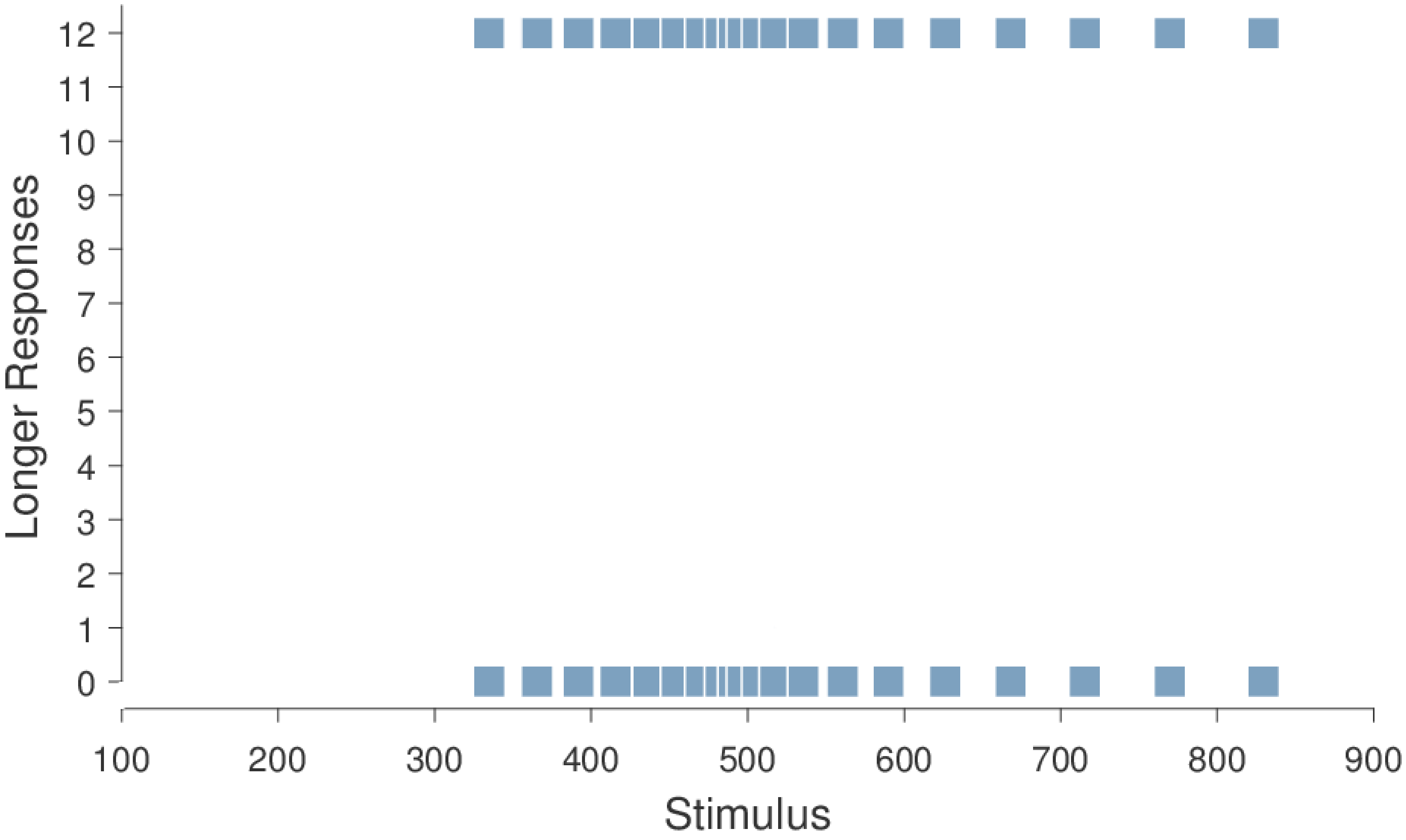
Prior predictive with flat prior



Prior on psychophysical function with other flat prior

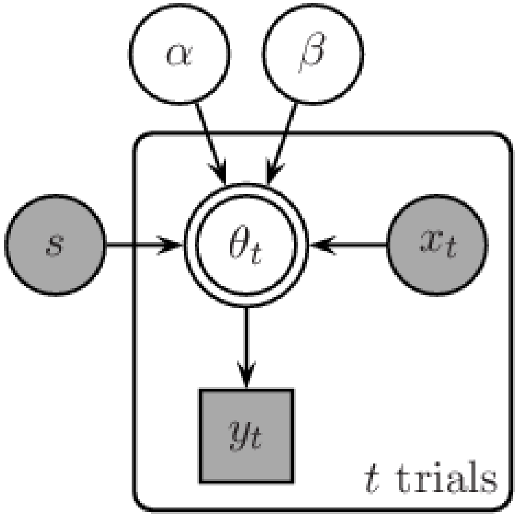


Prior predictive with other flat prior



Graphical model for inference

- To make inferences based on data, the data node is now observed
- Bayes rule automatically now implies a joint posterior distribution in the parameter space
  - Computational approximation of samples from this joint posterior found by JAGS (BUGS, Stan, …)



$$\alpha \sim \text{Gaussian}(0, 1/50^2)$$

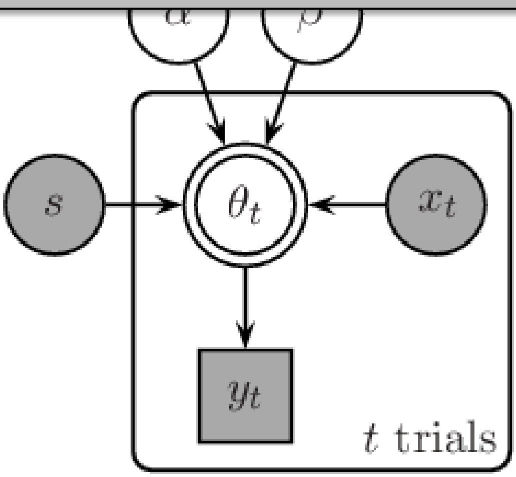
$$\beta \sim \text{TruncatedGaussian}_+(0, 1/100^2)$$

$$\theta_t = 1/\left(1 + \exp\left(-\tfrac{x_t - s - \alpha}{\beta}\right)\right)$$

$$y_t \sim \text{Bernoulli}(\theta_t)$$

Graphical model for inference

```
model{
 # Likelihood
 for (trial in 1:nTrials){
  theta[trial] = 1/(1+exp(-(stimulus[trial]-standard-alpha)/beta))
  y[trial] ~ dbern(theta[trial])
 }
 # Priors
 alpha ~ dnorm(0,1/50^2)
 beta ~ dnorm(0,1/100^2)T(0,)
}
```



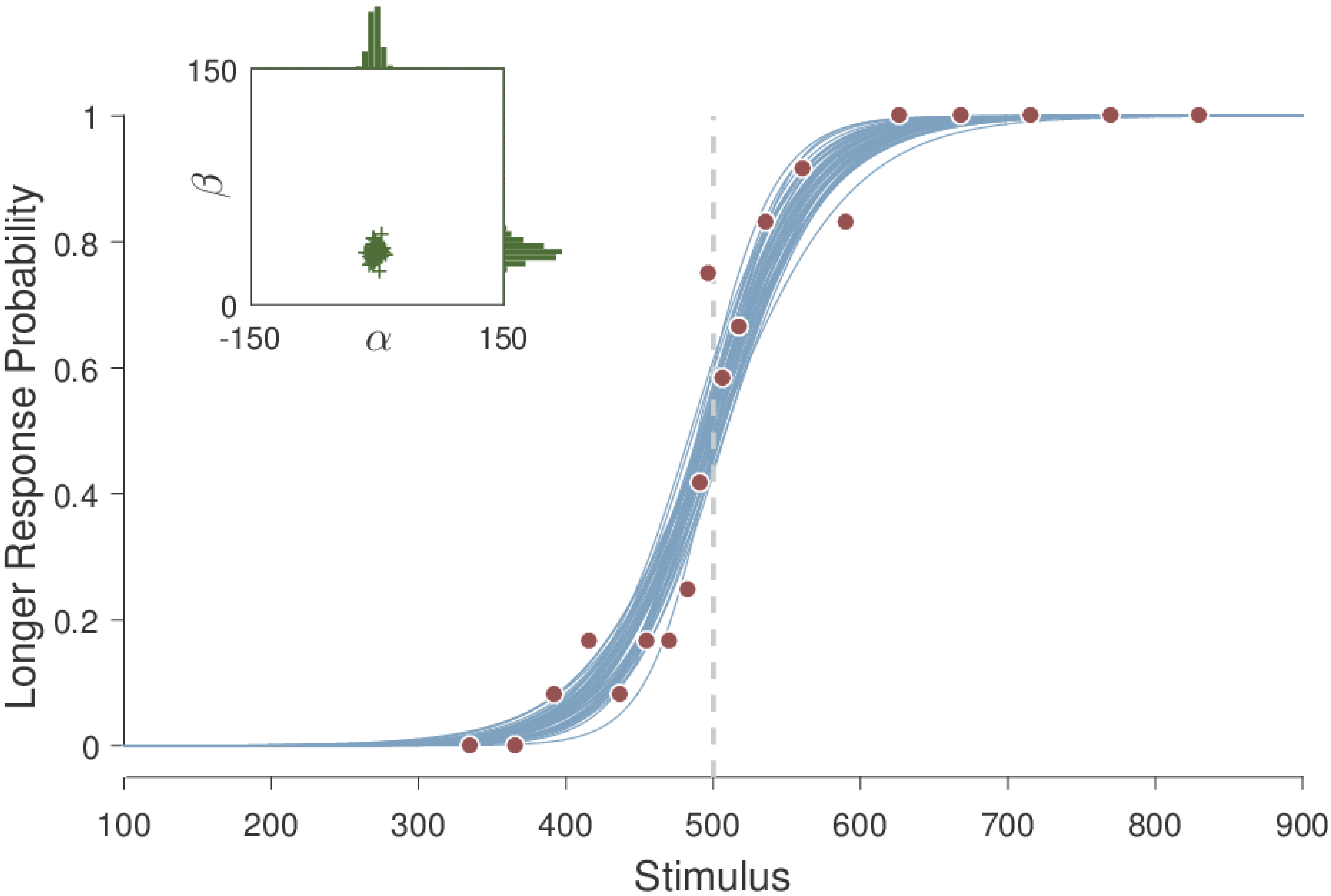
$$\alpha \sim \text{Gaussian}(0, 1/50^2)$$

$$\beta \sim \text{TruncatedGaussian}_+(0, 1/100^2)$$

$$\theta_t = 1/\left(1 + \exp\left(-\frac{x_t - s - \alpha}{\beta}\right)\right)$$

