ML4H Workshop

NeurIPS 2020

https://arxiv.org/abs/2011.01725

Hierarchical partial-pooling in multi-group settings

Vincent Valton

NIHR UCLH Fellow - Postdoctoral Researcher

Neuroscience and Mental Health

Institute of Cognitive Neuroscience

University College London

Background

Modelling cognition & behaviour

Computational Neuroscience/Psychiatry → Modelling cognition & behaviour in healthy & patient pop.

Healthy Controls

Patients

Diagnosis

- Not based on biological substrate
- Based solely on symptoms experienced (Questionnaire based)

Background

Computational Neuroscience/Psychiatry

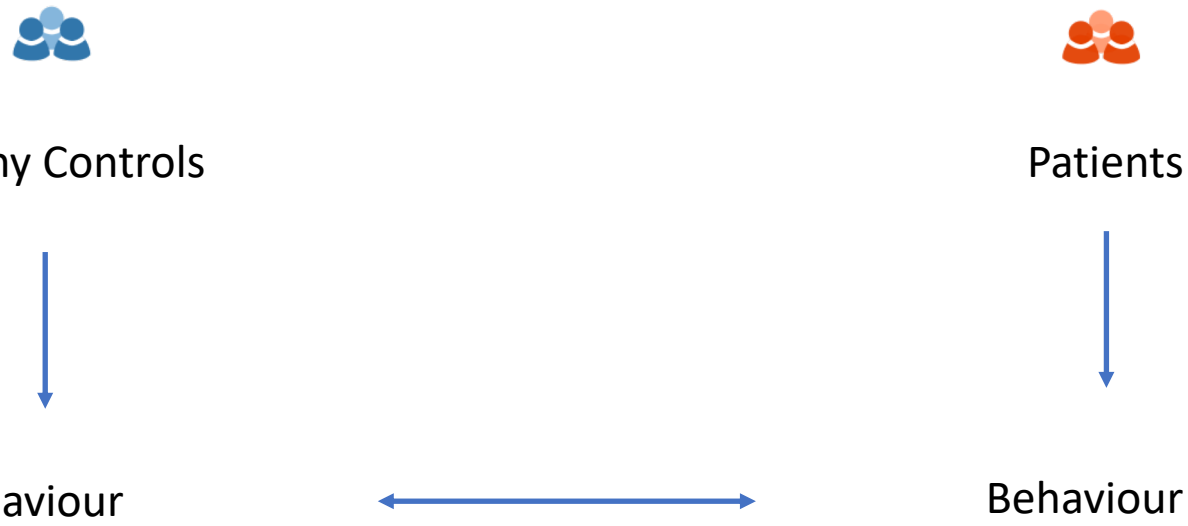
Diagnostic Criteria for Major Depressive Disorder and Depressive Episodes

DSM-IV Criteria for Major Depressive Disorder (MDD)

- Depressed mood or a loss of interest or pleasure in daily activities for more than two weeks.
- Mood represents a change from the person's baseline.
- Impaired function: social, occupational, educational.
- Specific symptoms, at least 5 of these 9, present nearly every day:
    1. **Depressed mood or irritable** most of the day, nearly every day, as indicated by either subjective report (e.g., feels sad or empty) or observation made by others (e.g., appears tearful).
    2. **Decreased interest or pleasure** in most activities, most of each day
    3. **Significant weight change (5%) or change in appetite**
    4. **Change in sleep**: Insomnia or hypersomnia
    5. **Change in activity**: Psychomotor agitation or retardation
    6. **Fatigue or loss of energy**
    7. **Guilt/worthlessness:** Feelings of worthlessness or excessive or inappropriate guilt
    8. **Concentration**: diminished ability to think or concentrate, or more indecisiveness
    9. **Suicidality:** Thoughts of death or suicide, or has suicide plan

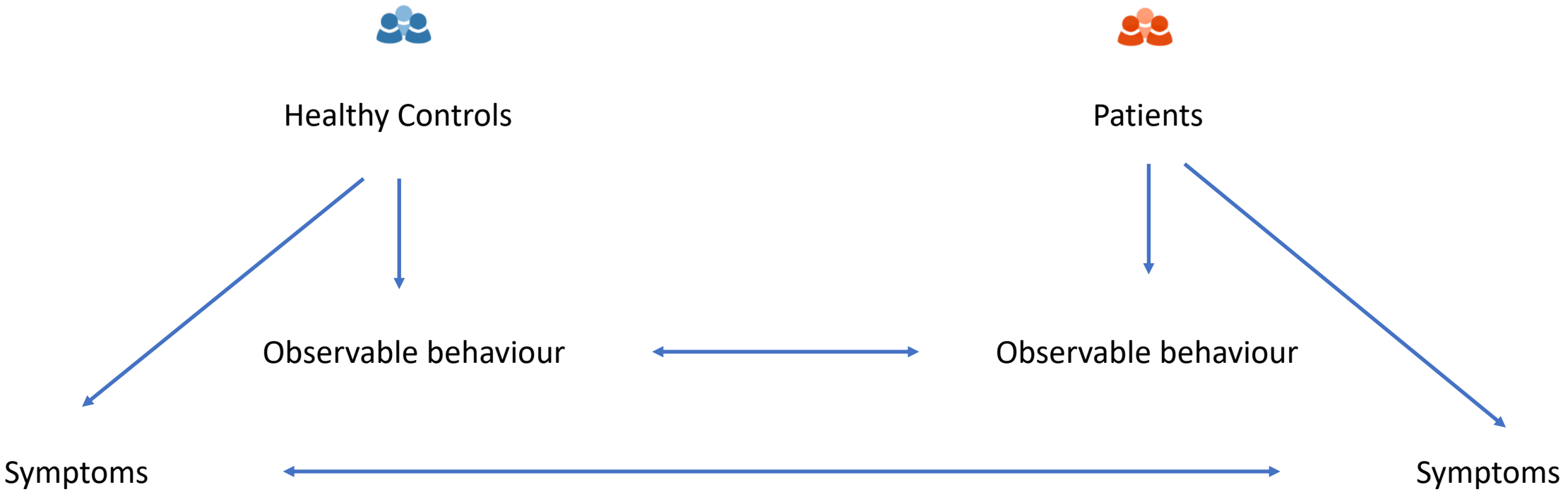
Background

Computational Neuroscience/Psychiatry



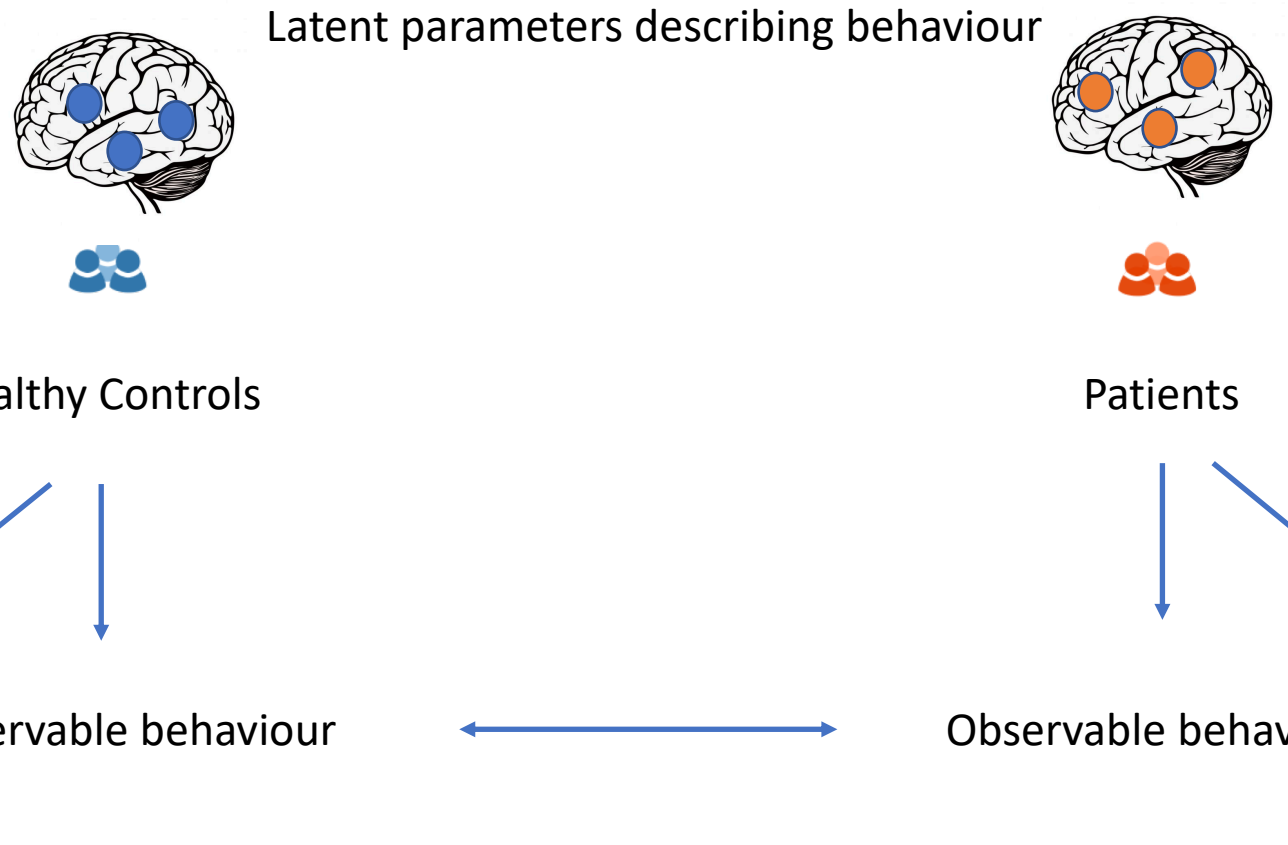
Background

Computational Neuroscience/Psychiatry



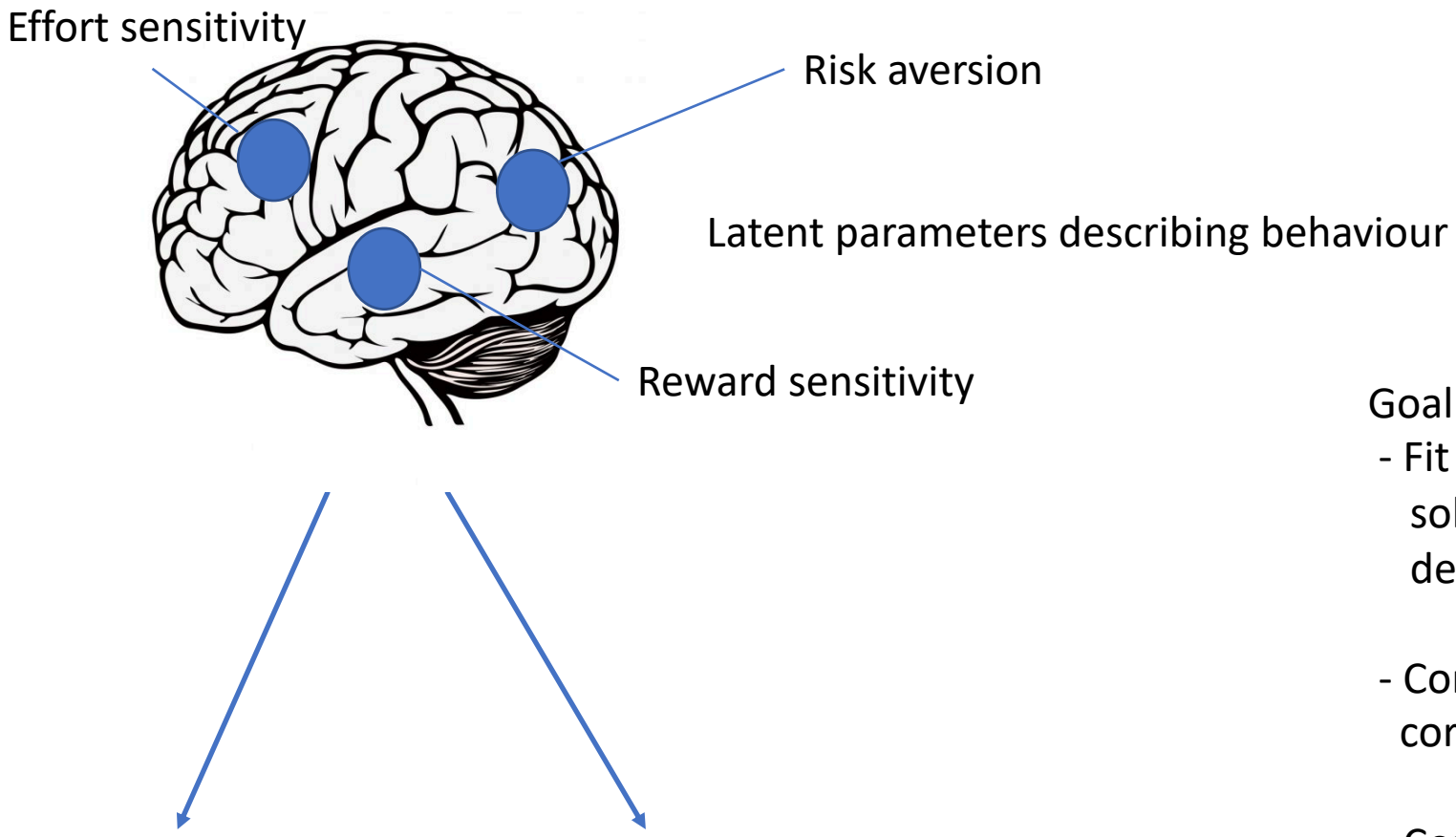
Background

Computational Neuroscience/Psychiatry



Background

Computational Neuroscience/Psychiatry



Goal:

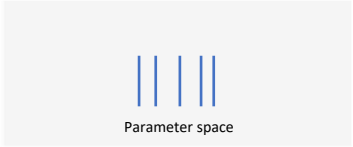
- Fit models of cognition to understand how people solve decision problems & extract parameters describing their behaviour

- Compare whether parameters extracted correlate with symptoms observed

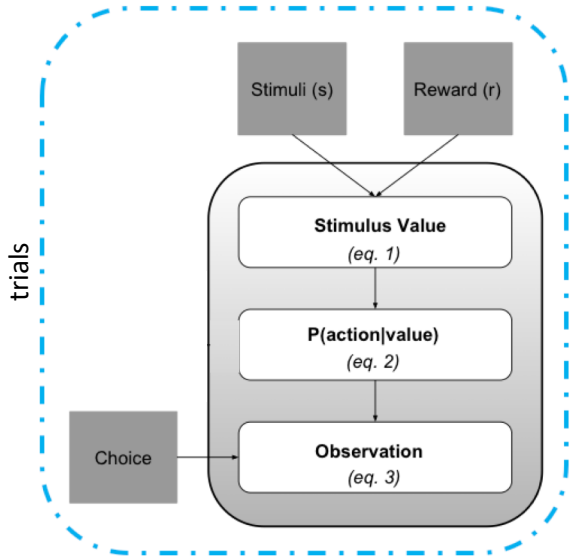
- Compare parameters extracted between Patients and Control groups

- Could be used for diagnosis

Background



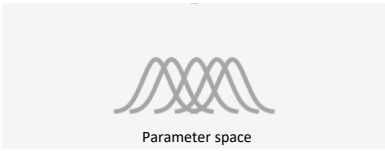
Maximum Likelihood Estimation



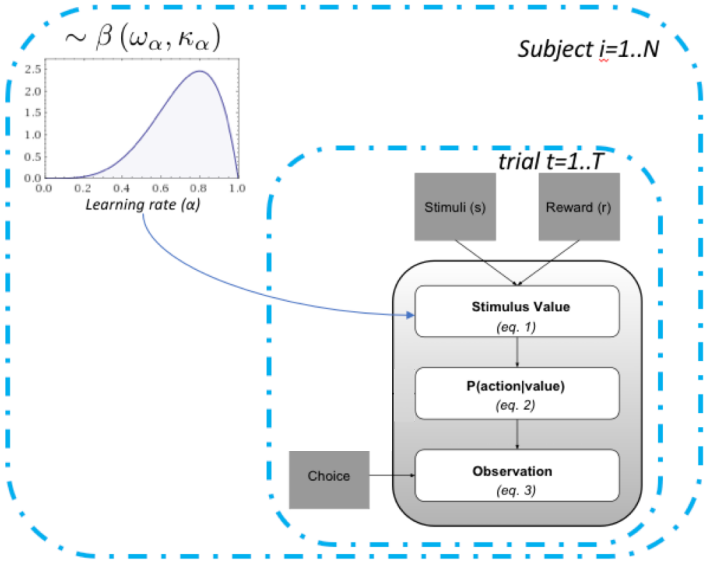
$eq.1 \quad V(s) = V(s) + \alpha(r - V(s))$

$eq.2 \quad softmax(eq.1|\tau)$

$eq.3 \quad categorical(eq.2)$



Sampling

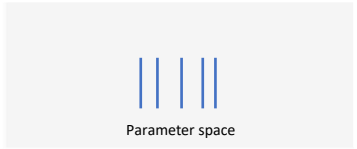


$eq.1 \quad V(s) = V(s) + \alpha(r - V(s))$

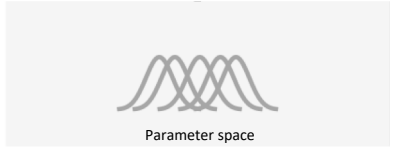
$eq.2 \quad softmax(eq.1|\tau)$

$eq.3 \quad categorical(eq.2)$

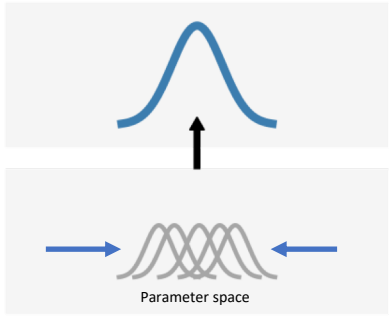
Background



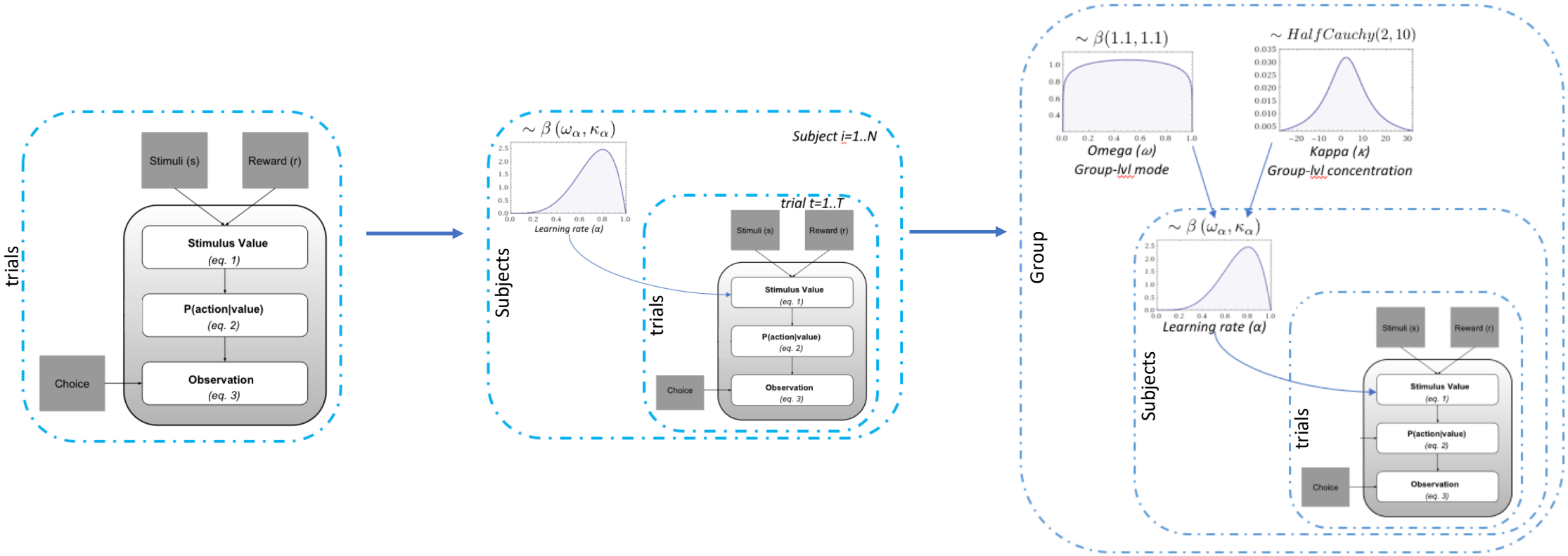
Maximum Likelihood Estimation



Sampling (Non-Hierarchical)



Sampling (Hierarchical partial pooling)



Problems

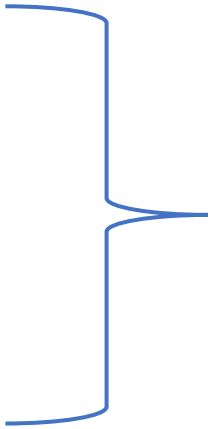
- Hierarchical modelling techniques exist, but no guidelines exist for best practice and approach for multi-group specifications
  - Issue: Different assumptions on model specification may lead to vastly different parameter estimates.
  - Issue: We don't know the extent of the problem because it hasn't been quantified.

- How does data quality affect parameter recovery in hierarchical setting?

