Bayesian models for response times in cognitive experiments

Peter F. Craigmile

Department of Mathematics and Statistics

HUNTER

The City University of New York

http://craigmile.com/peter/

Functional data analysis working group (FDAWG),

Department of Biostatistics, Columbia University

March 5, 2024

Support: National Science Foundation

Collaborators

Sungmin Kim, Waymo

Deborah Kunkel, Clemson University

Kevin Potter, Massachusetts General Hospital

Zhifei Yan, Google

Mario Peruggia, The Ohio State University

Trisha Van Zandt, The Ohio State University

Response time (RT) – a measure of human performance

- Used as a window on psychological processes for almost two centuries.
- Forms the foundation for most work in cognitive psychology.
  - Used to formulate theories of brain function and cognitive processing.
  - Employed as a basis to evaluate training regimens, user interface design, vehicle operation, and task design.
  - Can be used to evaluate medical conditions, especially schizophrenia, learning disorders, and other psychological disorders.

Response time data

- RT sequence: Times taken by an individual to respond to a sequence of stimuli
- Can be analyzed as time series data
- RT data show interesting features:

  1. changes over time at different scale levels
  2. marginal distributions with heavy tails

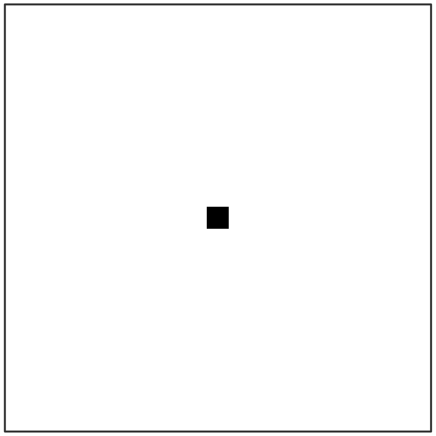
A dataset from Wagenmakers et al. [2004]

- A **"simple RT"** study with 6 subjects, to investigate **autocorrelation** structure across long sequences of trials
  - A randomly generated **response-stimulus interval (RSI)** is used to prevent anticipatory responses (extremely fast RTs)
  - Two different RSIs conditions:

    Short RSIs: $\sim$ U(550ms, 950ms)

    Long RSIs: $\sim$ U(1150ms, 1550ms)

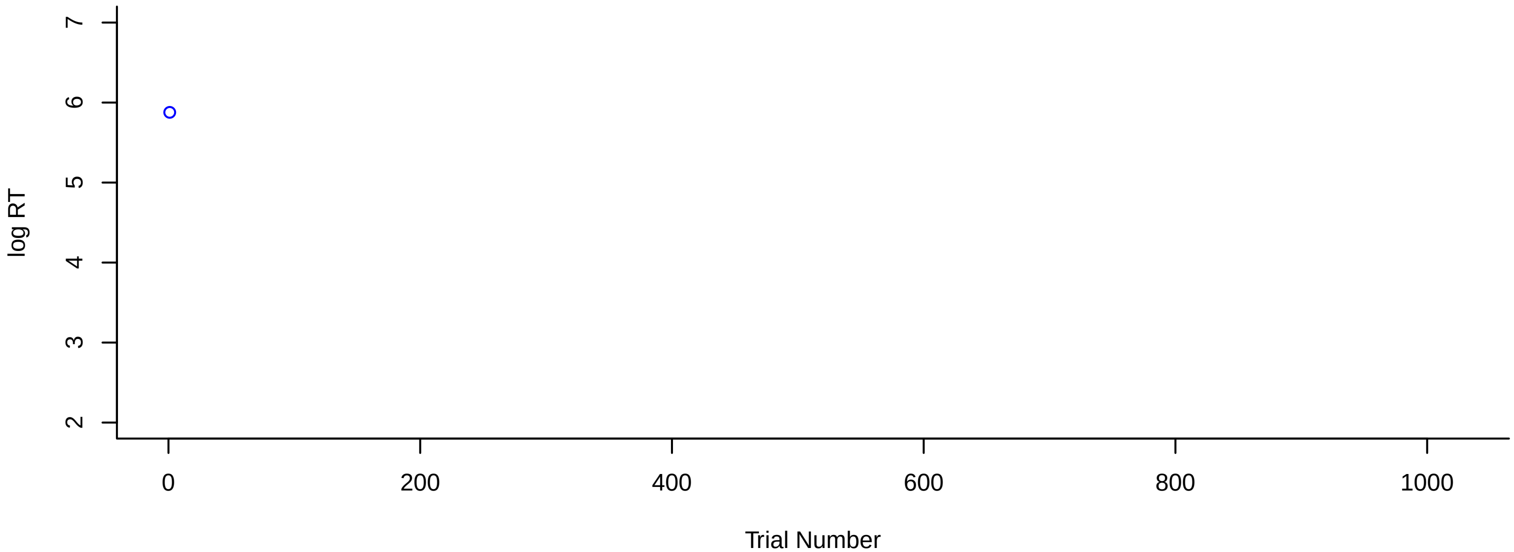
A dataset from Wagenmakers et al. [2004]



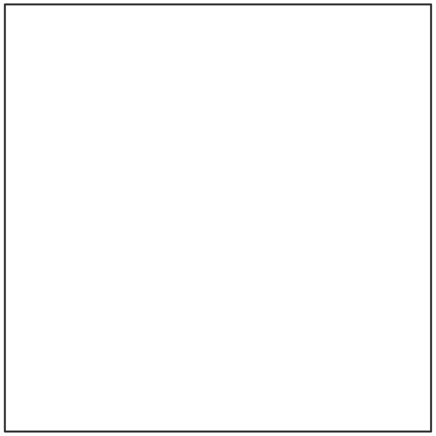
The subject sees the symbol on the screen

and presses the '/' key

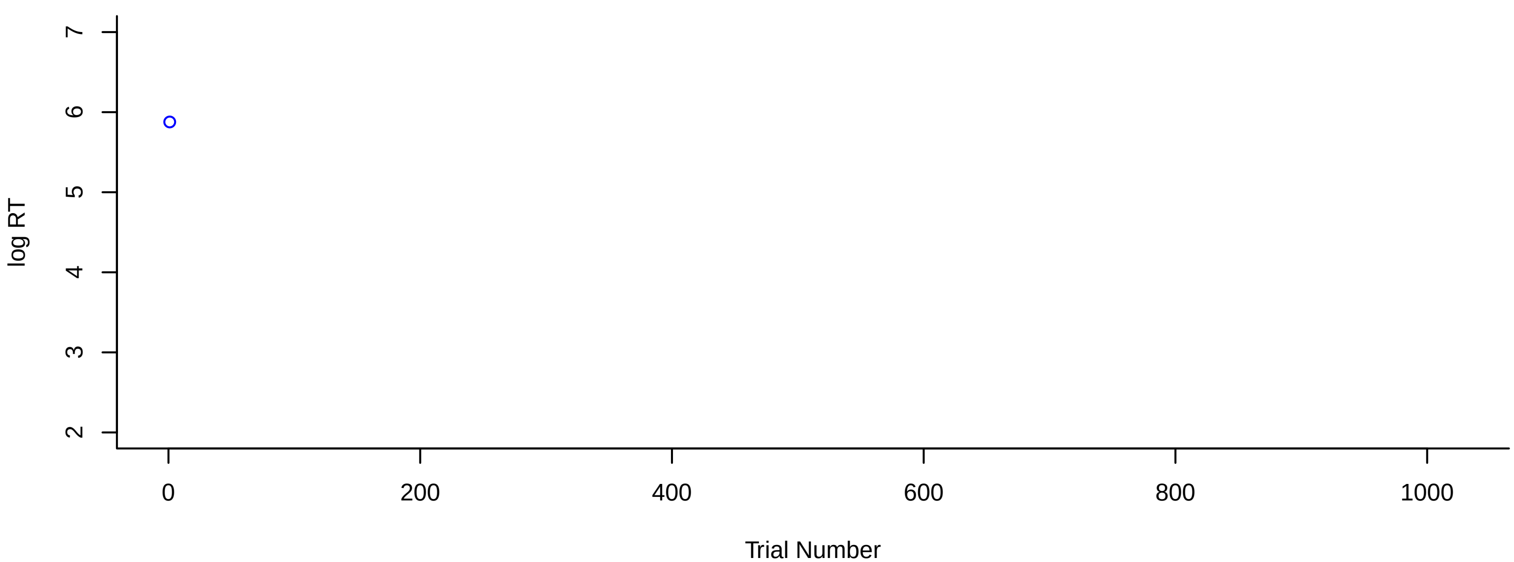
The time to respond is the first RT



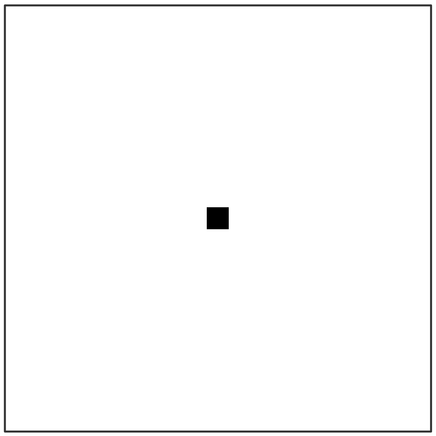
A dataset from Wagenmakers et al. [2004]



The subject then sees nothing for a randomly generated response-stimulus interval (RSI)



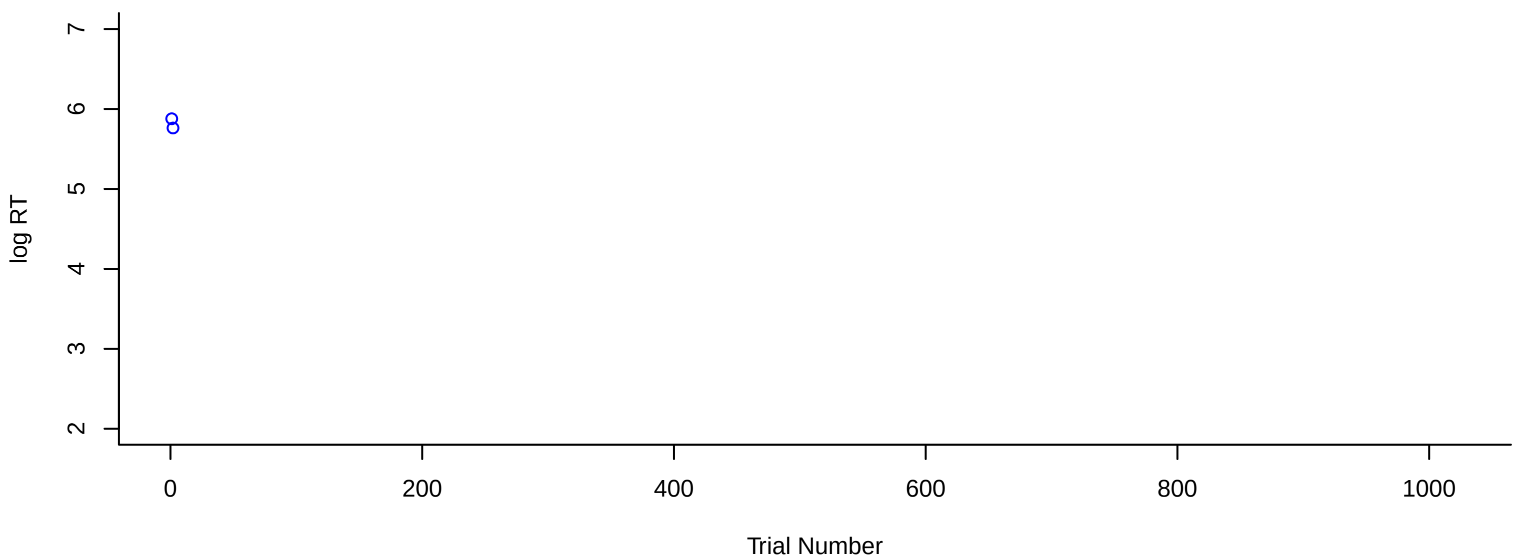
A dataset from Wagenmakers et al. [2004]