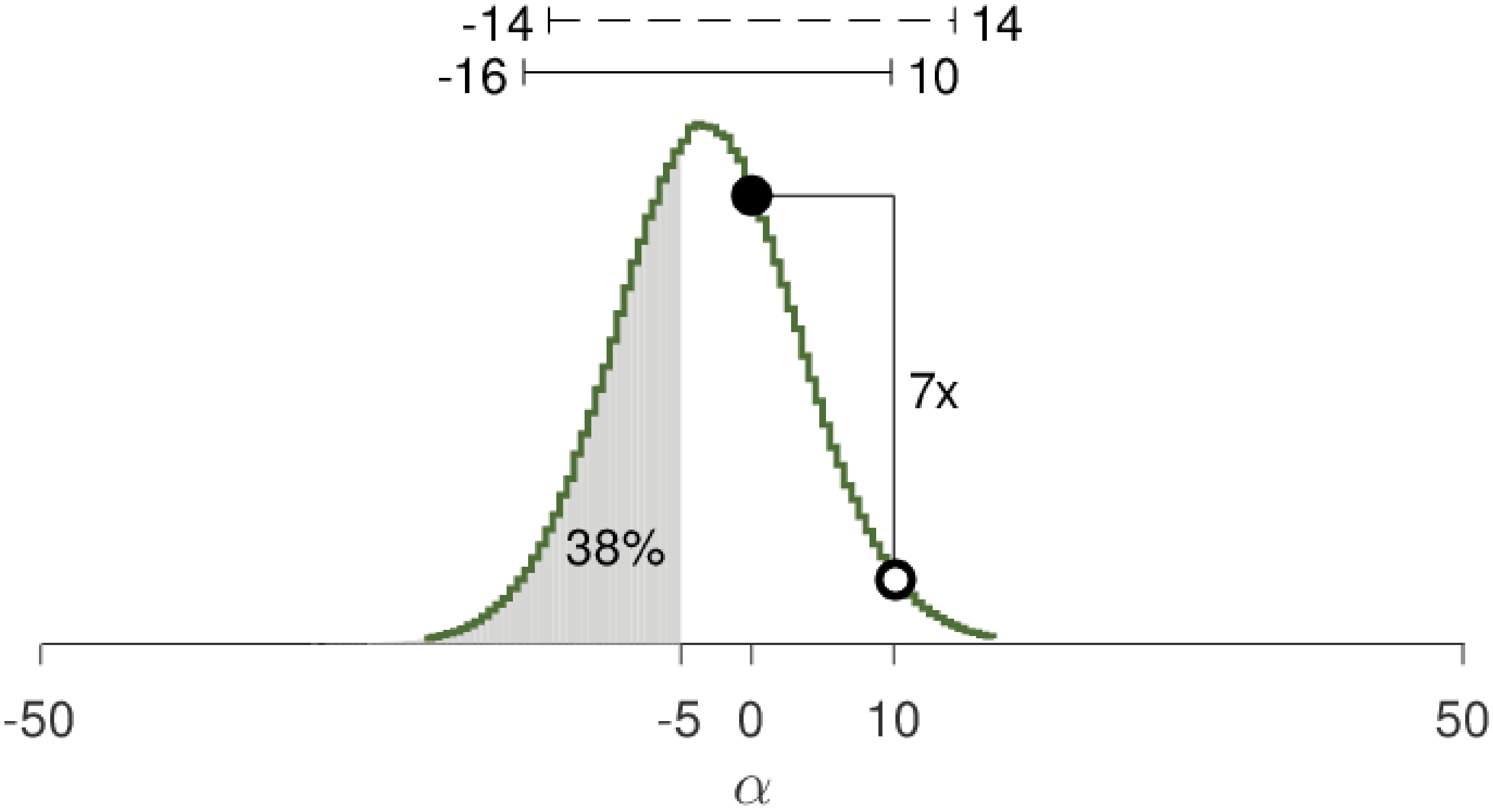
$$y_t \sim \text{Bernoulli}(\theta_t)$$

Inference for subject A



Interpretation of shift parameter posterior

- The posterior distribution provides a complete representation of uncertainty, that can be summarized by credible intervals, probabilities of ranges, relative densities, point estimates, …



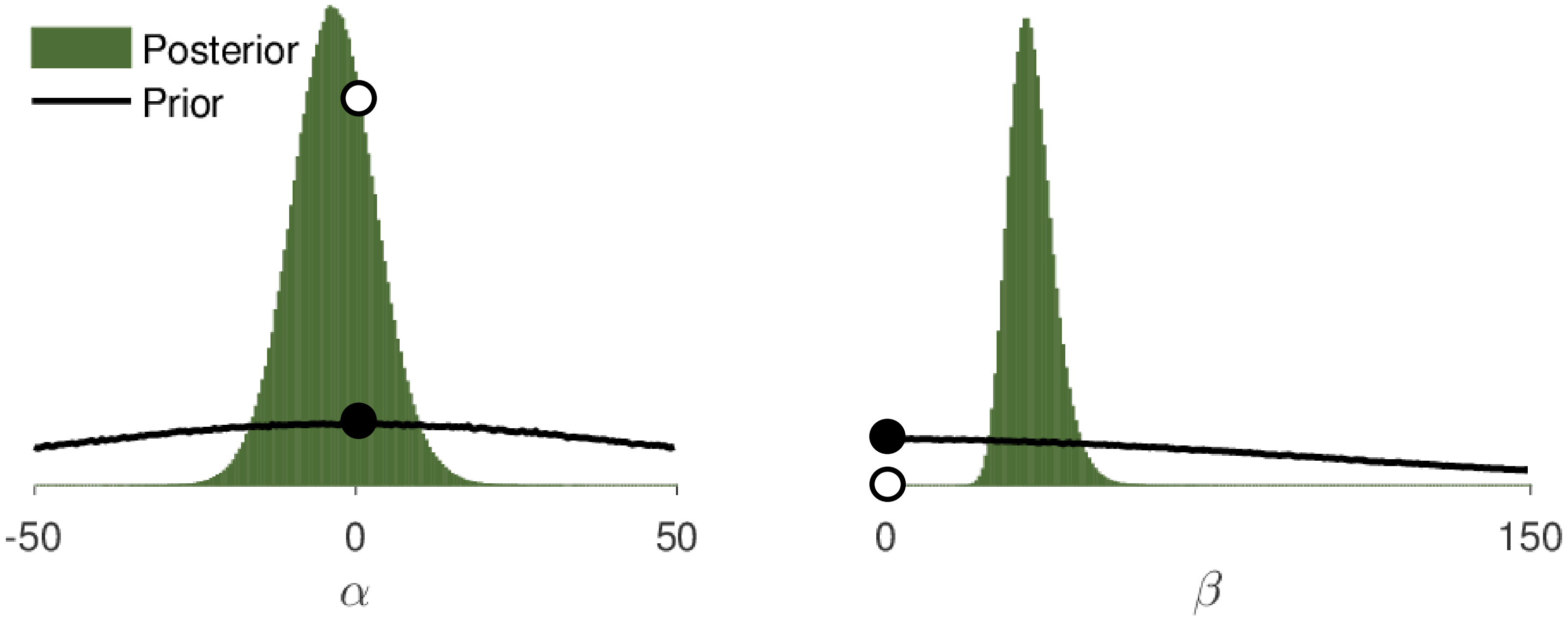
Model selection and parameter inference

- Proponents of Bayesian methods sometimes try to use parameter inference to make model selection decisions
  - e.g., testing whether a parameter value is “credibly different” from 0, or has bounds within a “region of practical equivalence (ROPE)”
- Our view is this is conceptually wrong (because the parameter inference is conditional on the model being tested, and so has already been assumed true) and perilous in practice (which credible interval?)
- Probability theory leads to Bayes factors as a measure to compare models

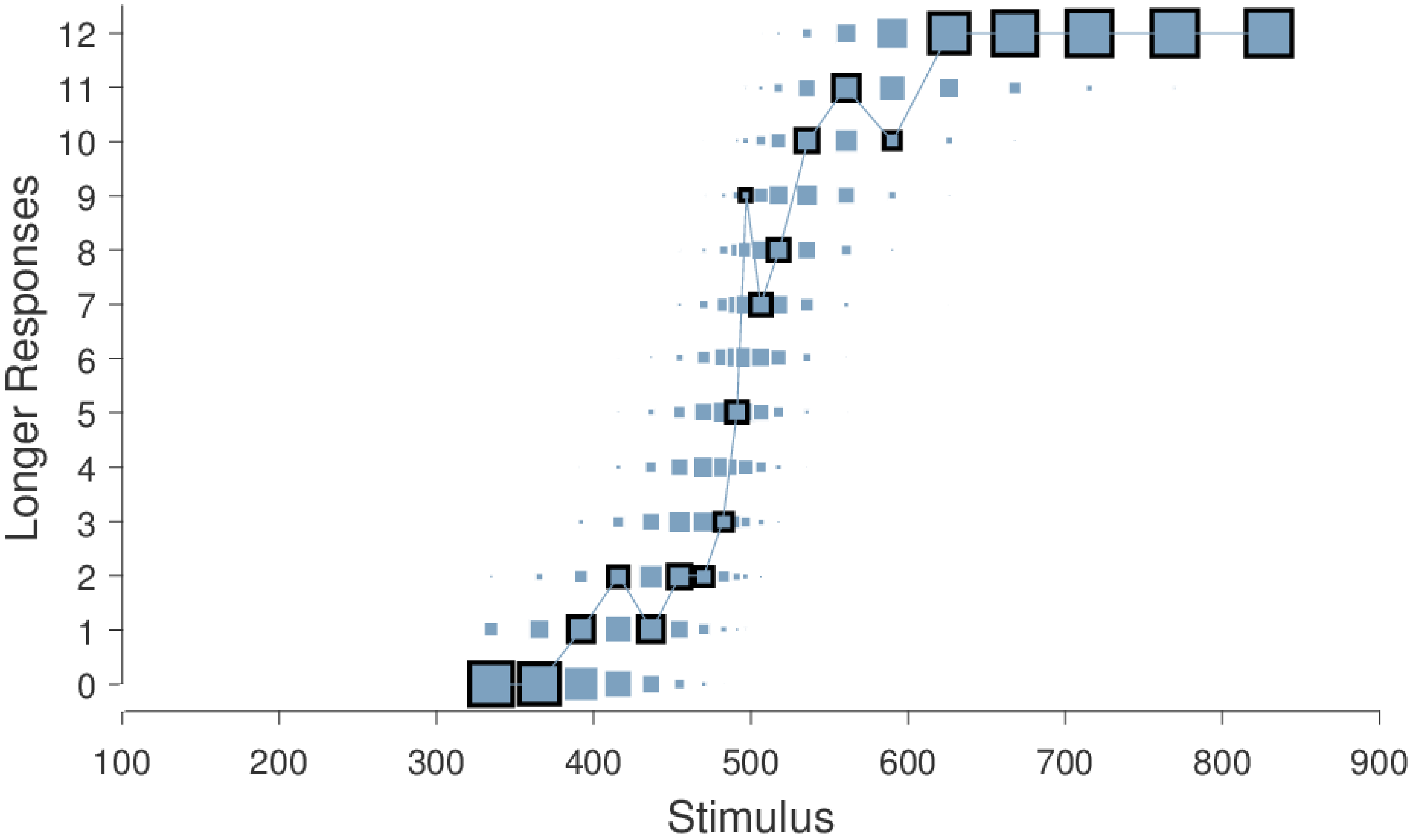
$$BF_{ab} = \frac{p\left(y \mid M_a\right)}{p\left(y \mid M_b\right)}$$

Model testing

- For nested models, the Savage-Dickey method estimates of Bayes factors from prior and posterior distributions
  - e.g., BF about 7 in favor of no-bias for shift, and massively in favor of non-step-function for scale
- Conceptually, Bayes factors are possible for non-nested models, but often harder to estimate in practice



Posterior predictive



Prediction and description

- The agreement between the observed data and posterior predictive distribution is often called the “goodness-of-fit” or “fit” in both Bayesian and other modeling approaches

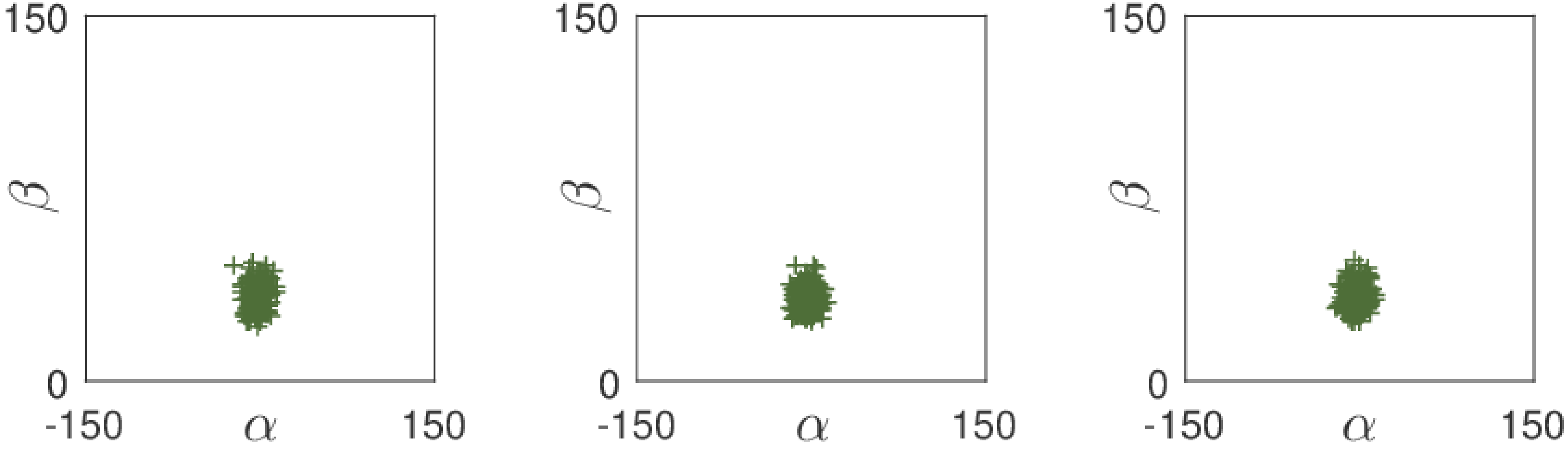
- We think these terms are very unhelpful (Roberts & Pashler, 2000), especially when researchers call fitted values “predictions” in their papers