Questions - Aims

https://arxiv.org/abs/2011.01725

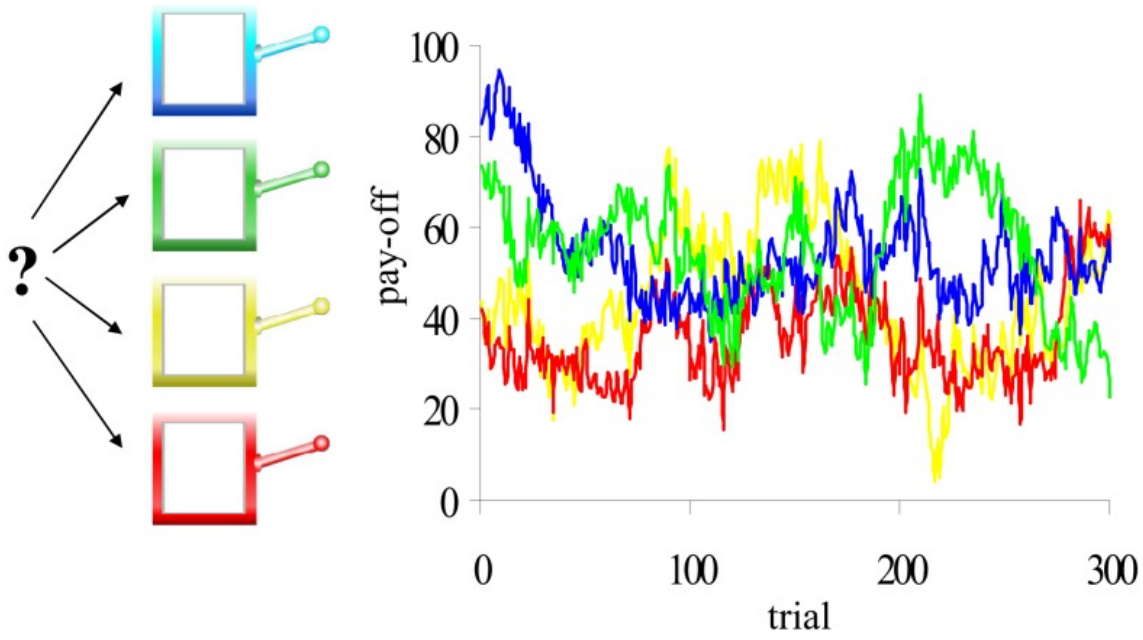
- 1 - Compare different hierarchical model specifications for group studies
- 2 - Are some model specifications more/less resilient to data quality issues?



Guidelines for future studies

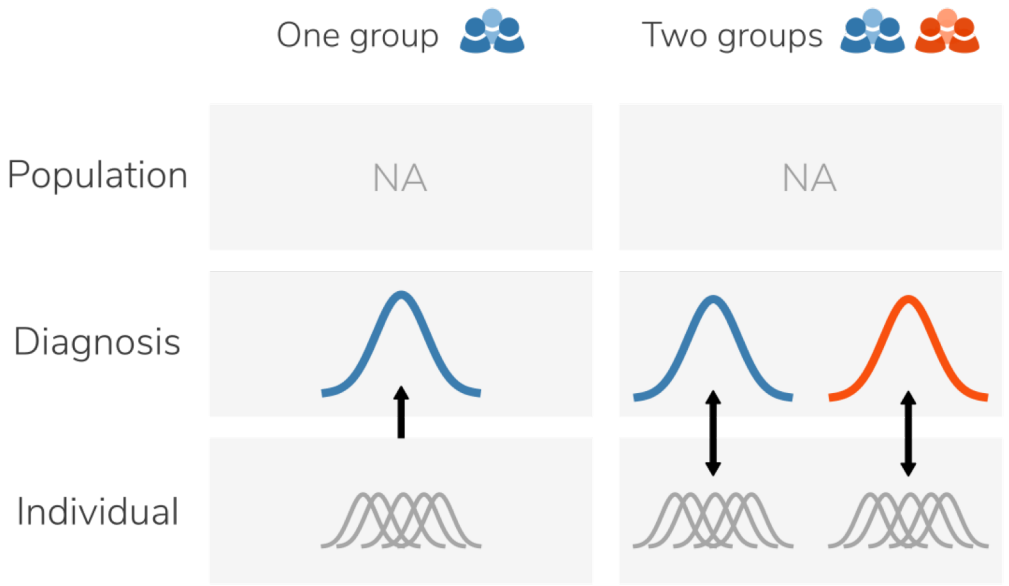
Task

- Task used in psychology/psychiatry research
  - Reinforcement Learning task
- No asymptotic performance (stationary env.)
  - roving punishment and reward probabilities
- Good record of parameter recovery (enough trials, enough stimuli)
  - 4 Arm bandit w. 300 trials
- Assume only 2 groups at this stage (for simplicity)
  - Could be 2 different populations (Controls vs Patients)
  - Or different manipulation in same population (Drug vs Placebo)

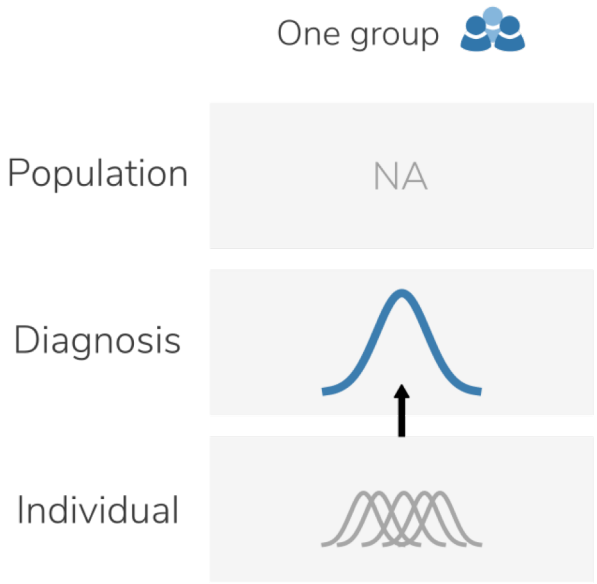


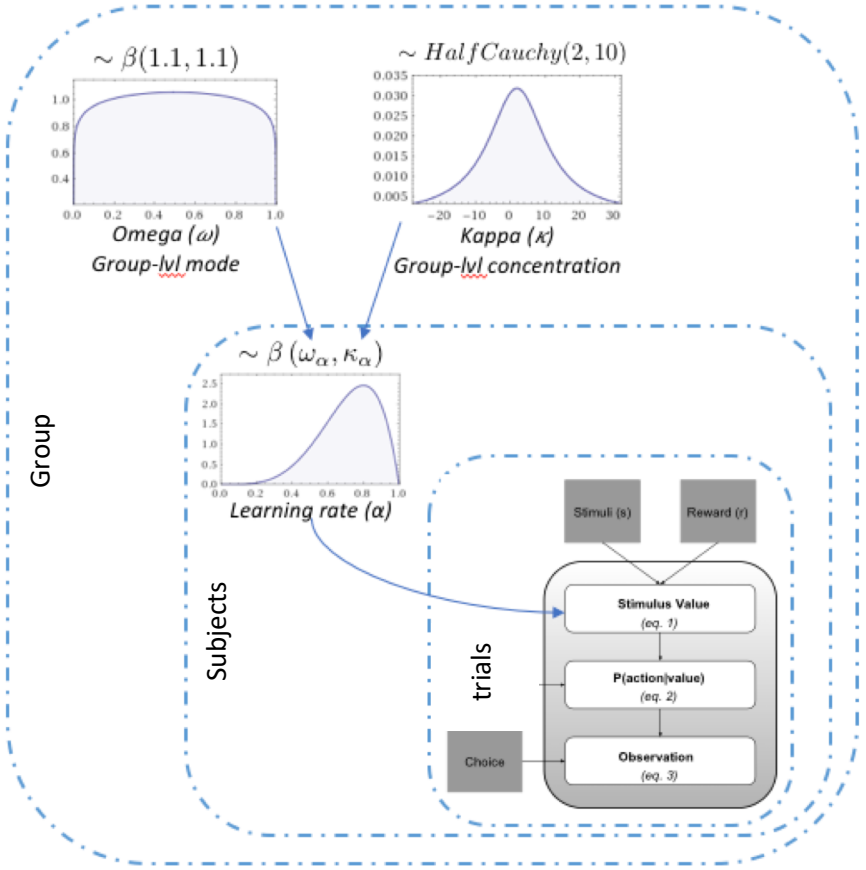
1 – Compare model specifications

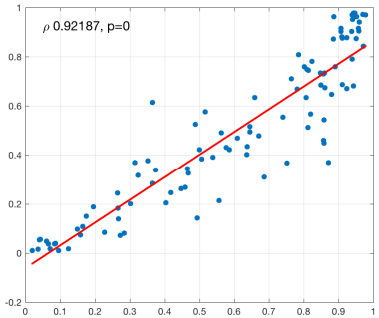
Model assumptions

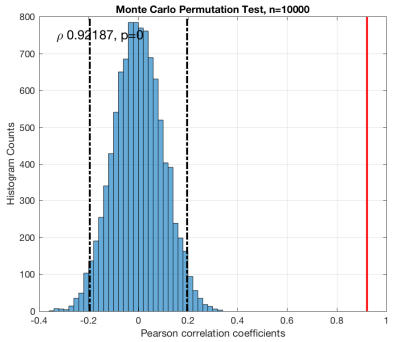


Model assumptions (model 1)







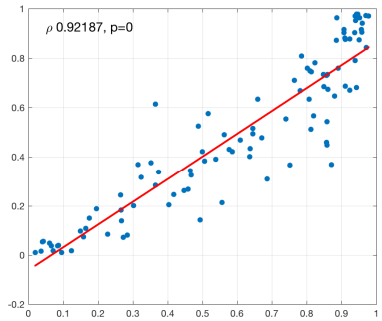


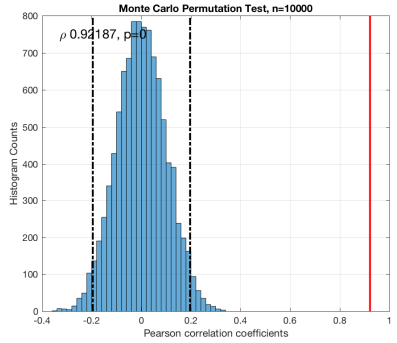
$$eq.1 \quad V(s) = V(s) + \alpha(r - V(s))$$

$$eq.2 \quad softmax(eq.1|\tau)$$

$$eq.3 \quad categorical(eq.2)$$

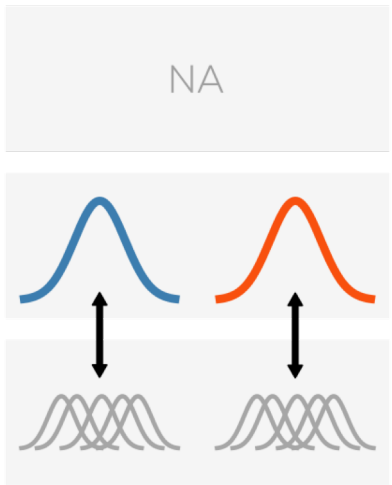
Model assumptions

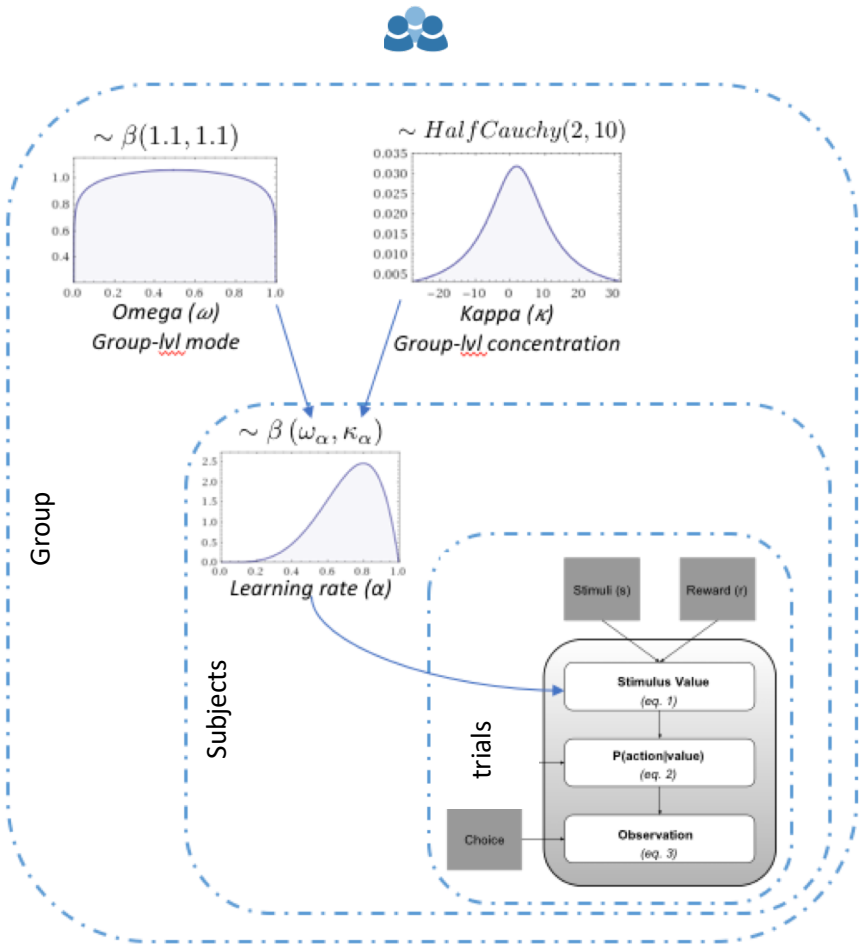


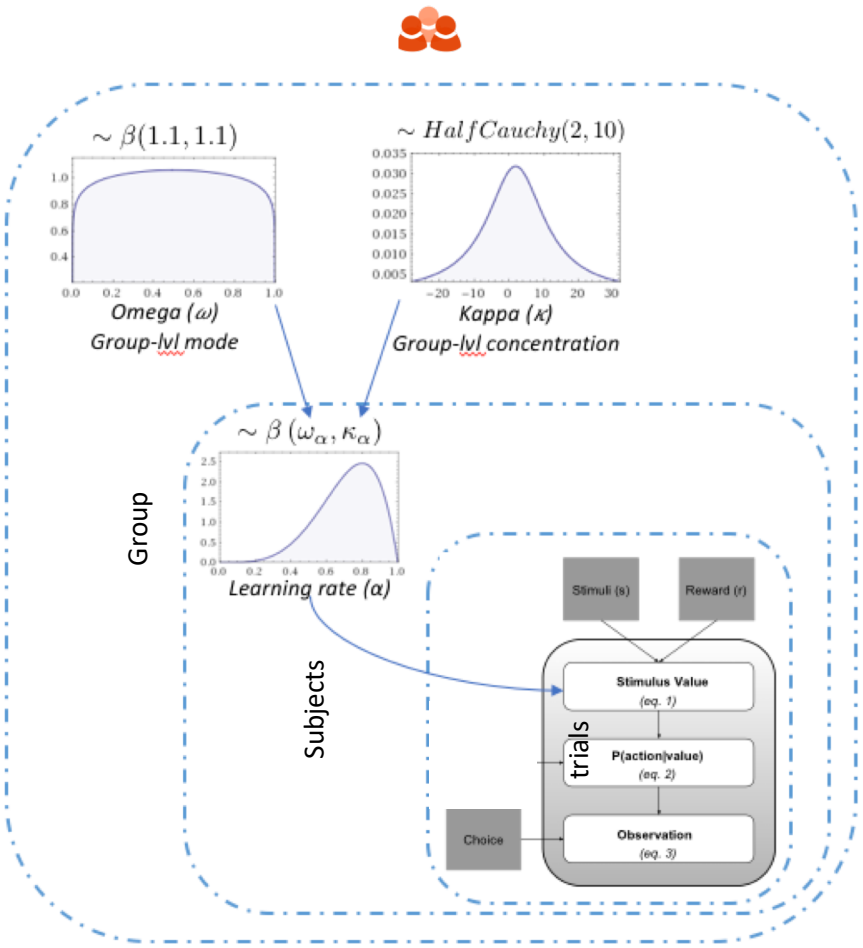


Two groups

NA



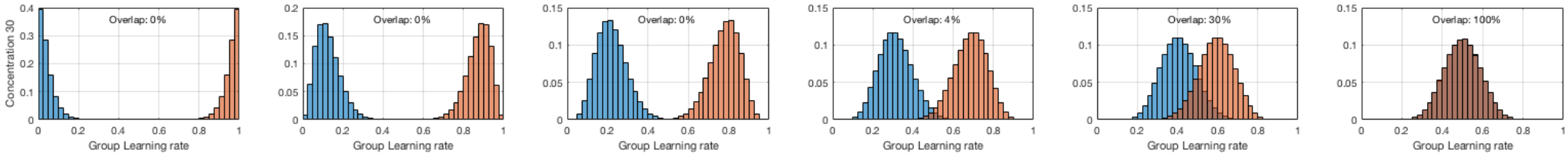




Simulation study –
Synthetic data generation

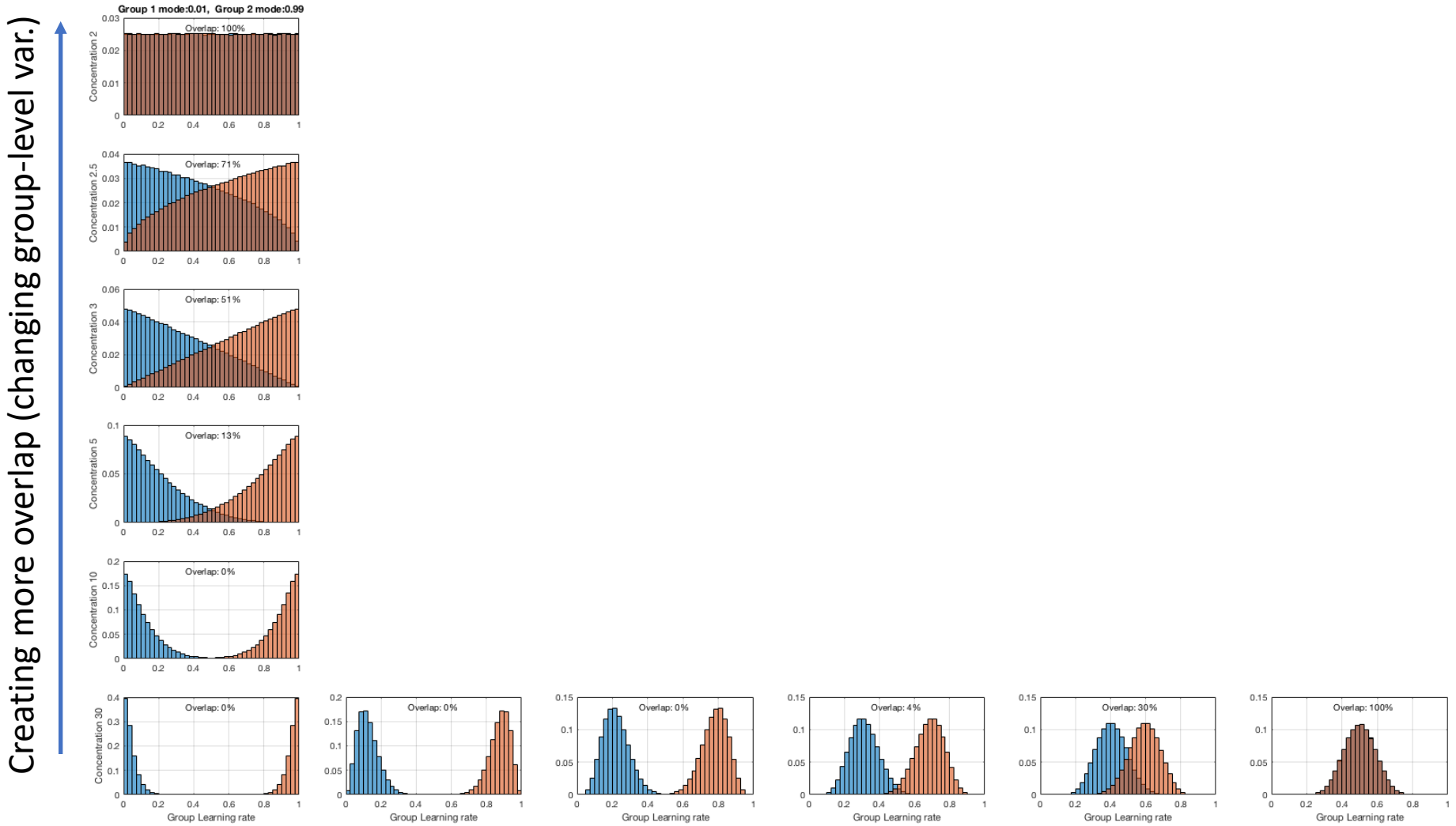
(Dataset 1: Single parameter to estimate (learning rate)

Creating more overlap (changing group-level mode)