The subject sees the symbol again

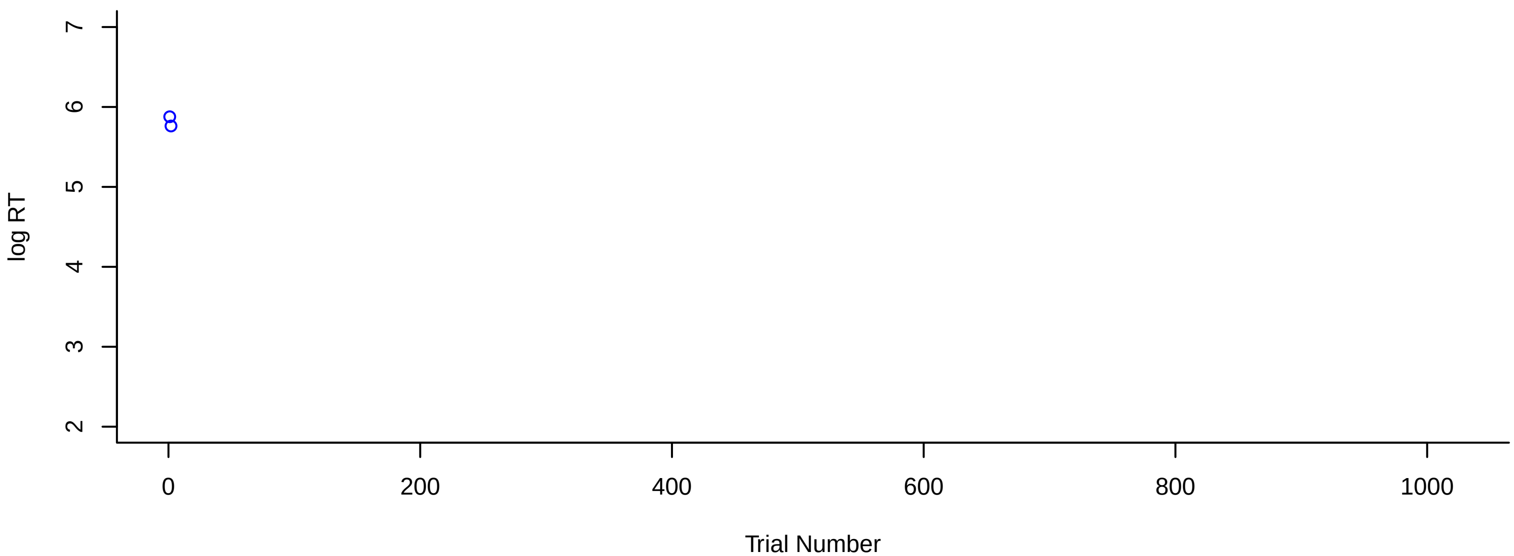
and presses the '/' key

The time to respond now is the second RT.



A dataset from Wagenmakers et al. [2004]

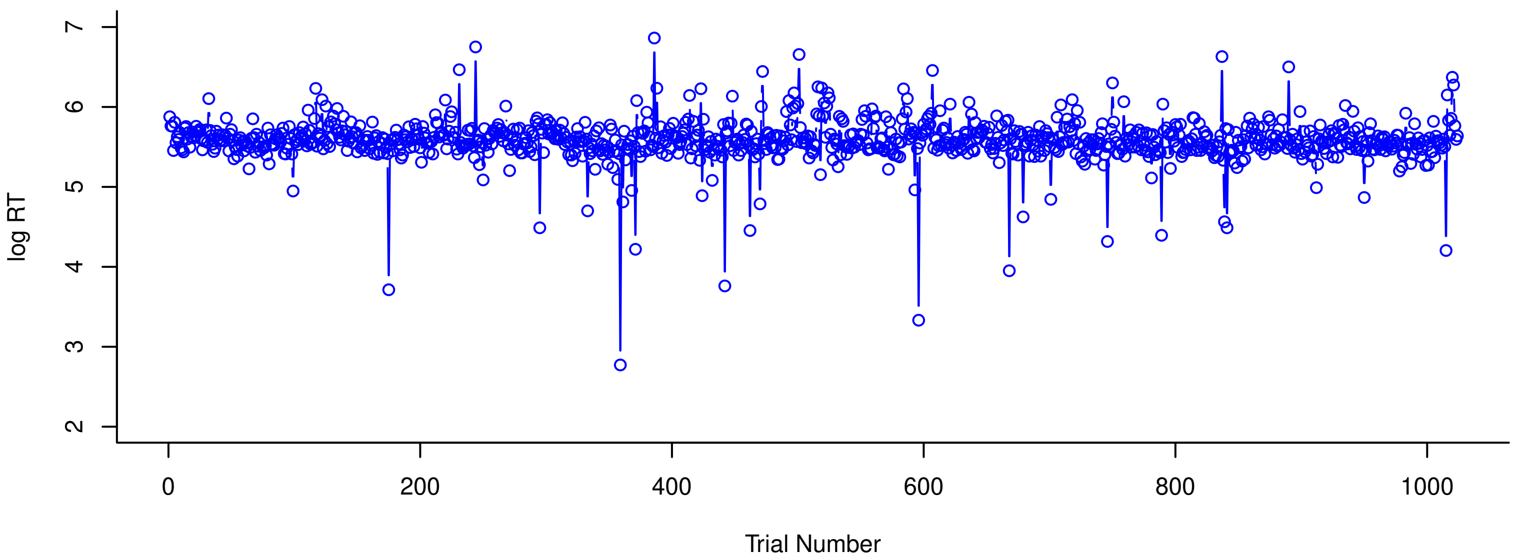
Then again nothing for another randomly generated response-stimulus interval (RSI)

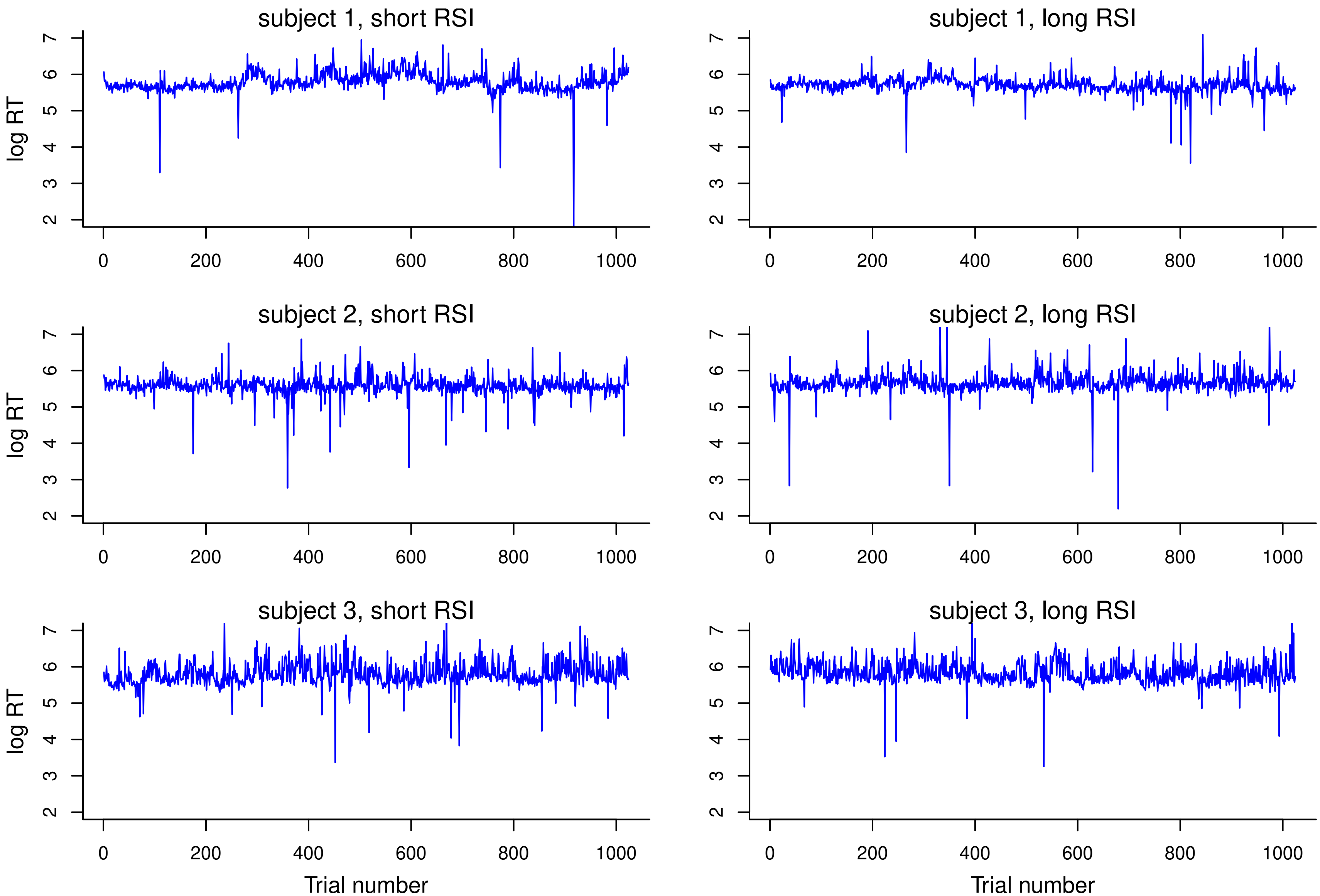


A dataset from Wagenmakers et al. [2004]

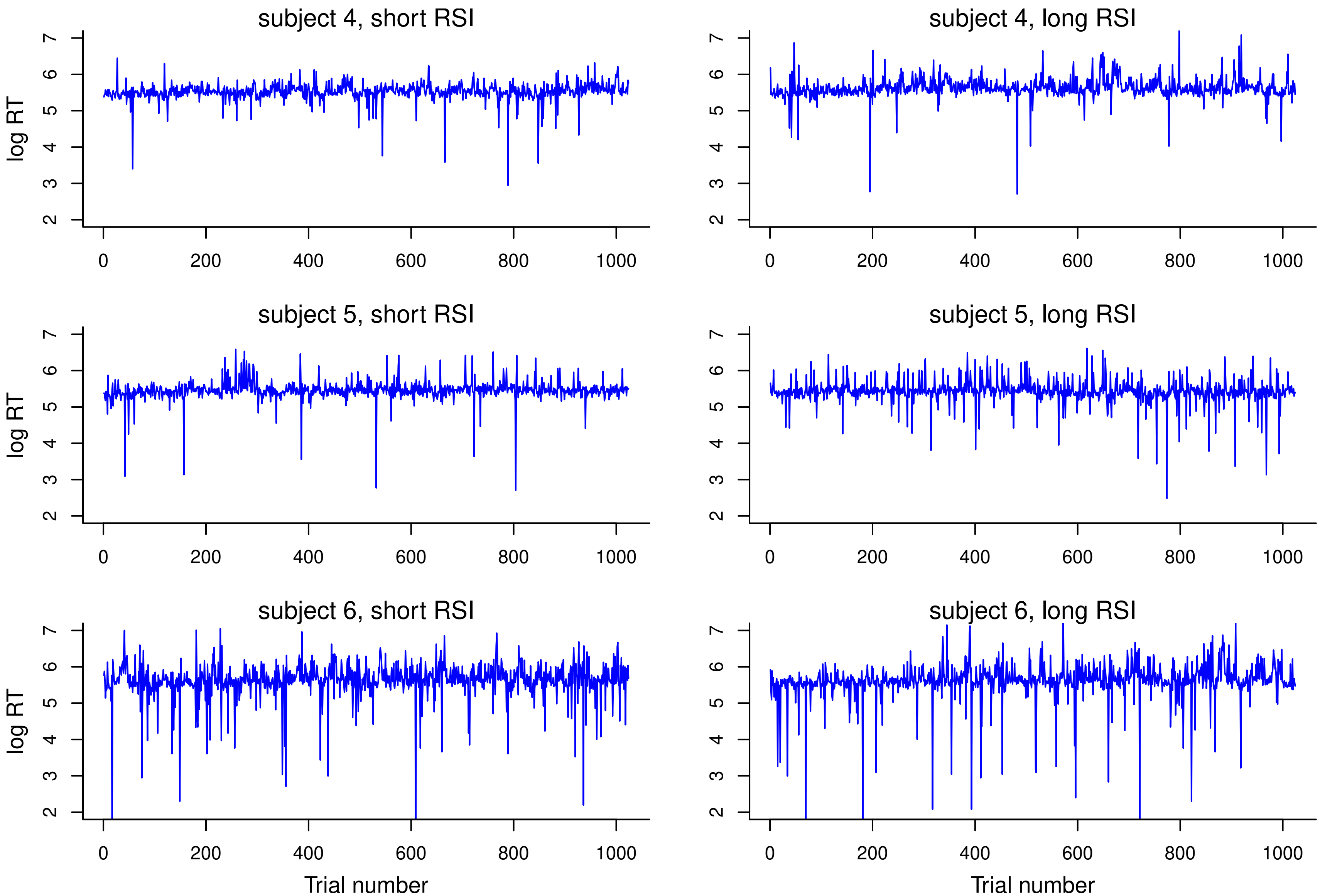
We continue in the same way until we collect 1024+24 trials per subject and RSI condition.