- The idea of fit leads to bad but pervasive practices: “[t]o formally test their theory, mathematical psychologists rely on their model's ability to fit behavioral data” (Turner et al, 2016)

- The emphasis should be on the model’s ability to **predict** the data (via the prior predictive, on which Bayes factors are based) and not its ability to re-describe the data once it has already seen it

Testing sensitivity to priors

- It is true that the inferences depend on the choice of priors
  - This is desirable, since they are modeling assumptions
- If there is a range of possible priors consistent with an (imprecise, under-developed) theory, it is important to test the robustness or sensitivity of inferences to the range
- Posterior distributions for 3 other priors in a reasonable range are shown



Testing sensitivity to the likelihood

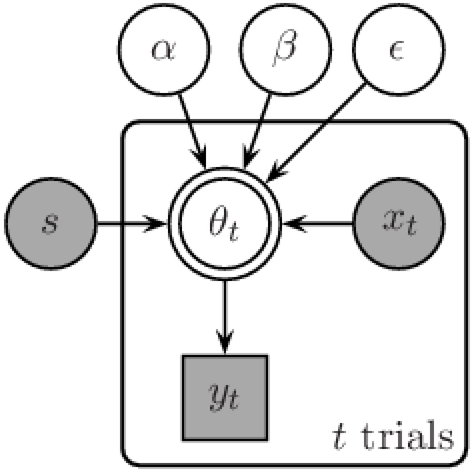
- It is also true that the inferences depend on the choice of likelihood function
  - This is also desirable, since they are modeling assumptions
- Researchers who are very concerned about the influence of priors seem not to worry about the influence of equally arbitrary choices in the likelihood
  - These sorts of sensitivities should be tested in the same way, and for the same reasons, as priors

Learner (1983, p. 37) had it right

- The difference between a fact and an opinion for purposes of decision making and inference is that when I use opinions, I get uncomfortable. I am not too uncomfortable with the opinion that error terms are normally distributed because most econometricians make use of that assumption.
- This observation has deluded me into thinking that the opinion that error terms are normal may be a fact, when I know deep inside that normal distributions are actually used only for convenience.
- In contrast, I am quite uncomfortable using a prior distribution, mostly I suspect because hardly anyone uses them.
- If convenient prior distributions were used as often as convenient sampling distributions, I suspect that I could be as easily deluded into thinking that prior distributions are facts as I have been into thinking that sampling distributions are facts.

Sensitivity to a form of sequential dependency

- One strong assumption in the current model is the independence of trials
  - There is no influence from one trial to the next
- To test the sensitivity of inferences to this assumption, consider a modified model that allows for a simple sequential dependency



$$\alpha \sim \text{Gaussian}(0, 1/50^2)$$

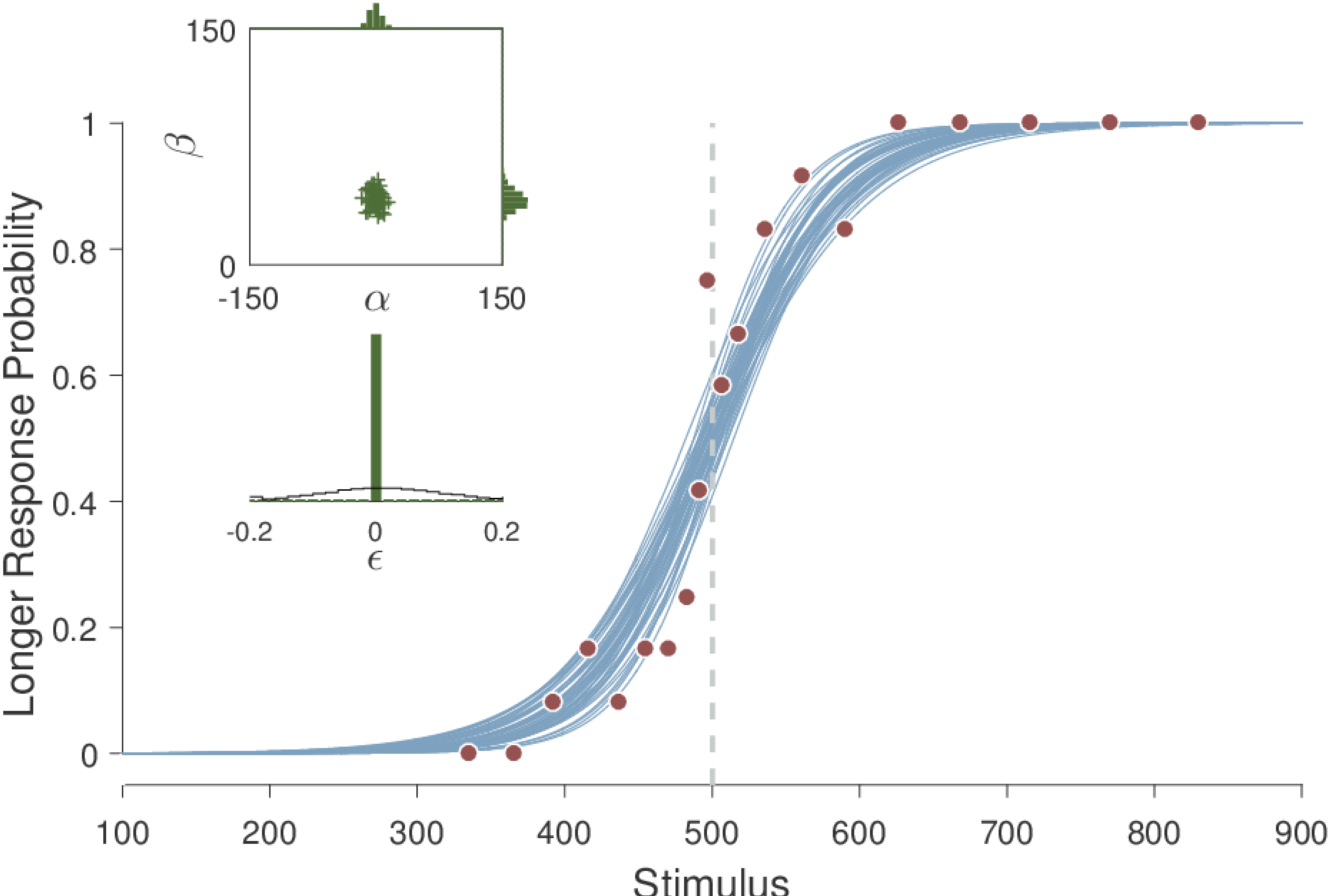
$$\beta \sim \text{TruncatedGaussian}_+(0, 1/100^2)$$

$$\epsilon \sim \text{Gaussian}(0, 100)$$

$$\theta_t = \begin{cases} 1/\left(1 + \exp\left(-\frac{x_t - s - \alpha}{\beta}\right)\right) + \epsilon & \text{if } y_{t-1} = 0, t > 1 \\ 1/\left(1 + \exp\left(-\frac{x_t - s - \alpha}{\beta}\right)\right) - \epsilon & \text{if } y_{t-1} = 1, t > 1 \end{cases}$$

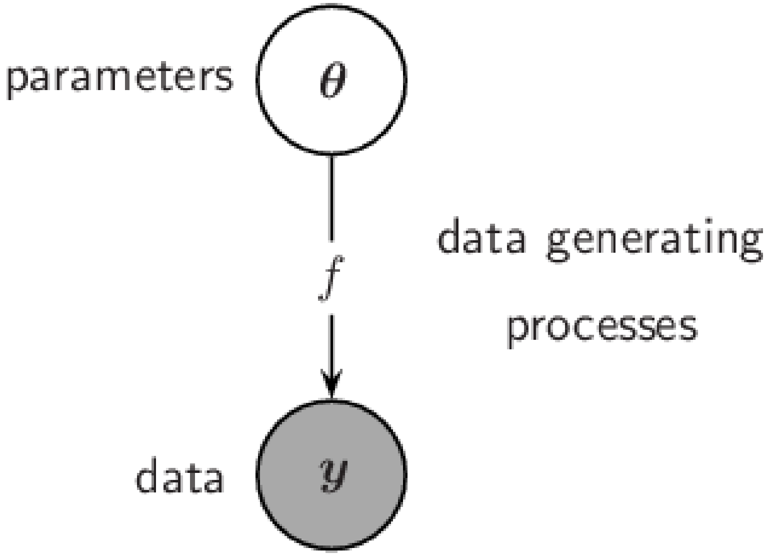
$$y_t \sim \text{Bernoulli}(\theta_t)$$

Inferences for sequential dependency model



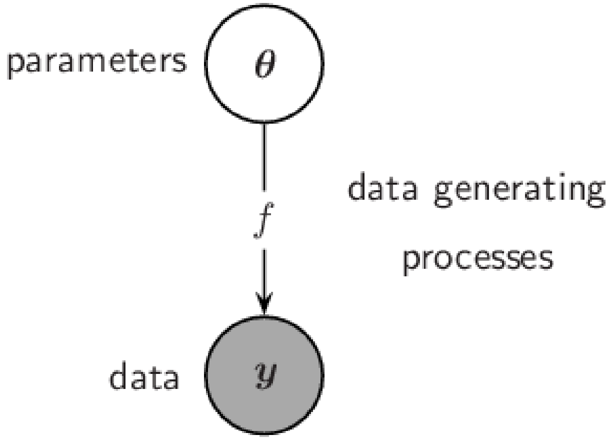
Flexibility of Bayesian methods

- Beyond the conceptual coherence and completeness, the great advantage of Bayesian methods is they allow cognitive that are more complicated than the standard one to be considered

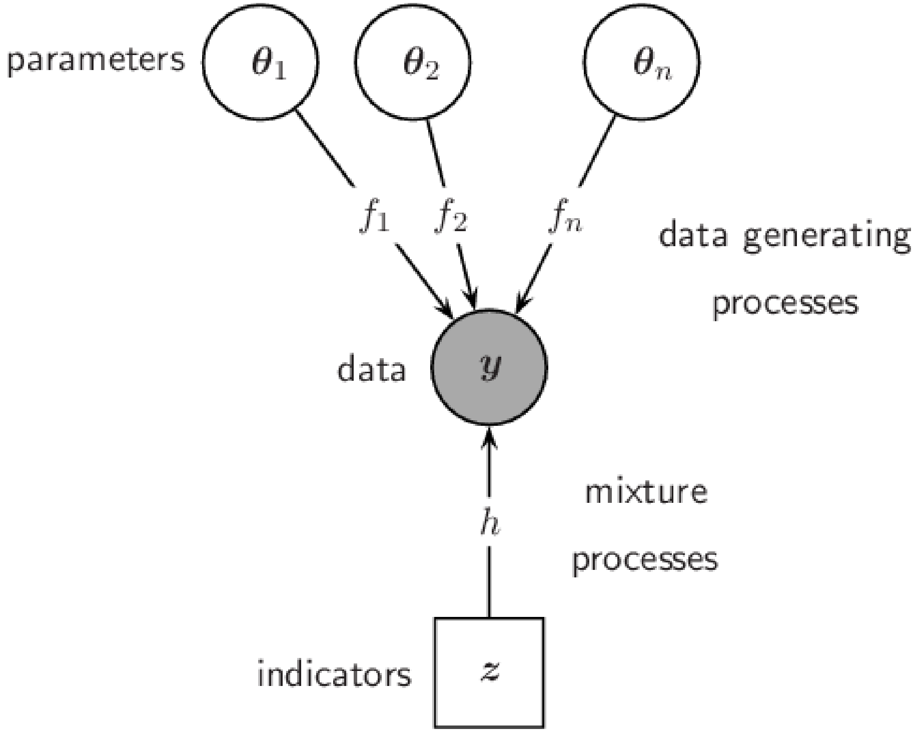


Latent-mixture modeling

- Latent-mixture models extend the standard approach by allowing behavioral data to be generated as a mixture of multiple different processes and controlling parameters