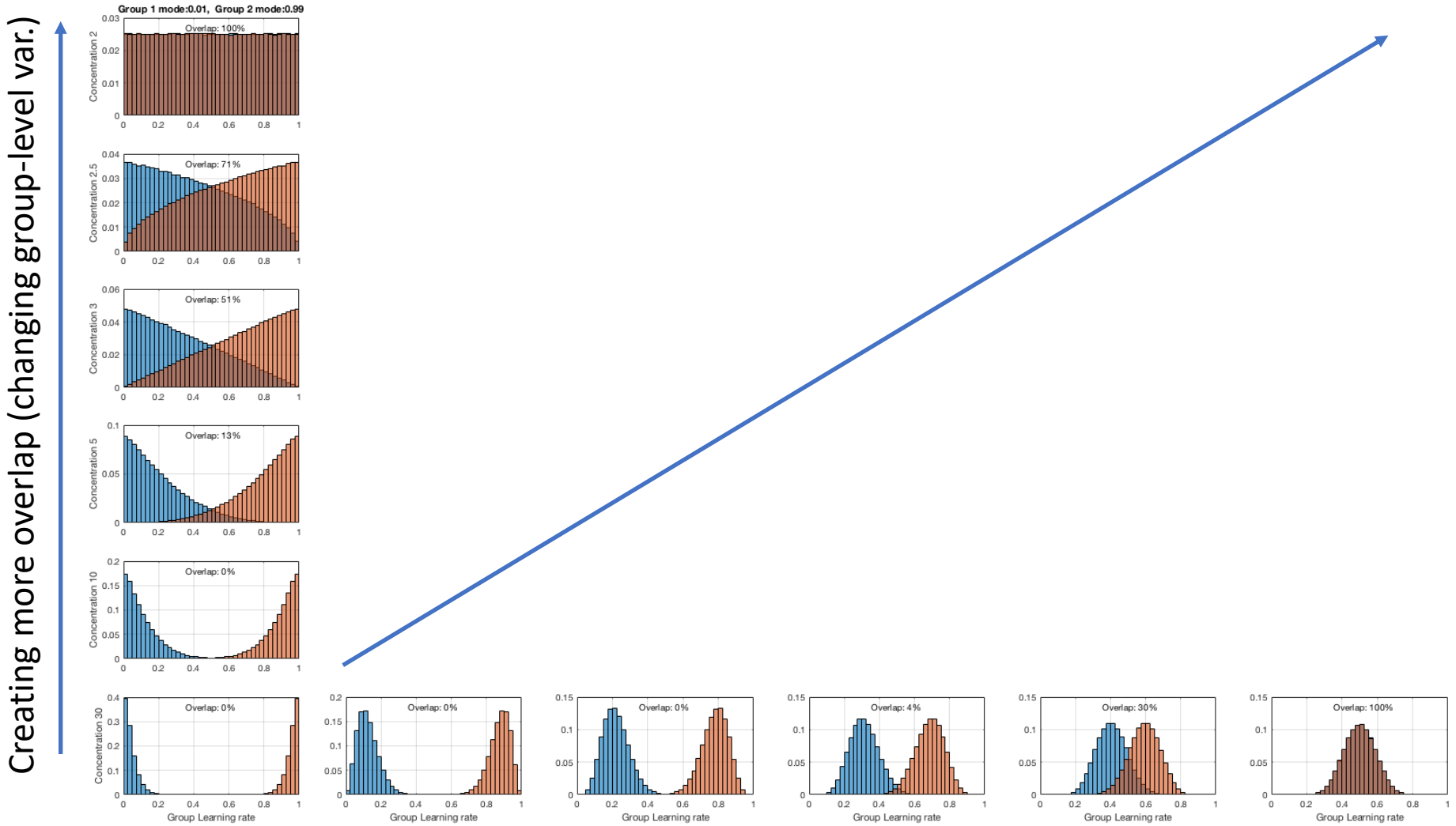
Simulation study –
Synthetic data generation

(Dataset 1: Single parameter to estimate (learning rate)



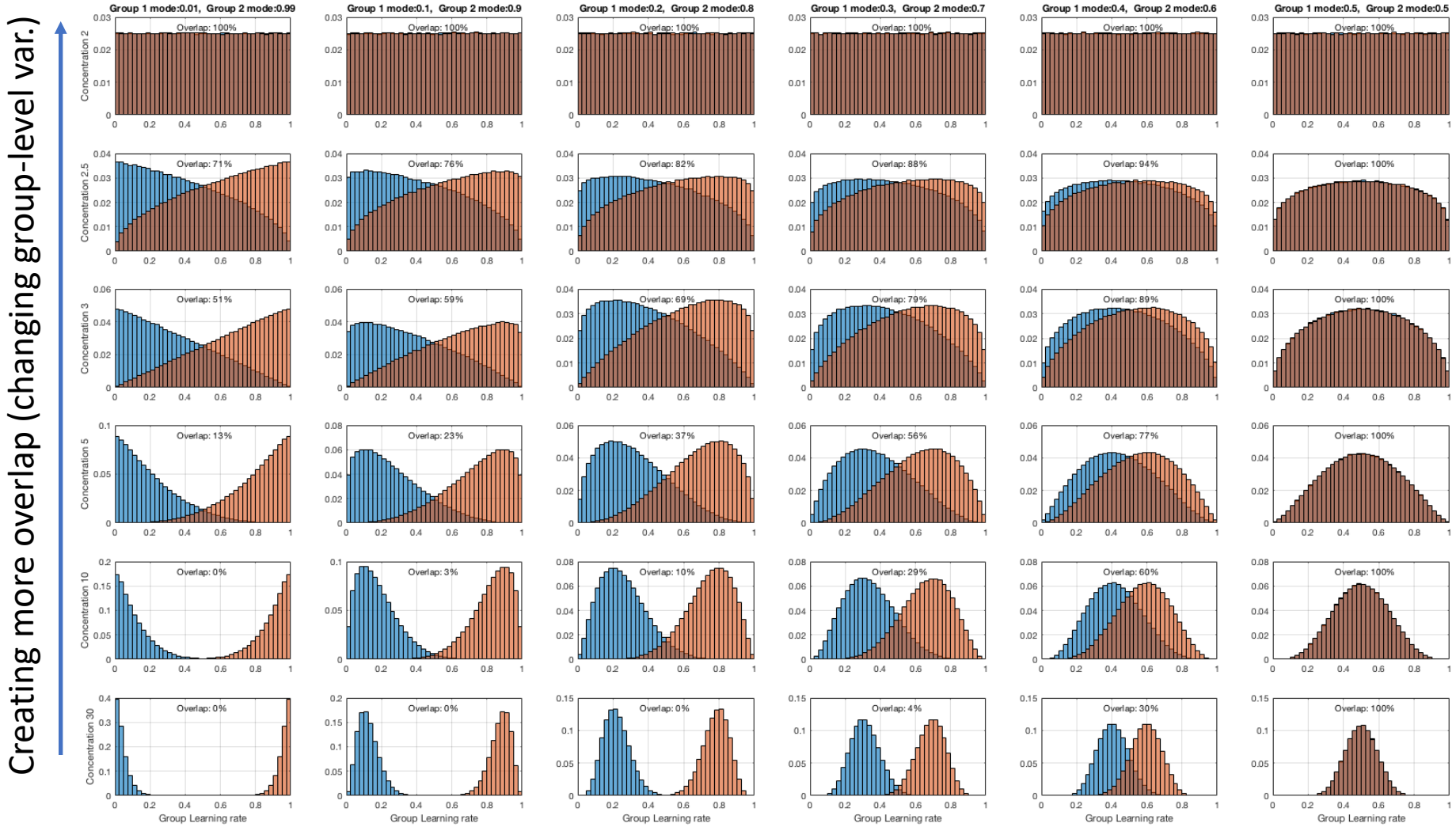
Simulation study –
Synthetic data generation

(Dataset 1: Single parameter to estimate (learning rate)



Simulation study –
Synthetic data generation

(Dataset 1: Single parameter to estimate (learning rate)



Creating more overlap (changing group-level var.)

Creating more overlap (changing group-level mode)

2 – Study impact of data quality on recovery for different model specifications

Are some model specifications more sensitive to data quality?

Study the effect of trial number and sample size on group level recovery

Data quality		# Participants	
# Trials		N=15	N=50
	T=40	Poor	Med.
	T=200	Med.	High

Does the number of trials affect parameter recovery ?

Study the effect of trial number and sample size on group level recovery

Data quality		# Participants	
# Trials		N=15	N=50
	T=40	Poor	Med.
	T=200	Med.	High

Does the number of trials affect parameter recovery ?

Does the number of participants affect parameter recovery ?

Study the effect of trial number and sample size on group level recovery

How bad is parameter recovery when using suboptimal combination of # Participants & # Trials

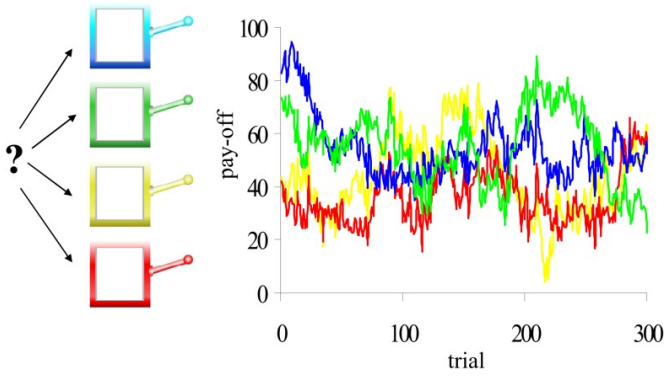
(e.g. typical of fMRI studies)

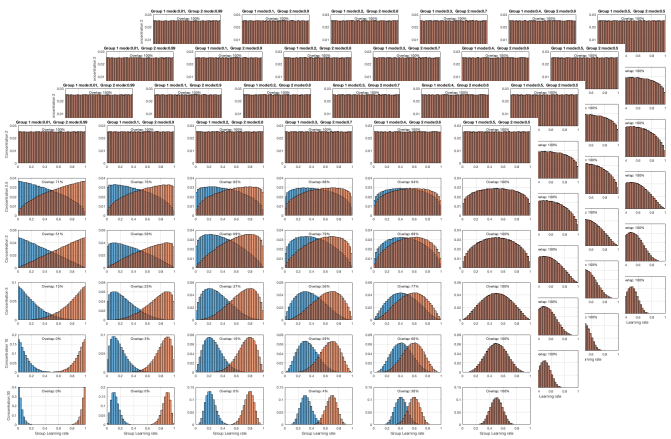
Data quality		# Participants	
		N=15	N=50
# Trials	T=40	Poor	Med.
	T=200	Med.	High

Does the number of trials affect parameter recovery ?

Does the number of participants affect parameter recovery ?

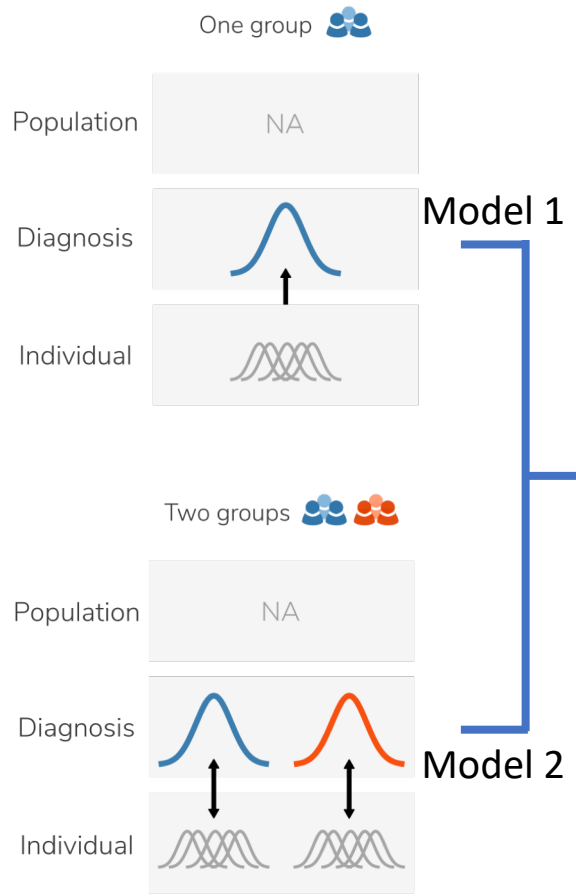
Reinforcement Learning task





Generate synthetic data
for case-control studies

Modelling assumptions



Extract

Effect Size (ES)

Metrics

- False Positive Rate
- False Negative Rate
- F1-score
- Effect Size Recovery error

Perturbations

Data quality		# Participants	
		N=15	N=50
# Trials	T=40	Poor	Med.
	T=200	Med.	High

1 – Initial approach :

Solve model specification using Model comparison

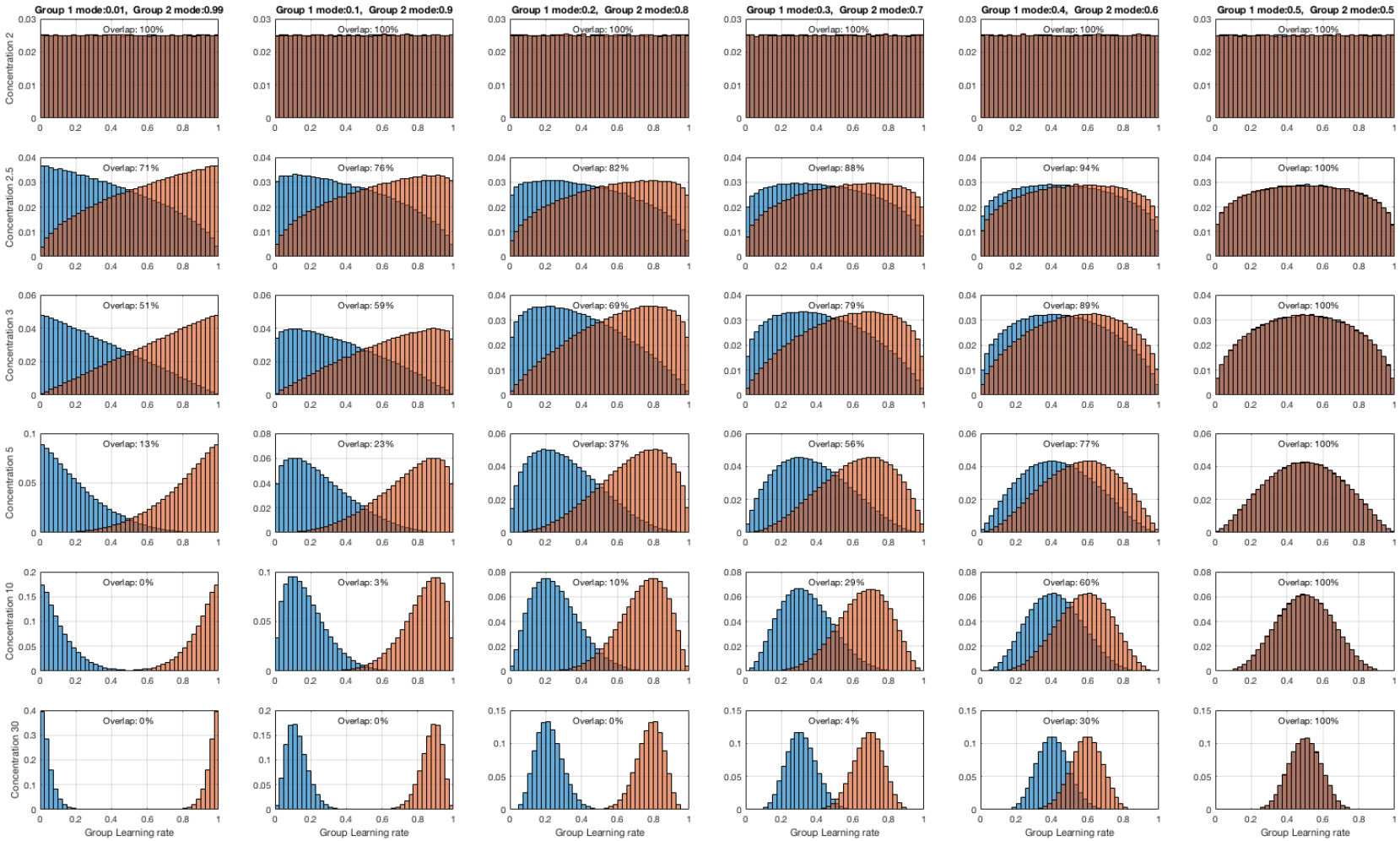
Model Comparison

One group

Two groups

Model 1

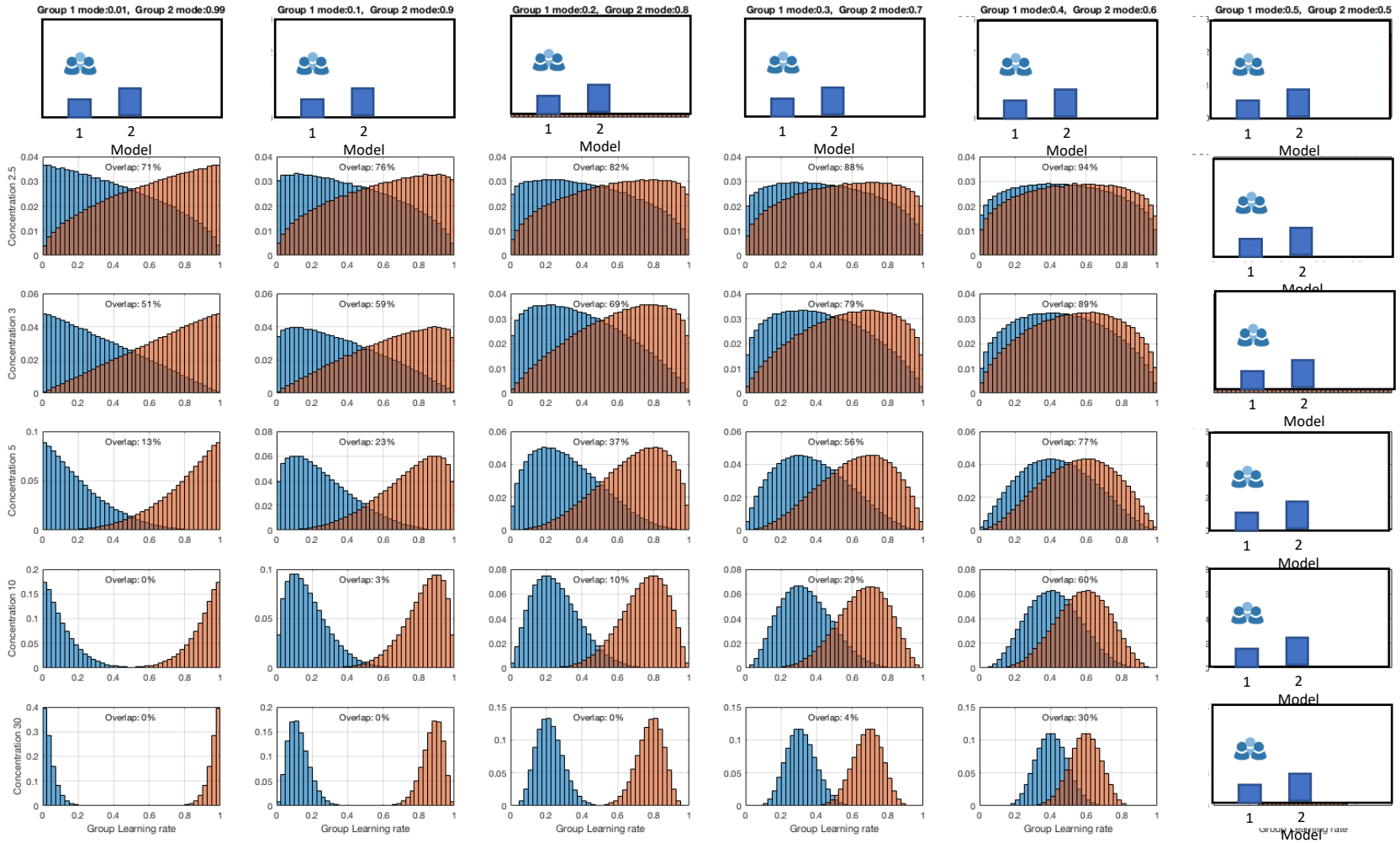
Model 2



Model Comparison (Predictions)