The first 24 trials are discarded.





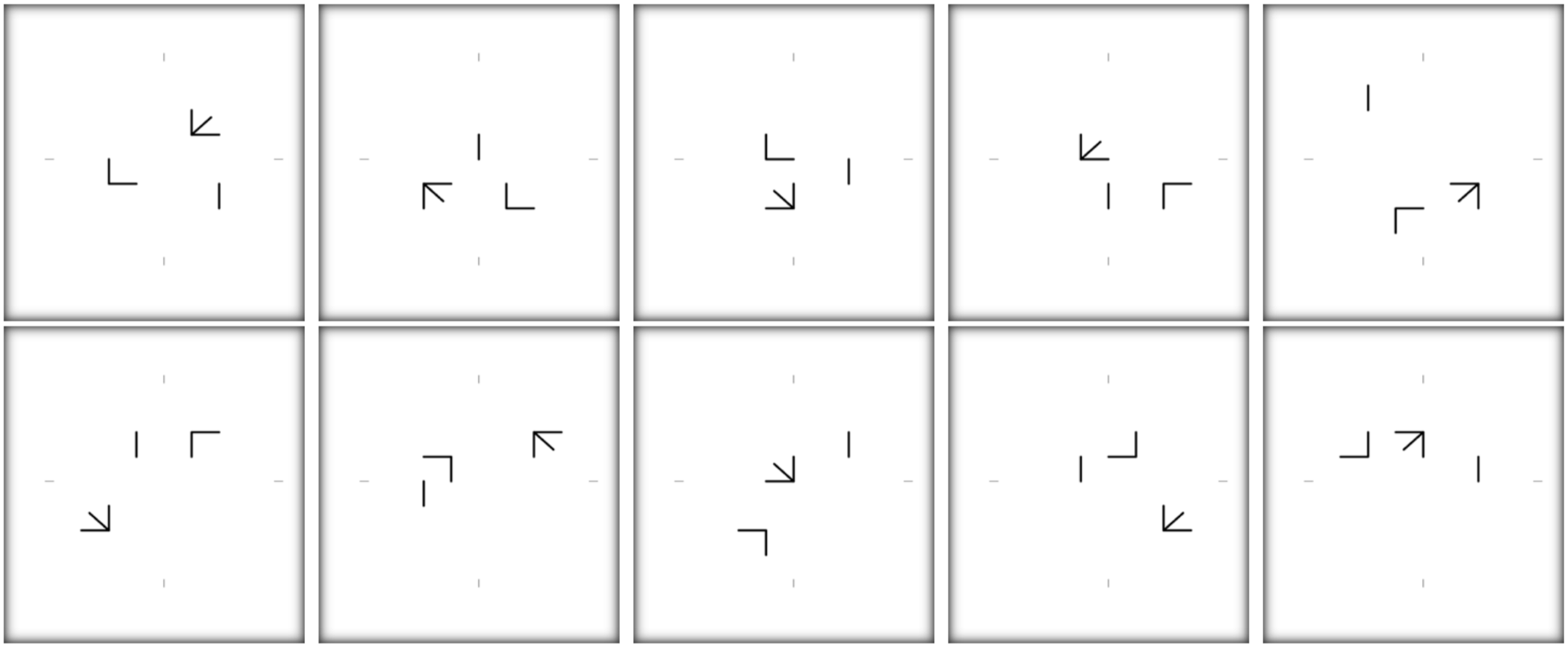
subject 1, short RSI
subject 1, long RSI
subject 2, short RSI
subject 2, long RSI
subject 3, short RSI
subject 3, long RSI
log RT
2
3
4
5
6
7
0
200
400
600
800
1000
Trial number


subject 4, short RSI
subject 4, long RSI
subject 5, short RSI
subject 5, long RSI
subject 6, short RSI
subject 6, long RSI
log RT
Trial number
0
200
400
600
800
1000
2
3
4
5
6
7

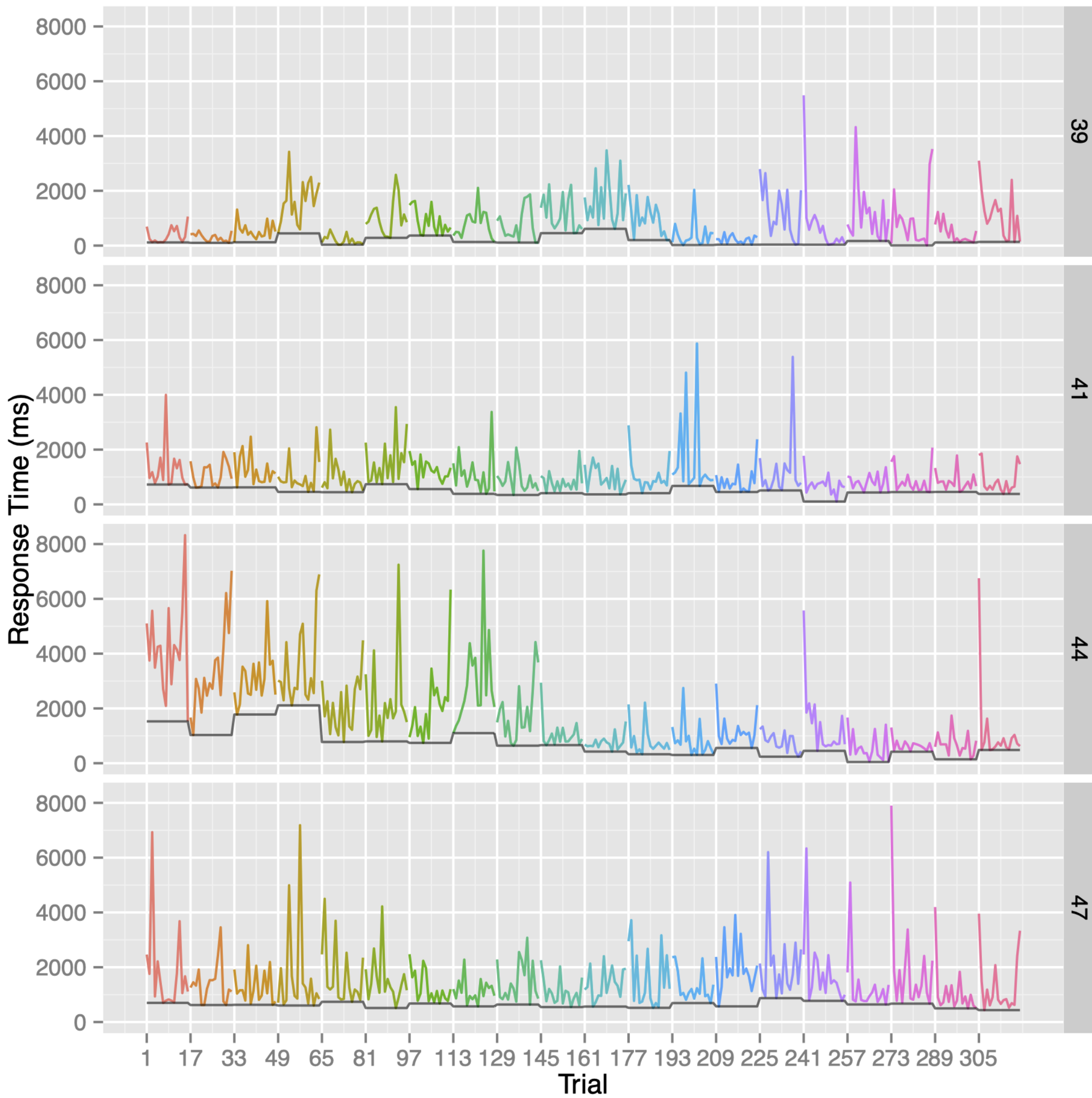
A Recognition Memory Experiment

- Twenty subjects drawn from the research pool of introductory psychology students at The Ohio State University.
- We presented each subject with 21 study lists of pictures in which four pictures were presented 1, 2, 3, or 4 times, with two additional initial pictures and two additional final pictures included as buffer items to control for any potential primacy and recency effects.
- The order of presentation of the four pictures and their repetitions was randomized across lists but not subjects.
- The eight study pictures were presented along with eight new pictures at test, when the subject had to identify each picture as old or new.
- Their RTs and the responses were recorded and the last 20 lists were analyzed.
- This was a hard task because the pictures were difficult to discriminate.