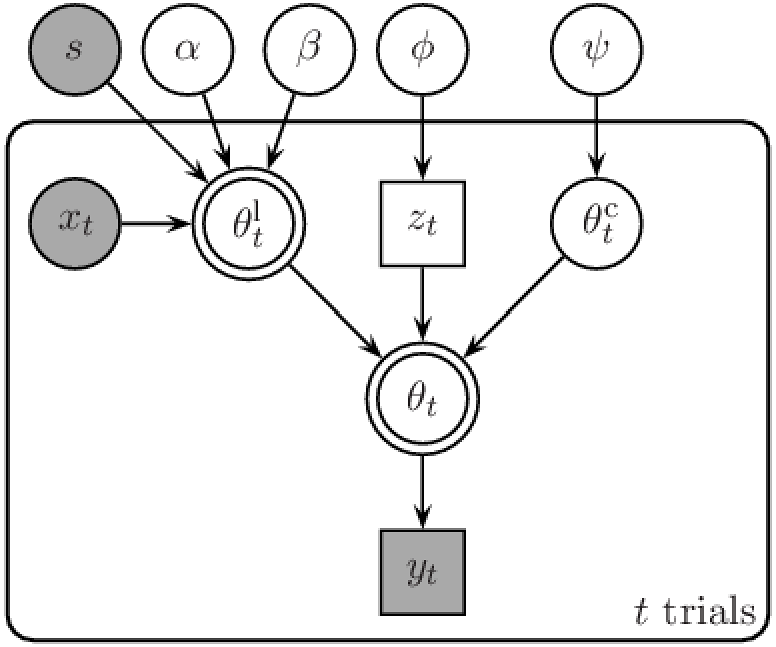


Latent Mixture



Contaminant model

- A latent-mixture model that allows the to use either the logistic model or a simple contaminant process on each trial
  - A discrete model indicator parameter *z* for each trial controls which process is used



$$\alpha \sim \text{Gaussian}(0, 1/50^2)$$

$$\beta \sim \text{TruncatedGaussian}_+(0, 1/100^2)$$

$$\theta^{\mathrm{l}}_t = 1/\left(1 + \exp\left(-\frac{x_t - s - \alpha}{\beta}\right)\right)$$

$$\psi \sim \text{Uniform}(0, 1)$$

$$\theta^{\mathrm{c}}_t \sim \text{Bernoulli}(\psi)$$

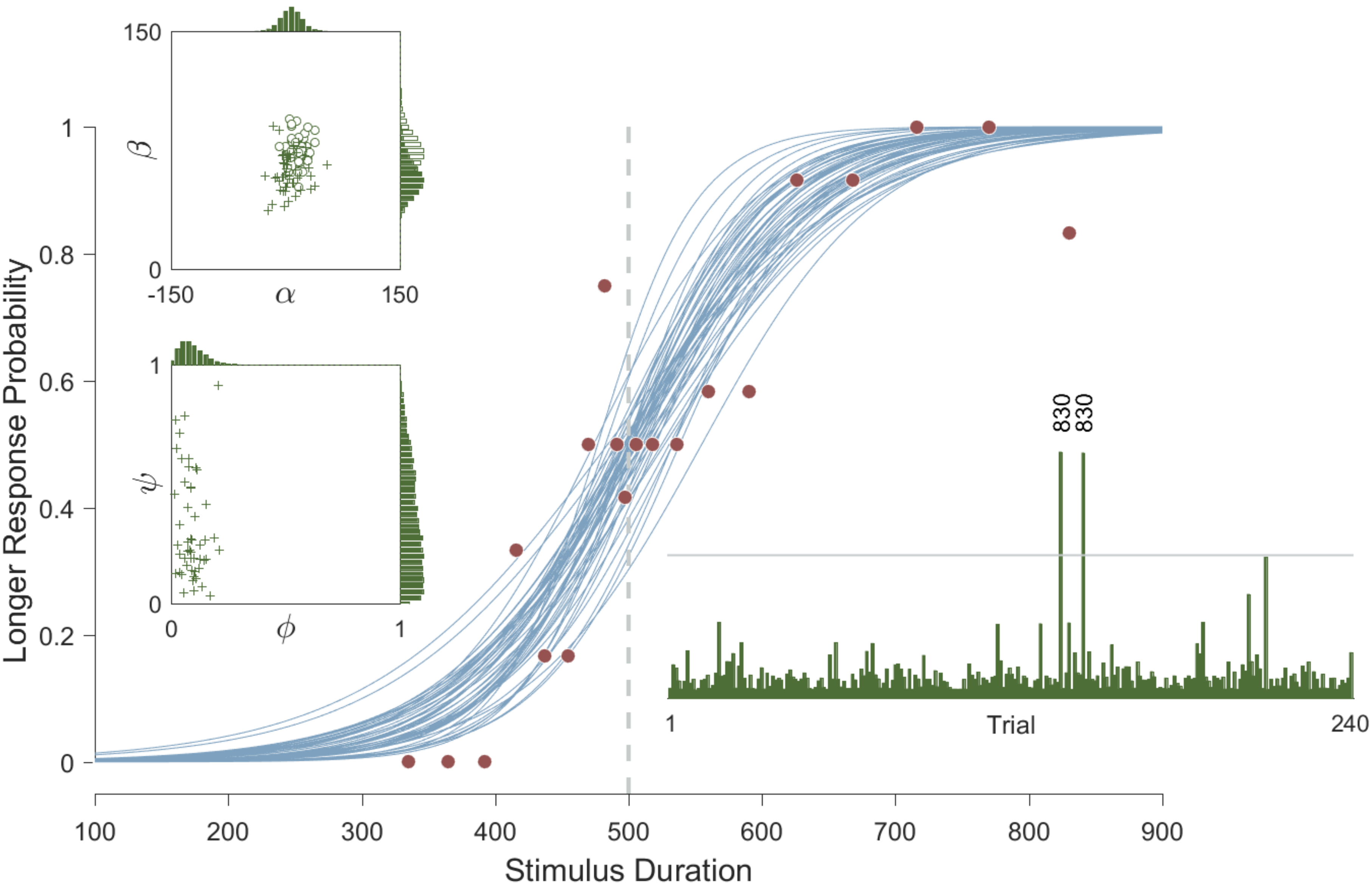
$$\phi \sim \text{Uniform}(0, 1)$$

$$z_t \sim \text{Bernoulli}(\phi)$$

$$\theta_t = \begin{cases} \theta^{\mathrm{l}}_t & \text{if } z_t = 0 \\ \theta^{\mathrm{c}}_t & \text{if } z_t = 1 \end{cases}$$

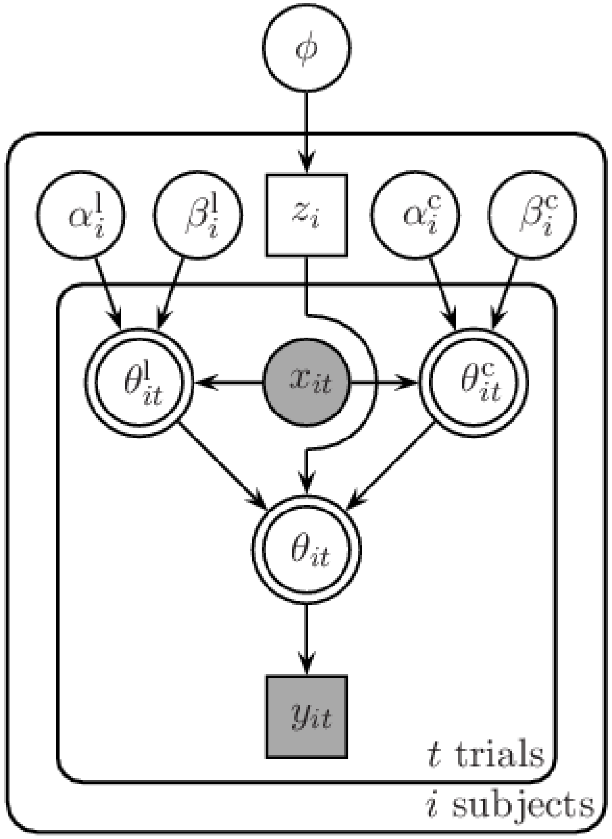
$$y_t \sim \text{Bernoulli}(\theta_t)$$

Inferences of contaminant model



Latent mixture logistic and Cauchy model

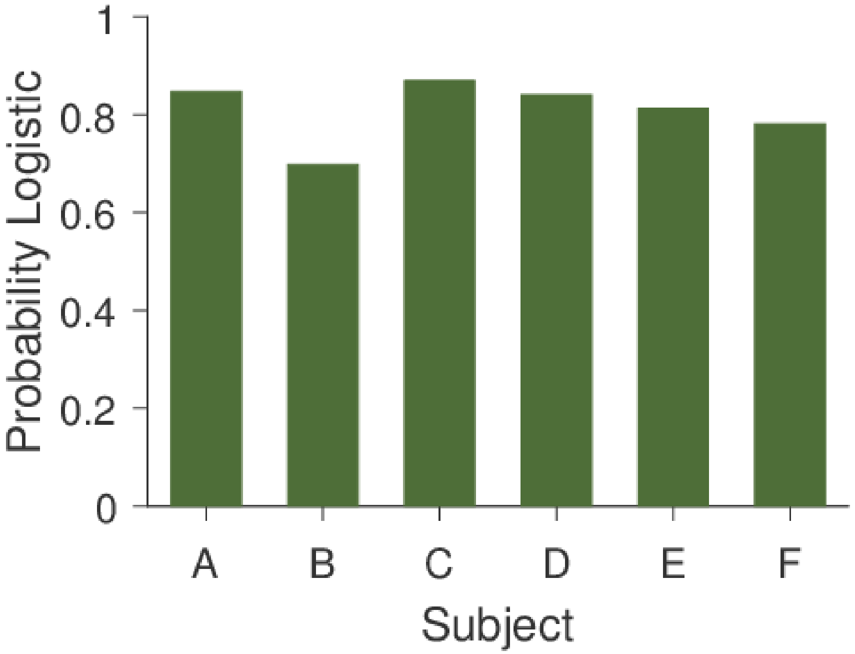
- A latent-mixture model that allows each subject to use either the logistic or Cauchy model
- Can be thought of as a form of model selection at the subject level, sensitive to all forms of model complexity