One group Two groups Common population

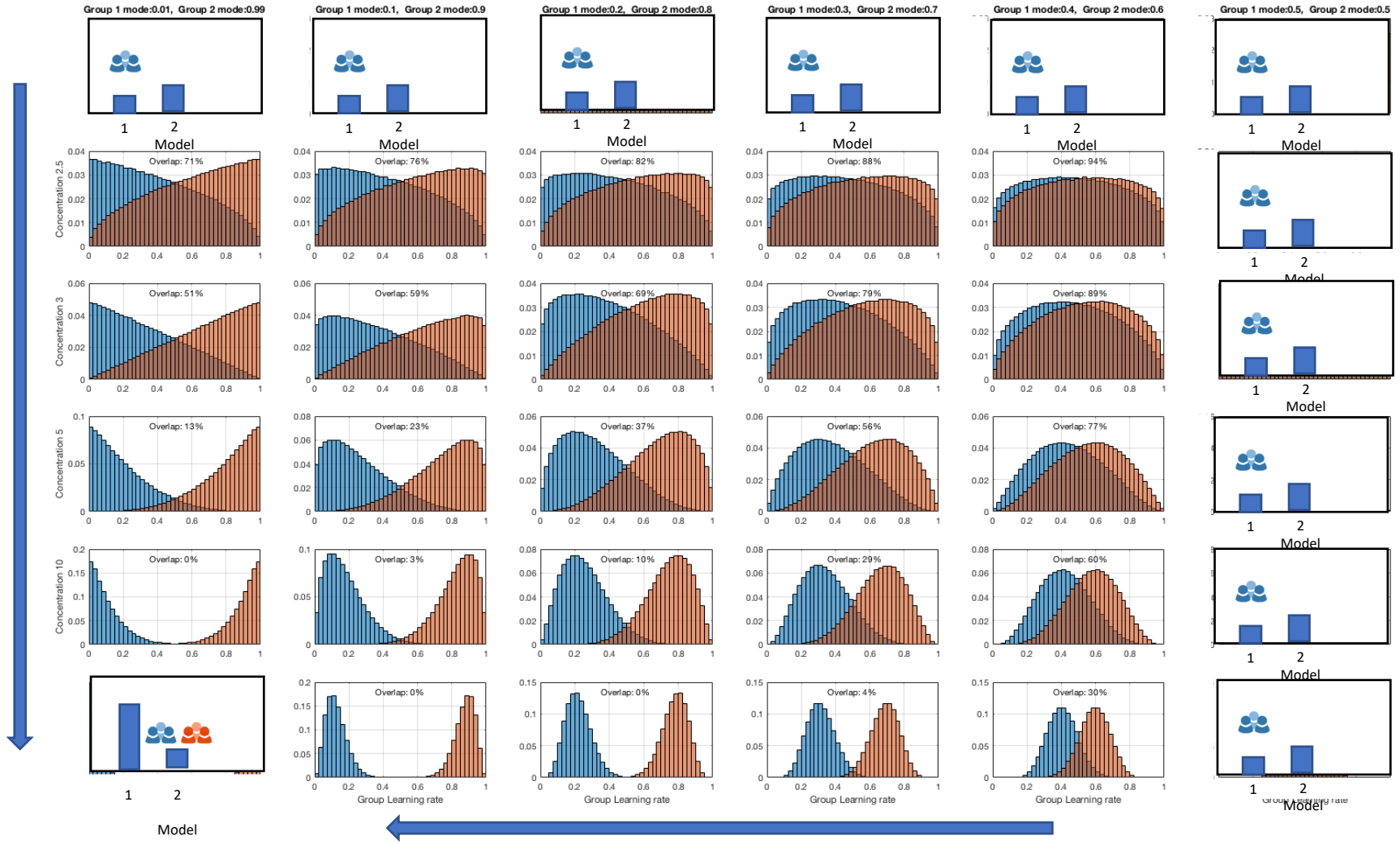
Model 1 Model 2 Model 3



One group Two groups Common population

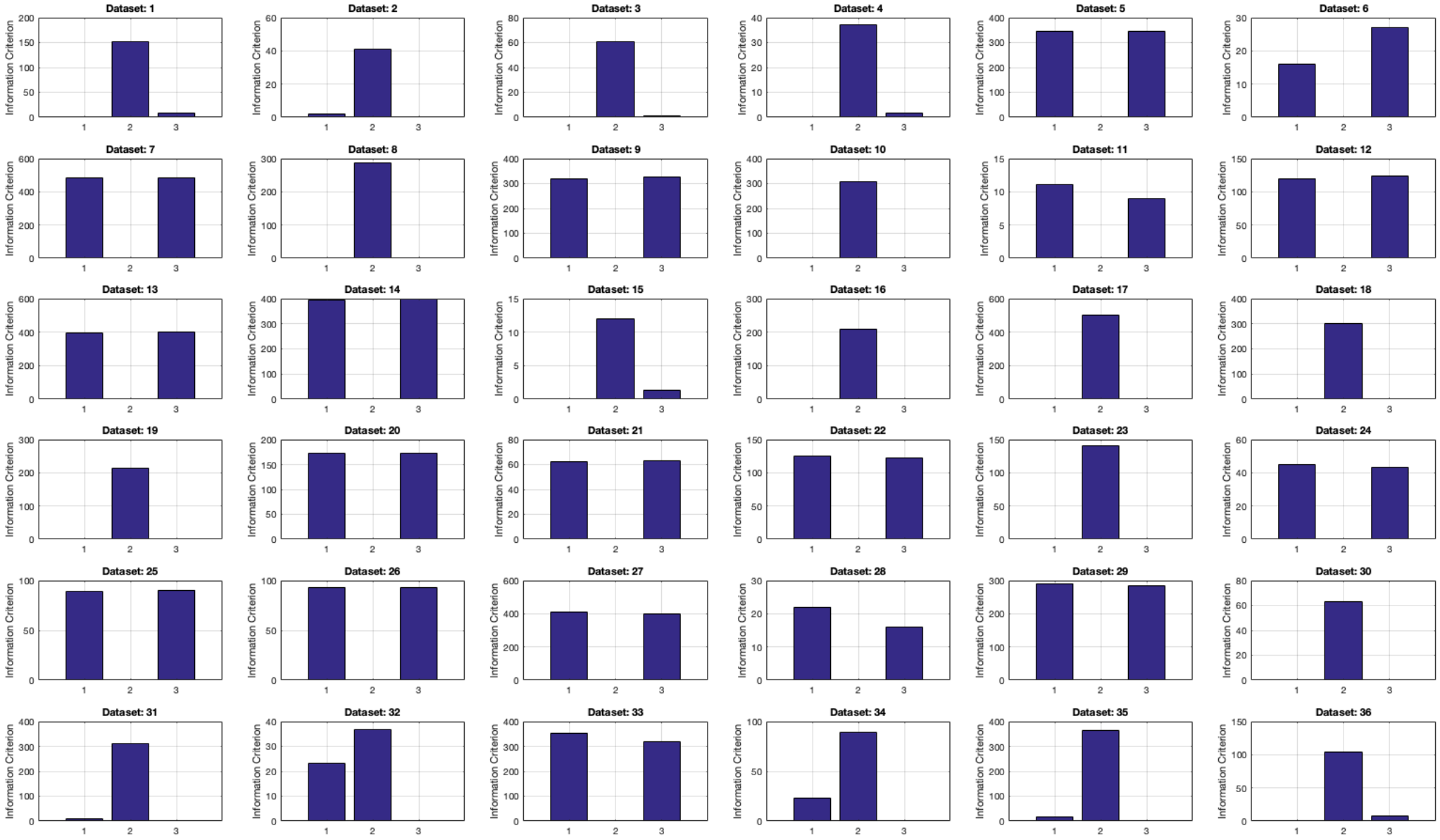
Model Comparison (Predictions)

Model 1 Model 2 Model 3

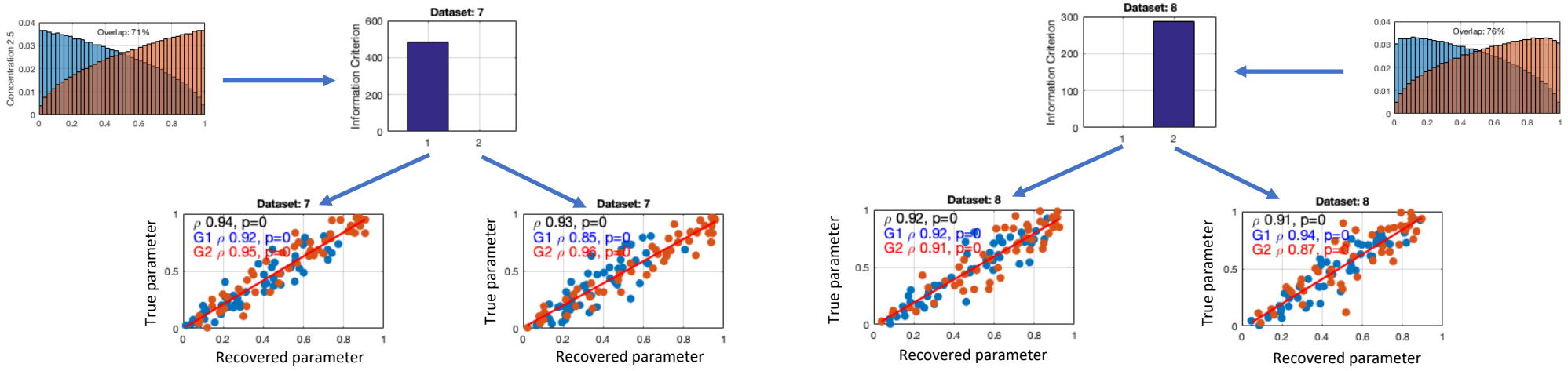


Model comparison (WAIC)

Watanabe-Akaike information Criterion (WAIC)



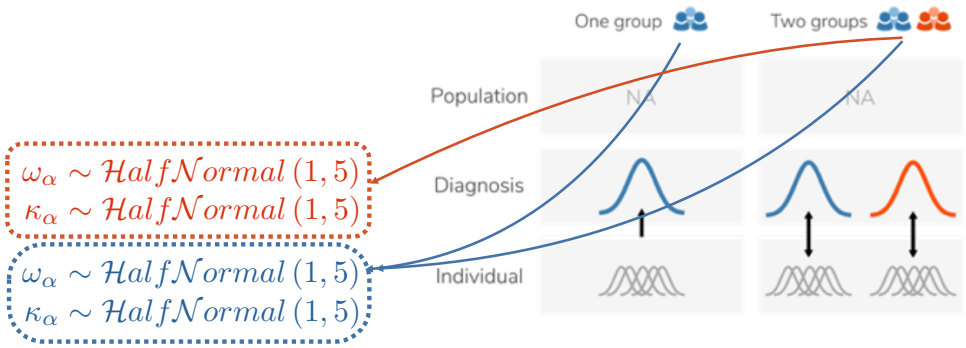
Model comparison (WAIC)



2 – Can we recover Effect Sizes?

Reinforcement Learning model specifications

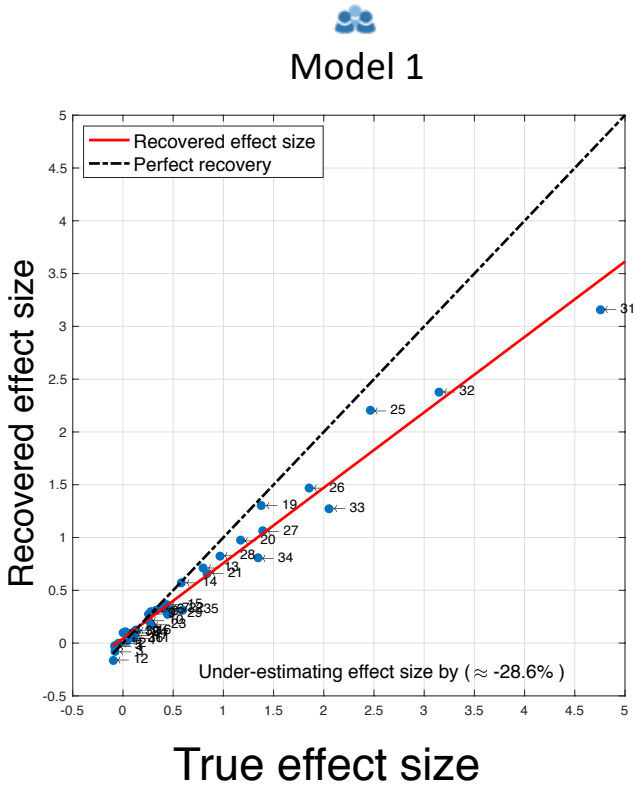
Reinforcement Learning Model **Prior specifications**

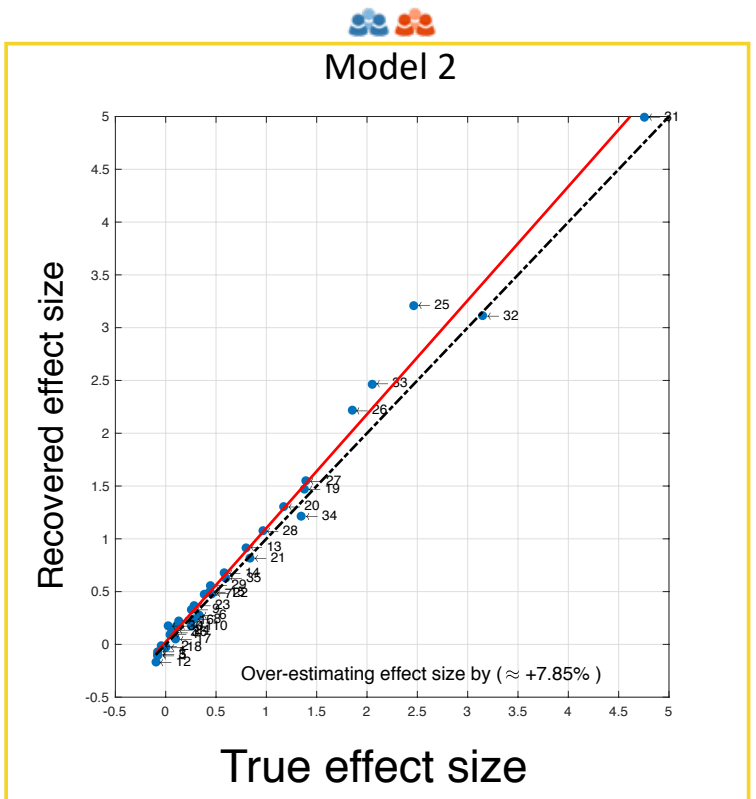


$$\alpha \sim beta\,(\omega_\alpha, \kappa_\alpha)$$
$$\tau \sim beta\,(\omega_\tau, \kappa_\tau)$$
$$V\,(s) = V\,(s) + \alpha\,(r - V\,(s))$$
$$P\,(a|\tau, V) = softmax\,(V|\tau)$$
$$choice \sim categorical\,(P\,(a|\tau, V))$$

Results: Effect size recovery (using sampling)

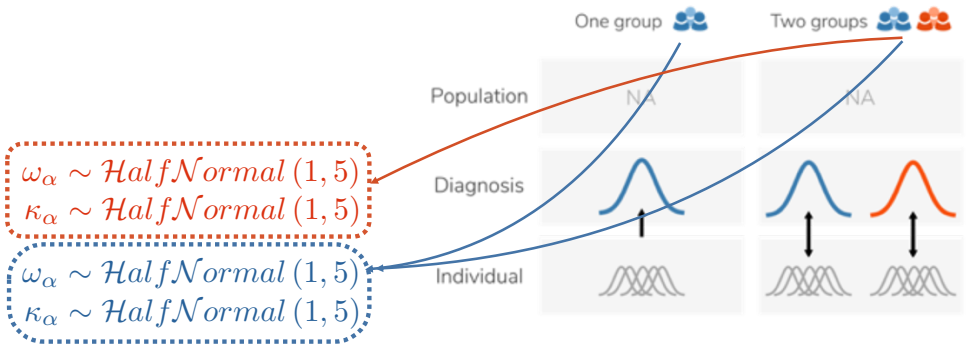
Model 1 underestimates effect sizes, Model 2 overestimates effect sizes (#datasets= 36, high data quality)





Reinforcement Learning model specifications

Reinforcement Learning Model **Prior specifications**



$$\alpha \sim beta\,(\omega_\alpha, \kappa_\alpha)$$
$$\tau \sim beta\,(\omega_\tau, \kappa_\tau)$$
$$V\,(s) = V\,(s) + \alpha\,(r - V\,(s))$$
$$P\,(a|\tau, V) = softmax\,(V|\tau)$$
$$choice \sim categorical\,(P\,(a|\tau, V))$$

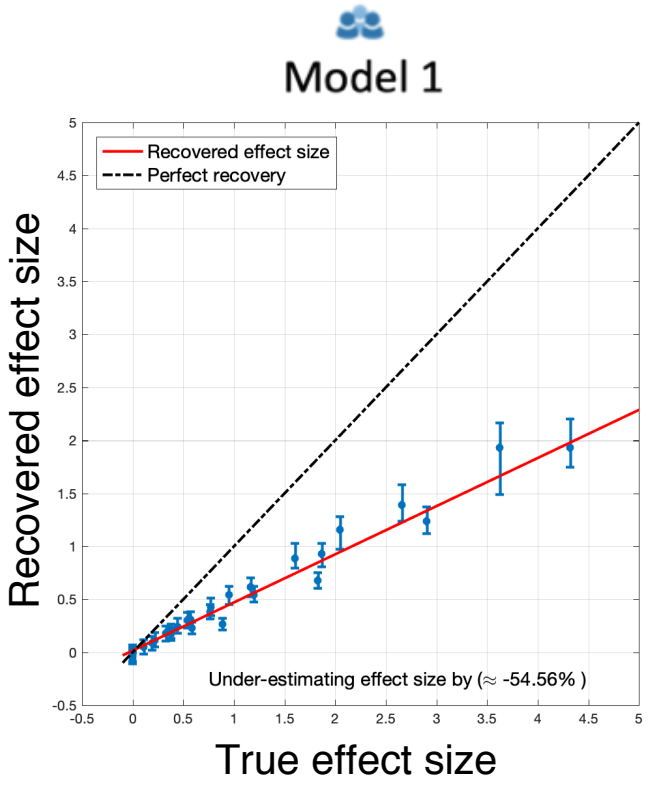
Robustness: F1-Score, and data perturbations

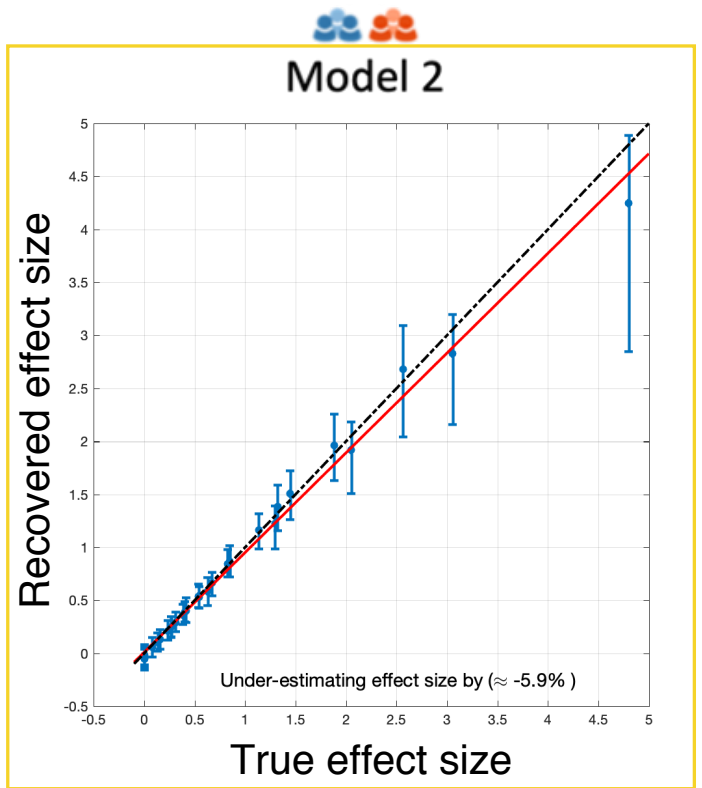
Table 1: Model accuracy: False positive rate, false negative rate, F1-Score, and 95% CI

50 subj. 200 trials	False Pos. Rate (%)	False Neg. Rate (%)	F1-Score (%)
Model 1	0.48 [± 0.07]	6.03 [± 0.24]	96.73 [± 0.23]
Model 2	2.66 [± 0.16]	**1.75 [± 0.13]**	**98.26 [± 0.17]**

Robustness: False Pos., False Neg. (using VB)

Model 1 vastly underestimates effect sizes, Model 2 slightly underestimate effect sizes (#datasets= 36,000 — high data quality)



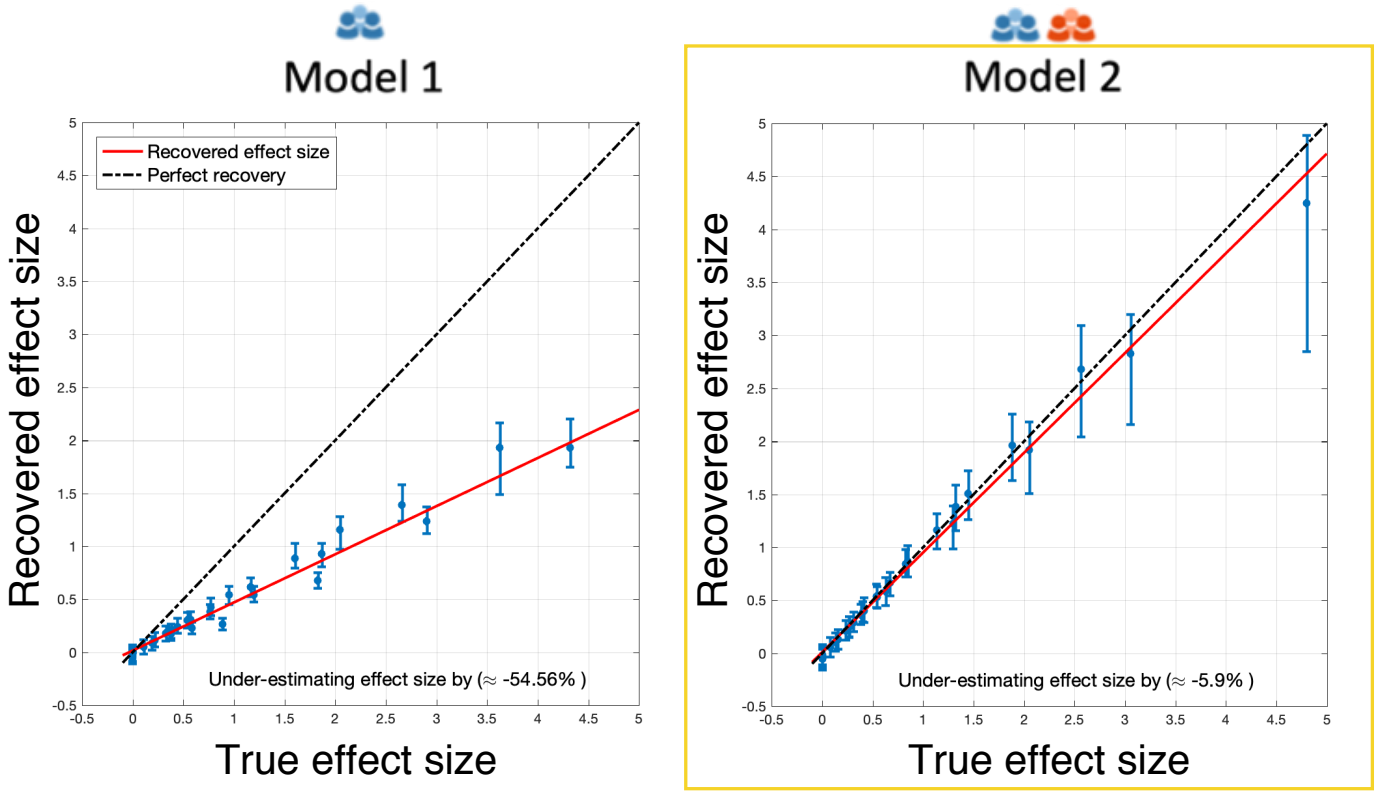


Data Perturbations

Data quality		# Participants	
# Trials		N=15	N=50
	T=40	Poor	Med.
	T=200	Med.	High

