NTT

APRIS 2023

Asia Pacific Conference on Robot IoT System Development and Platform (APRIS)
2nd November 2023

Personalized affective computing

How to combat model explainability, uncertainties in subjective judgment and response biases in estimating individual’s affect

Shiro Kumano

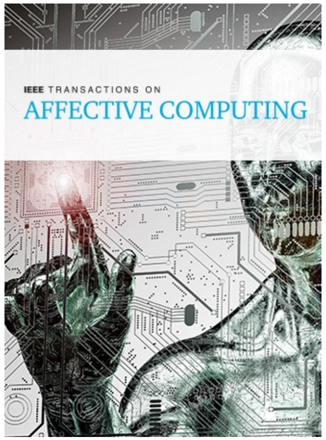
NTT Communication Science Laboratories

What is Affective Computing?

[Picard 97]

- To build machines understand human emotions,
- To make machines behave emotionally, and
- To develop machines having emotions.

An interdisciplinary field for incorporating emotion state recognition, understanding, simulation, and stimulation into computer system design.



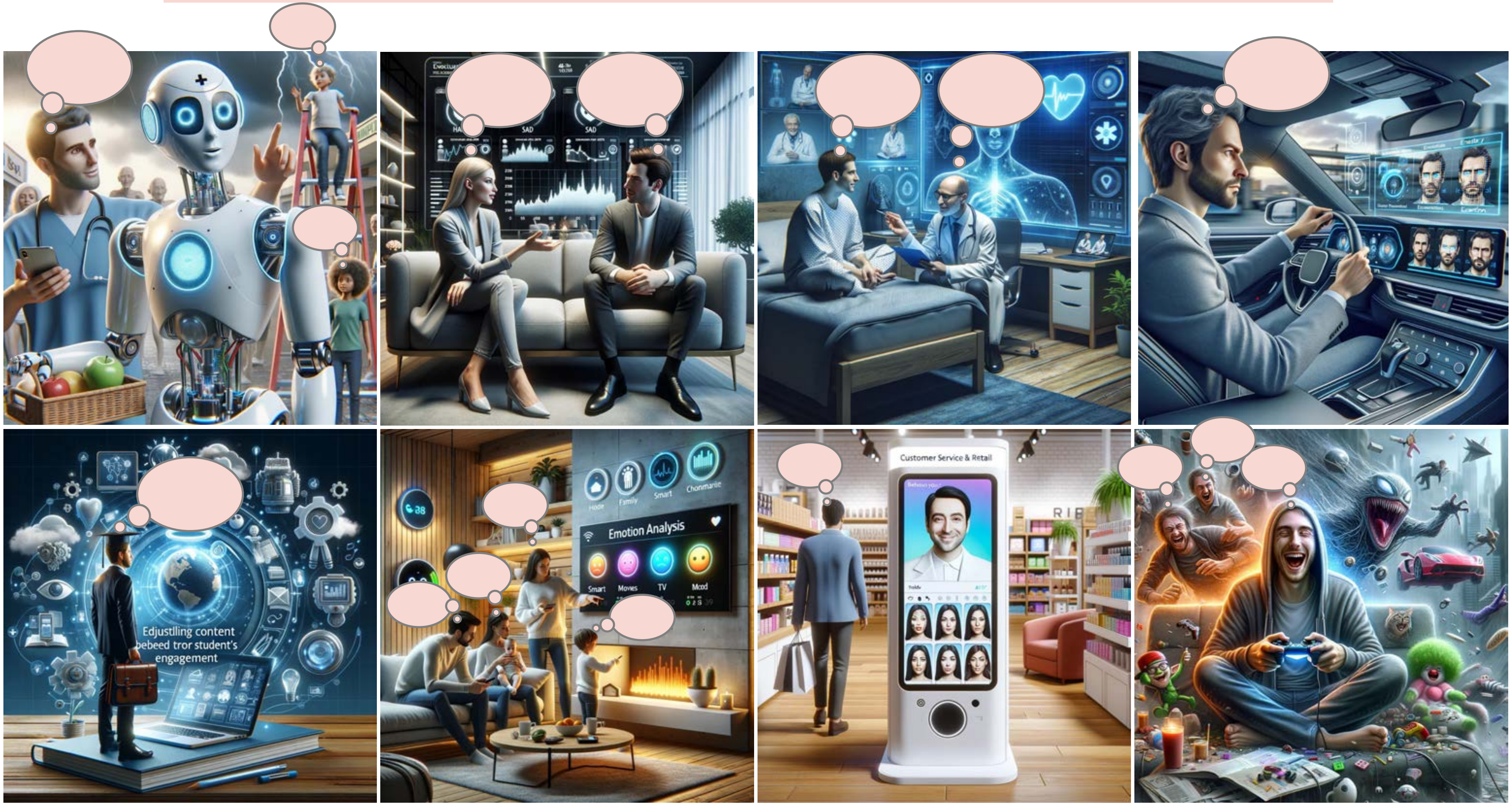
IEEE Trans. Affective Computing

IF=13.99

Flagship conference: ACII, Affective Computing and Intelligent Interaction

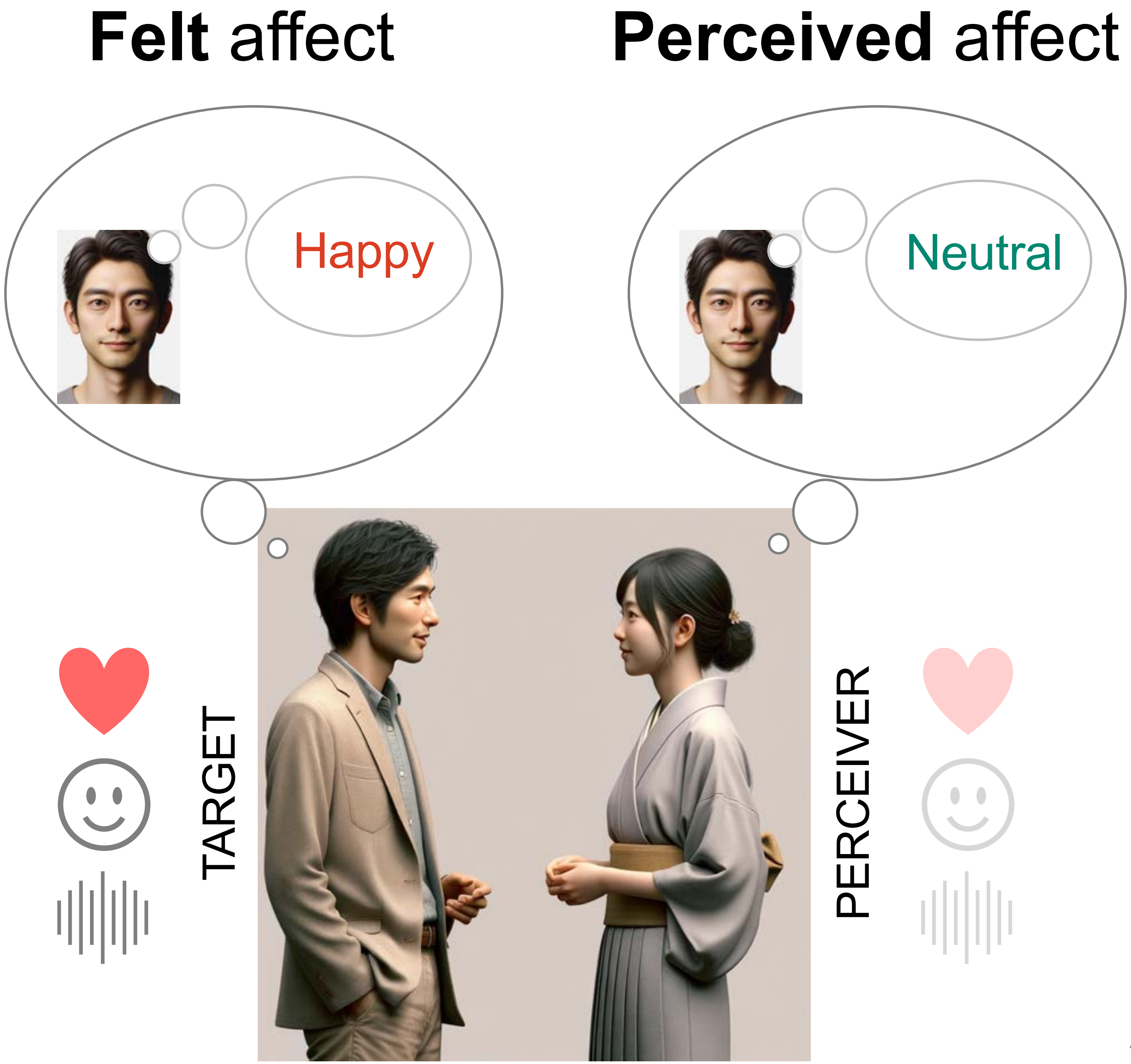
Potential applications of affect sensing

Subjective experiences matter in many domains



All were drawn by DALL-E 3

Felt affect versus perceived affect



All were drawn by DALL-E 3

Basic steps of supervised learning for subjective emotion prediction

Preprocessing / Feature extraction → Training → Evaluation

Output

Happy

DALL-E 3

No preprocessing / Standardization

Input

Feature extraction

Likelihood maximization / Loss minimization

Accuracy, F1, AUC
MAE, RMSE, …

Issue 1: Ethical, Legal, and Social Issues (ELSI)