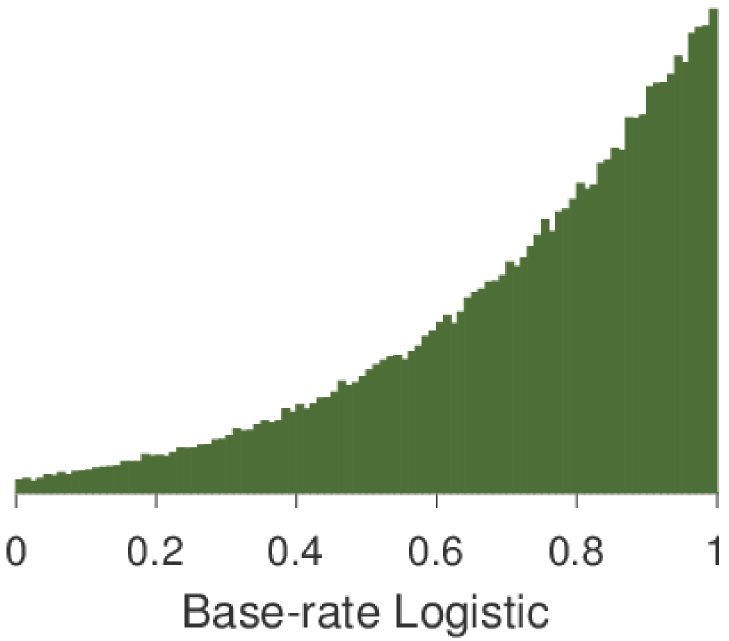


$$
\begin{aligned}
\alpha_i^{\text{l}} &\sim \text{Gaussian}(0, 1/50^2) \\
\beta_i^{\text{l}} &\sim \text{TruncatedGaussian}_{+}(0, 1/100^2) \\
\theta_{it}^{\text{l}} &= 1/\left(1+\exp\left(-\tfrac{x_t - s - \alpha_i^{\text{l}}}{\beta_i^{\text{l}}}\right)\right) \\
\alpha_i^{\text{c}} &\sim \text{Gaussian}(0, 1/50^2) \\
\beta_i^{\text{c}} &\sim \text{TruncatedGaussian}_{+}(0, 1/100^2) \\
\theta_{it}^{\text{c}} &= \arctan\left(\tfrac{x_t - s - \alpha_i^{\text{c}}}{\beta_i^{\text{c}}}\right)/\pi + 0.5 \\
\phi &\sim \text{Uniform}(0, 1) \\
z_i &\sim \text{Bernoulli}(\phi) \\
\theta_{it} &= \begin{cases} \theta_{it}^{\text{l}} & \text{if } z_i = 0 \\ \theta_{it}^{\text{c}} & \text{if } z_i = 1 \end{cases} \\
y_{it} &\sim \text{Bernoulli}(\theta_{it})
\end{aligned}
$$

Inferences for visual condition

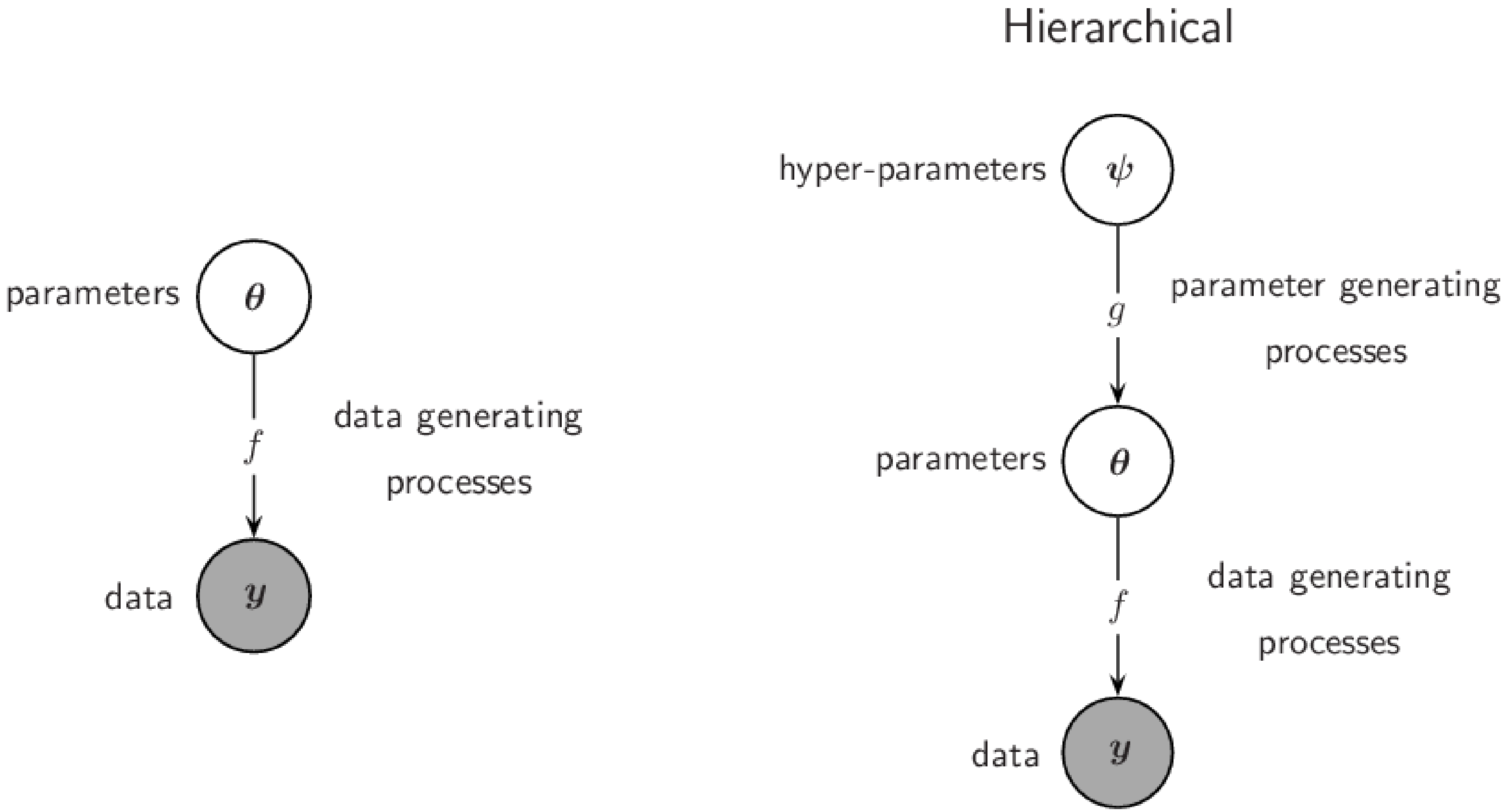
- Inferences for the six subjects show the posterior probability of using each model
- Also inferred is the base-rate of model use





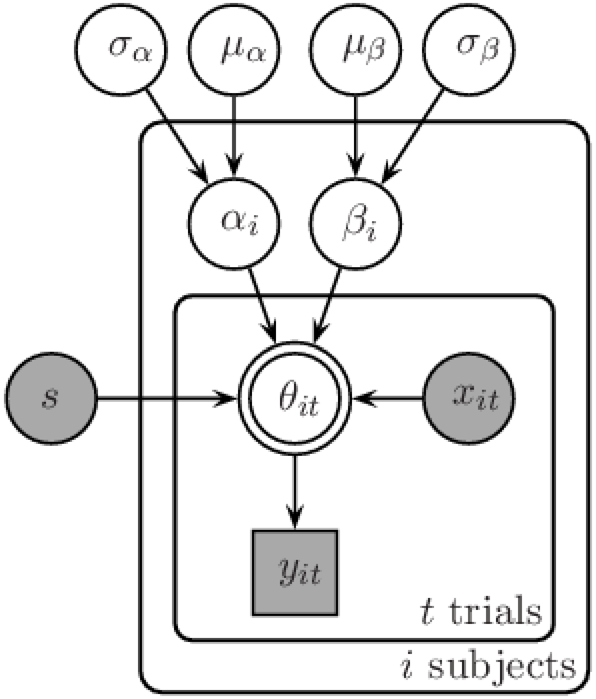
Hierarchical modeling

- Hierarchical models extend the standard approach by including a modeling account of how the basic model parameters themselves are generated



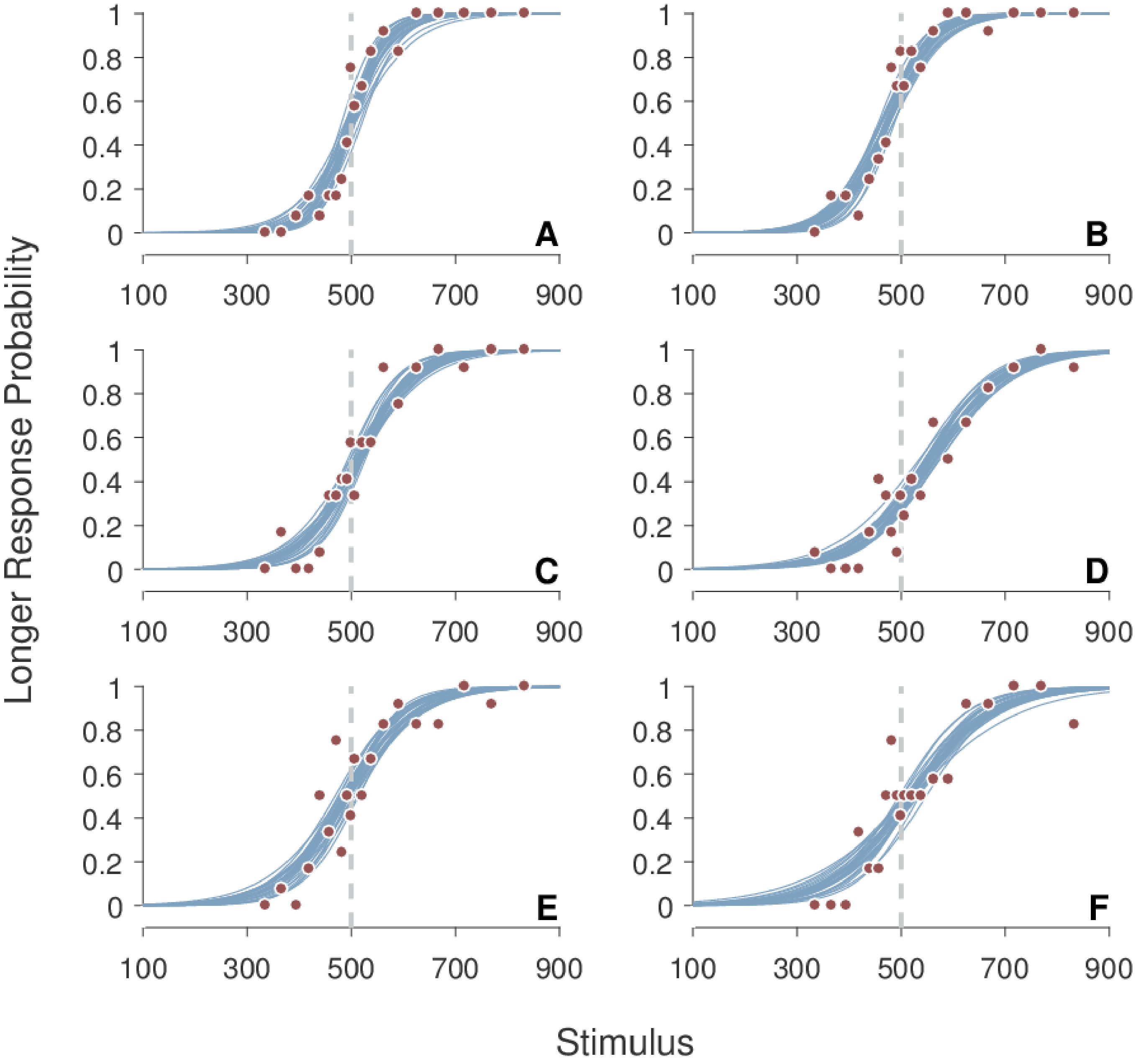
Hierarchical model

- A hierarchical model of multiple subjects that allows for structured individual differences between them
  - models the group distributions that generates the individual-level parameters