Robustness: False Pos., False Neg. (using VB)

Model 1 vastly underestimates effect sizes, Model 2 slightly underestimate effect sizes (#datasets= 36,000 — high data quality)



Robustness: F1-Score, and data perturbations

Table 1: Model accuracy: False positive rate, false negative rate, F1-Score, and 95% CI

50 subj. 200 trials	False Pos. Rate (%)	False Neg. Rate (%)	F1-Score (%)
Model 1	0.48 [± 0.07]	6.03 [± 0.24]	96.73 [± 0.23]
Model 2	2.66 [± 0.16]	**1.75 [± 0.13]**	**98.26 [± 0.17]**

Table 2: E.S. error (%) - positive and negative sign denote over/under estimation

Data perturb.	Effect size recovery error (%) *50 subj. 200 trials*	*15 subj. 200 trials*	*50 subj., 40 trials*	*15 subj. 40 trials*
Model 1	-28.60	-17.74	-63.81	-64.46
Model 2	**+7.85**	**+9.41**	**+28.12**	**+29.09**

Acknowledgements

NEURAL INFORMATION PROCESSING SYSTEMS

ML4H Workshop

NeurIPS 2020

https://arxiv.org/abs/2011.01725

Funding

Supported by wellcome trust

NIHR | National Institute for Health Research

NIHR | University College London Hospitals Biomedical Research Centre

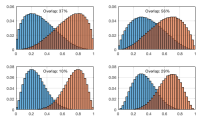
Collaborators

- O. Robinson (ICN, UCL)

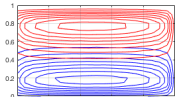
- T. Wise (MPC, UCL)

Data quality		# Participants	
		N=15	N=50
# Trials	T=40	Poor	Med.
	T=200	Med.	High

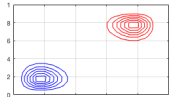
Datasets



Dataset 1



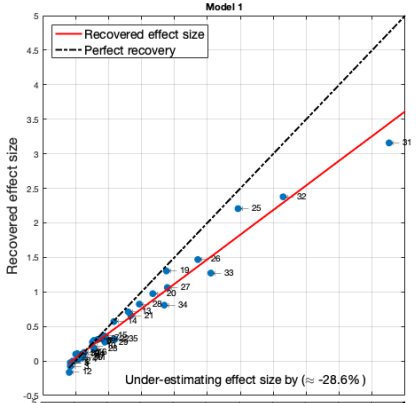
Dataset 2

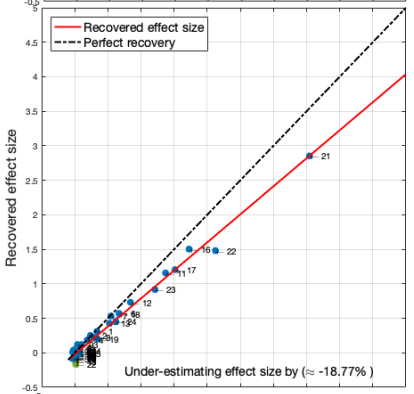


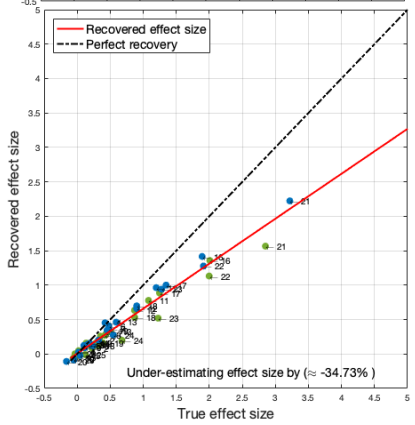
Dataset 3

Models

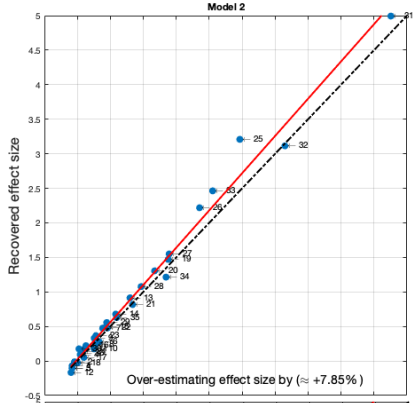
Model 1

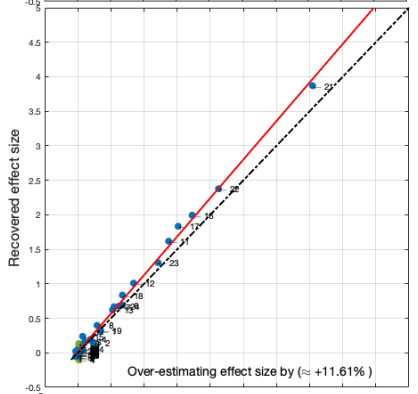


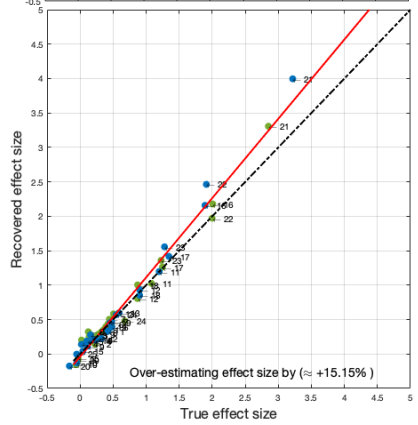




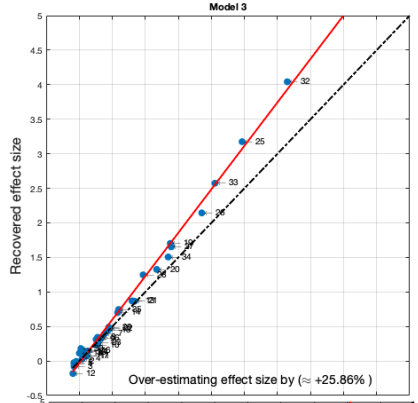
Model 2

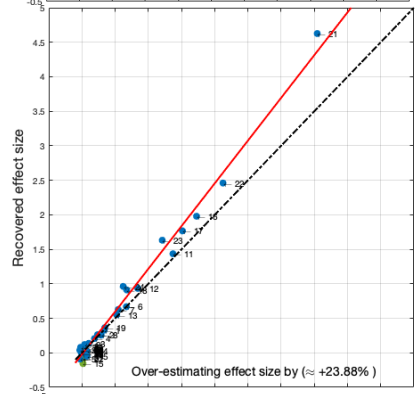


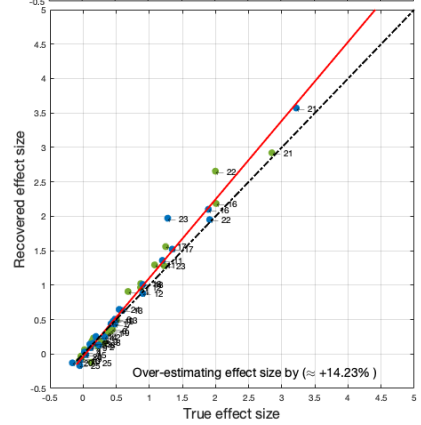




Model 3

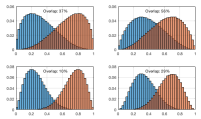




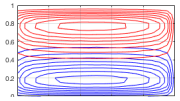


Data quality		# Participants	
		N=15	N=50
# Trials	T=40	Poor	Med.
	T=200	Med.	High

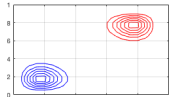
Datasets



Dataset 1



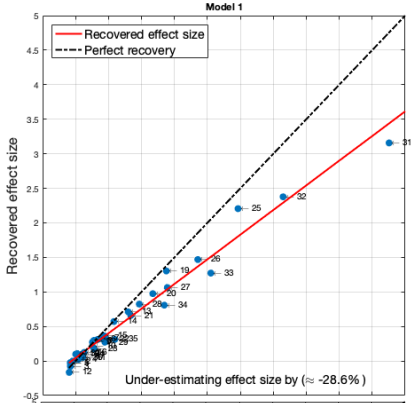
Dataset 2

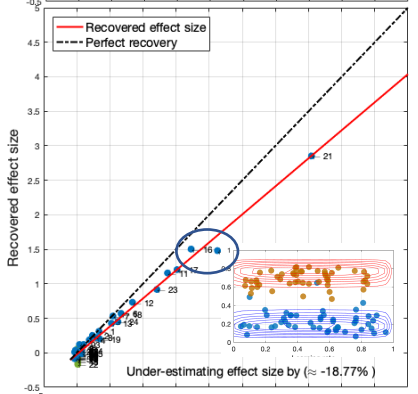


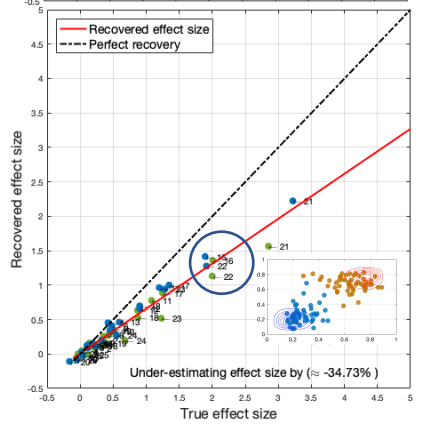
Dataset 3

Models

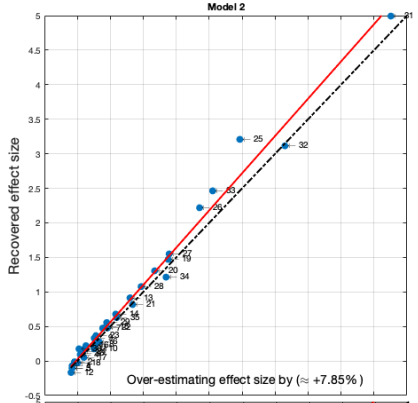
Model 1

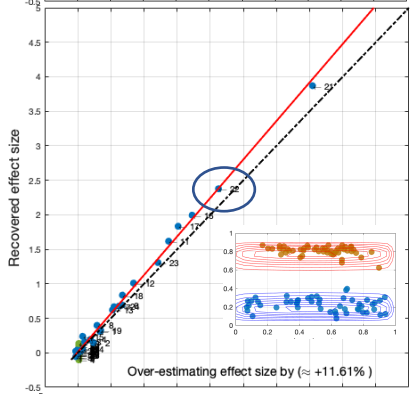


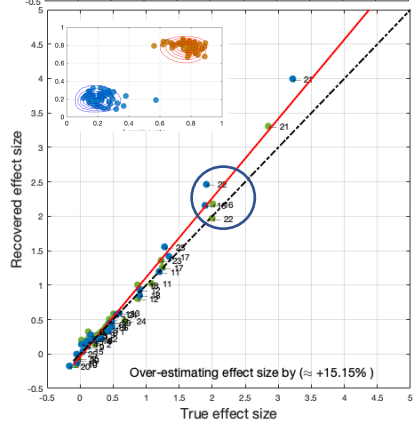




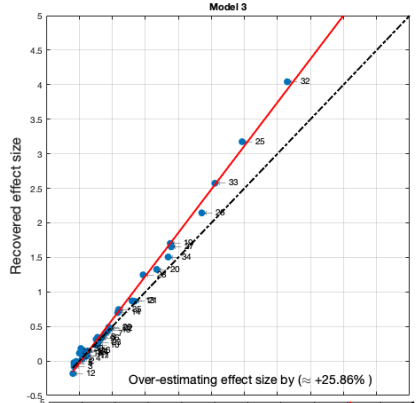
Model 2

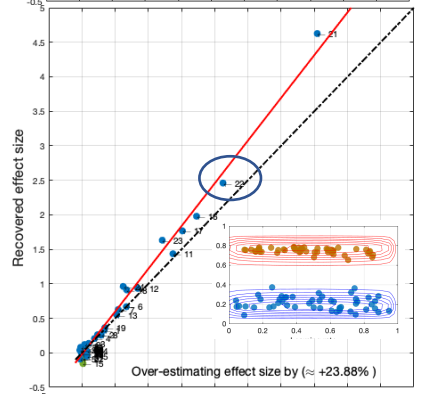


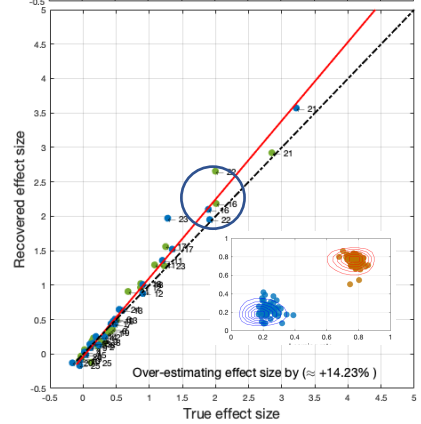




Model 3







Guidelines

- Artificially pooling participants together appear to:
  - Underestimate true effect sizes (sometimes vastly)
  - Can lead to abnormal parameter recovery
  - Can lead to increased false negative discovery

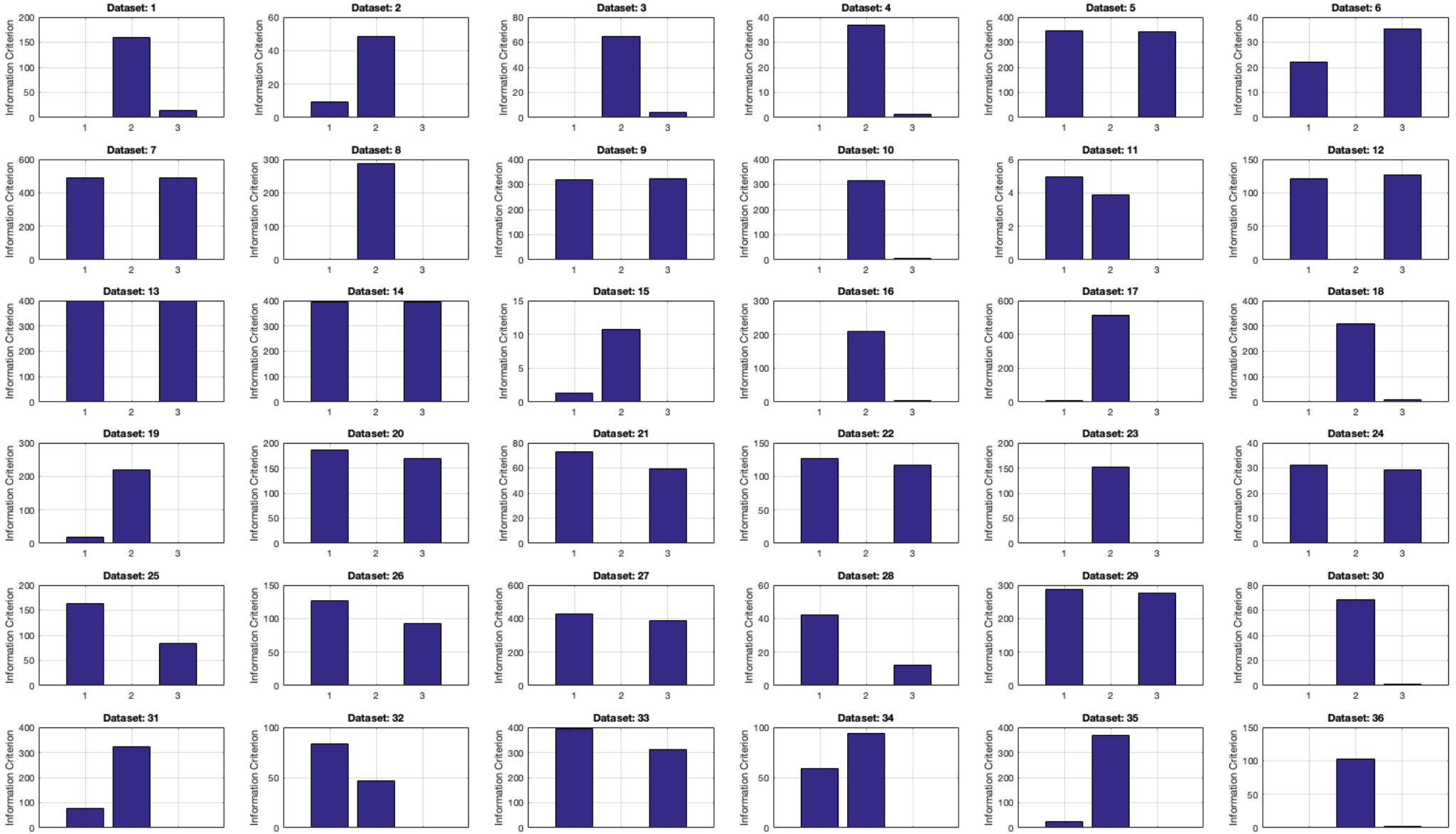
    -> huge potential impact

- Partial-pooling by group/manipulation/diagnostic appear to:
  - Overestimate true effect sizes (albeit slighty)
  - Can lead to increased false positive discovery

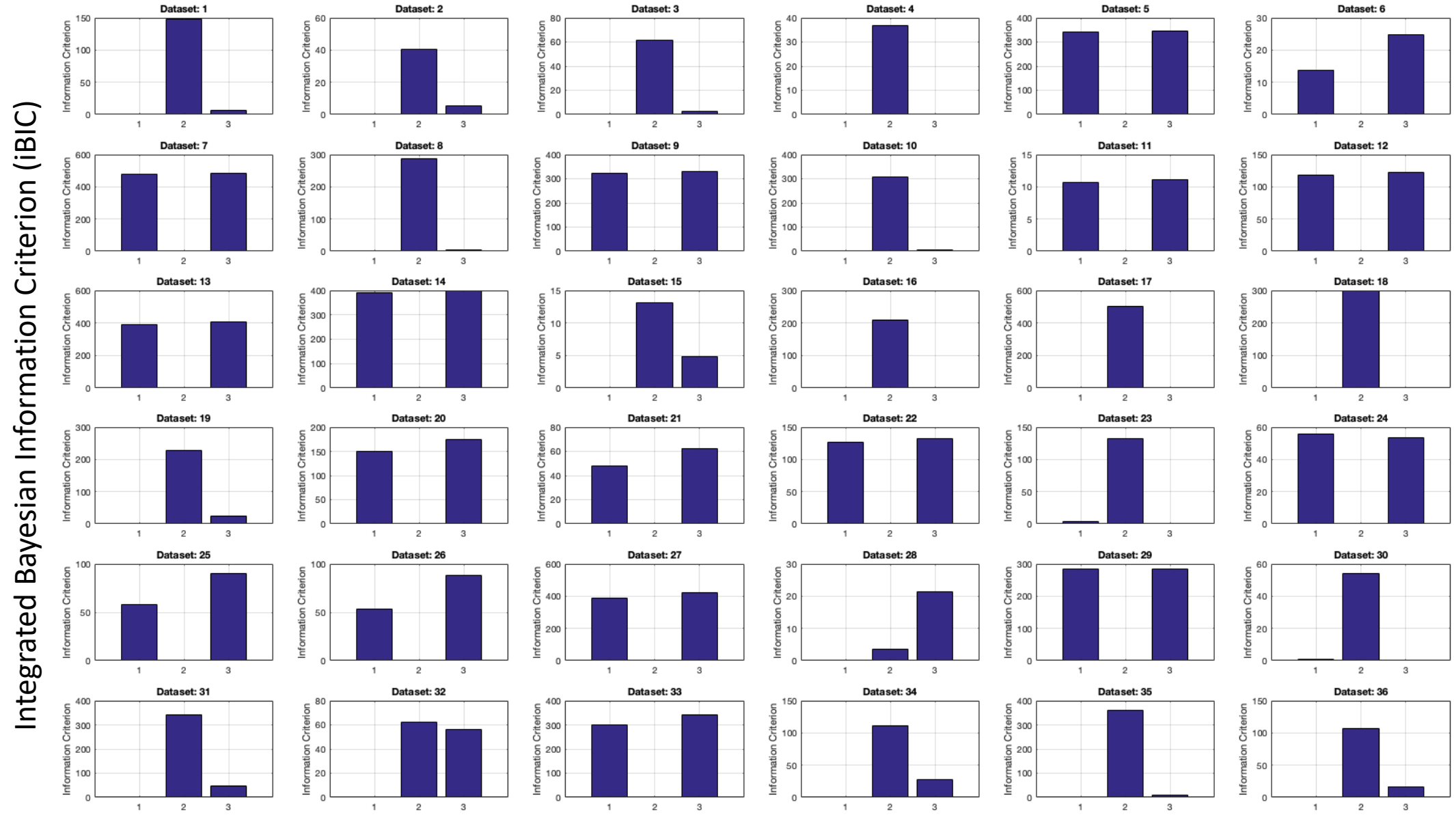
    -> can rectify this with replication studies

Model comparison (LOO)

PSIS-Leave One Out

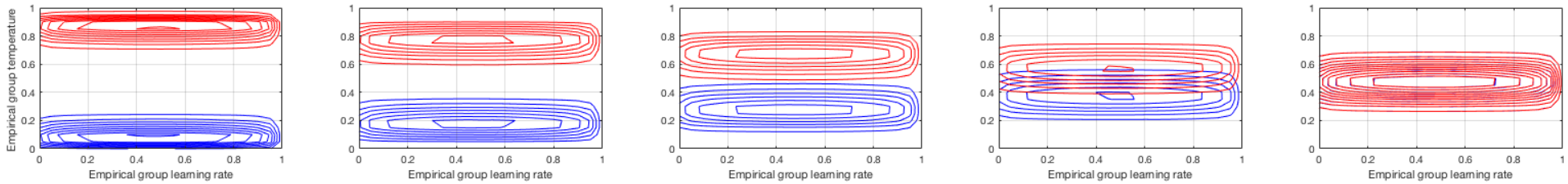


Model comparison (iBIC)