Preprocessing / Feature extraction → Training → Evaluation

Output

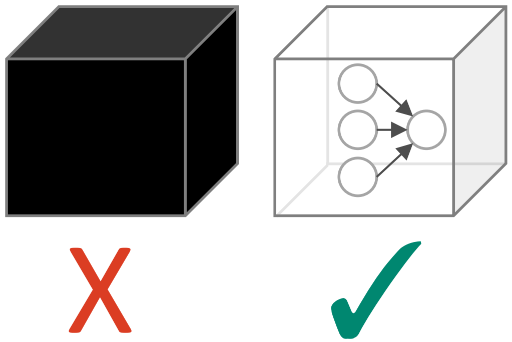
No preprocessing / Standardization

Input

Feature extraction

Likelihood maximization / Loss minimization

Accuracy, F1, AUC
MAE, RMSE, …



Transparent & explainable models are needed

Issue 2: Uncertainty in subjective judgment

Preprocessing / Feature extraction → Training → Evaluation

Output

Happy

No preprocessing / Standardization

Likelihood maximization /

Accuracy, F1, AUC
MAE, RMSE, …

Input

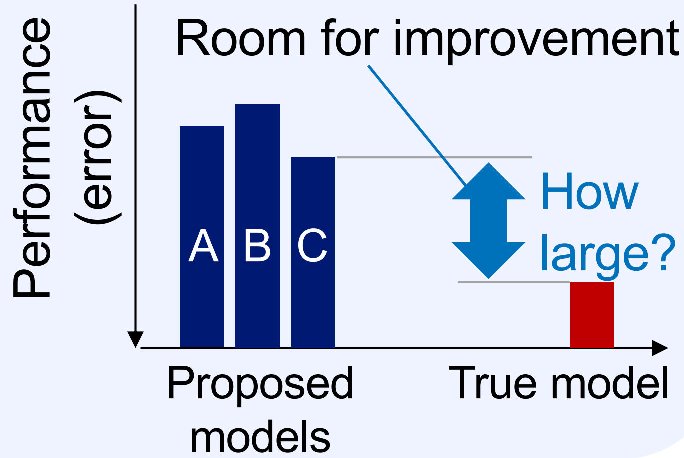
Fea

Same person may answer differently next time → **Unknown upper-bound performance ☹**

Happy ⟷ Neutral

1st time

2nd time



Aleatoric & epistemic uncertainties

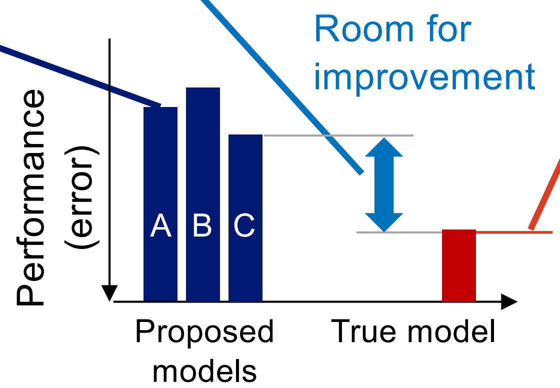
Uncertainty theory [Smith and Gal, 2018]

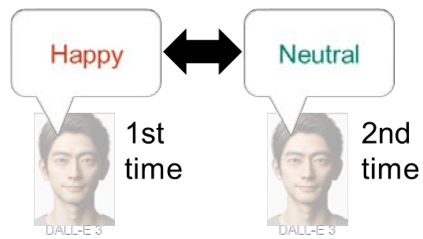
Predictive uncertainty = Epistemic uncertainty + Aleatoric uncertainty

Observed model performance

Imperfection of model

Uncertainty in data itself





If aleatoric uncertainty $\neq 0$, no model can achieve perfect performance.

Problem of previous uncertainty prediction methods

Predict aleatoric uncertainty not using data itself but using the model, using MC-dropout or model ensembles.

Issue 3: Response biases

Preprocessing / Feature extraction → Training → Evaluation

Output

DALL-E 3

No preprocessing / Standardization

Input

Feature extraction

Likelihood maximization / Loss minimization

Accuracy, F1, AUC
MAE, RMSE, …

Response biases

(Paulhus 1991, Baumgartner & Steenkamp 2001, Wetzel et al. 2016)

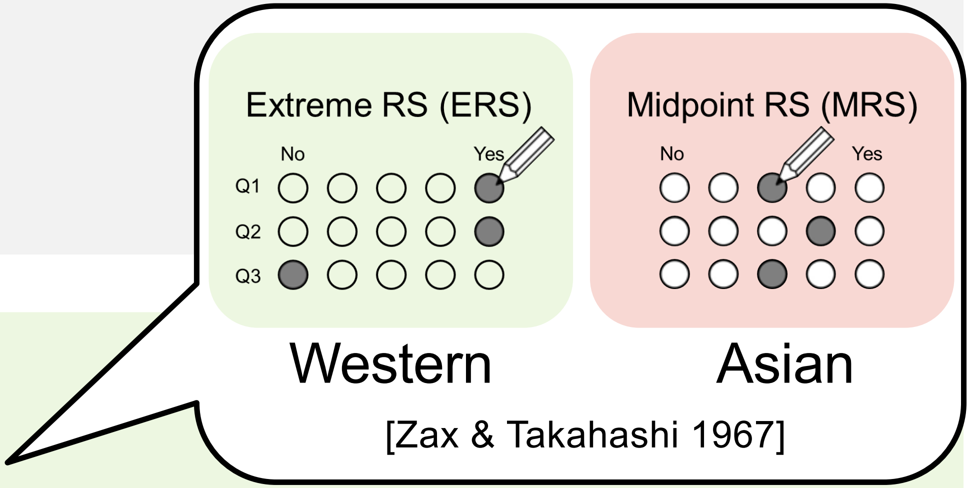
Degrade validity of correlation- and variance-based analyses

(Dolnicar & Grun 2009)

Task-dependent biases

- Socially desirable responding
- Halo effect
- Leniency/severity

…



Task-independent biases

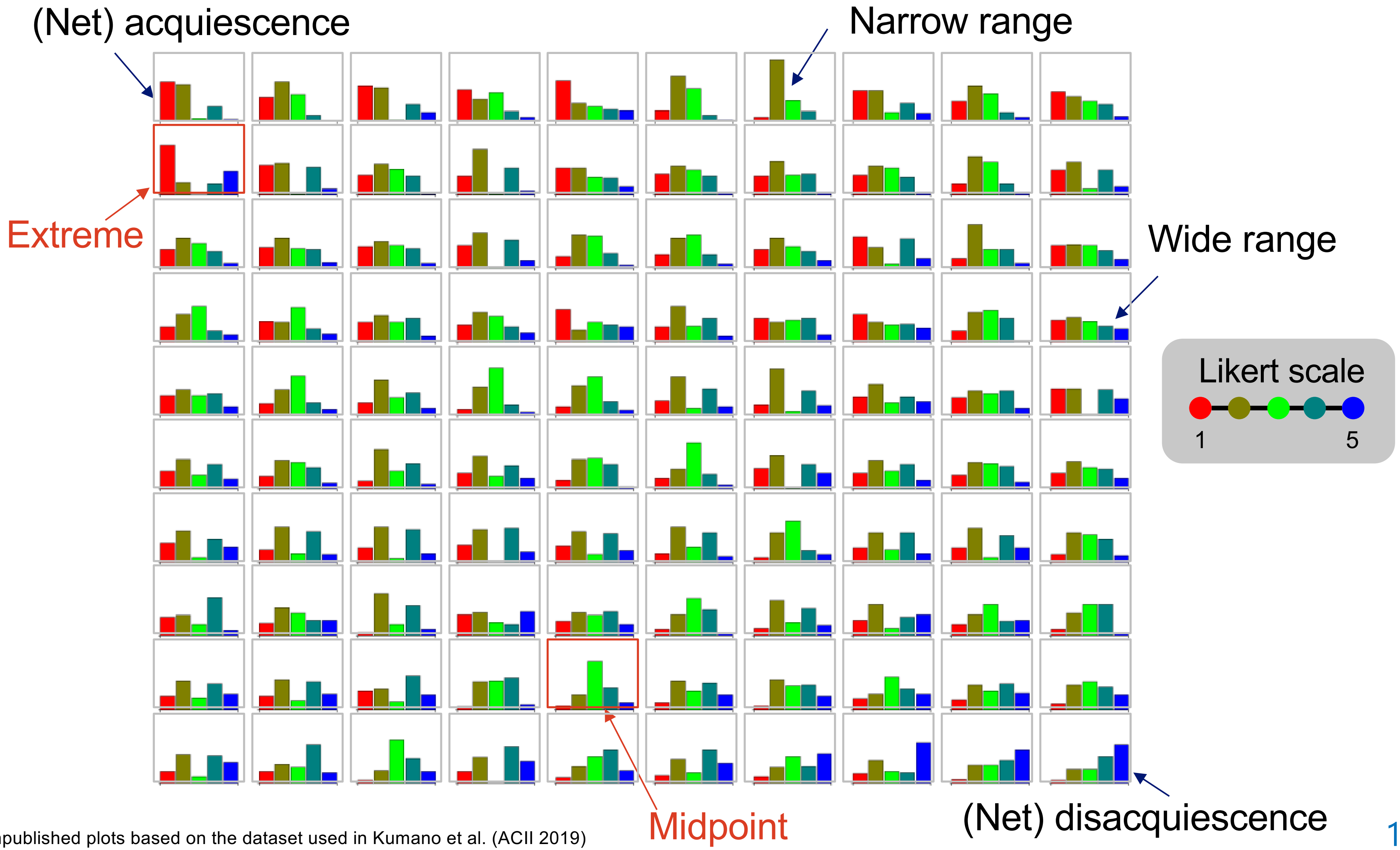
- Response style (RS)