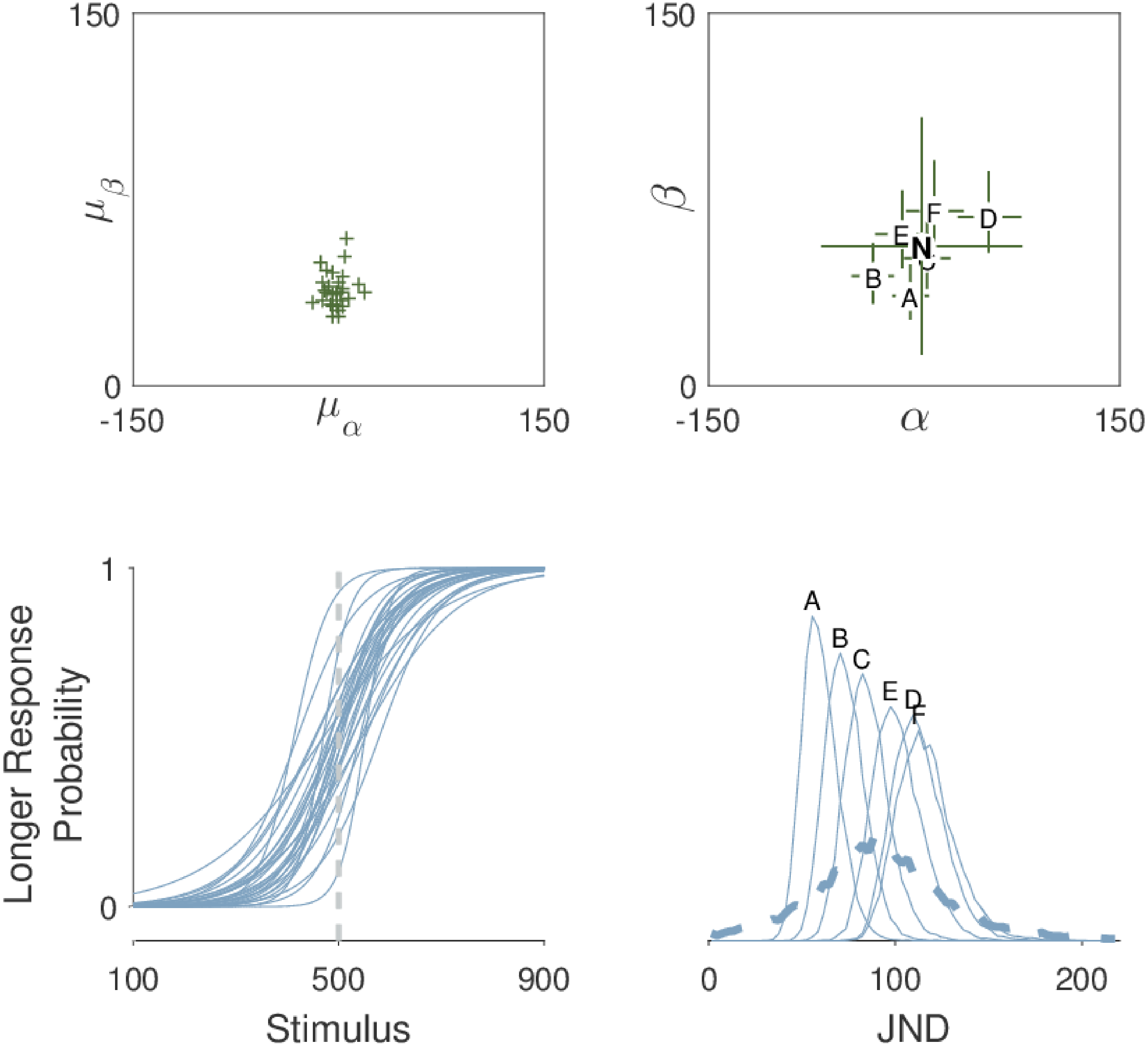


$$\mu_\alpha \sim \text{Gaussian}(0, 1/50^2)$$
$$\sigma_\alpha \sim \text{Uniform}(0, 50)$$
$$\mu_\beta \sim \text{TruncatedGaussian}_+(0, 1/100^2)$$
$$\sigma_\beta \sim \text{Uniform}(0, 100)$$
$$\alpha_i \sim \text{Gaussian}(\mu_\alpha, 1/\sigma_\alpha^2)$$
$$\beta_i \sim \text{TruncatedGaussian}_+(\mu_\beta, 1/\sigma_\beta^2)$$
$$\theta_{it} = 1/\left(1 + \exp\left(-\frac{x_t - s - \alpha_i}{\beta_i}\right)\right)$$
$$y_{it} \sim \text{Bernoulli}(\theta_{it})$$

Inferences of hierarchical model

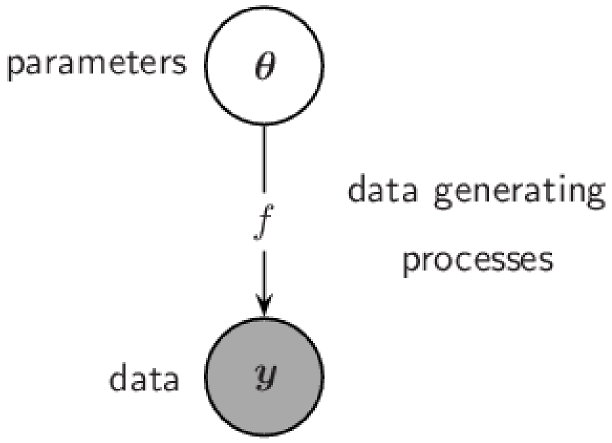


Generalization of hierarchical model to new subject

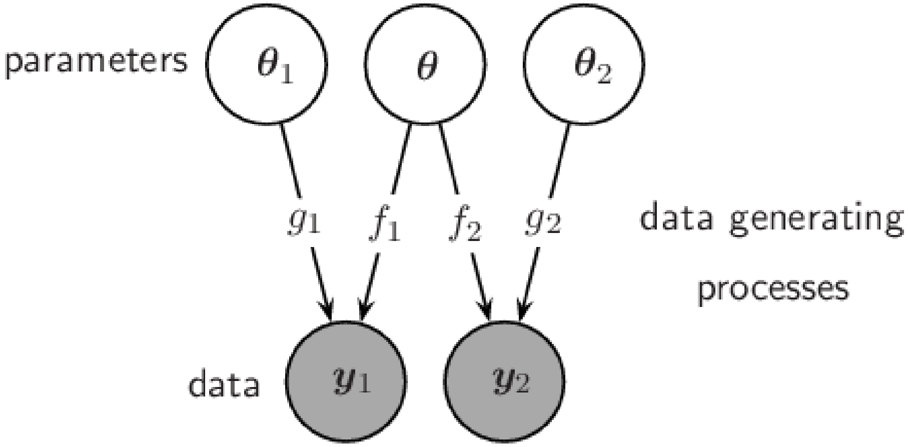


Common-cause modeling

- Common-cause models extend the standard approach by allowing the same psychological variables to influence multiple sorts of observed behavior

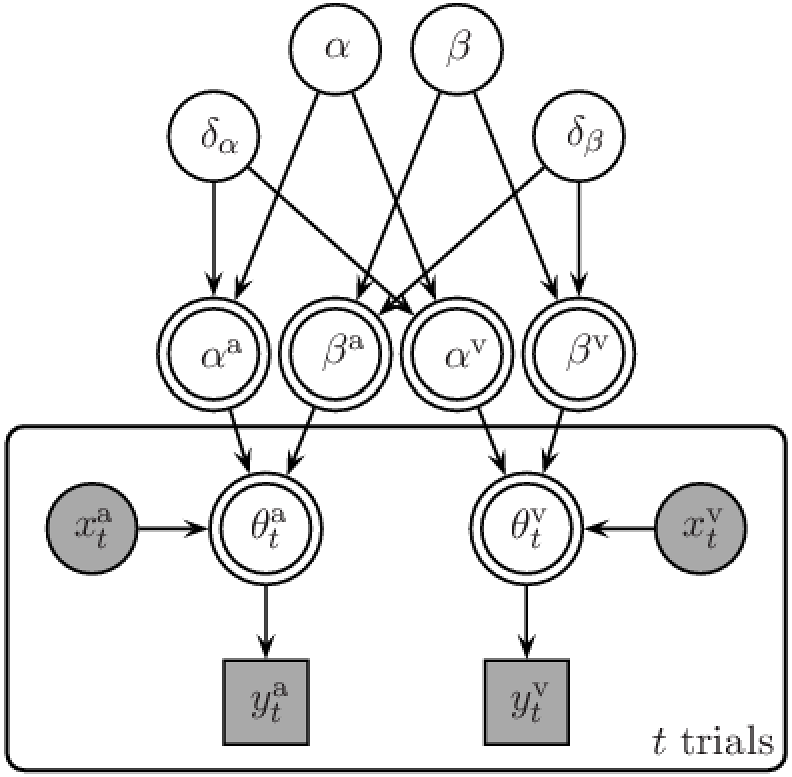


Common Cause



Effect of visual and auditory stimuli on shift and scale

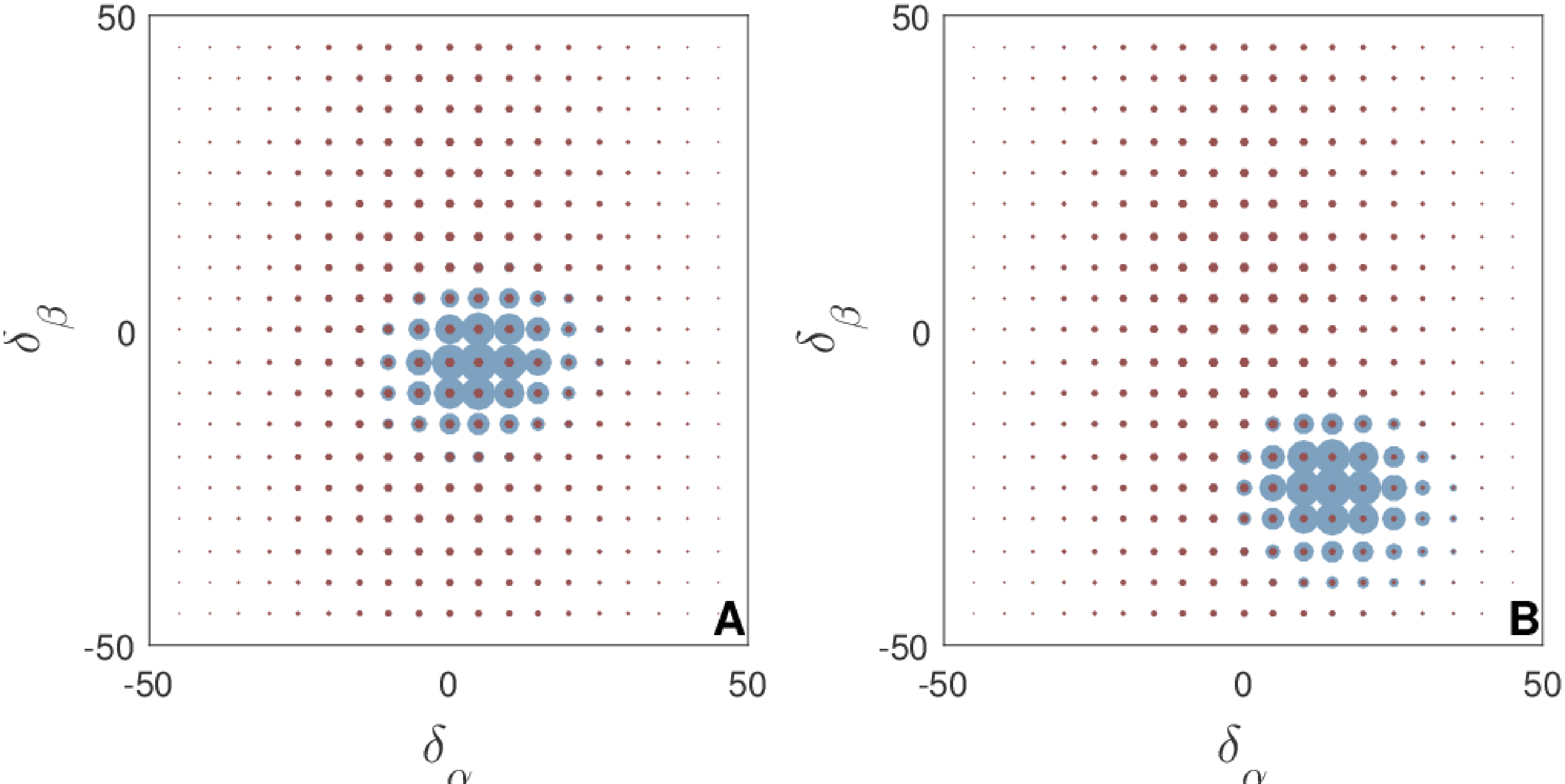
- A model that assumes effect size differences in shift and scale with the change from auditory to visual stimuli



$$\begin{aligned}
\alpha &\sim \text{Gaussian}(0, 1/50^2)\\
\beta &\sim \text{TruncatedGaussian}_+(0, 1/100^2)\\
\delta_\alpha &\sim \text{Gaussian}(0, 1/20^2)\\
\delta_\beta &\sim \text{Gaussian}(0, 1/40^2)\\
\alpha^{\mathrm{a}} &= \alpha + \tfrac{1}{2}\delta_\alpha\\
\alpha^{\mathrm{v}} &= \alpha - \tfrac{1}{2}\delta_\alpha\\
\beta^{\mathrm{a}} &= \beta + \tfrac{1}{2}\delta_\beta\\
\beta^{\mathrm{v}} &= \beta - \tfrac{1}{2}\delta_\beta\\
\theta_t^{\mathrm{a}} &= 1/\left(1 + \exp\left(-\frac{x_t^{\mathrm{a}} - s - \alpha^{\mathrm{a}}}{\beta^{\mathrm{a}}}\right)\right)\\
\theta_t^{\mathrm{v}} &= 1/\left(1 + \exp\left(-\frac{x_t^{\mathrm{v}} - s - \alpha^{\mathrm{v}}}{\beta^{\mathrm{v}}}\right)\right)\\
y_t^{\mathrm{a}} &\sim \text{Bernoulli}(\theta_t^{\mathrm{a}})\\
y_t^{\mathrm{v}} &\sim \text{Bernoulli}(\theta_t^{\mathrm{v}})
\end{aligned}$$