Simulation study – Synthetic data generation

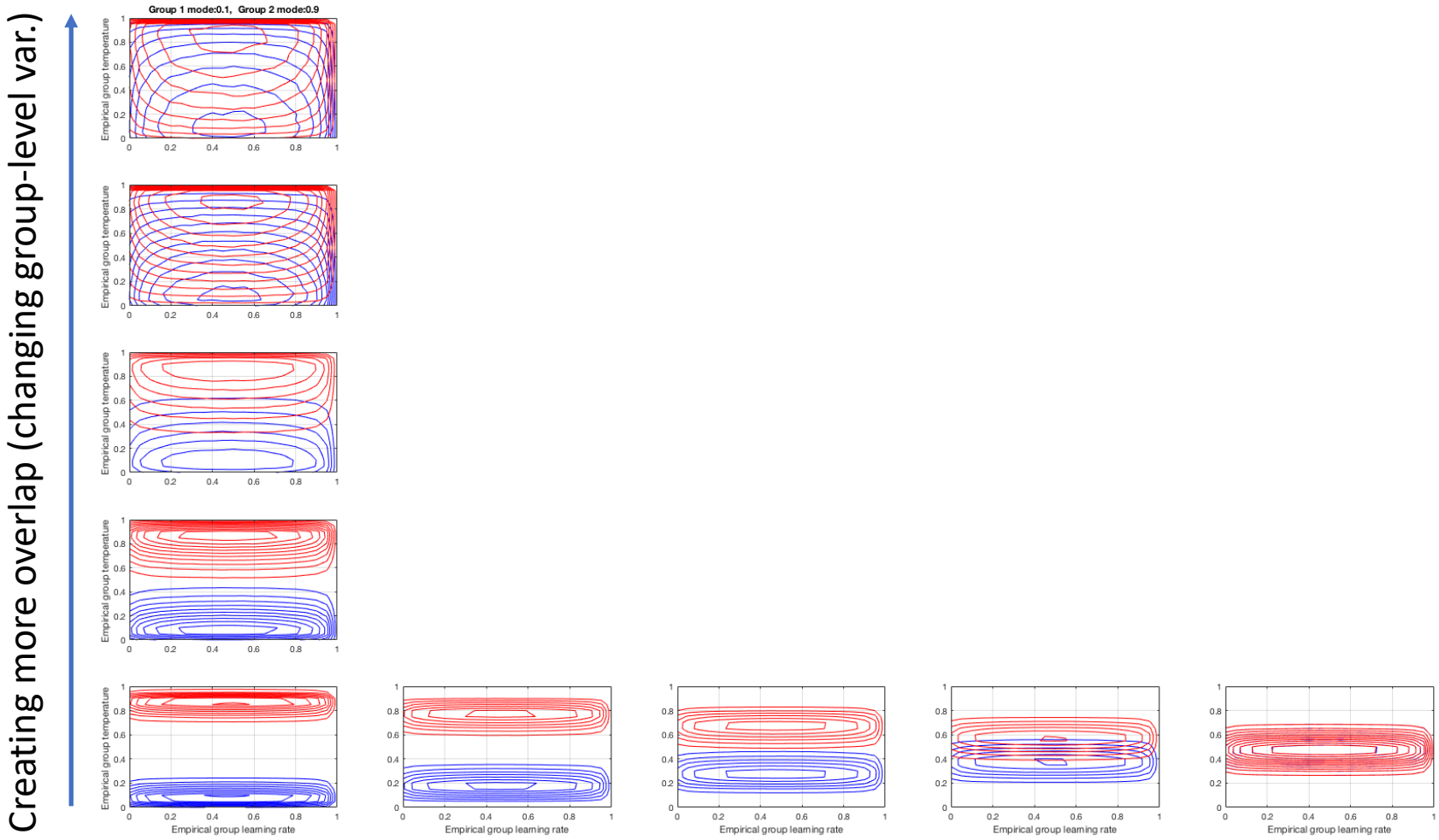
Dataset 2: Two parameter estimation:
- 1 parameter overlapping (learning rate)
- 1 parameter differing (temperature)



Creating more overlap in Temperature parameter (changing group-level mode)

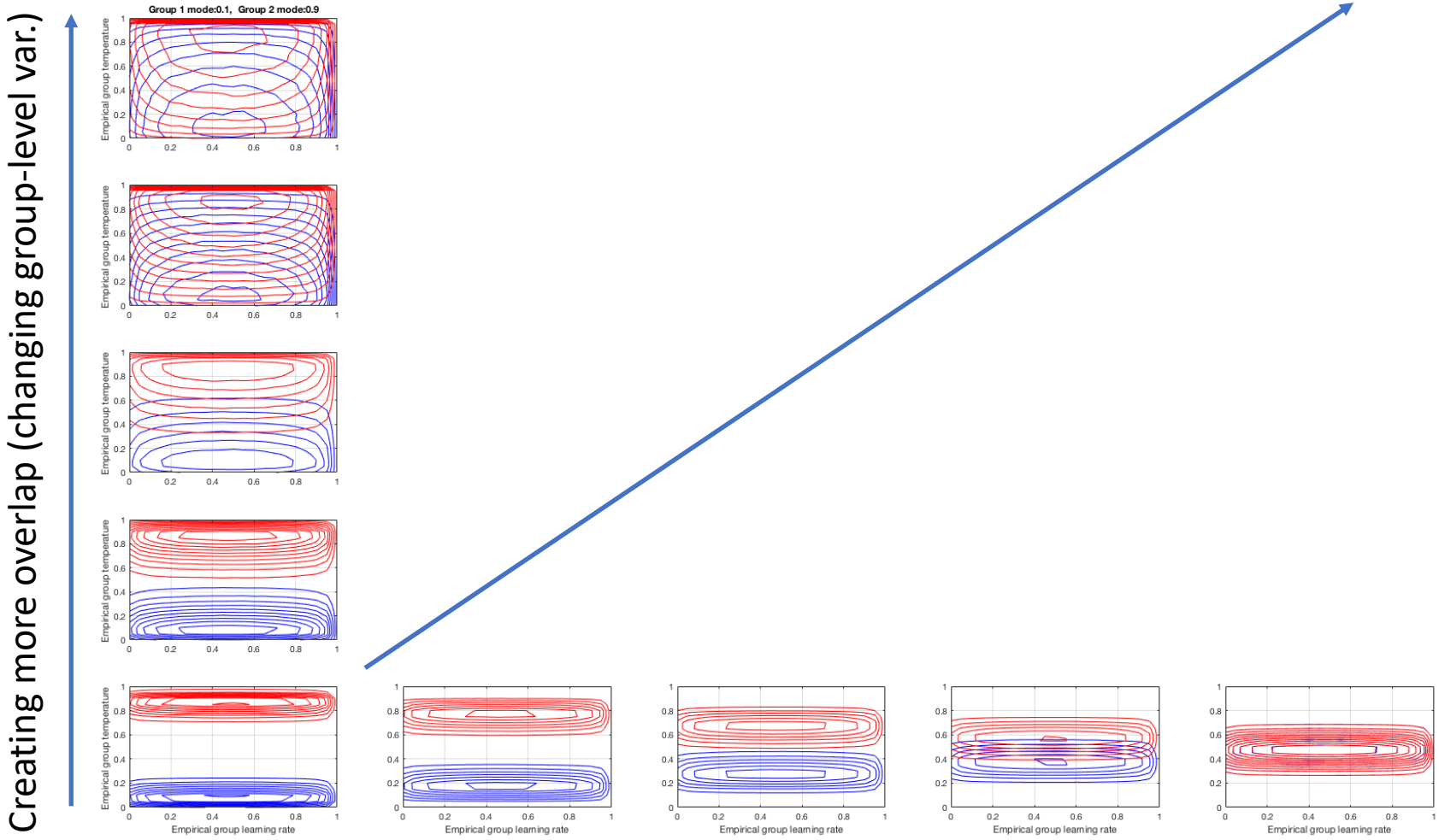
Simulation study – Synthetic data generation

Dataset 2: Two parameter estimation:
- 1 parameter overlapping (learning rate)
- 1 parameter differing (temperature)



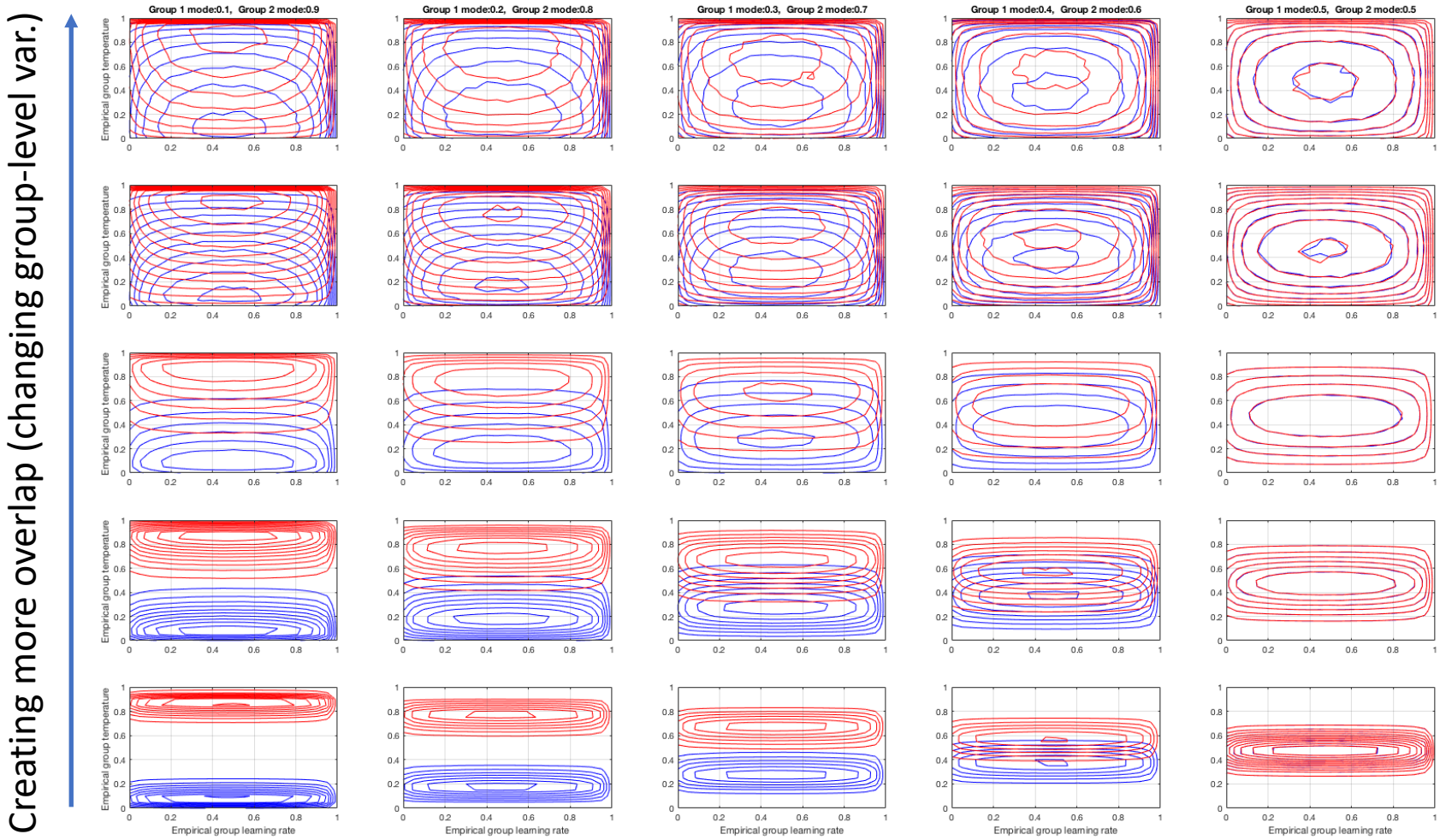
Simulation study – Synthetic data generation

Dataset 2: Two parameter estimation:
- 1 parameter overlapping (learning rate)
- 1 parameter differing (temperature)



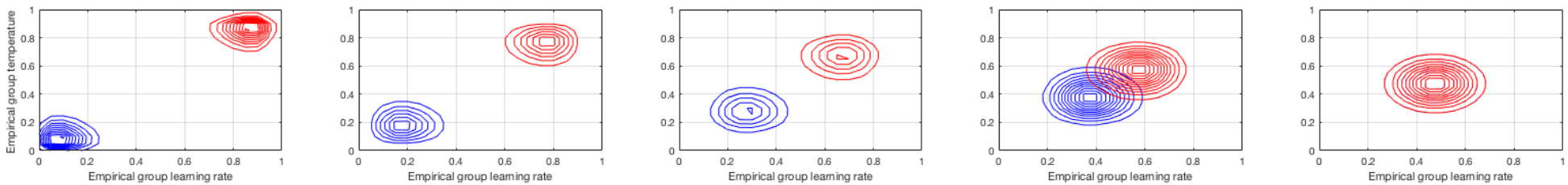
Simulation study – Synthetic data generation

Dataset 2: Two parameter estimation:
- 1 parameter overlapping (learning rate)
- 1 parameter differing (temperature)



Simulation study – Synthetic data generation

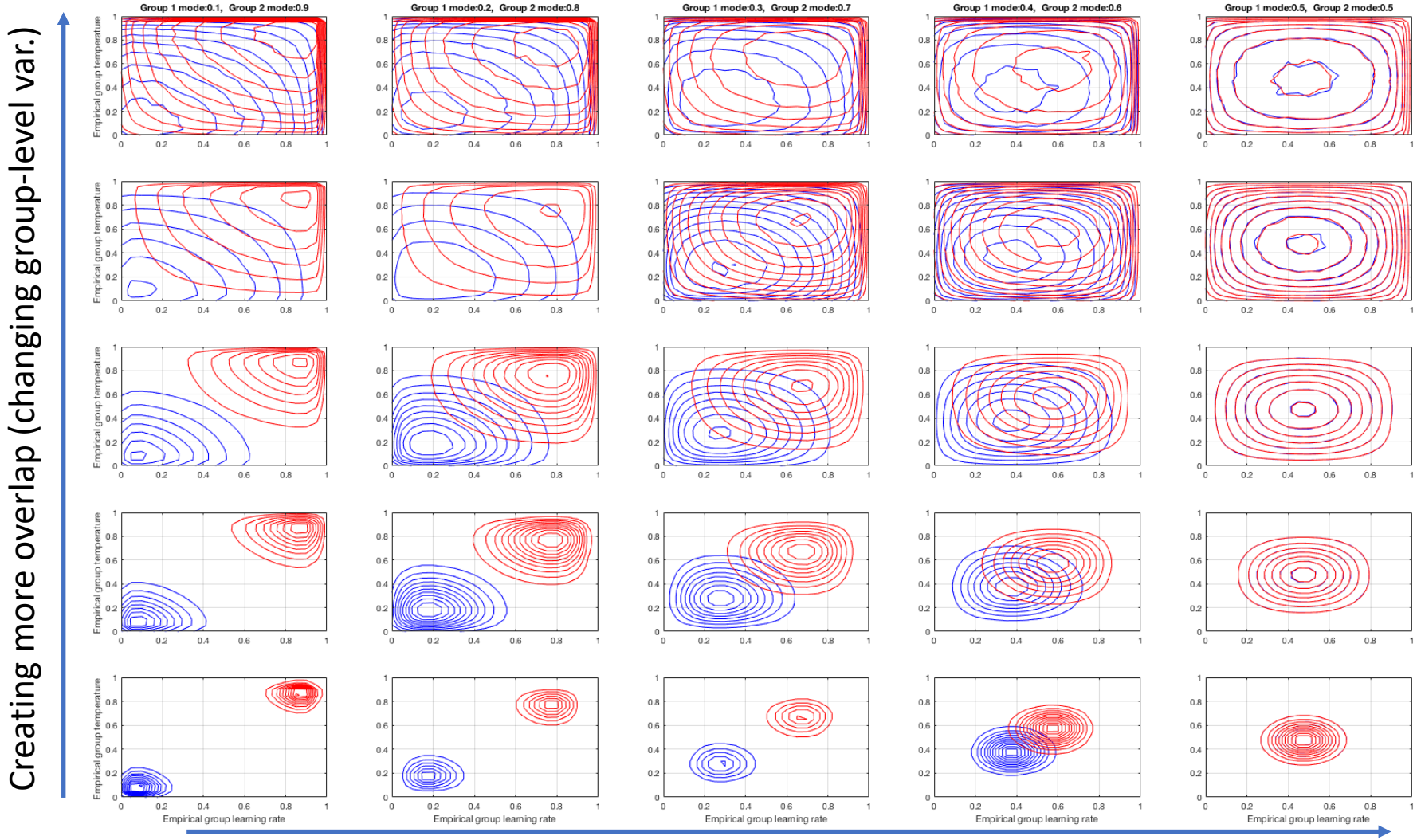
Dataset 3: Two parameter estimation: 2 parameters differing (Temperature & Learning rate)



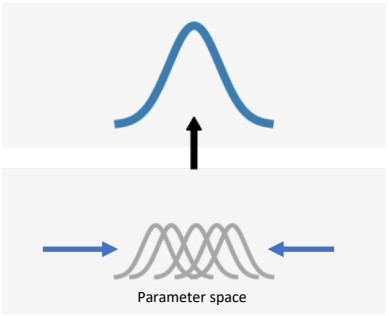
Creating more overlap in all parameter space (changing group-level mode)

Simulation study – Synthetic data generation

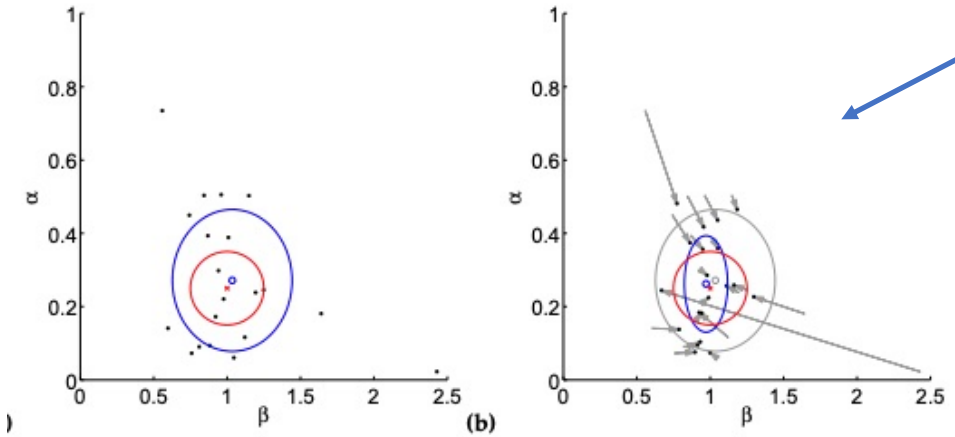
Dataset 3: Two parameter estimation: 2 parameters differing (Temperature & Learning rate)

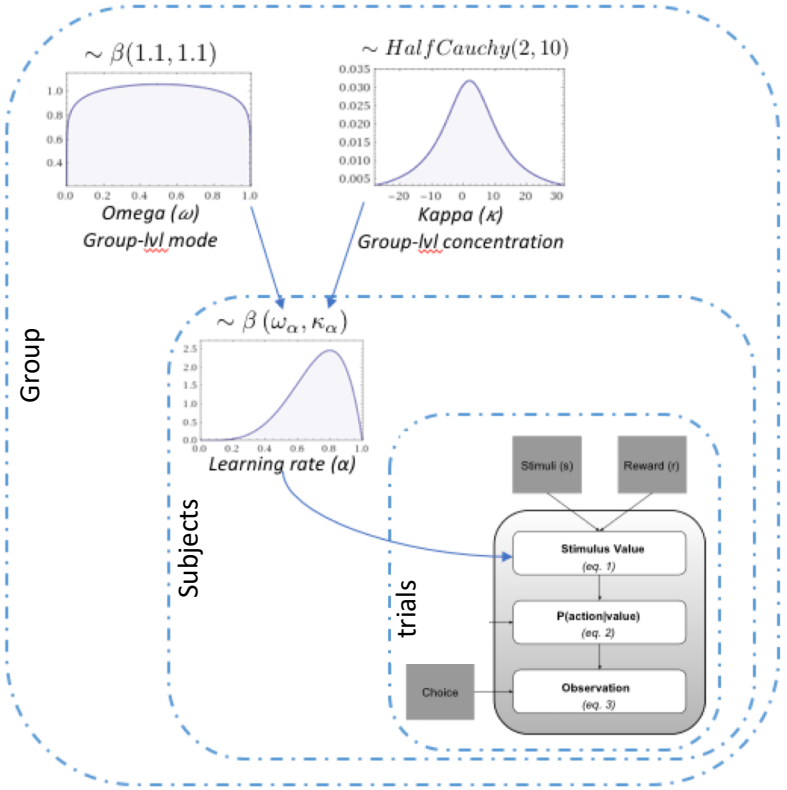


Background



Sampling (Hierarchical partial pooling)