  Tendency to choose specific categories regardless of content
- Random responding

…

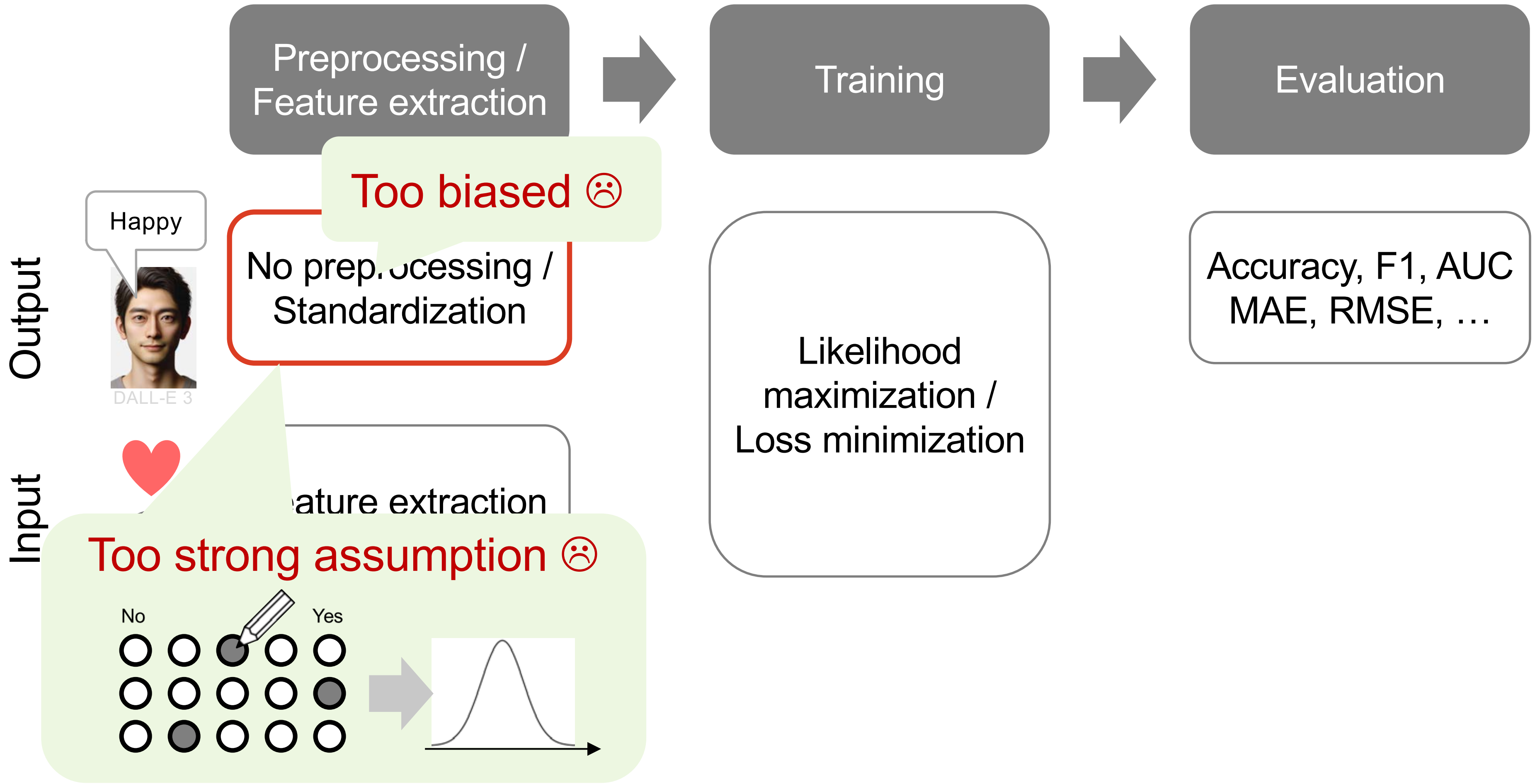
Rating frequency histogram per respondent

Simple standardization for each person does not make sense.

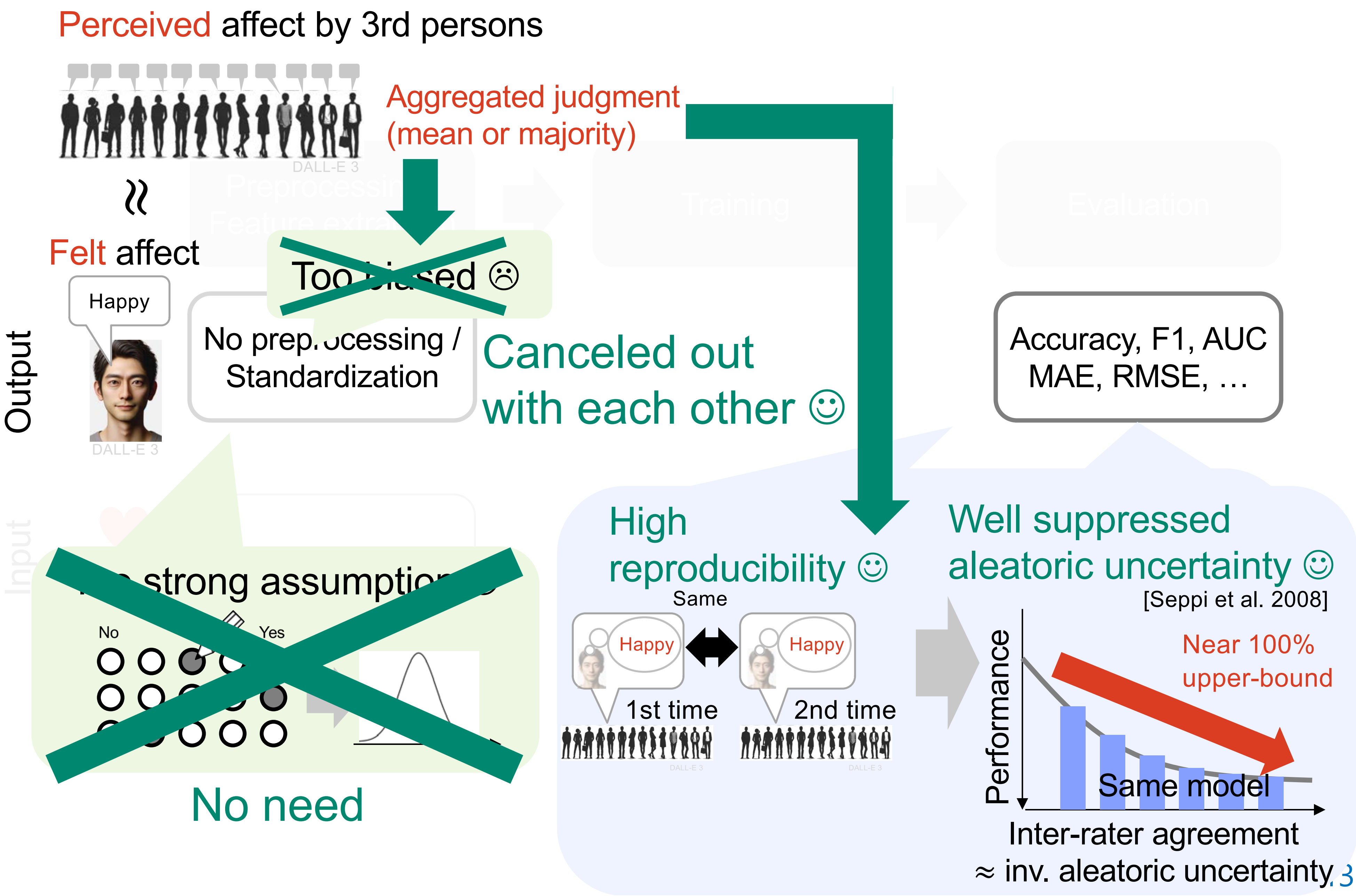


Unpublished plots based on the dataset used in Kumano et al. (ACII 2019)