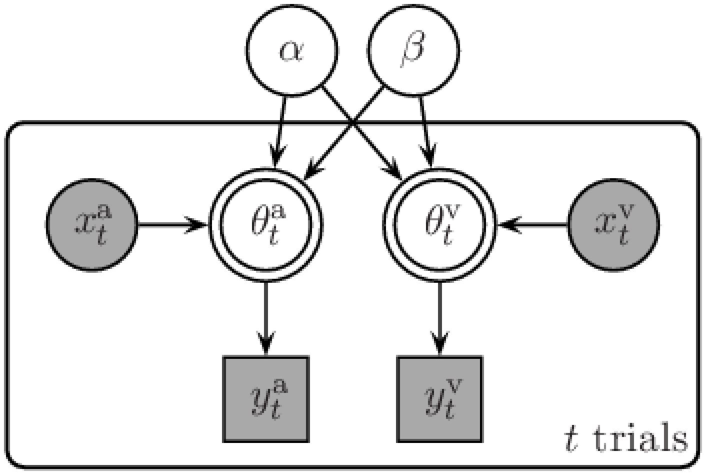
Inferences for two subjects

- The joint posterior for the two effect size parameters estimates a Bayes factor comparing no-effect vs effect models
  - and allows inferences about the size of the effects, if there is evidence they exist, as for subject B



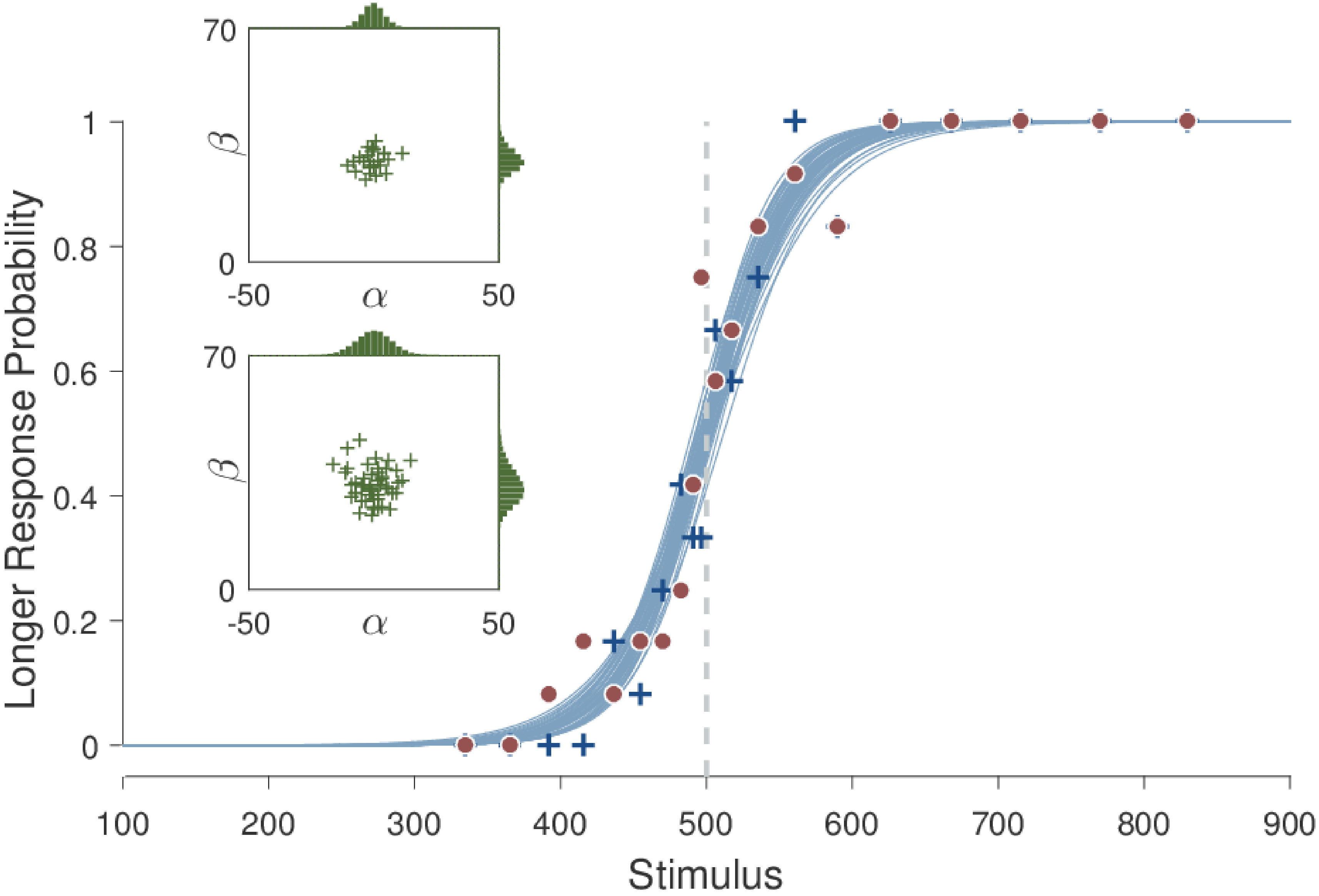
Common-cause model

- A common-cause model that assumes the same shift and scale parameters, via the same psychophysical function, apply to both modalities



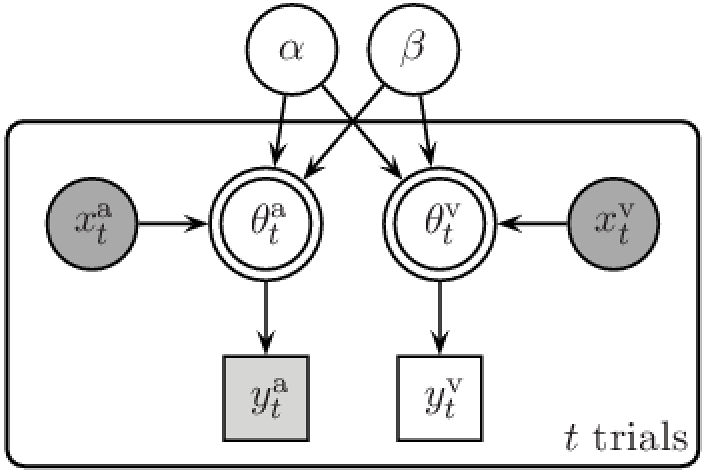
$$\begin{aligned}
\alpha &\sim \text{Gaussian}(0, 1/50^2)\\
\beta &\sim \text{TruncatedGaussian}_+(0, 1/100^2)\\
\theta_t^{\mathrm{a}} &= 1/\left(1+\exp\left(-\tfrac{x_t^{\mathrm{a}}-s-\alpha}{\beta}\right)\right)\\
\theta_t^{\mathrm{v}} &= 1/\left(1+\exp\left(-\tfrac{x_t^{\mathrm{v}}-s-\alpha}{\beta}\right)\right)\\
y_t^{\mathrm{a}} &\sim \text{Bernoulli}(\theta_t^{\mathrm{a}})\\
y_t^{\mathrm{v}} &\sim \text{Bernoulli}(\theta_t^{\mathrm{v}})
\end{aligned}$$

Inferences of common-cause model



Prediction and generalization

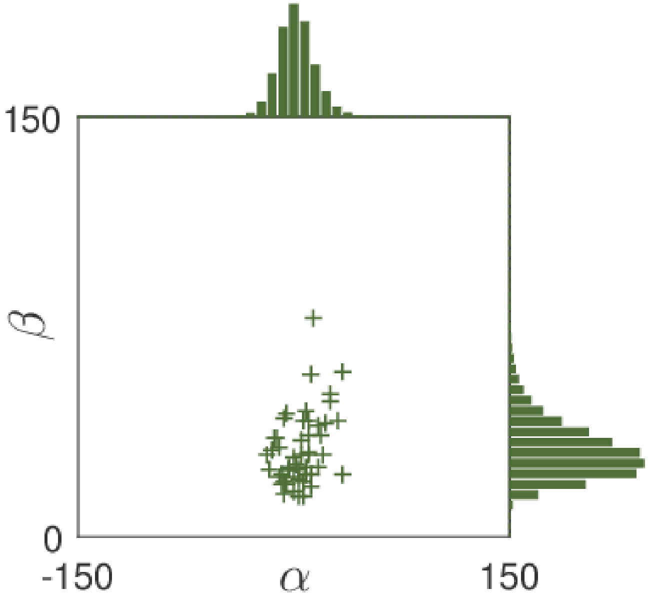
- An extension of the common-cause model, observing only the first 60 out of 240 trials in the auditory condition, to allow
  - **prediction** of the remaining 180 auditory trials
  - **generalization** to a different (related) task, for all pof the visual trials

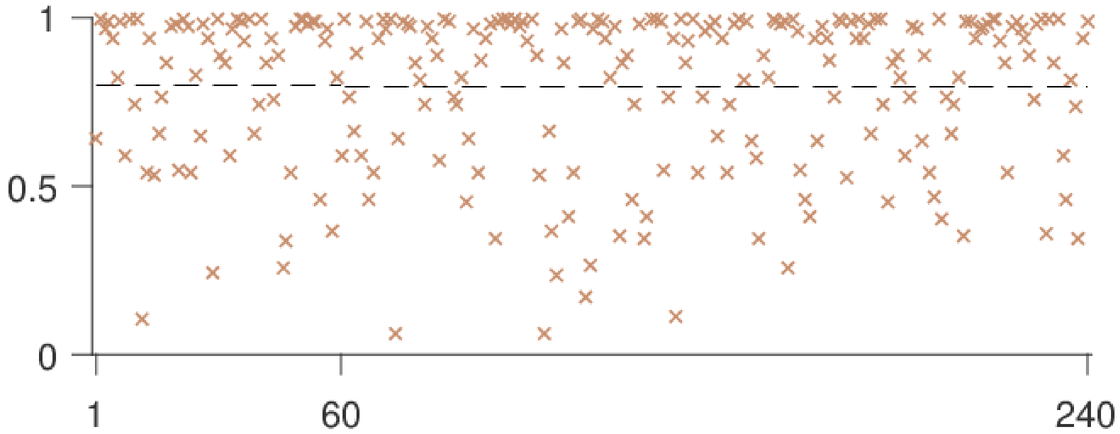


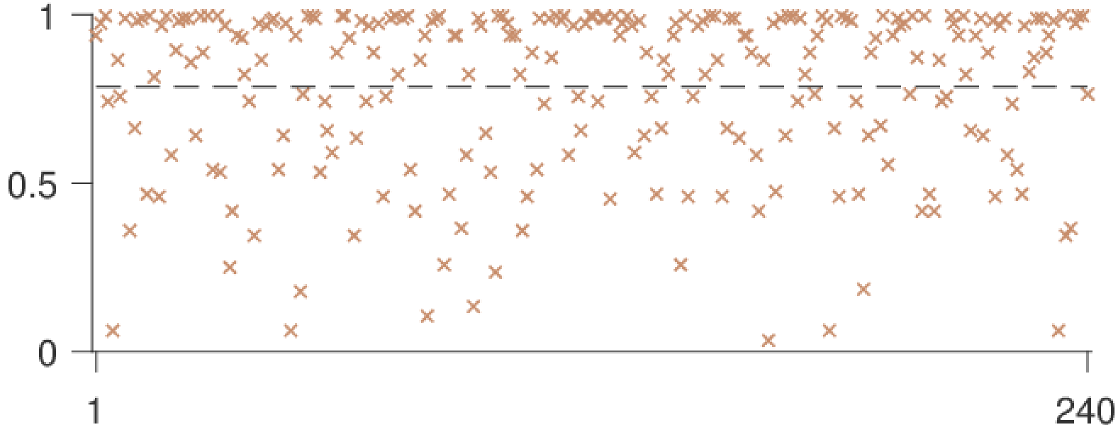
$$\begin{aligned}
\alpha &\sim \mathrm{Gaussian}(0, 1/50^2)\\
\beta &\sim \mathrm{TruncatedGaussian}_+(0, 1/100^2)\\
\theta_t^{\mathrm{a}} &= 1/\left(1+\exp\left(-\tfrac{x_t^{\mathrm{a}}-s-\alpha}{\beta}\right)\right)\\
\theta_t^{\mathrm{v}} &= 1/\left(1+\exp\left(-\tfrac{x_t^{\mathrm{v}}-s-\alpha}{\beta}\right)\right)\\
y_t^{\mathrm{a}} &\sim \mathrm{Bernoulli}(\theta_t^{\mathrm{a}})\\
y_t^{\mathrm{v}} &\sim \mathrm{Bernoulli}(\theta_t^{\mathrm{v}})
\end{aligned}$$