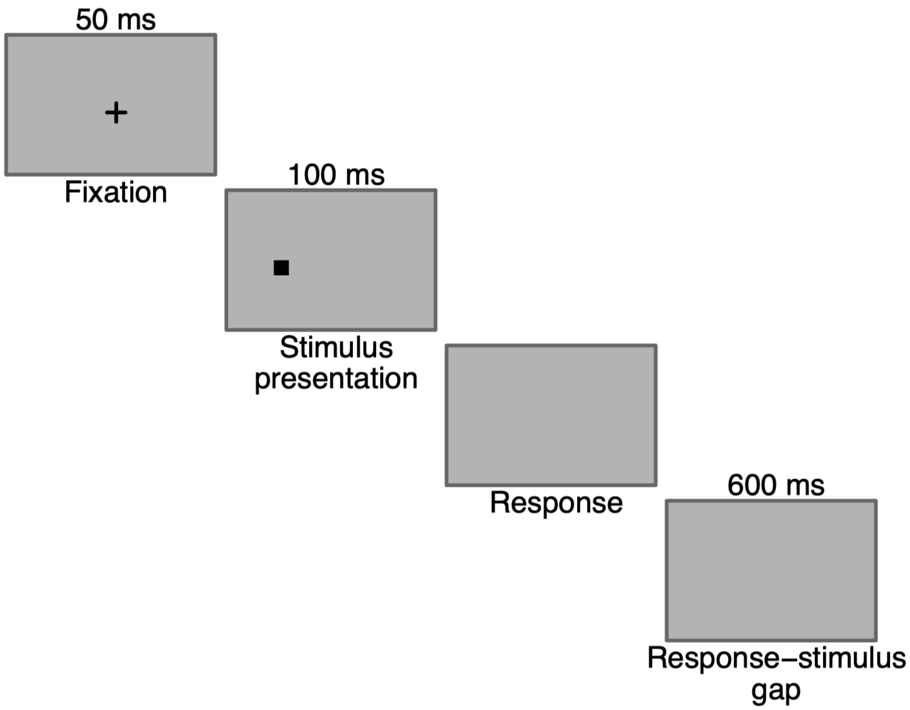
Sample images

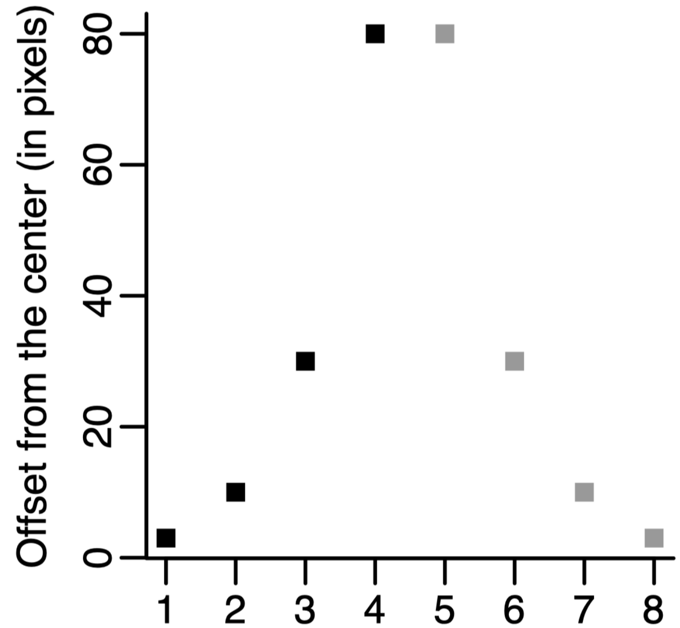


Sample TS plots

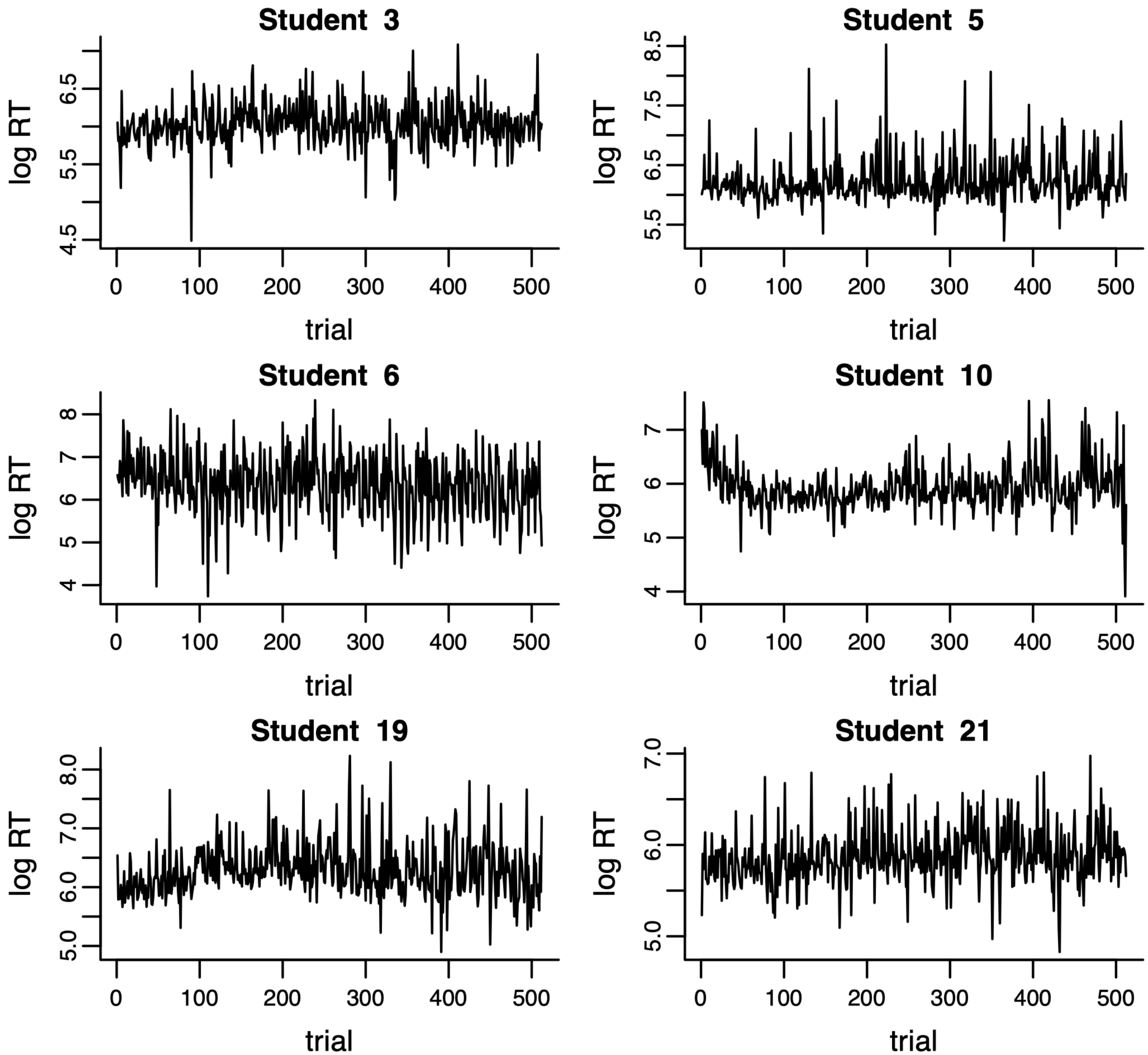


Left or right?





Sample TS plots



Modeling RT series - main features

- Models for RT sequences should capture the following features:

  1. Smooth changes (trends) in RT levels due to learning effects, fatigue, etc.
  2. Local sequential dependencies.
  3. Upper and lower tail behavior (extremes).

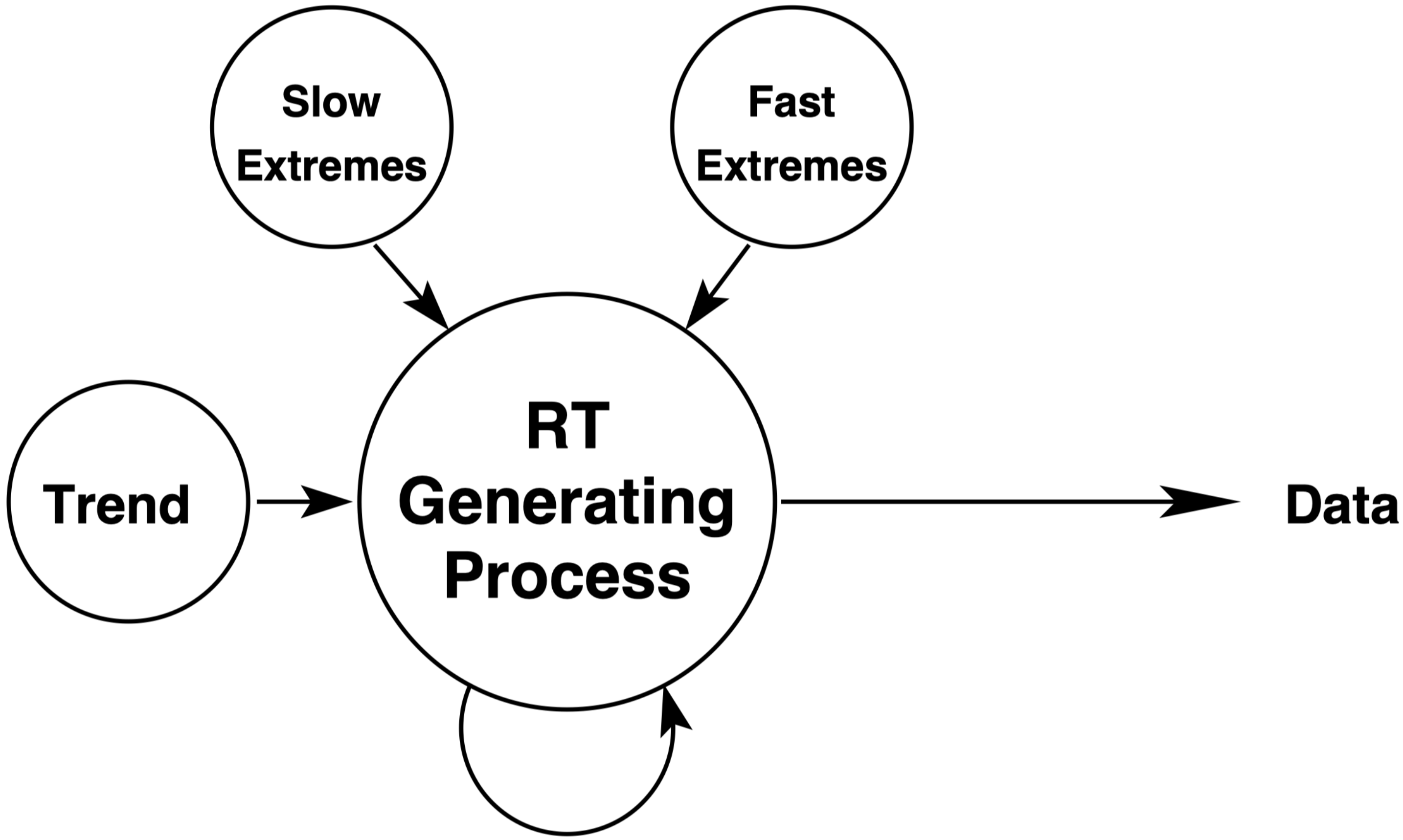
- Should account for the possibility that a subject may respond without actually performing the cognitive task (e.g., guessing).

- Should not discard RTs that are too fast or too slow (as is common practice).

- Should provide a framework for performing RT analysis that does not compromise model realism and serial dependence structure.

- Should be based on a hierarchical structure that accommodates differences among subjects.

A framework for analysis



Possible strategies

- Descriptive approach

  1. Peruggia et al. [2002]
  2. Rouder et al. [2003]
  3. Craigmile et al. [2010]

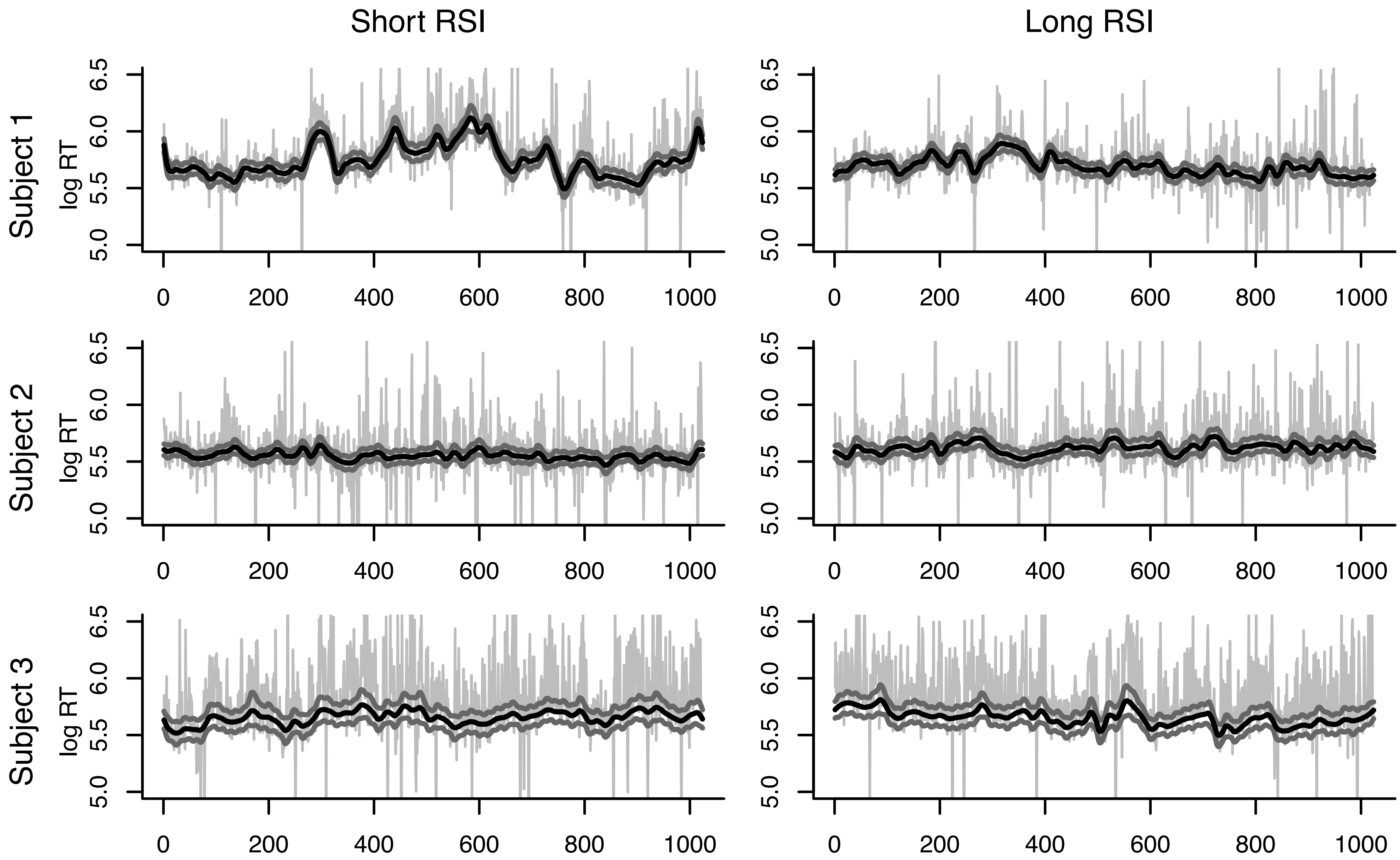
- Theoretical approach

  1. Logan [1988], Logan [1995], Kunkel et al. [2021]
  2. Race models [Van Zandt et al., 2000, Kim et al., 2017]
  3. Diffusion models [Ratcliff, 1978, Kunkel et al., 2019]