Issue 3: Response biases



Wisdom of crowd for perceived affect



Individual's affective judgment

Perceived affect by 3rd persons

~~Aggregated judgment (mean or majority)~~

Output

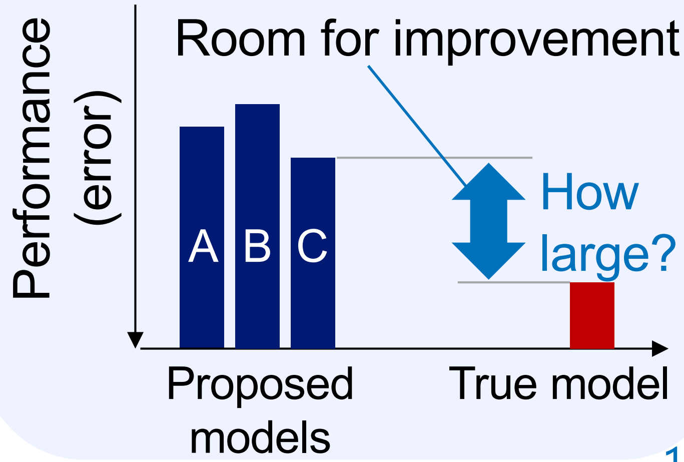
No preprocessing / Standardization

Too biased ☹

Likelihood maximization / Loss minimization

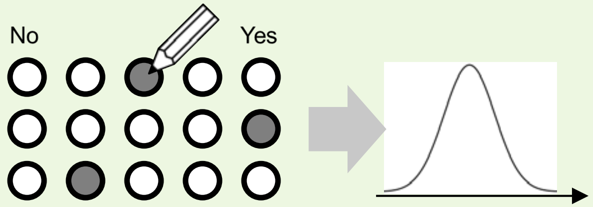
Accuracy, F1, AUC
MAE, RMSE, …

Unknown upper-bound performance ☹



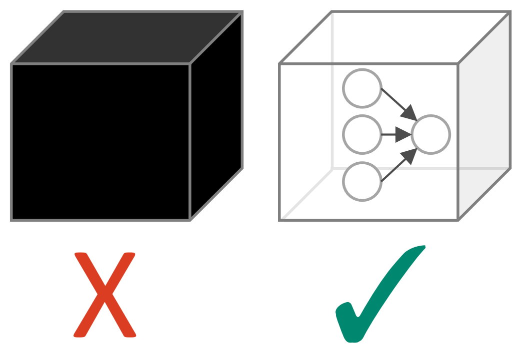
Input

Too strong assumption ☹

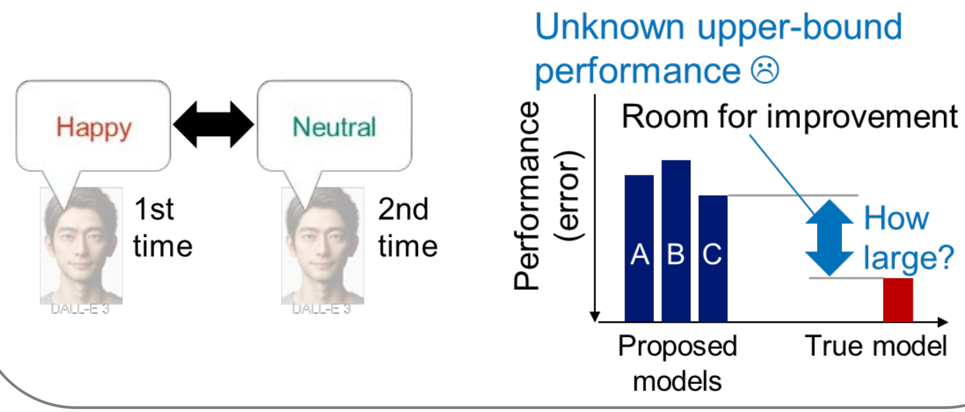


Interim summary of 3 issues in personalized affective computing

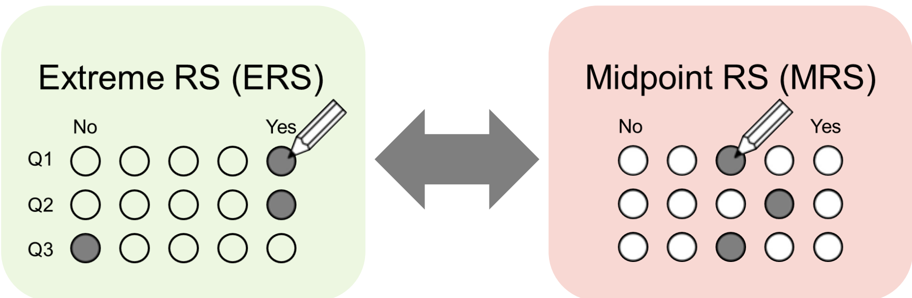
1. Model explainability



2. Uncertainties in subjective data & unknown upper-bound

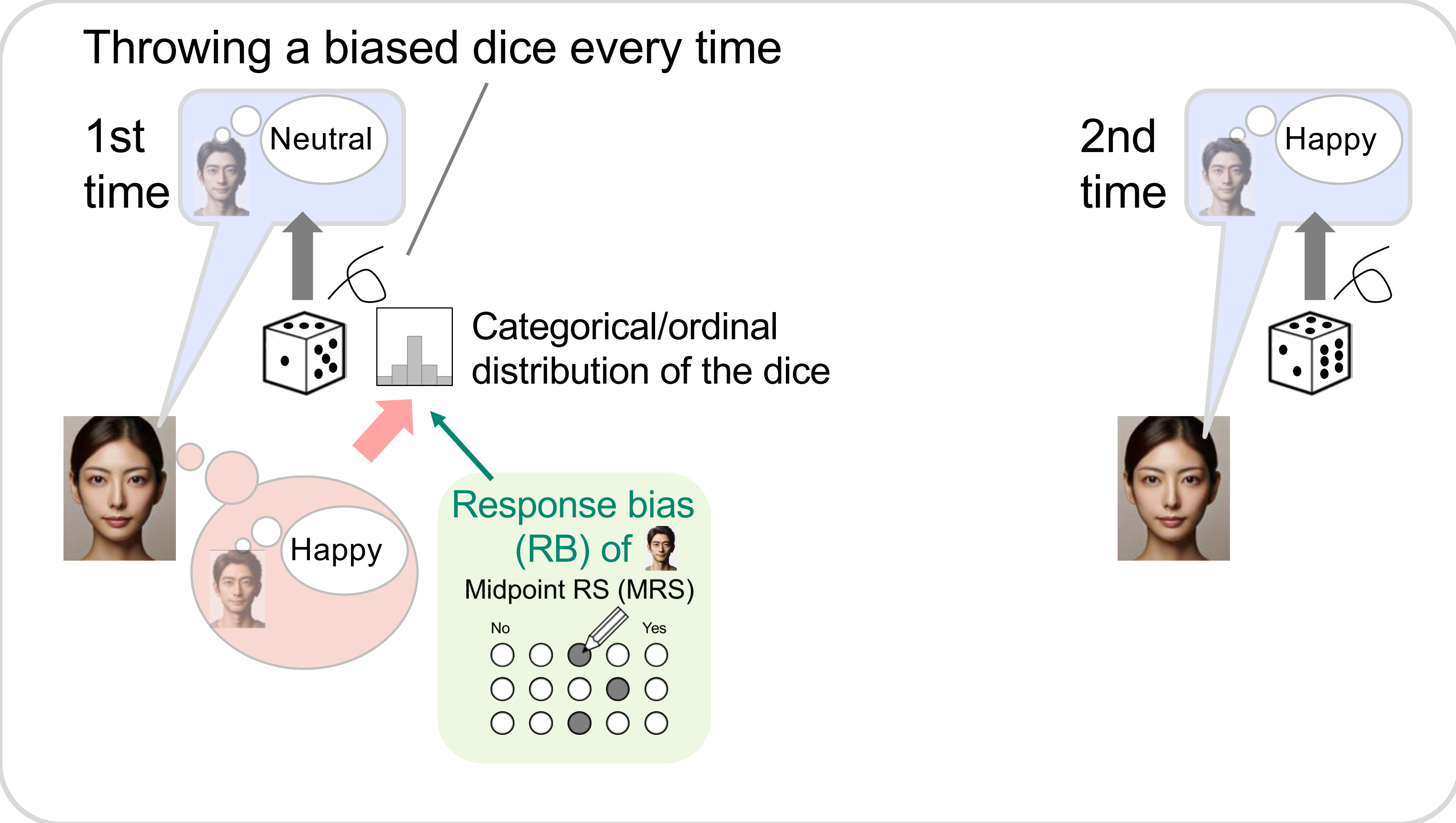


3. Response biases in subjective judgment



Our basic assumption to approach to the issues

Judgments are made following a stochastic process (categorical/ordinal distribution) including response styles



Deep explanatory item-response model for prediction of individual's affective rating

[presented at ACII 2021]

Yang Zhou, Tsukasa Ishigaki and Shiro Kumano

TOHOKU UNIVERSITY

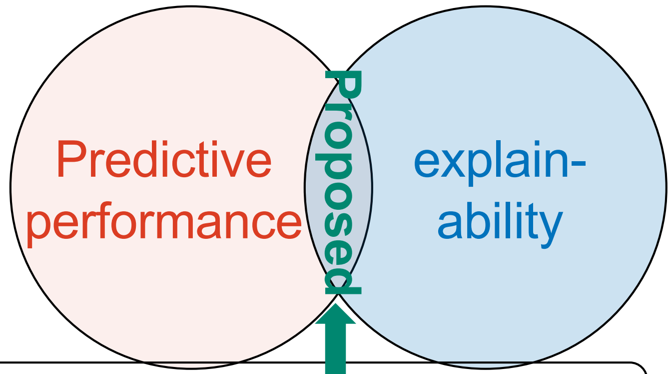
NTT

Balance predictive performance & explainability

Our approach

Balancing performance and explainability through the integration of deep learning (DL) and item-response theory (IRT).

Trade-off [Arrieta et al. 2020]

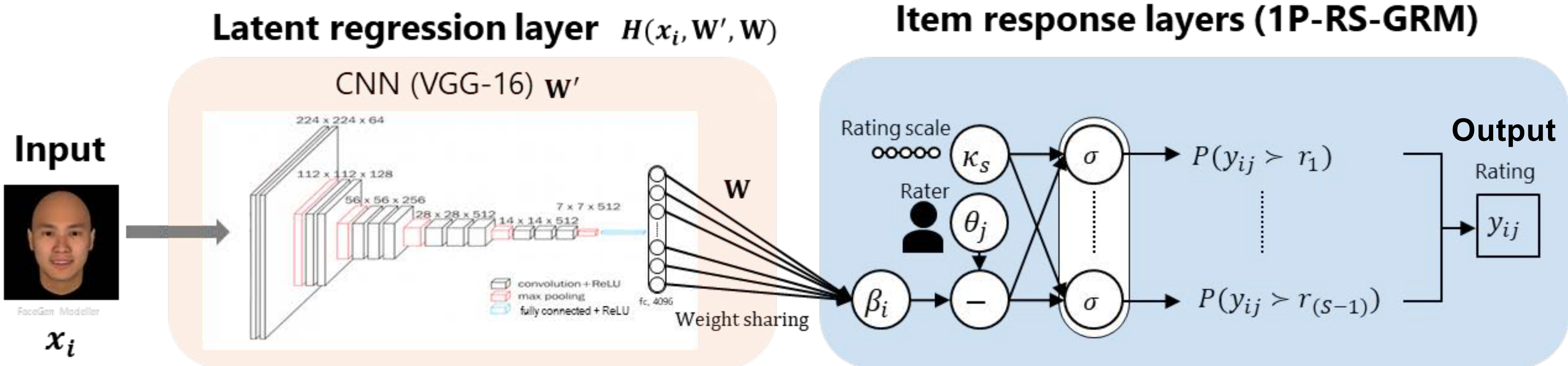


	Predictive performance	explain-ability
Deep learning Black box	✓	X
Item-response theory White box	X	✓