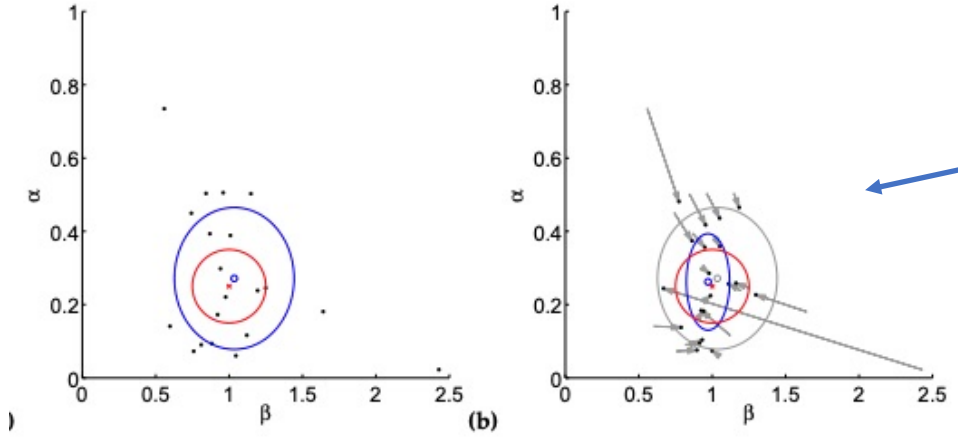


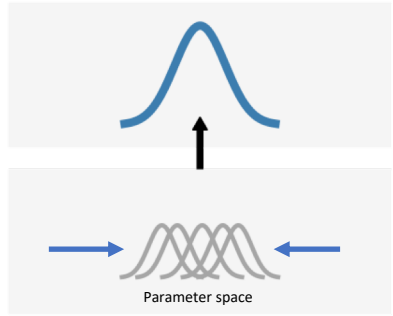


Daw 2010. “Trial-by-trial analysis of behavior”

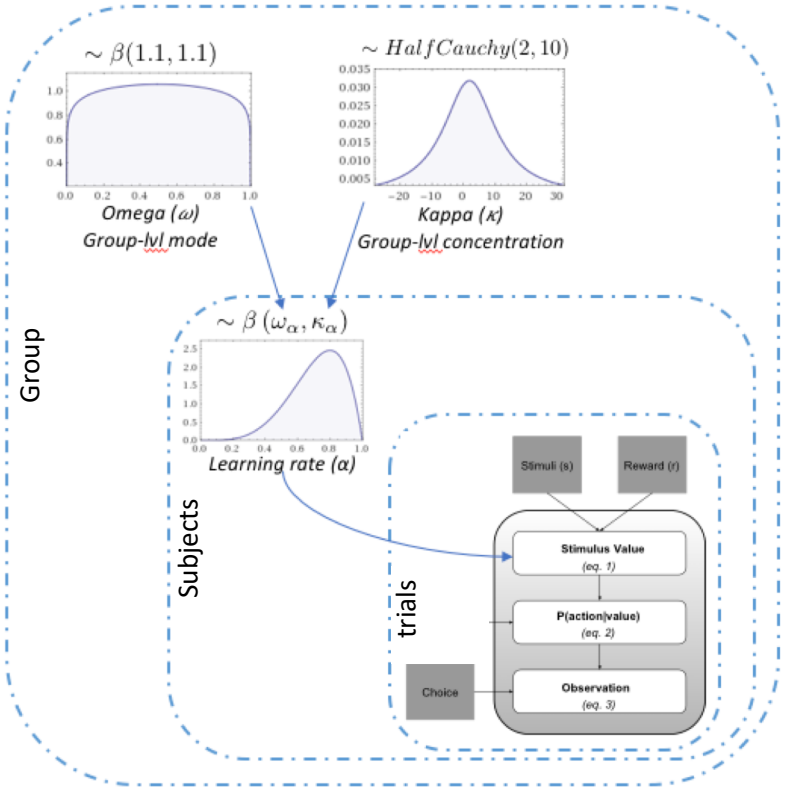
Background

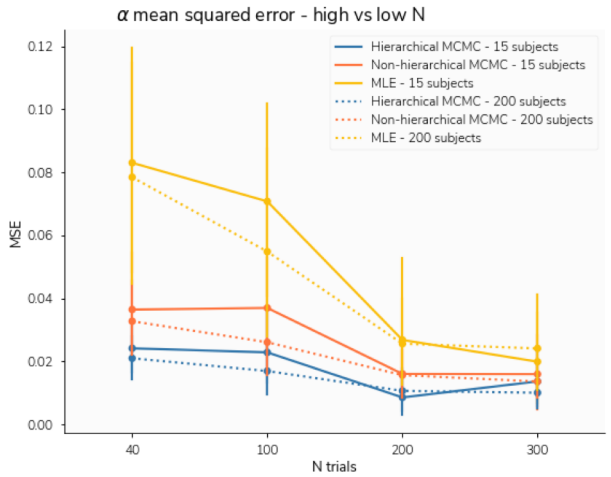
N. Daw 2010

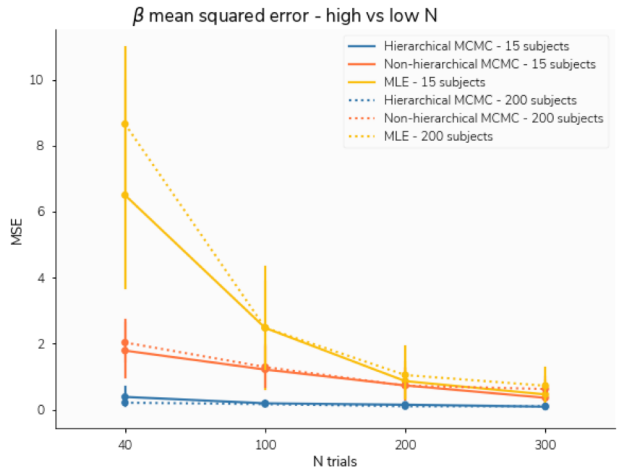




Sampling (Hierarchical partial pooling)

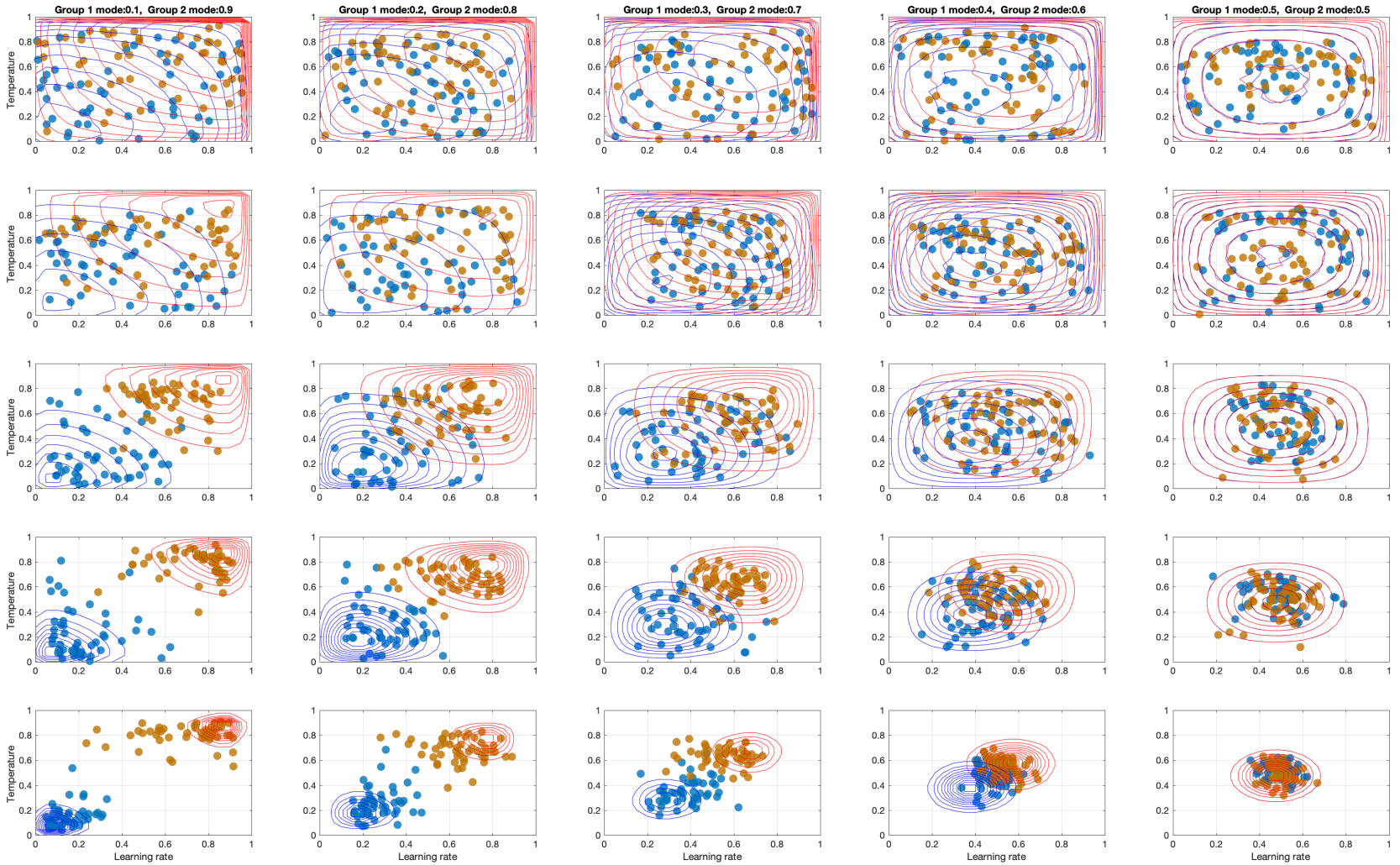






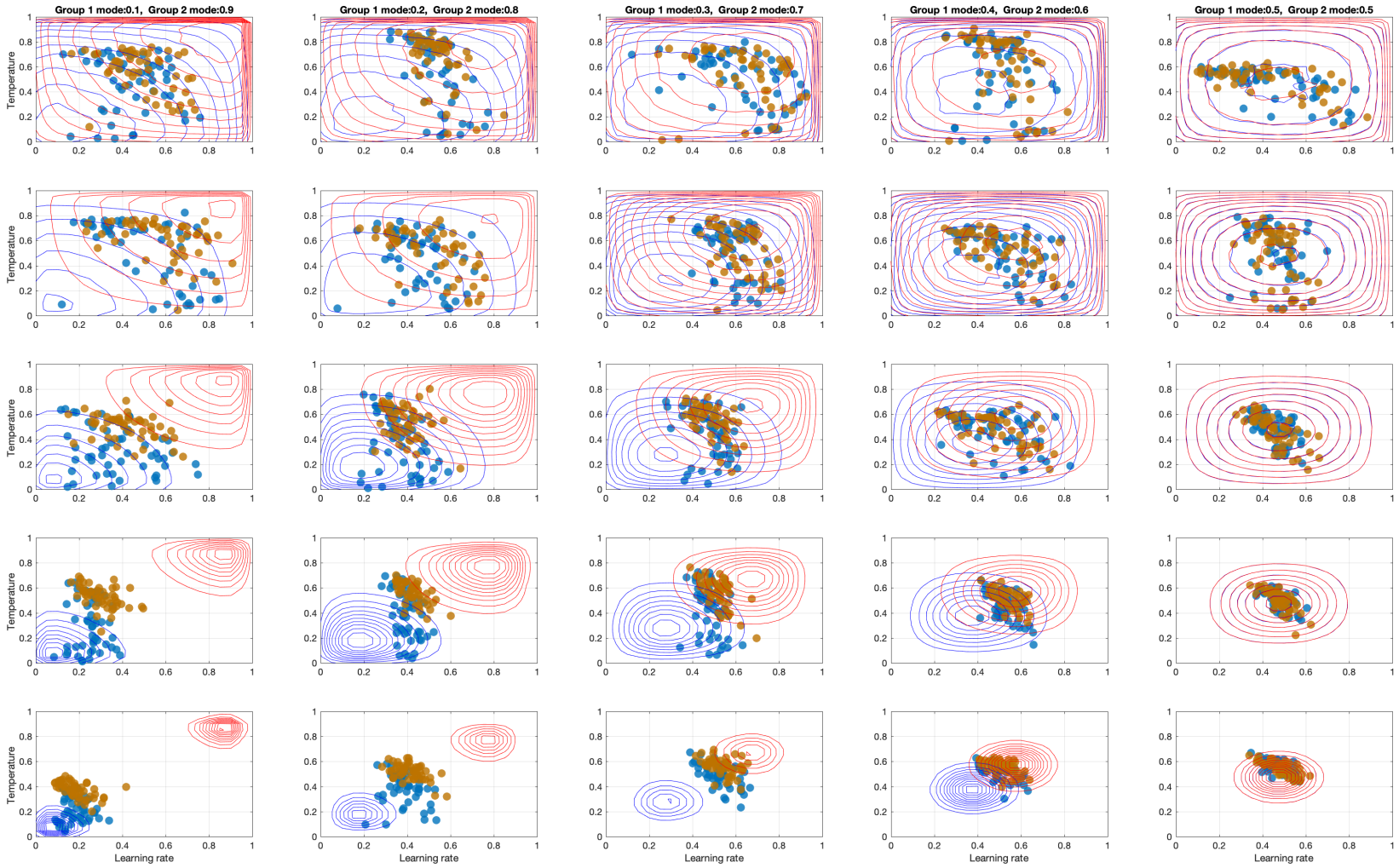
Model 1

Data quality		# Participants	
		N=15	N=50
# Trials	T=40	Poor	Med.
	T=200	Med.	High



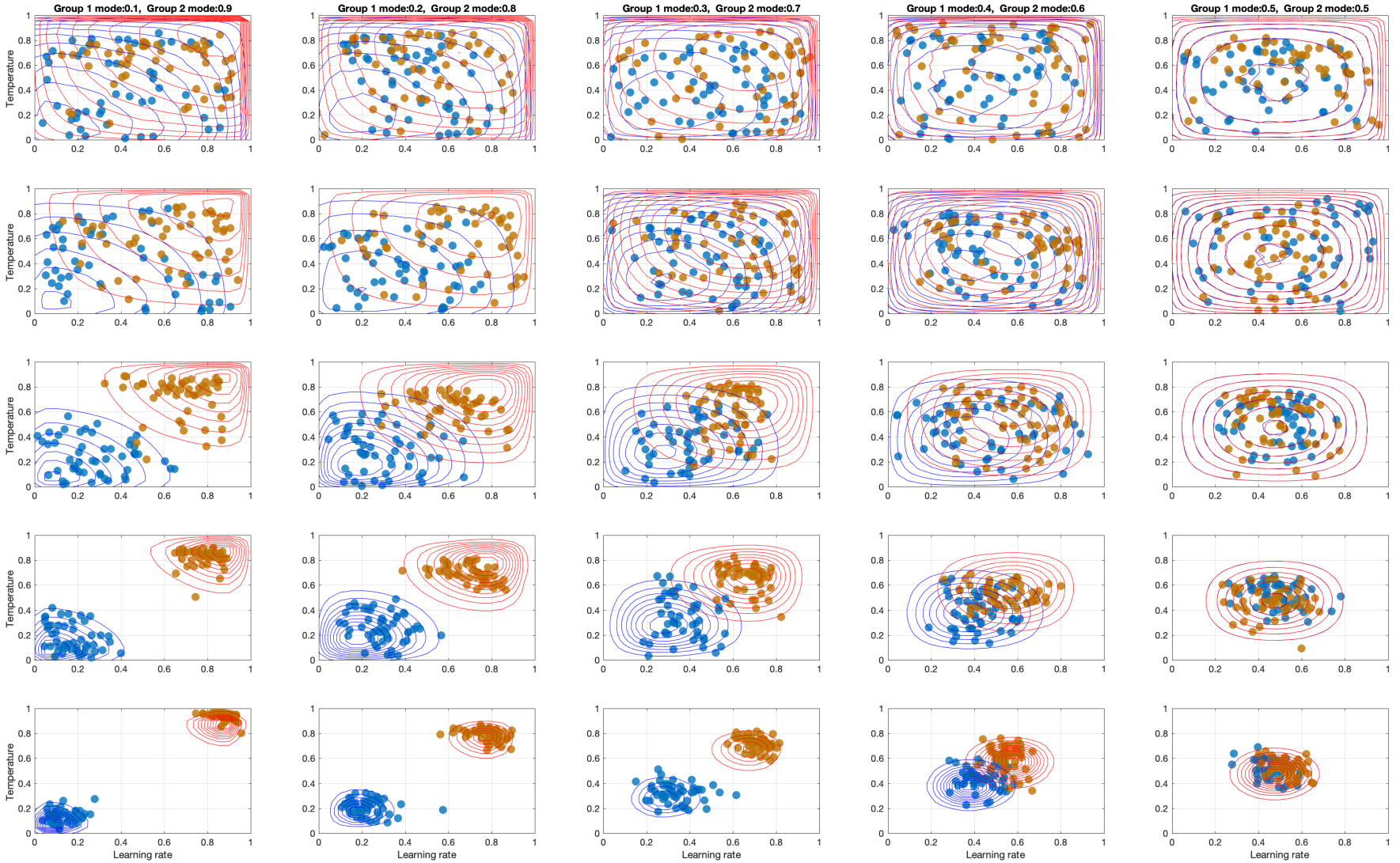
Model 1

Data quality		# Participants	
		N=15	N=50
# Trials	T=40	Poor	Med.
	T=200	Med.	High



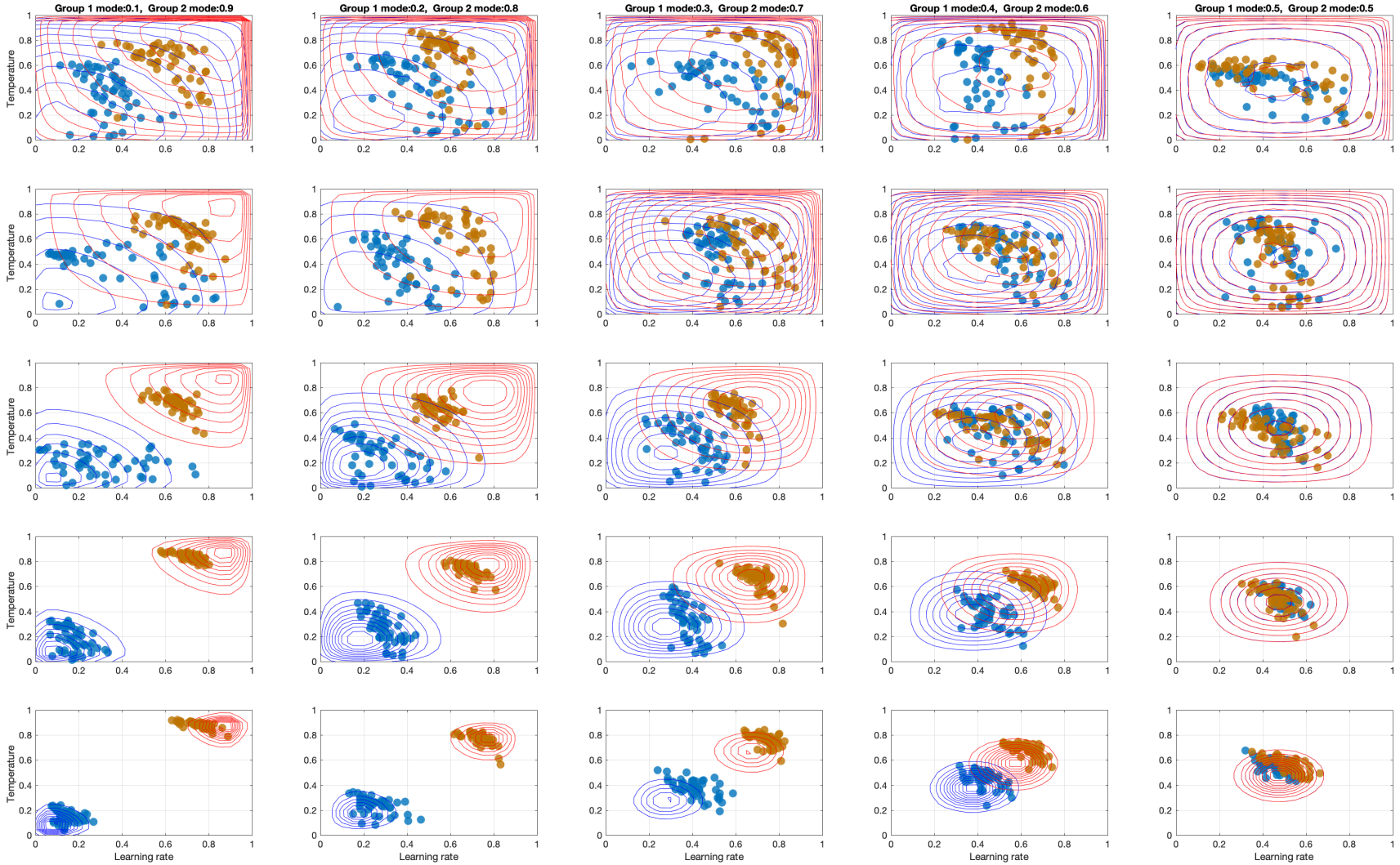
Model 2

Data quality		# Participants	
		N=15	N=50
# Trials	T=40	Poor	Med.
	T=200	Med.	High



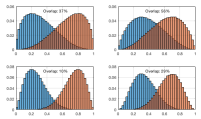
Model 2

Data quality		# Participants	
# Trials		N=15	N=50
	T=40	Poor	Med.
	T=200	Med.	High

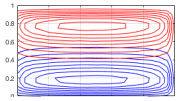


Data quality		# Participants	
		N=15	N=50
# Trials	T=40	Poor	Med.
	T=200	Med.	High