Prediction and generalization inference

- If the most likely alternative becomes a forced-choice, about 80% of the predictions and generalizations turn out to be right



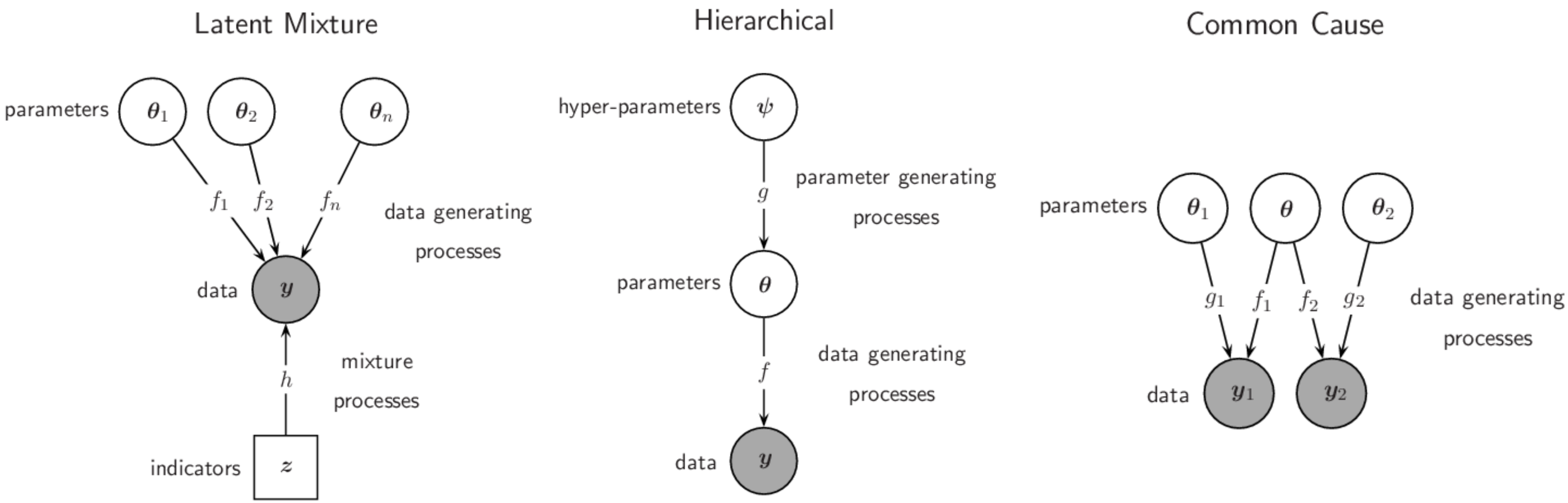




Conclusion

Theoretical freedom …

- Statistical methods for relating models to data are fundamental in science, since they allow
  - parameter inference, data prediction, and model evaluation
- Bayesian methods allow theorists to develop, evaluate, and use **richer** generative models of how psychological variables and processes generate behavior



… with rigorous assessment …

- The coherent and general foundations of Bayesian statistical methods allow them to “scale up” to these theoretically highly expressive models
- Bayesian methods allow rigorous evaluation in terms of observation, and will naturally reign in models that do not describe and predict data well
  - Models that formalize good additional theoretical assumptions will be simpler, and Bayesian evaluation is sensitive to this

… and flexible inferences

- Bayesian methods allow for inferences about
  - The uncertainty about parameters, or joint uncertainty about multiple parameters, or the uncertainty of one parameter conditional on the value of another, …
  - Inferences about both continuous values and discrete values, especially in latent mixtures
  - Inferences and prediction about partially observed parameters or data
  - Inferences and predictions about missing, partially observed or entirely unobserved parameters or data
  - Evaluation of any model against any other model, based on available information
  - …

Bayesian methods afford **theoretical freedom** with **rigorous assessment** and **flexible inferences**

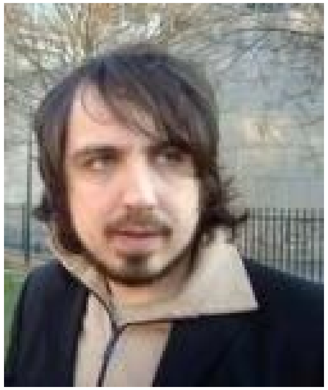
Room for improvement

- Better modeling of individual differences
  - Current models are largely statistical rather than psychological
  - This is fine as a place to start, but not to finish, and integration with psychometric theories is needed
- We need to construct informative priors as a matter of course
  - Who wants to read a paper by researchers who knew nothing about the key variables before they started
- Common-cause modeling should be everywhere
  - Important psychological variables should impact multiple behaviors, and we should break out of a “one model for one task” stovepipe

References and acknowledgments

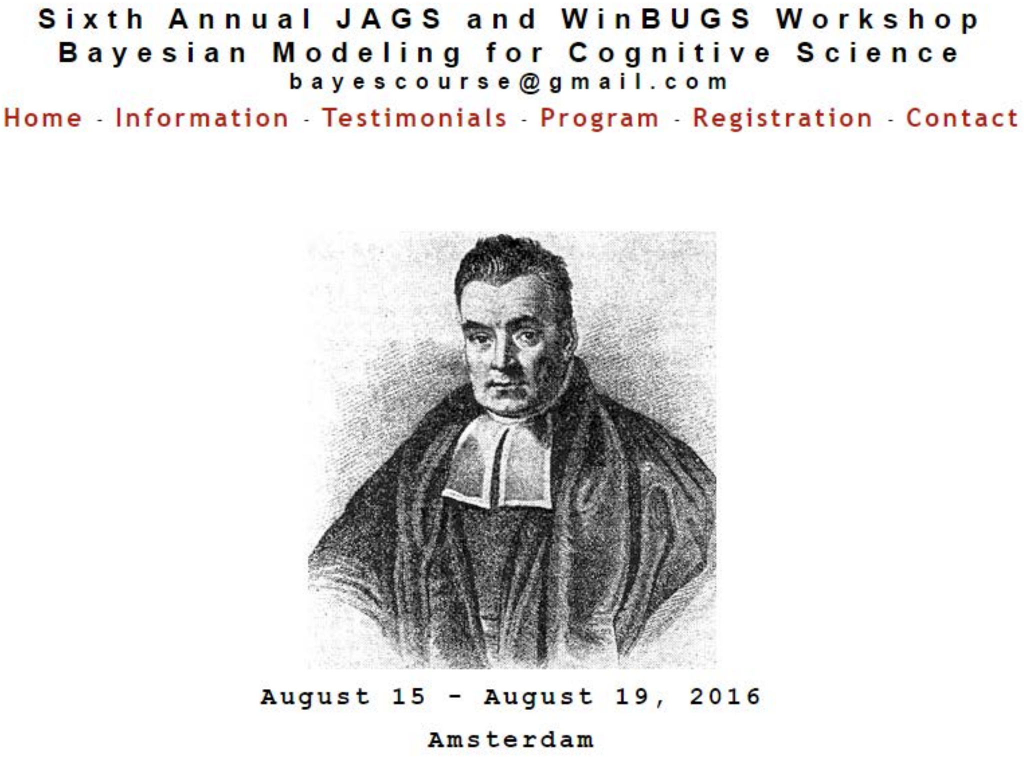
- Lee, M.D. (accepted). Bayesian methods in cognitive modeling. *The Stevens' Handbook of Experimental Psychology and Cognitive Neuroscience, Fourth Edition.*
  - Draft chapter, and all the code and data, available on the Open Science Framework [https://osf.io/zur8m/]

EJ Wagenmakers



Wolf Vanpaemel

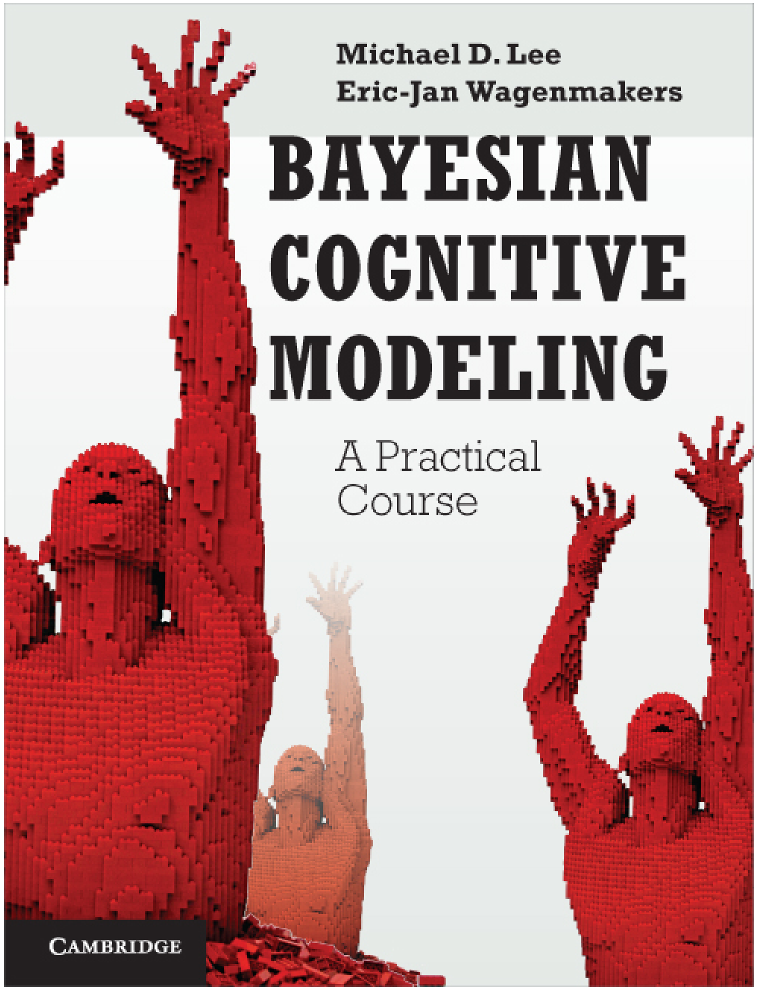
www.bayesmodels.com



Sixth Annual JAGS and WinBUGS Workshop
Bayesian Modeling for Cognitive Science
bayescourse@gmail.com

Home - Information - Testimonials - Program - Registration - Contact

August 15 - August 19, 2016
Amsterdam



Michael D. Lee
Eric-Jan Wagenmakers

BAYESIAN COGNITIVE MODELING

A Practical Course

CAMBRIDGE