Deep learning + Item-response theory → Proposed

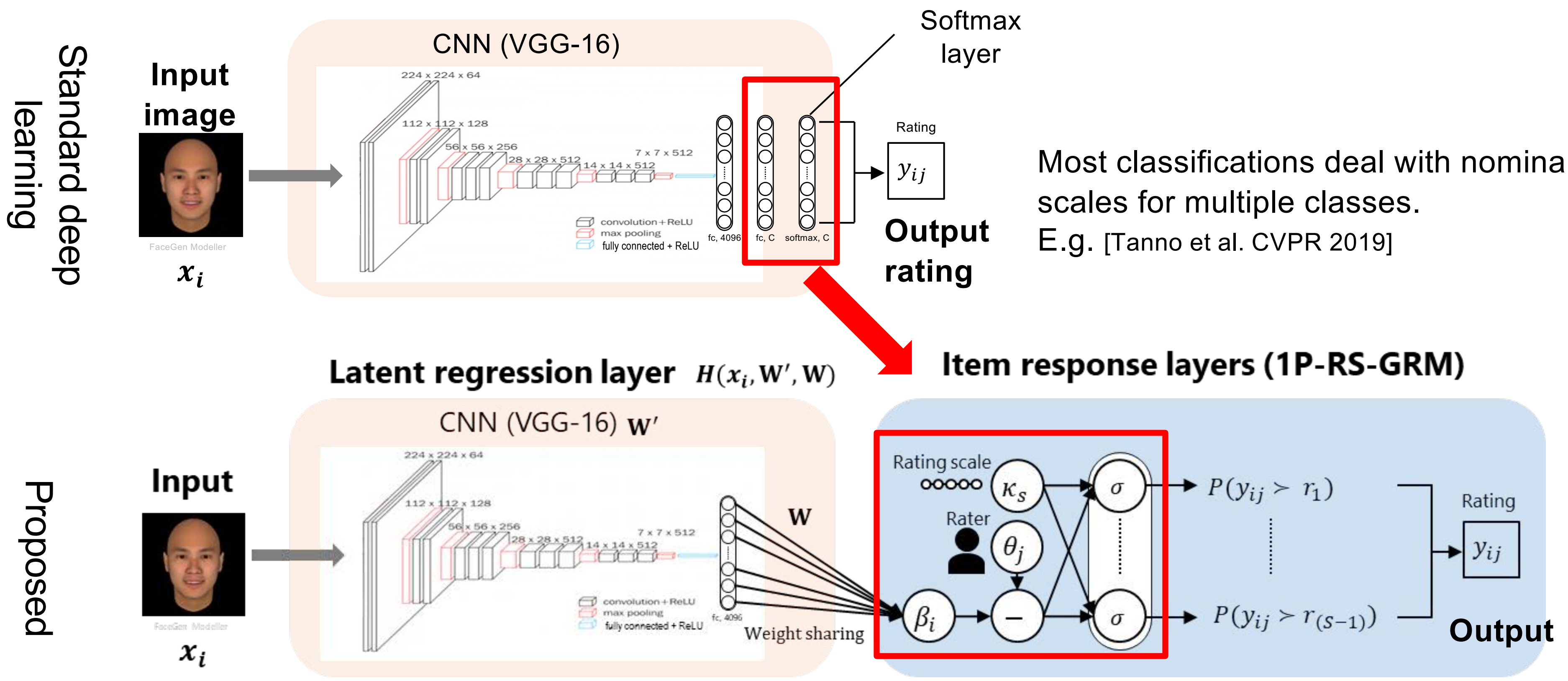
Deep learning

Item response theory



Proposed method from DL perspective

By constraining the upper layers of DNN with IRT, we obtain the explainability of how the output is determined at the upper layers

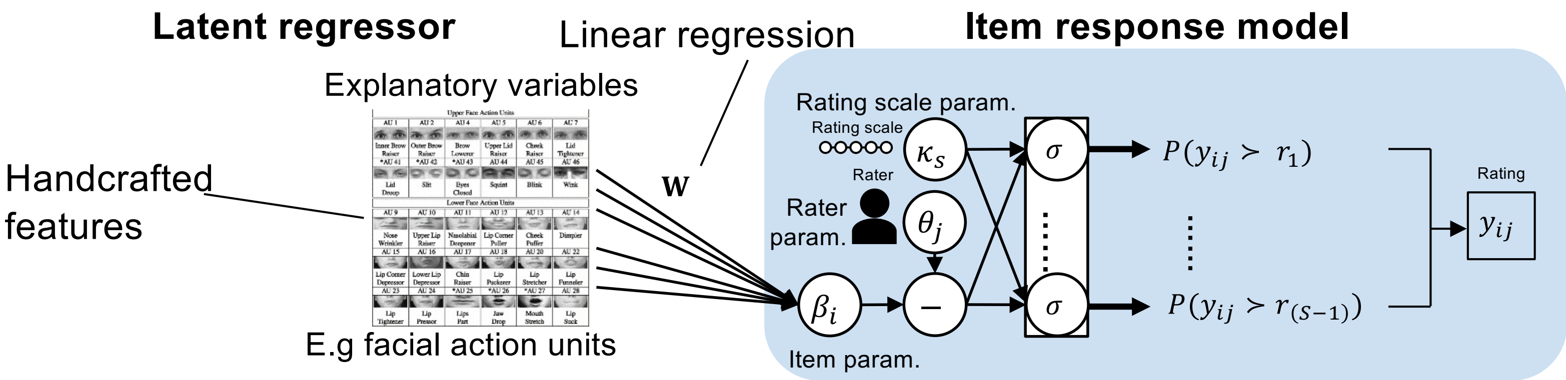


Most classifications deal with nominal scales for multiple classes.
E.g. [Tanno et al. CVPR 2019]

Proposed method from IRT perspective

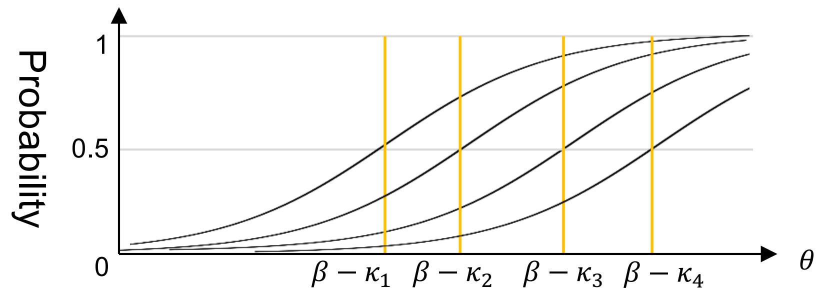
By performing end-to-end deep-regression of item parameters in an explanatory item response model, the estimation performance is enhanced.

Standard explanatory IRT



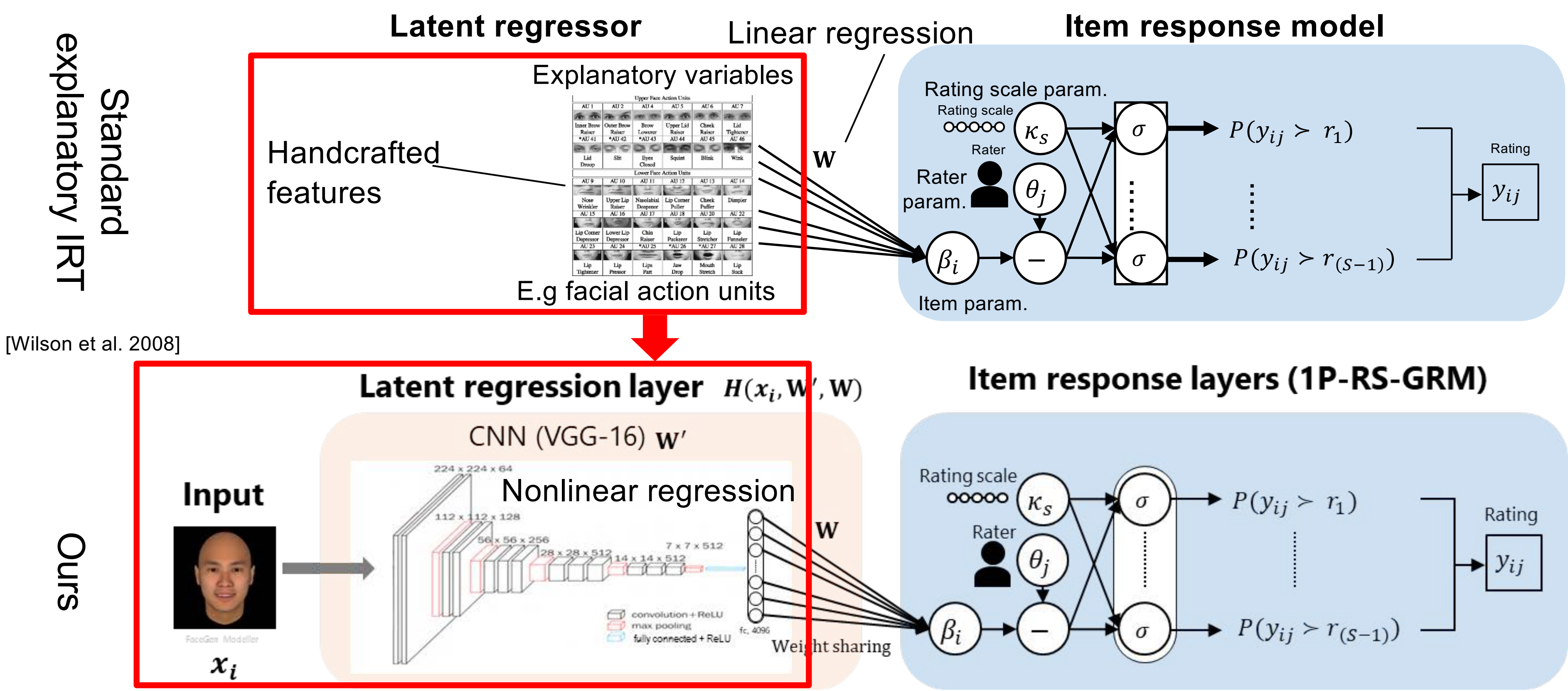
$$P(y_{ij} \succ r_s) = \sigma(\theta_j - \beta_i - \kappa_s), \qquad \kappa_1 \leq \cdots \leq \kappa_{S-1}$$

Item characteristic curve



Proposed method from IRT perspective

By performing end-to-end deep-regression of item parameters in an explanatory item response model, the estimation performance is enhanced.