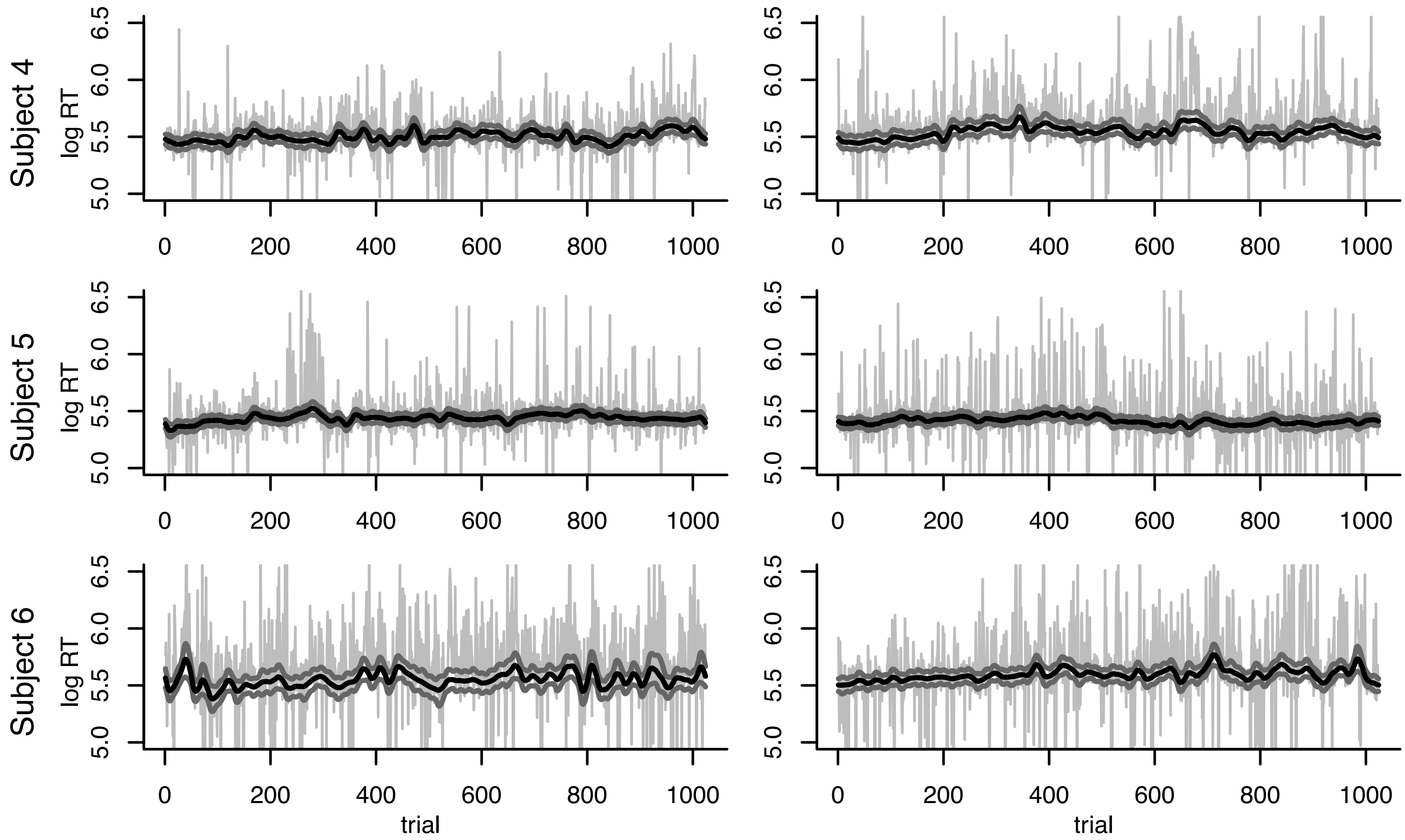
Craigmile, Peruggia and Van Zandt (2010a)

- Mixture model with several components:
  - Long-range trend to describe slow fluctuations in the mean.
  - Autoregressive component to capture local dependence
  - Components to model the heavy upper and lower tails.
- Model components are motivated by goodness-of-fit and predictive measures, rather than cognitive theories.
- Should trend be removed or modeled?
  - Remove it: Craigmile et al. [2011]
  - Model it: Craigmile et al. [2010]
  - For non-Gaussian, dependent series, standard methods to estimate trend will leave behind high-frequency components that will distort estimates of autocorrelation.

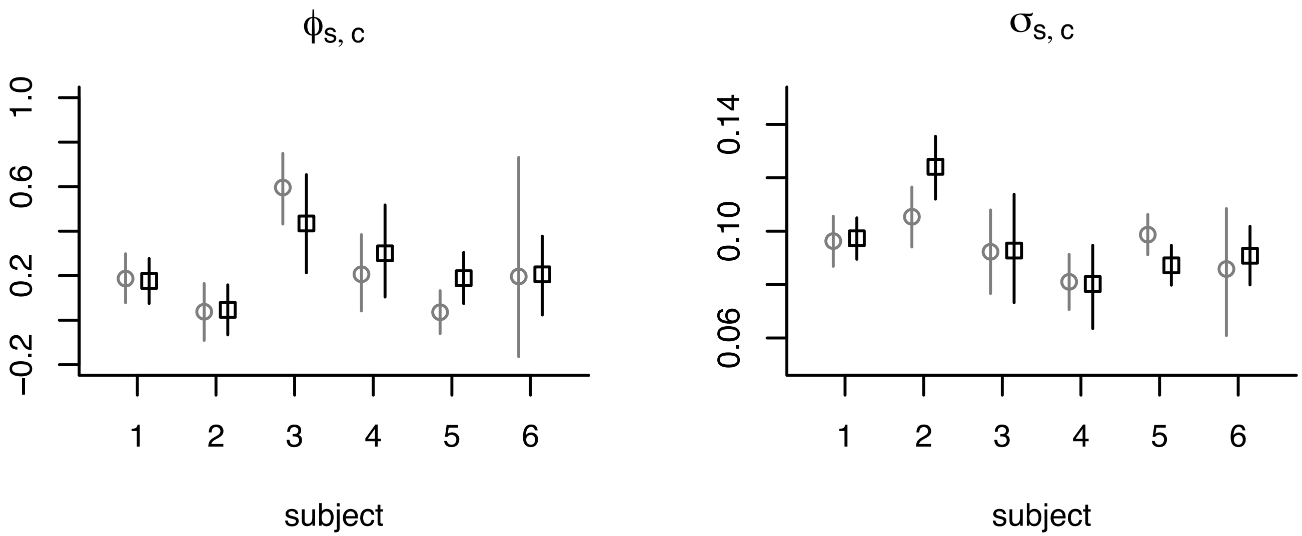
Results via MCMC: fitted trends for subjects 1, 2, and 3



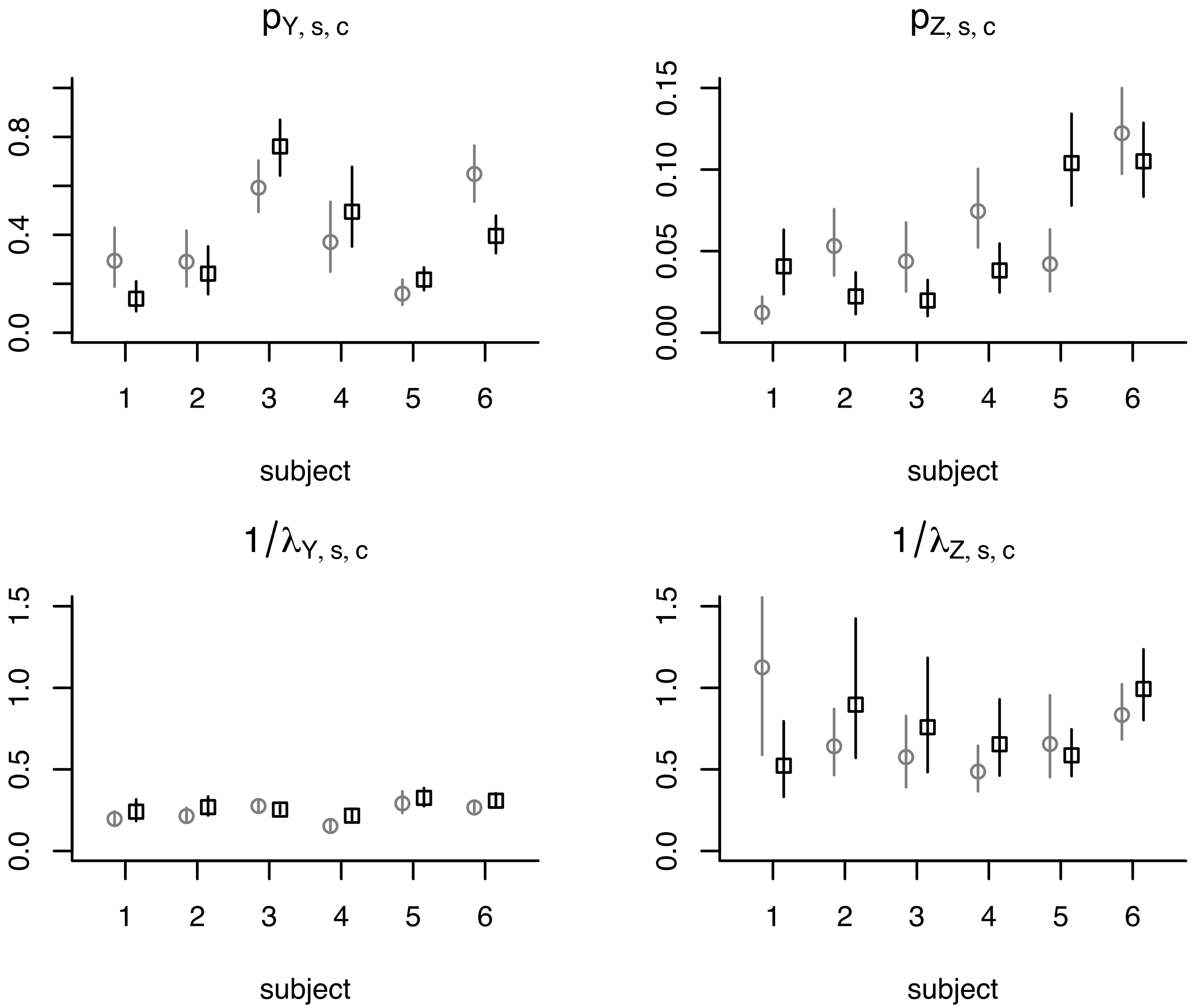
Results: fitted trends for subjects 4, 5, and 6