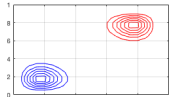
Datasets



Dataset 1



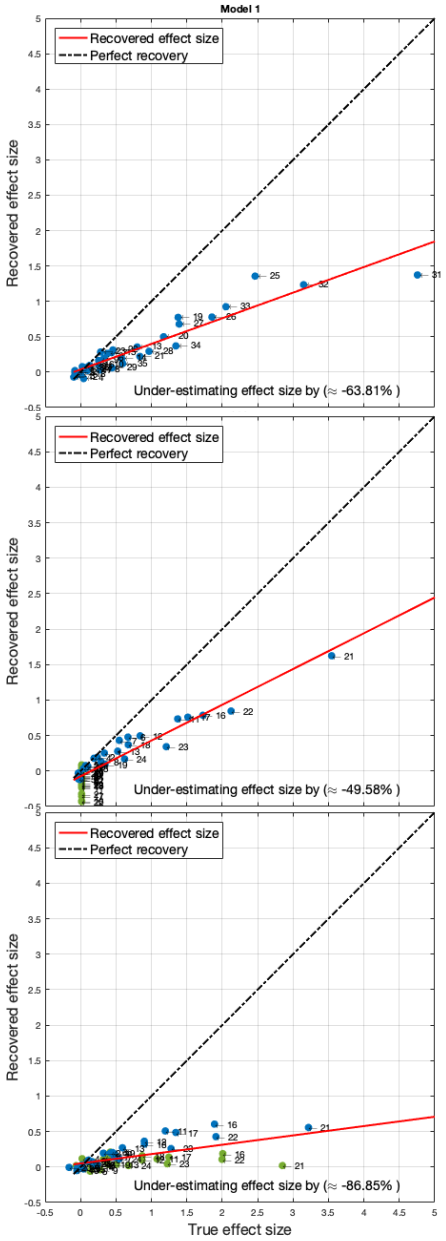
Dataset 2



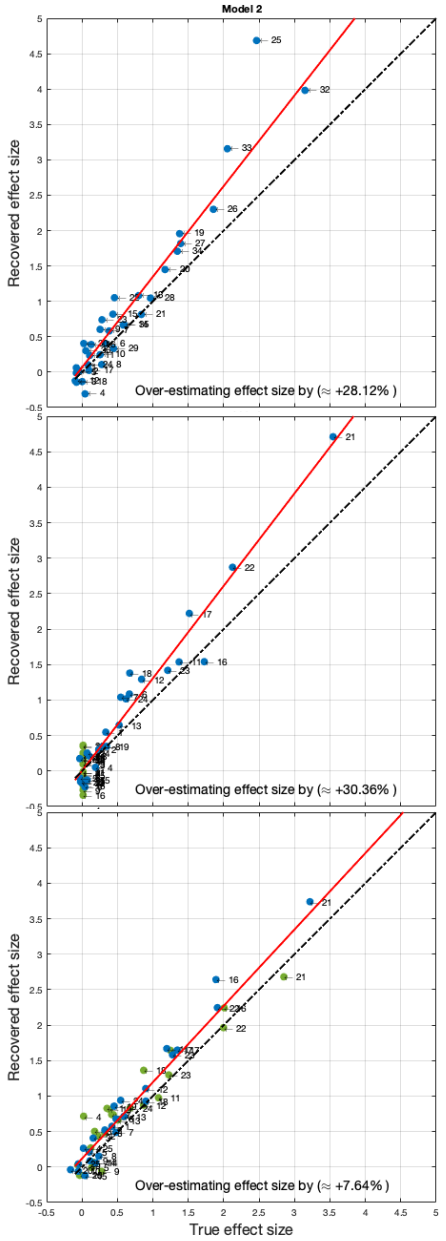
Dataset 3

Models

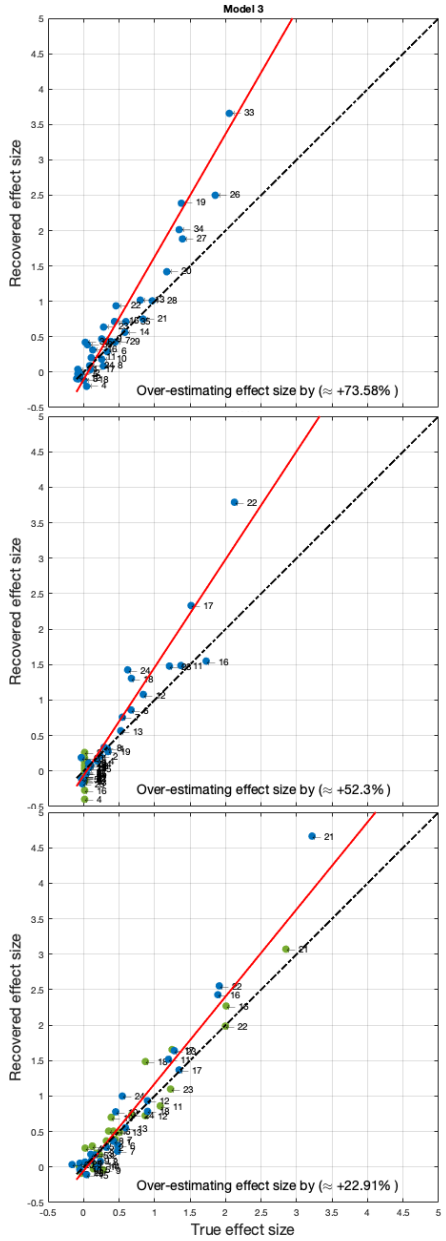
Model 1



Model 2

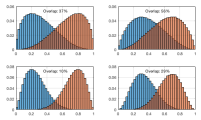


Model 3

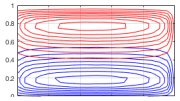


Data quality		# Participants	
		N=15	N=50
# Trials	T=40	Poor	Med.
	T=200	Med.	High

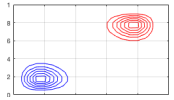
Datasets



Dataset 1



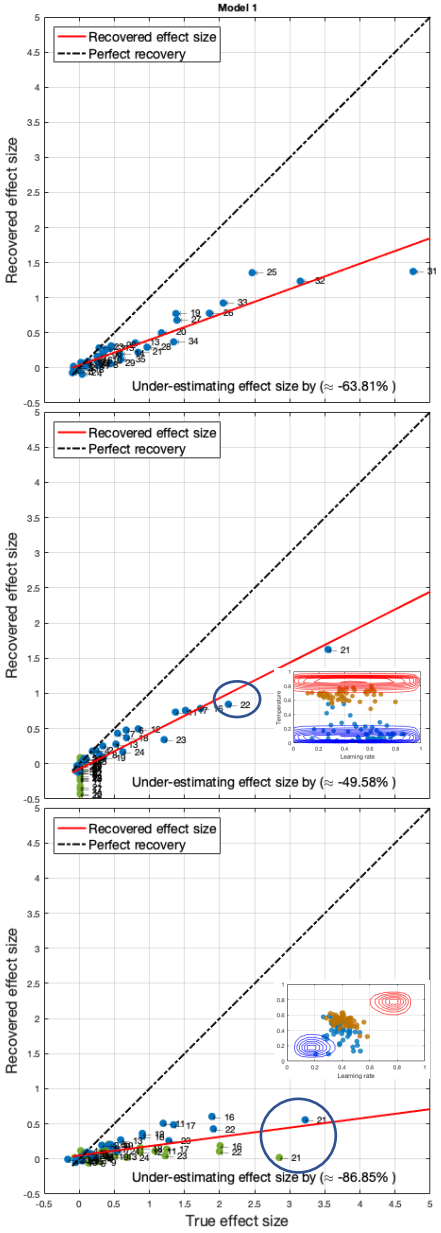
Dataset 2



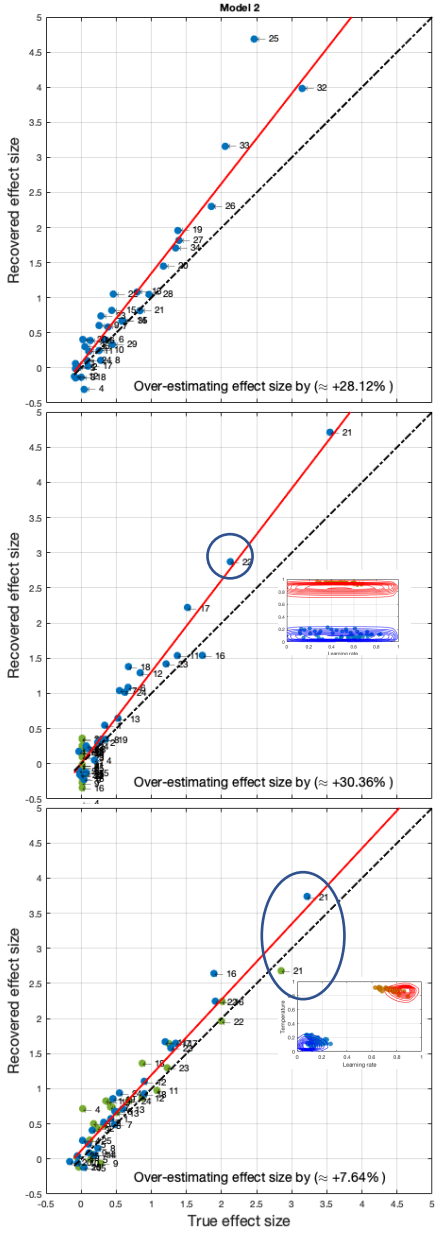
Dataset 3

Models

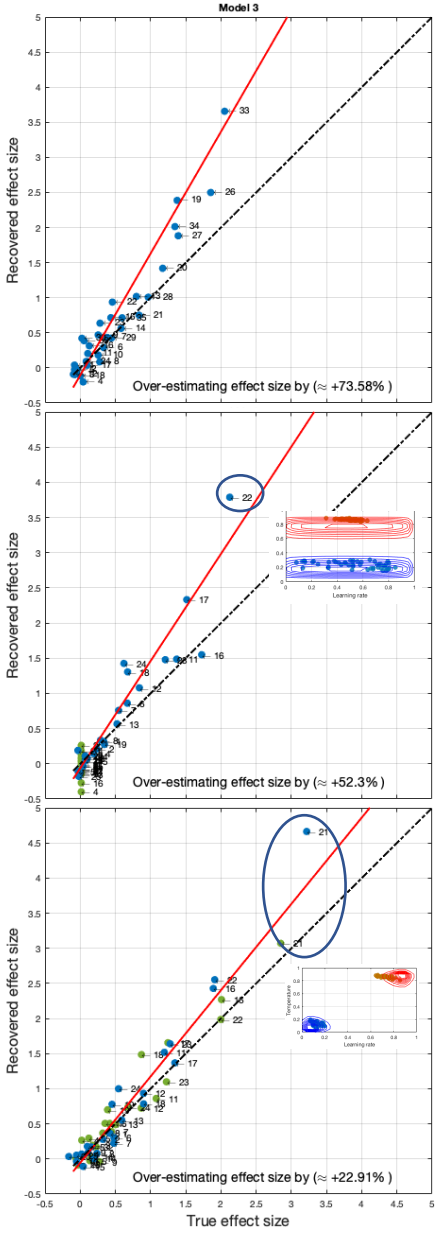
Model 1



Model 2

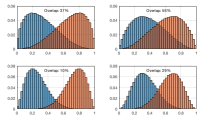


Model 3

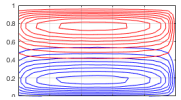


Data quality		# Participants	
		N=15	N=50
# Trials	T=40	Poor	Med.
	T=200	Med.	High

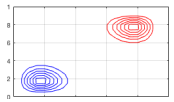
Datasets



Dataset 1



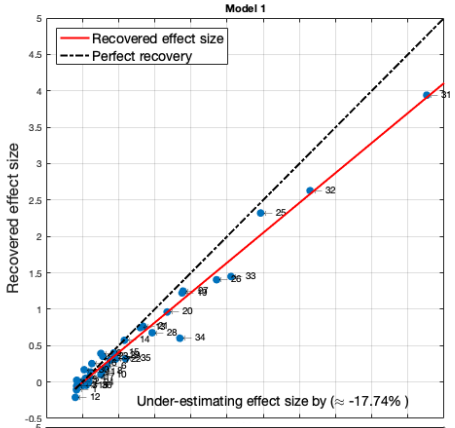
Dataset 2

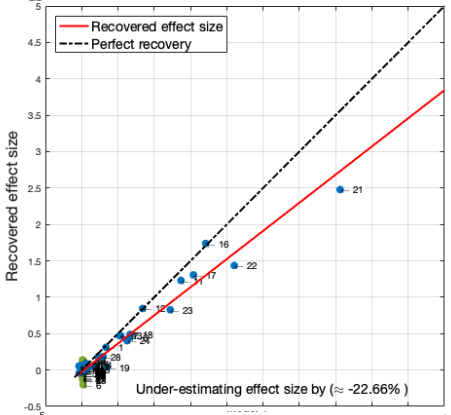


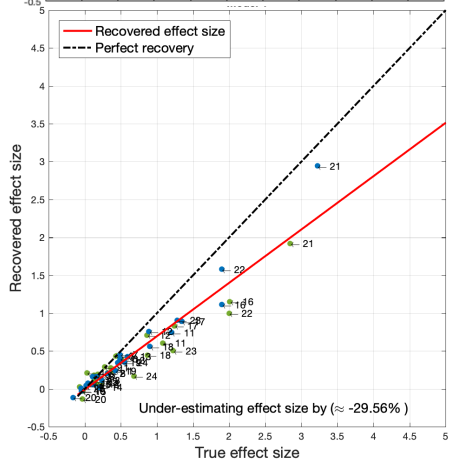
Dataset 3

Models

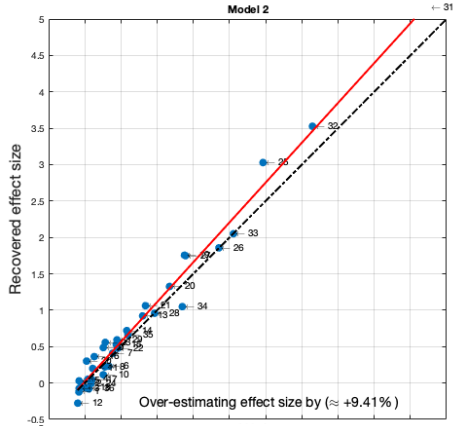
Model 1

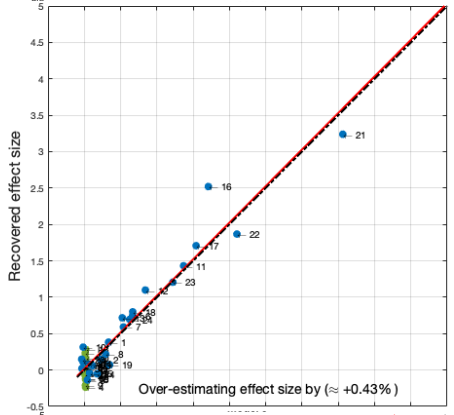


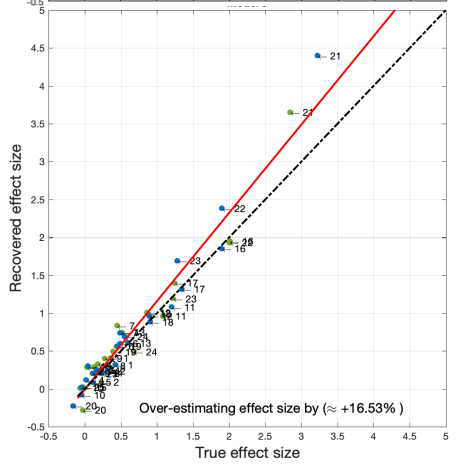




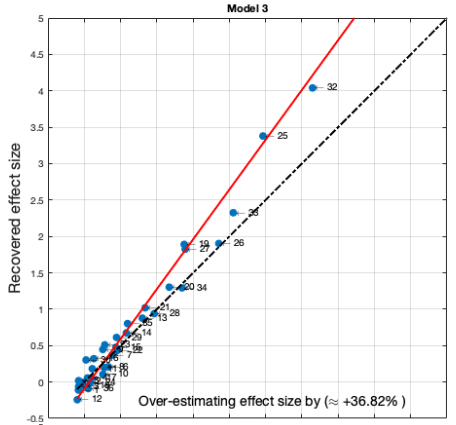
Model 2

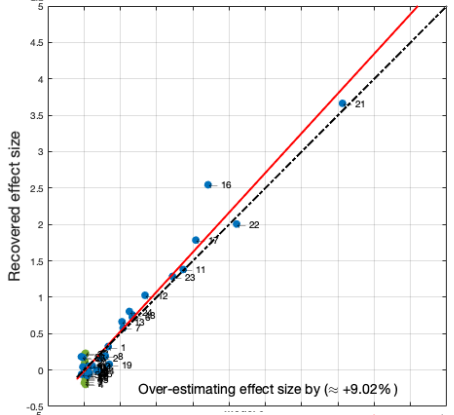


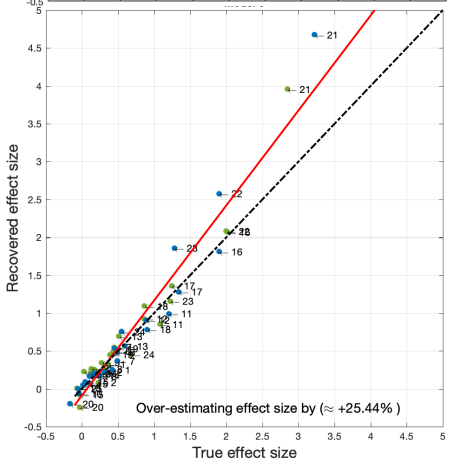




Model 3

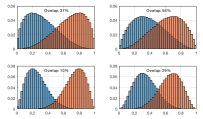




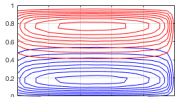


Data quality		# Participants	
		N=15	N=50
# Trials	T=40	Poor	Med.
	T=200	Med.	High

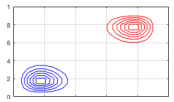
Datasets



Dataset 1



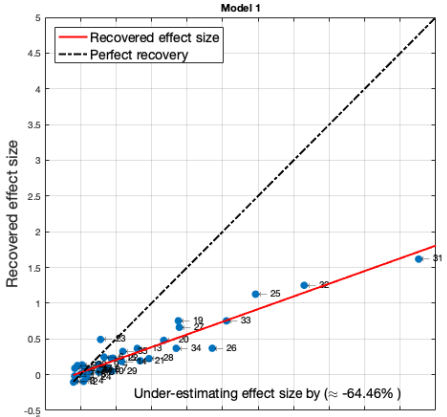
Dataset 2

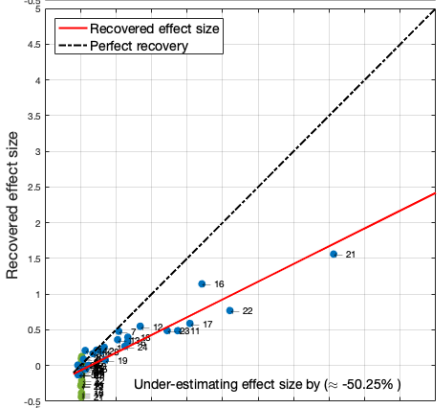


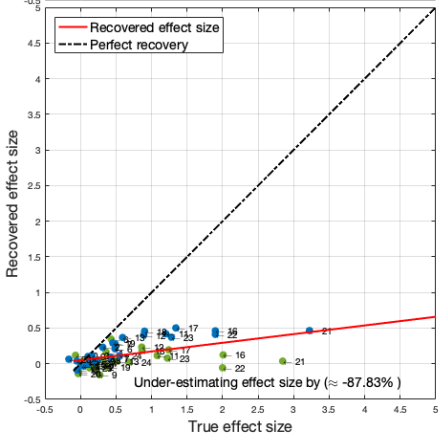
Dataset 3

Models

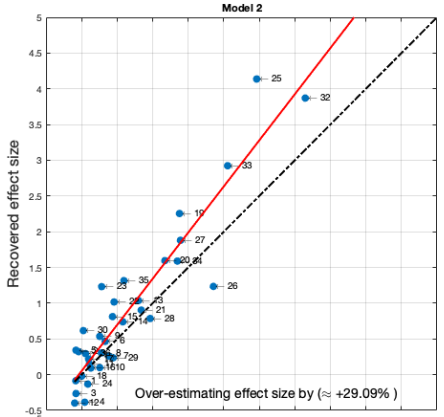
Model 1

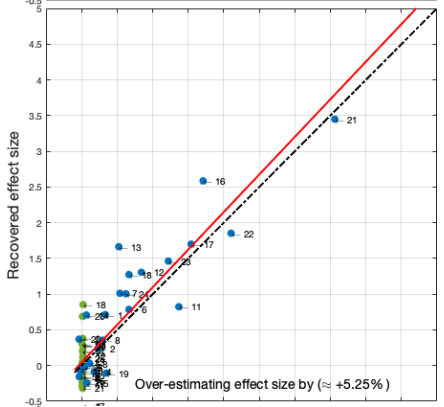


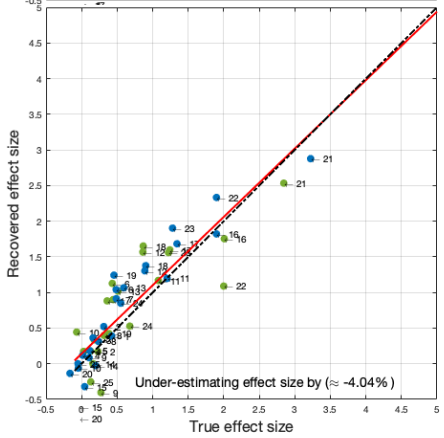




Model 2







Model 3