Instead of a rough linear regression from high-level features determined by humans as explanatory variables for the item parameter β, it is estimated from a refined nonlinear regression from the item itself (image).

Experimental results

The performance superiority and inferiority of the three models, as theoretically expected, were confirmed.

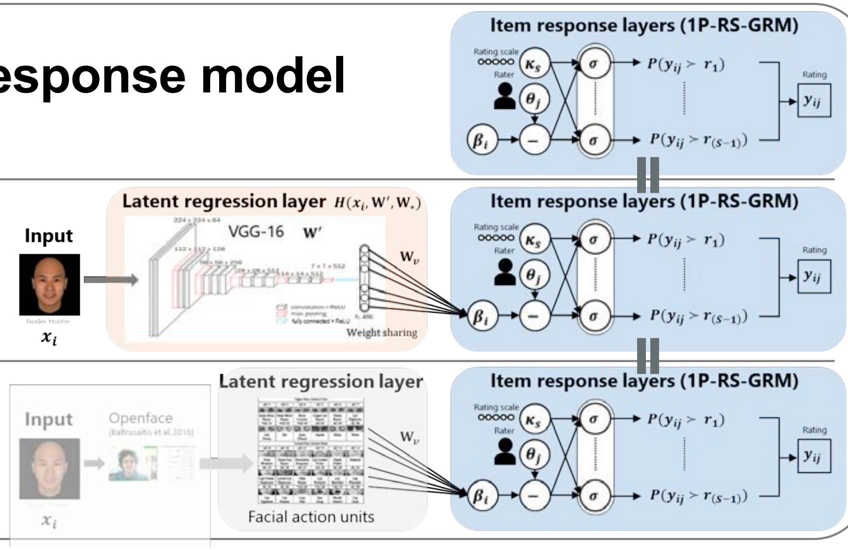
Performance

Explainability

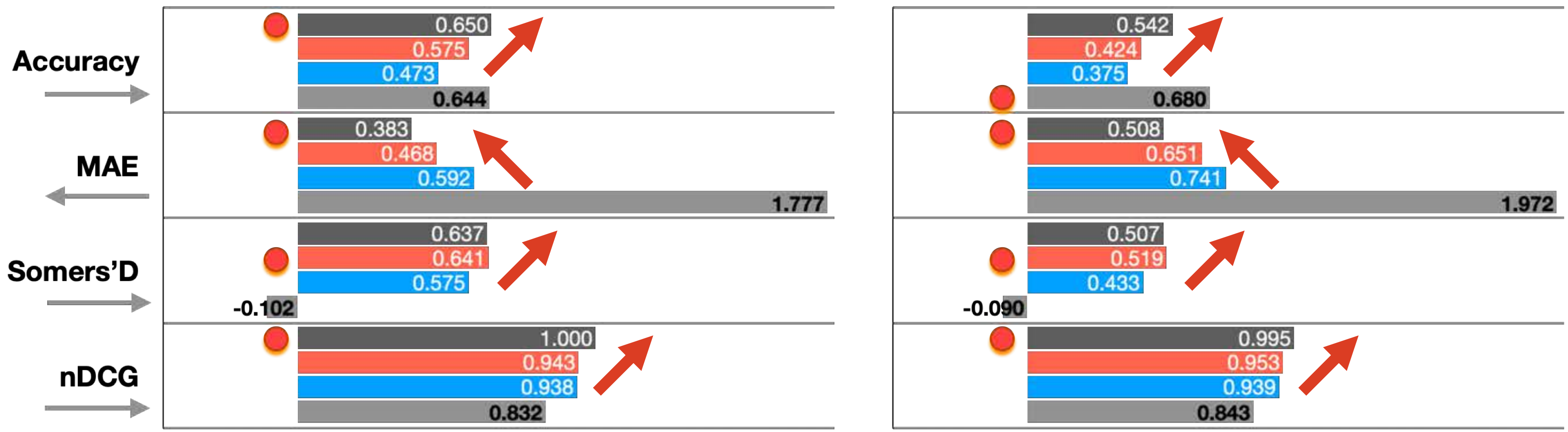
Standard item response model
1P-RS-GRM

Our model
CORAL-RS-GRM

Standard explanatory IRT
AU-RS-GRM



Rank-SVM



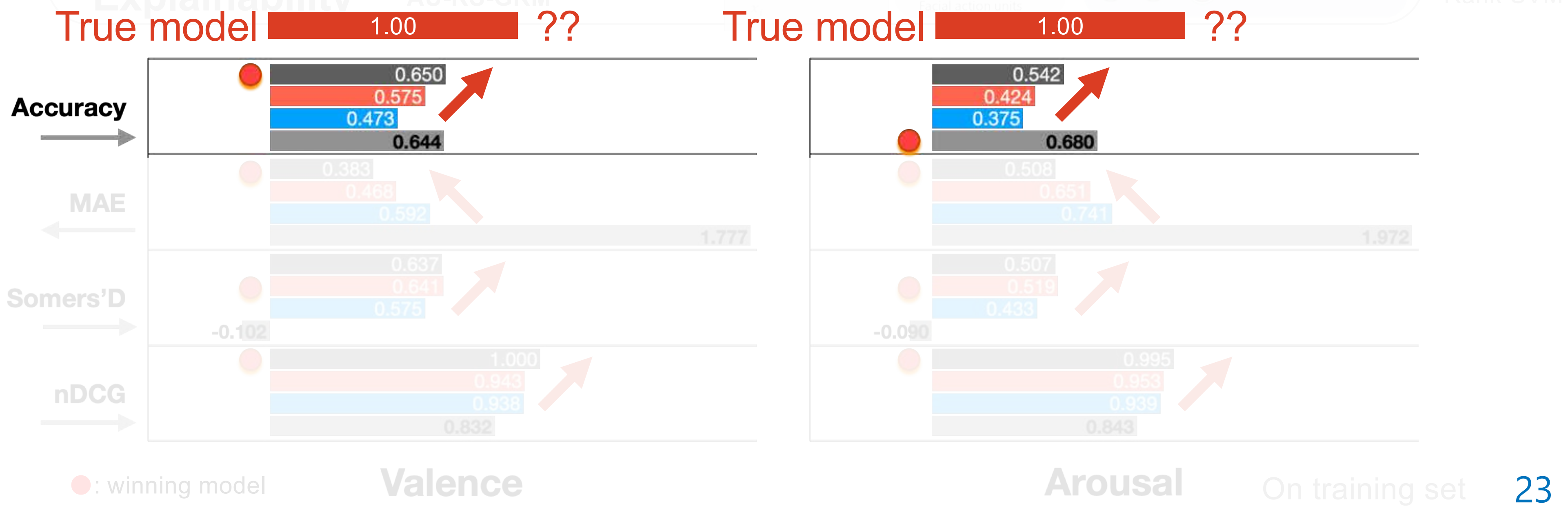
●: winning model

Valence **Arousal**

On training set

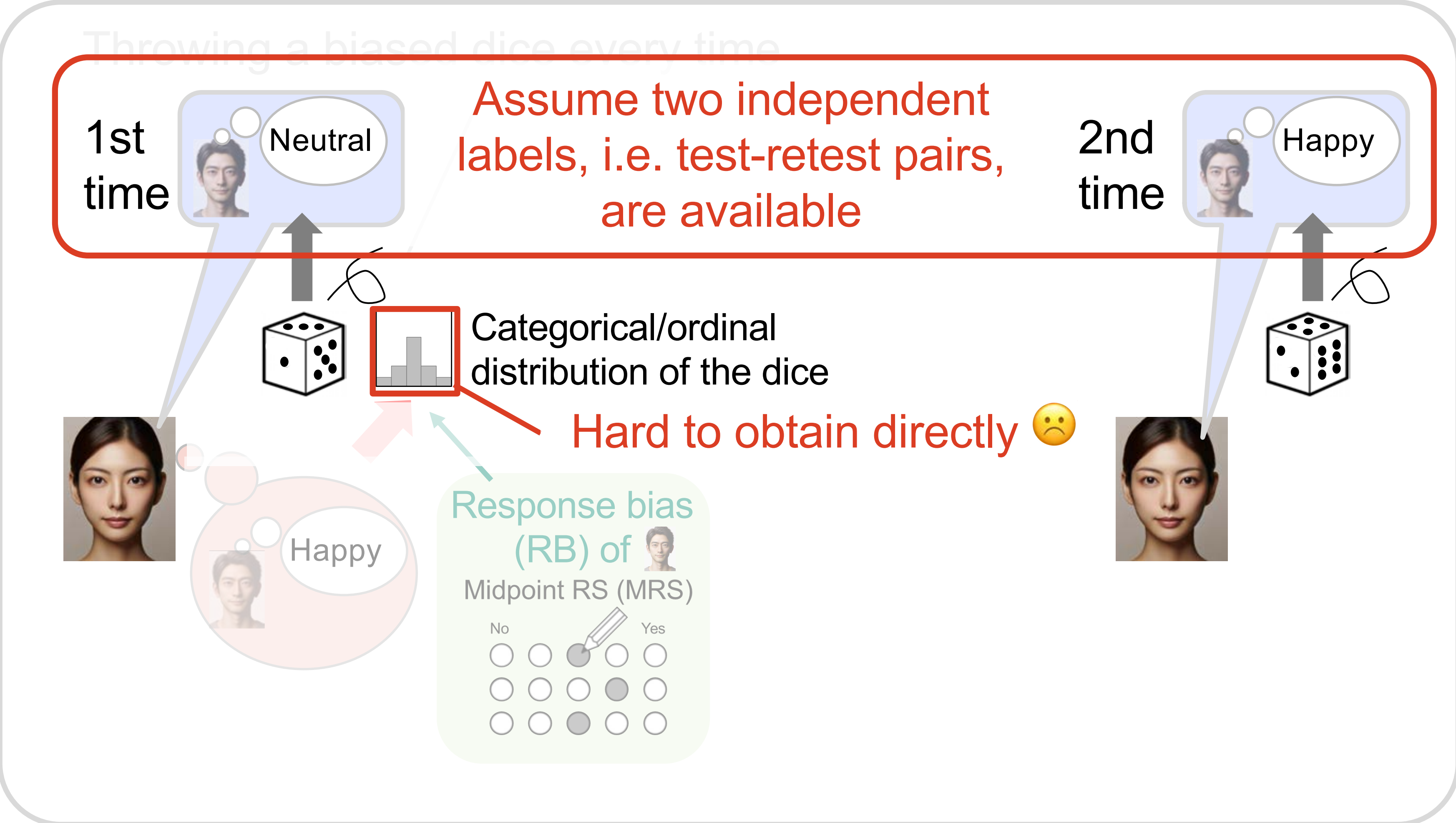
Issue 2: unknown upper-bound performance