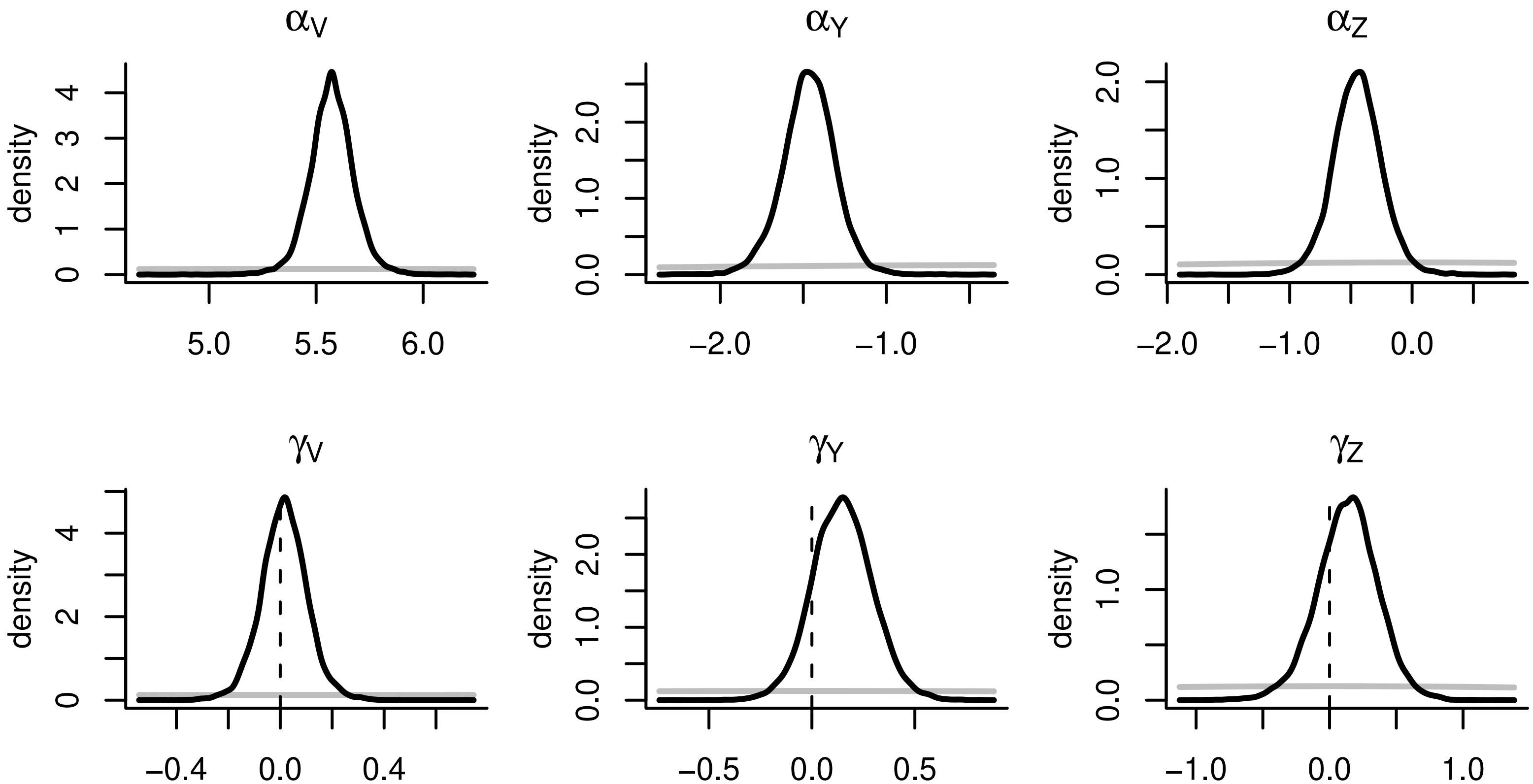
Posterior summaries for the AR(1) process



Posterior summaries for the tail parameters



Learning across subjects and conditions



Theoretical approach: background

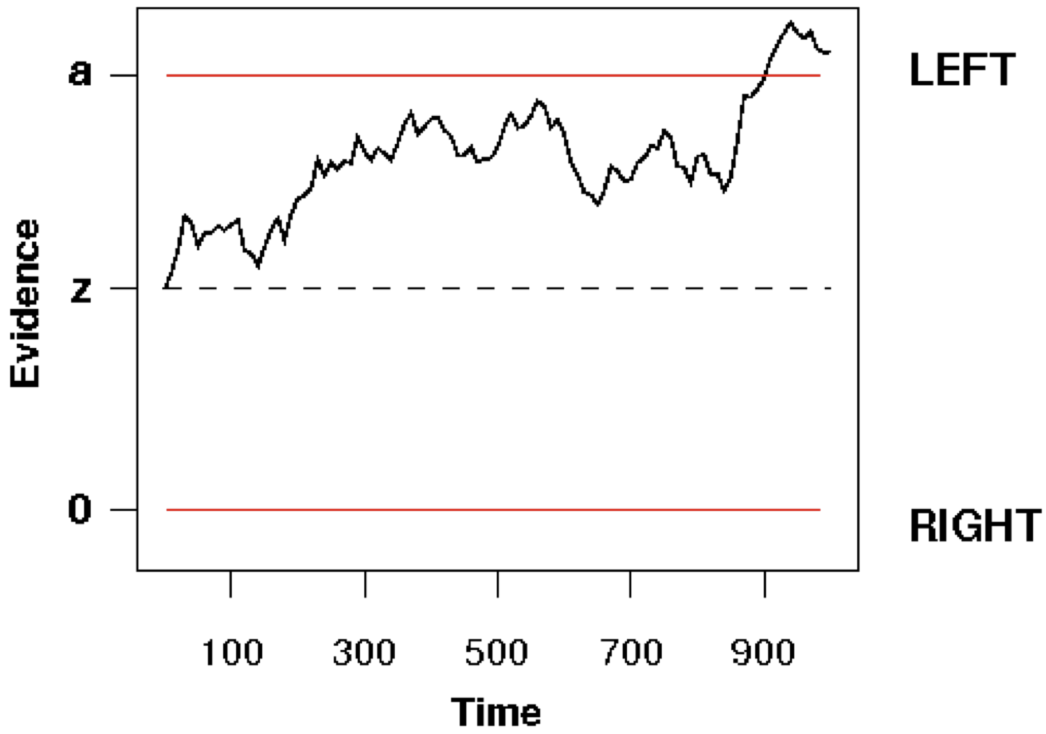
- The basic framework is related to Logan's theory of automaticity [Logan, 1988, 1995].
- The model subsumes some latent process that is a proxy for the brain activity yielding the RTs.
- Each exposure to a given stimulus results in a memory trace which can be used for later memory retrieval.
- Subsequent exposures to that stimulus trigger a race among traces and the observed RT is the shortest time (the winning time) among the times for all the traces that were laid down.

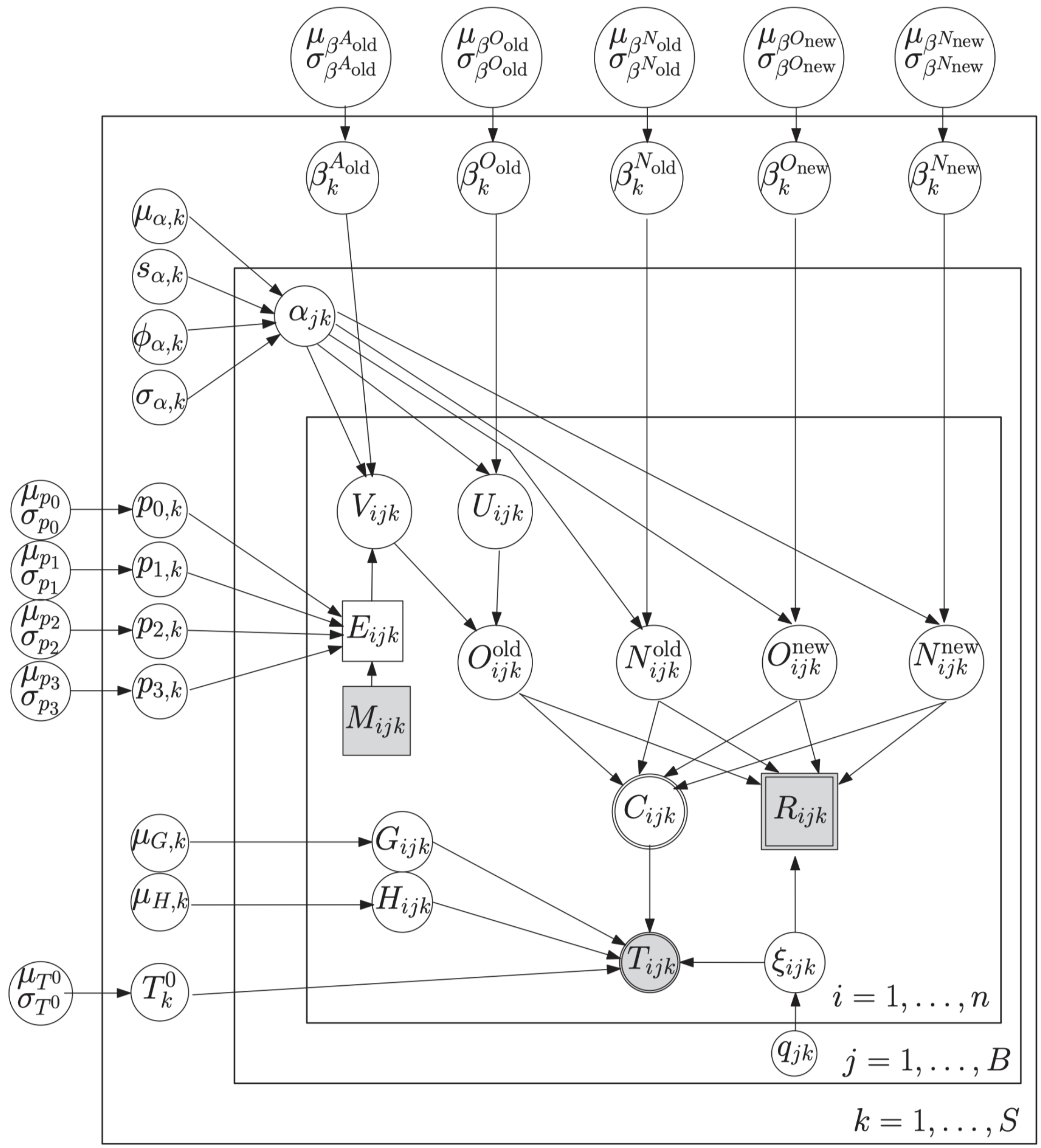
Interesting questions

- Logan's framework makes these fundamental assumptions:
  - A new memory trace is laid down each time a given stimulus is presented.
  - Traces operate independently.
  - A good model for the RTs is the Weibull distribution.
- Interesting questions:
  - Could traces be added probabilistically?
  - Could there be dependence between the various traces?
  - Could distributions other than a Weibull be used to model RTs?
- Our 2017 JASA paper [Kim et al., 2017] partially answers these questions.
  - Traces appear to be added probabilistically and not independently.
  - Gamma racers provide a better fit than models based on Weibull distributed RTs.