Since all are relative measures,
Upper-bound (how close the current model is to true model) is not clear



Our approach to the issues

Assume choices are made following a stochastic process (categorical/ordinal distribution) including response styles



Collision Probability Matching Loss for Disentangling Epistemic Uncertainty from Aleatoric Uncertainty

[Presented at AISTATS2023]

[1]Hiromi Narimatsu, [2]Mayuko Ozawa, [1]Shiro Kumano

[1]NTT Communication Science Laboratories

[2]Ritsumeikan University

Summary of proposed method

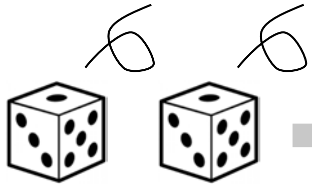
Dices based according to:

True distribution p (unobtainable directly)

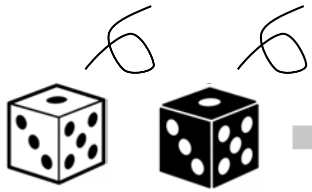
Predictive distribution q

Number = category or rating

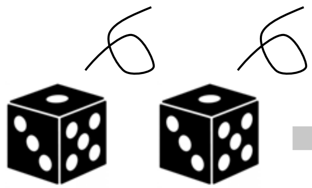
Probability distribution	Measure
True distribution p only	True collision probability (CP) of target's cognition $\sum_c p_c^2$ ≈ Test-retest reliability
True dist. p & predictive dist. q	Cross-CP $\sum_c p_c q_c$ ≈ Mean data likelihood $E[q_y]$
Predictive distribution q only	Predictive CP $\sum_c q_c^2$ Directly obtainable



Probability of getting the same number



Probability of getting the same number



Probability of getting the same number

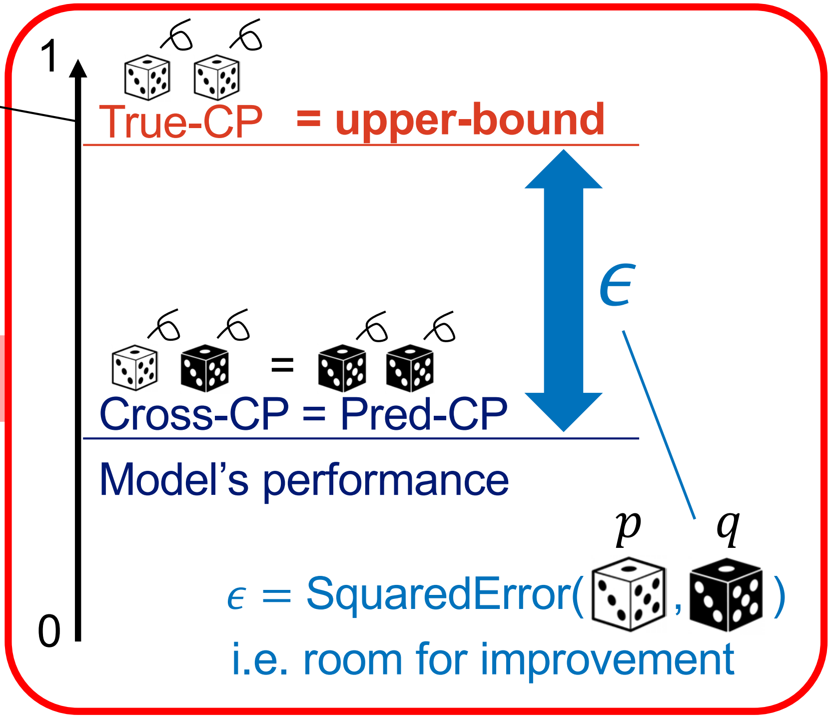
Collision probability = 2nd-order Renyi entropy

Summary of proposed method

True distribution p

Predictive distribution q

Probability distribution	Measure
True distribution p only	True collision probability (CP) of target's cognition $\sum_c p_c^2$ ≈ Test-retest reliability
	IV
True dist. p & predictive dist. q	Cross-CP $\sum_c p_c q_c$ ≈ Mean data likelihood $E[q_y]$
Proposed CP matching constraint	$\|$
Predictive distribution q only	Predictive CP $\sum_c q_c^2$ Directly obtainable



Relationship with confidence calibration

True dist. p Predictive dist. q

Prob. dist. | **Distribution-based measure** | **Max-class-based measure**

Proposed method is as an extended method of CC by replacing max-class-based to distribution-based

True collision probability (CP)

$\sum_c p_c^2$

Approx. directly measurable

True (human) confidence

[Jogan & Stocker 2014, Li & Ma 2020]

True distribution p only

Proposed method

Confidence calibration (CC)