Thresholds and evidence accumulation in a diffusion model

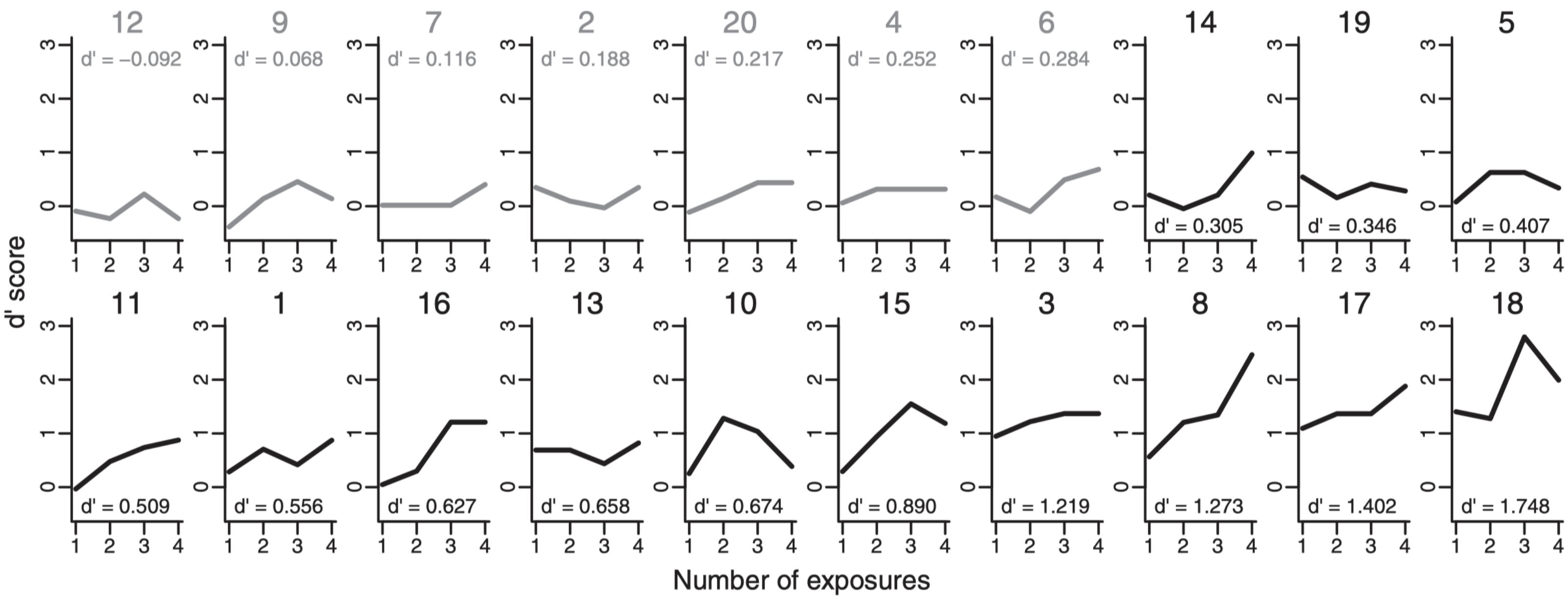
$$X(t+s) - X(s) \sim N(\mu t, \sigma^2 t), \quad t > 0$$




$\mu_{\beta^{A_{\rm old}}}$ $\sigma_{\beta^{A_{\rm old}}}$
$\mu_{\beta^{O_{\rm old}}}$ $\sigma_{\beta^{O_{\rm old}}}$
$\mu_{\beta^{N_{\rm old}}}$ $\sigma_{\beta^{N_{\rm old}}}$
$\mu_{\beta^{O_{\rm new}}}$ $\sigma_{\beta^{O_{\rm new}}}$
$\mu_{\beta^{N_{\rm new}}}$ $\sigma_{\beta^{N_{\rm new}}}$
$\beta_k^{A_{\rm old}}$
$\beta_k^{O_{\rm old}}$
$\beta_k^{N_{\rm old}}$
$\beta_k^{O_{\rm new}}$
$\beta_k^{N_{\rm new}}$
$\mu_{\alpha,k}$
$s_{\alpha,k}$
$\phi_{\alpha,k}$
$\sigma_{\alpha,k}$
α_{jk}
μ_{p_0} σ_{p_0}
μ_{p_1} σ_{p_1}
μ_{p_2} σ_{p_2}
μ_{p_3} σ_{p_3}
$p_{0,k}$
$p_{1,k}$
$p_{2,k}$
$p_{3,k}$
V_{ijk}
U_{ijk}
E_{ijk}
M_{ijk}
$O_{ijk}^{\rm old}$
$N_{ijk}^{\rm old}$
$O_{ijk}^{\rm new}$
$N_{ijk}^{\rm new}$
C_{ijk}
R_{ijk}
$\mu_{G,k}$
$\mu_{H,k}$
G_{ijk}
H_{ijk}
μ_{T^0} σ_{T^0}
T_k^0
T_{ijk}
ξ_{ijk}
q_{jk}
$i = 1, \ldots, n$
$j = 1, \ldots, B$
$k = 1, \ldots, S$

Discriminability and the sensitivity index d'

As the number of exposures of a picture increases, its familiarity should increase, resulting in larger values of d' [e.g. Hirshman, 1995]

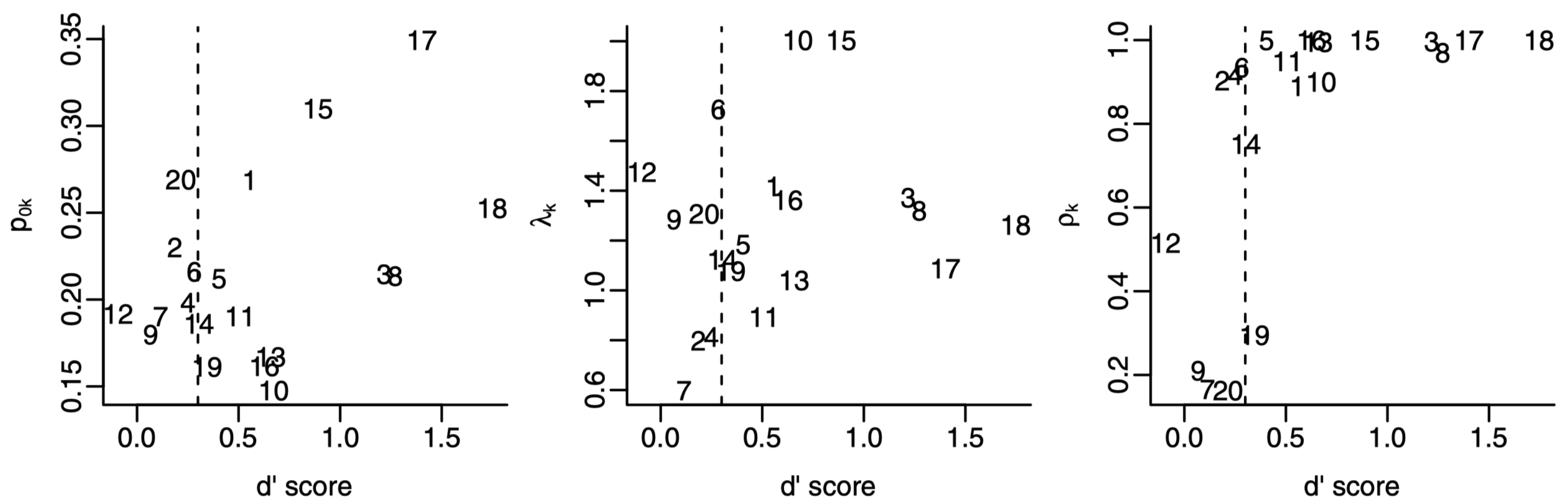


Designing our own measures of performance

1. Individual-level probability of adding a first trace, $p_{0,k}$

2. How well subjects extract information from a test picture by computing the ratio of the new and old accumulation process scale parameters when an old item was presented, λ_k

3. The proportion of non-subcognitive responses ρ_k



Psychometrics: take home messages

- Not everyone does what you tell them to do in the same way.
- When people perform a repetitive task over time, the way they perform that task changes over time.
- Models that can explain how a task is performed well should also be able to explain how a task is performed badly.
- Using sufficiently complex models allows us to better understand the cognitive process of interest, and removes the need for ad hoc methods of data preprocessing.

Statistical conclusions

- RT data have a rich dependence structure, and collapsing across experimental trials hides much that is interesting in RT data.
- A Bayesian approach can help isolate the different components of an RT series.
- Model diagnostics and comparison is an important component:
  - Examining posterior predictive distributions is key – but not just means!
  - Comparing marginal likelihood.
- Once separated, we can ask more intelligent questions about trend and dependence, including what kinds of processes may be responsible for each and how those processes are influenced by experimental effects.

Extensions: Hierarchical Hidden Markov Models for RT data

- We postulate the existence of three distinct RT distributions and we introduce **latent response modes** that correspond to these distributions:

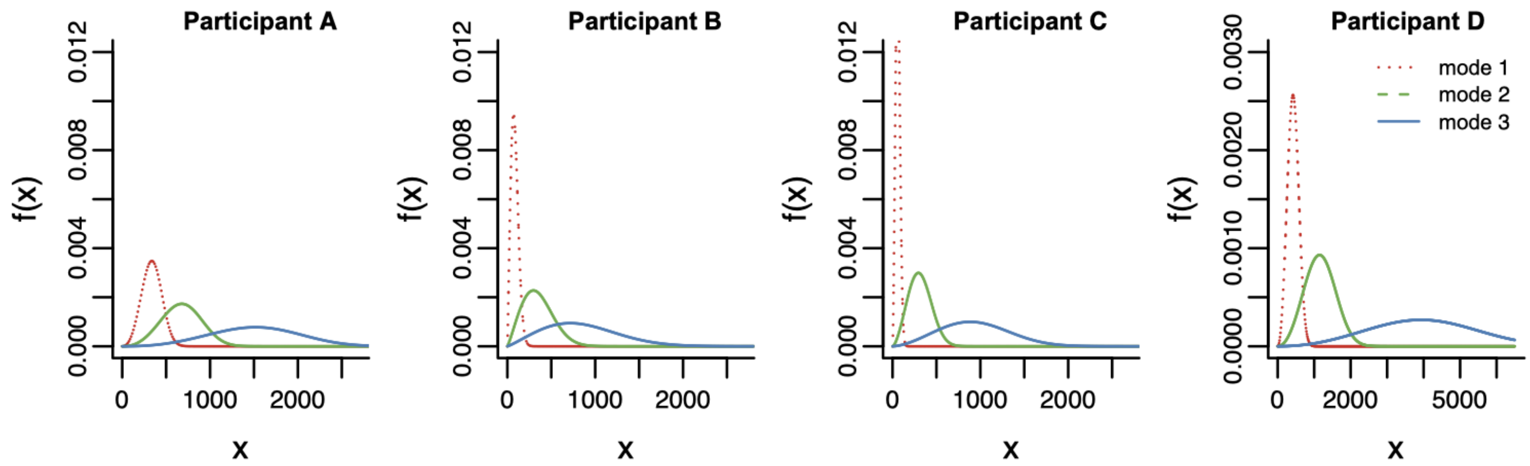
  1. A person in response mode 1 at trial t will generate fast RTs from the fast process.
  2. A person in response mode 2 at trial t will generate medium-length RTs.
  3. A person in response mode 3 at trial t will generate slow RTs.

- These three distributions may represent three modes of performance in a task, or they may indicate responses from subcognitive, cognitive, and supracognitive processes.

Hidden Markov Models for RT data: environments

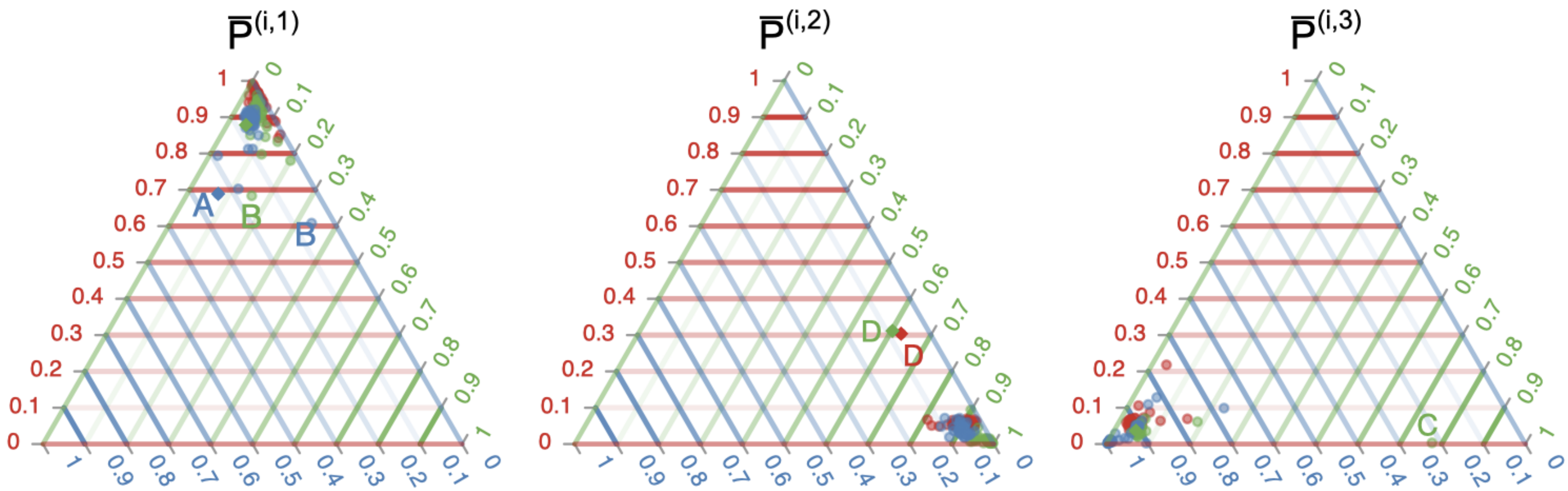
- We further posit the existence of several latent environments, each of which causes people to transition differently among the latent response modes.
- In practical applications, these latent environments may relate to external, unmeasured background conditions and circumstances influencing a person's performance, such as learning and fatigue.

Predictive densities by response mode



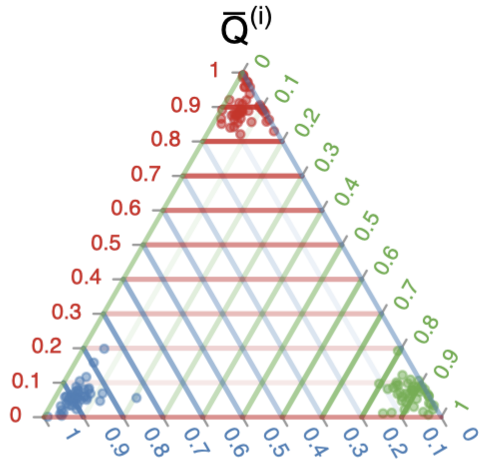
Monte Carlo estimates of the predictive densities for the three response modes. (The x-axis for Participant D is shown on a different scale.)

Transitions among response models



The panels show the posterior means of the rows of (from left) $\bar{P}^{(i,1)}$ $\bar{P}^{(i,2)}$ and $\bar{P}^{(i,3)}$. Rows 1, 2, and 3 are plotted in red, blue, and green, respectively.

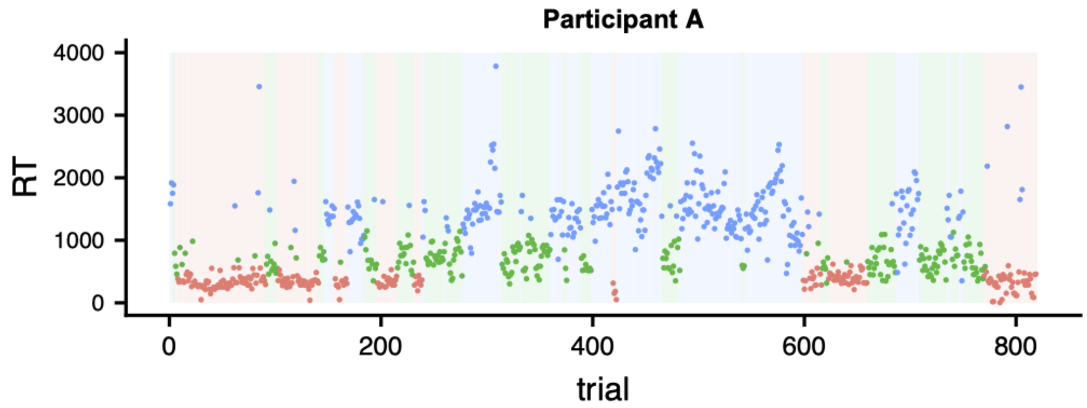
Transitions among environments



The posterior means of the rows of $\bar{Q}^{(i)}$ for each person.

Rows 1, 2, and 3 are plotted in red, blue, and green, respectively.

Participant effects



$\bar{Q}$

0.96	0.03	0.01
0.03	0.92	0.05
0.01	0.04	0.94

$\bar{P}1$

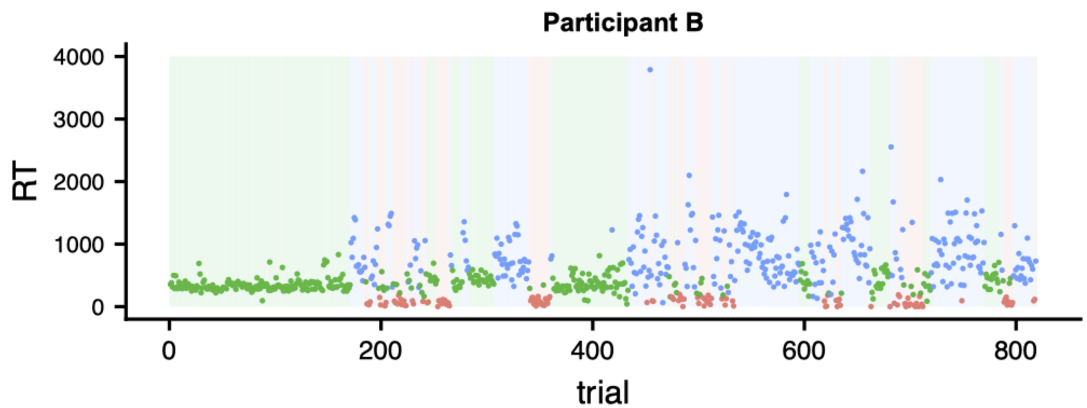
0.92	0.05	0.03
0.88	0.04	0.08
0.69	0.07	0.24

$\bar{P}2$

0.05	0.90	0.04
0.01	0.94	0.05
0.02	0.88	0.10

$\bar{P}3$

0.06	0.04	0.90
0.05	0.06	0.90
0.00	0.00	0.99



$\bar{Q}$

0.85	0.02	0.13
0.01	0.96	0.03
0.06	0.03	0.91

$\bar{P}1$

0.84	0.14	0.02
0.68	0.16	0.16
0.61	0.34	0.05

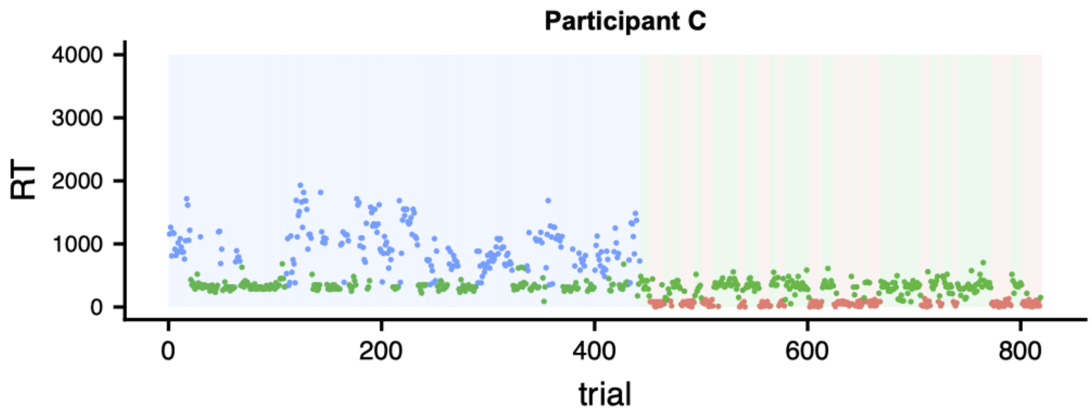
$\bar{P}2$

0.05	0.90	0.05
0.00	0.99	0.01
0.04	0.91	0.05

$\bar{P}3$

0.06	0.10	0.84
0.04	0.05	0.91
0.01	0.01	0.98

Participant effects



$\bar{Q}$

0.89	0.11	0.00
0.08	0.91	0.01
0.00	0.01	0.99

$\bar{P}1$

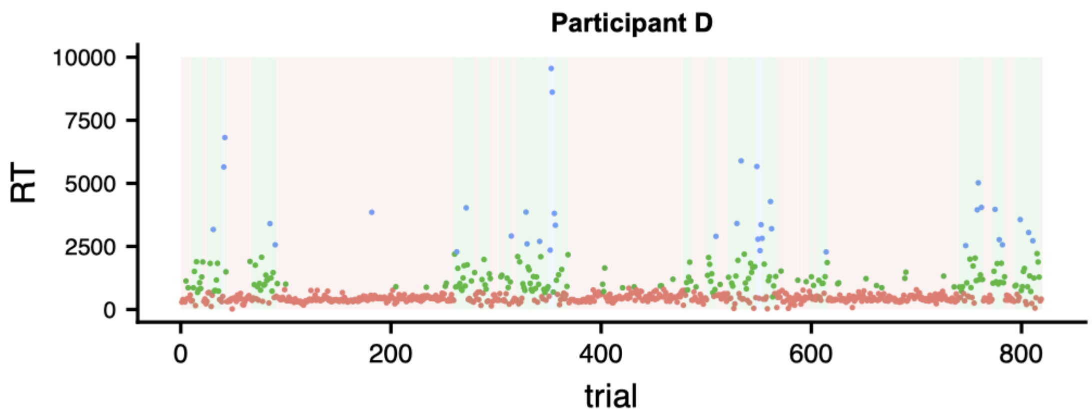
0.89	0.10	0.00
0.95	0.04	0.01
0.90	0.05	0.05

$\bar{P}2$

0.05	0.93	0.02
0.01	0.99	0.00
0.05	0.91	0.05

$\bar{P}3$

0.05	0.05	0.90
0.00	0.78	0.22
0.00	0.09	0.90



$\bar{Q}$

0.96	0.04	0.00
0.08	0.89	0.03
0.06	0.21	0.74

$\bar{P}1$

0.94	0.05	0.00
0.83	0.14	0.02
0.89	0.06	0.05

$\bar{P}2$

0.30	0.63	0.07
0.31	0.60	0.09
0.05	0.85	0.10

$\bar{P}3$

0.05	0.05	0.90
0.06	0.05	0.89
0.13	0.06	0.82

References

P. F. Craigmile, M. Peruggia, and T. Van Zandt. Hierarchical Bayes models for response time data. *Psychometrika*, 75:613–632, 2010.

P. F. Craigmile, M. Peruggia, and T. Van Zandt. Detrending response time series. In *Statistical Methods for Modeling Human Dynamics*, pages 213–240. Routledge, 2011.

E. Hirshman. Decision processes in recognition memory: criterion shifts and the list-strength paradigm. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 21:302, 1995.

S. Kim, K. Potter, P. F. Craigmile, M. Peruggia, and T. Van Zandt. A Bayesian race model for recognition memory. *Journal of the American Statistical Association*, 112:77–91, 2017.

D. Kunkel, K. Potter, P. F. Craigmile, M. Peruggia, and T. Van Zandt. A Bayesian race model for response times under cyclic stimulus discriminability. *The Annals of Applied Statistics*, 13:271–296, 03 2019.

D. Kunkel, Z. Yan, P. F. Craigmile, M. Peruggia, and T. Van Zandt. Hierarchical hidden markov models for response time data. *Computational Brain and Behavior*, 4:70–86, 2021.

G. D. Logan. Toward an instance theory of automatization. *Psychological Review*, 95:492–527, 1988.

G. D. Logan. The Weibull distribution, the power law, and the instance theory of automaticity. 102:751–756, 1995.

M. Peruggia, T. Van Zandt, and M. Chen. Was it a car or a cat I saw? An analysis of response times for word recognition. In *Case Studies in Bayesian Statistics*, volume 6, pages 319–334. Springer, New York, 2002.

R. Ratcliff. A theory of memory retrieval. *Psychological review*, 85:59, 1978.

J. N. Rouder, D. Sun, P. L. Speckman, J. Lu, and D. Zhou. A hierarchical Bayesian statistical framework for response time distributions. *Psychometrika*, 68:589–606, 2003.

T. Van Zandt, H. Colonius, and R. W. Proctor. A comparison of two response time models applied to perceptual matching. *Psychonomic Bulletin and Review*, 7:208–256, 2000.

E.-J. Wagenmakers, S. Farrell, and R. Ratcliff. Estimation and interpretation of $1/f$ noise in human cognition. *Psychonomic Bulletin and Review*, 11:579–615, 2004.