True dist. p & predictive dist. q

Cross-CP

$\sum_c p_c q_c$

$\approx$ Mean data likelihood $E[q_y]$

=

Model accuracy (MA) $p_{\hat{y}max}$

Matching of max predictive class with observed class

=

Predictive distribution q only

Predictive CP

$\sum_c q_c^2$

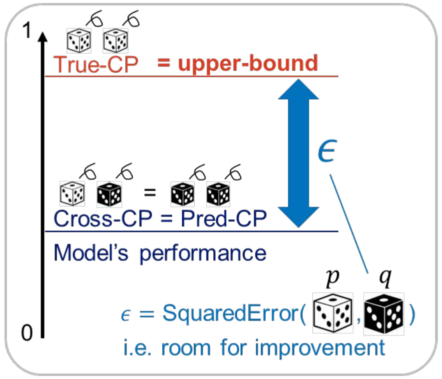
Directly obtainable

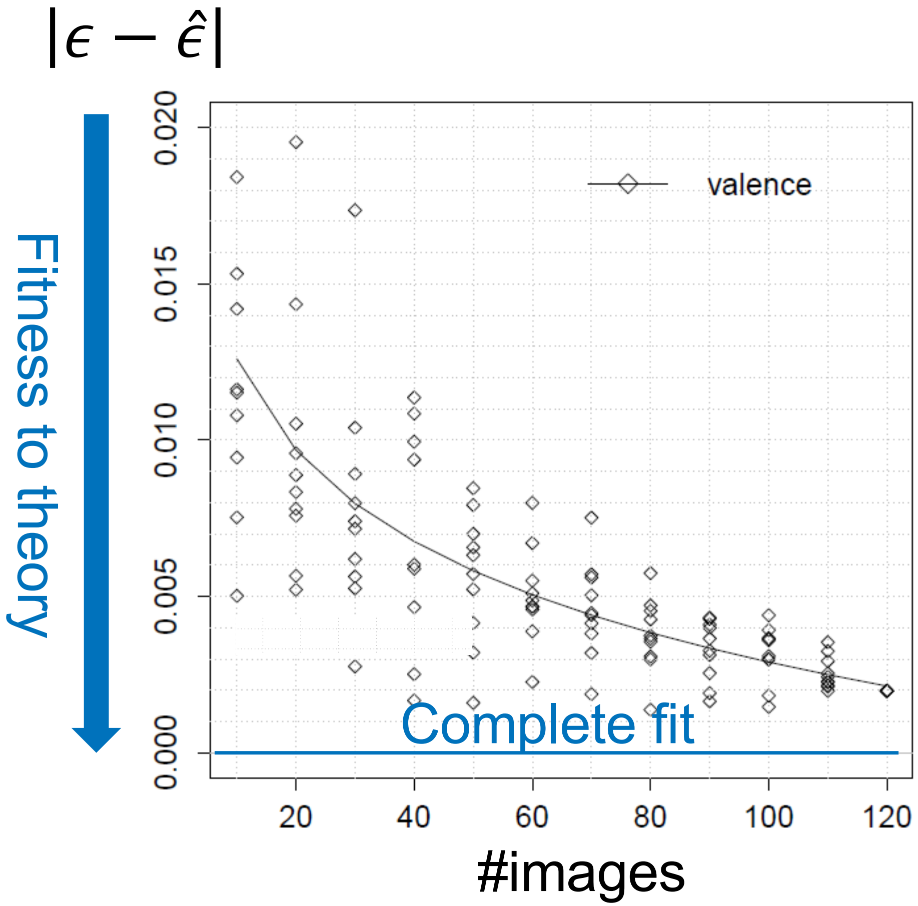
Model confidence (MC) $q_{\hat{y}max}$

Max predictive probability

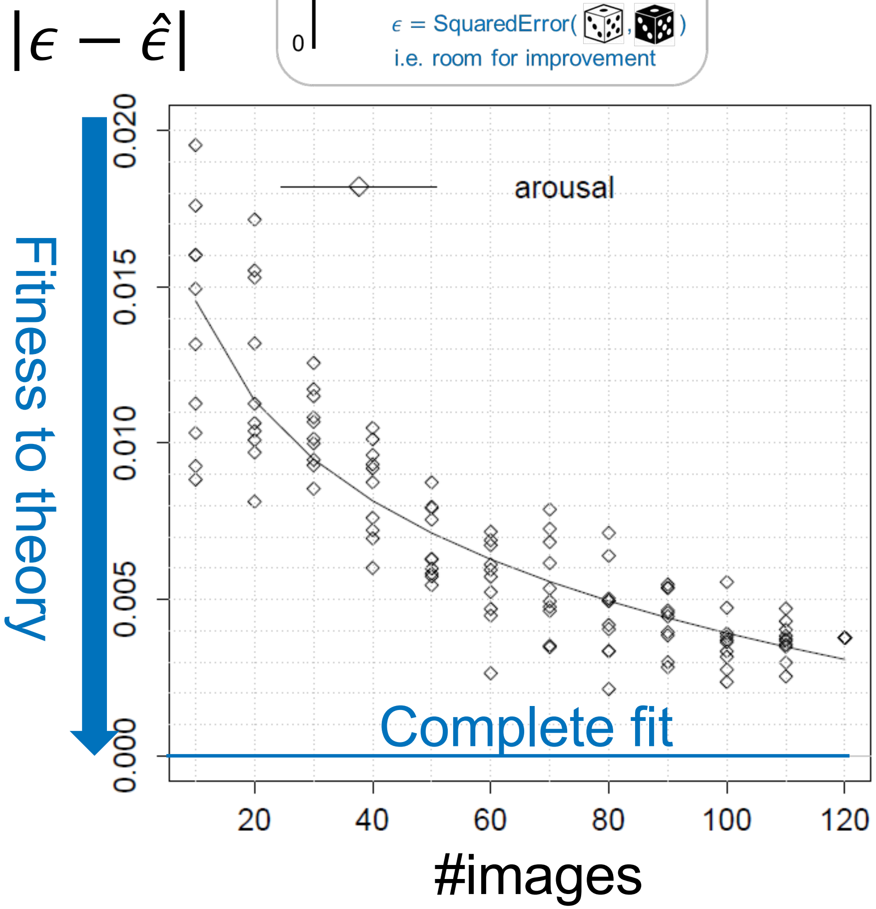
Collision probability = 2nd-order Renyi entropy

Results of simulation





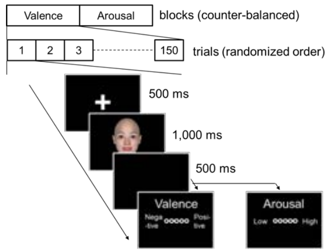
(a) Valence



(b) Arousal

Well fitted to theory with enough samples

Results on real data



Valence

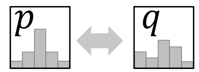
Proposed (CPM+CE losses)

Dist.-based	Max-class-based
True-CP 0.61	True conf. NA
Cross-CP 0.47	Model accuracy 0.60
Pred-CP 0.47	Model conf. 0.59

CE loss only

Dist.-based	Max-class-based
True-CP 0.61	True conf. NA
Cross-CP 0.61	M accuracy 0.62
Pred-CP 0.94	Model conf. 0.96

$\epsilon = 0.61 - 0.47 = 0.14 \longrightarrow$ Error in prob. per class = $\sqrt{0.14/5} = 0.17$



Arousal

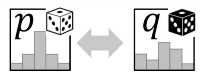
Proposed (CPM+CE losses)

Dist.-based	Max-class-based
True-CP 0.50	True conf. NA
Cross-CP 0.39	Model accuracy 0.50
Pred-CP 0.39	Model conf. 0.50

CE loss only

Dist.-based	Max-class-based
True-CP 0.50	True conf. NA
Cross-CP 0.55	M accuracy 0.54
Pred-CP 0.90	Model conf. 0.93

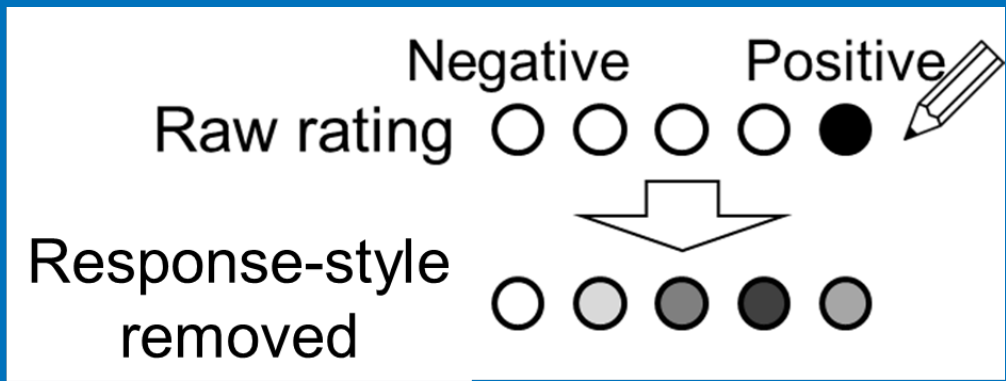
$\epsilon = 0.50 - 0.39 = 0.11 \longrightarrow$ Error in prob. per class = $\sqrt{0.11/5} = 0.15$



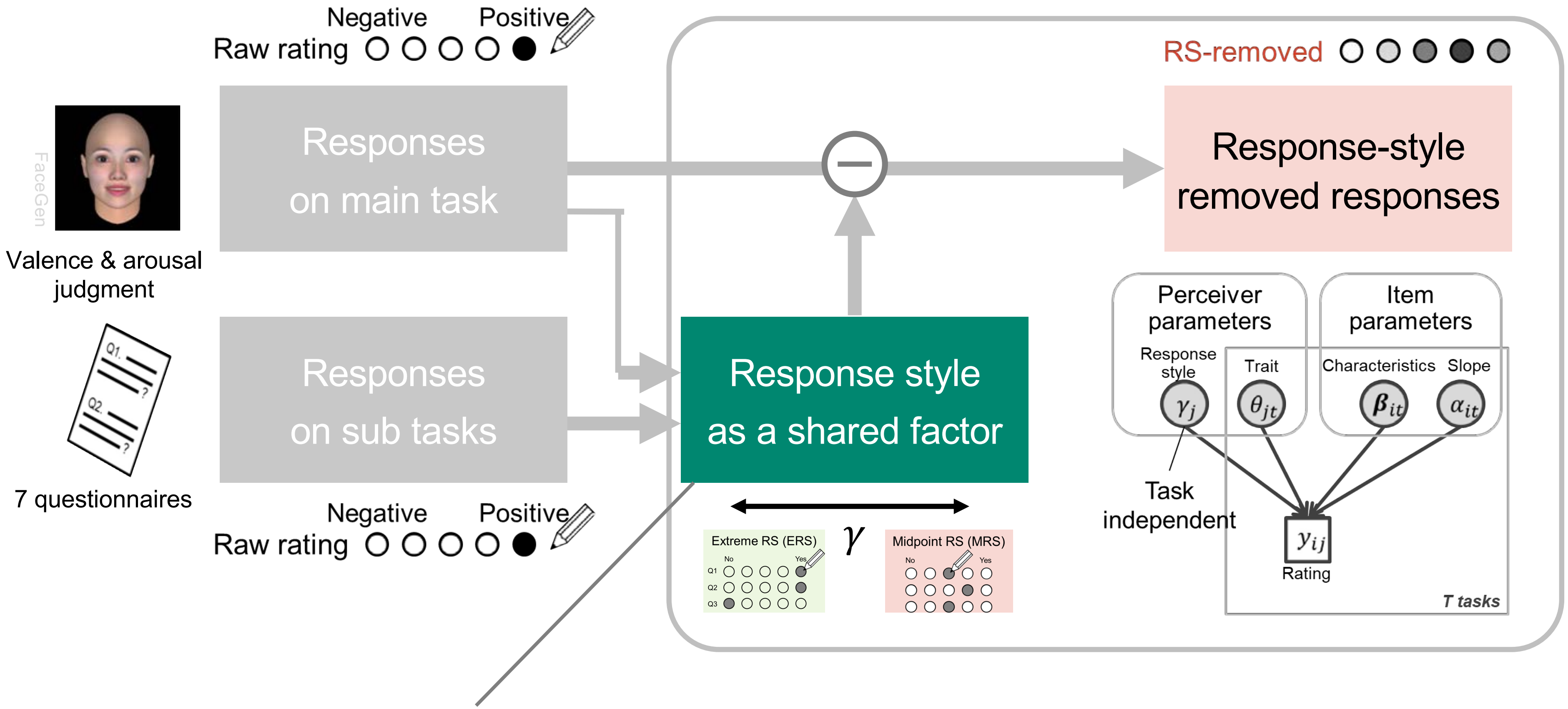
Removal of extreme/midpoint response styles on emotion judgment data

[presented at ACII 2019]

Shiro Kumano and Keishi Nomura



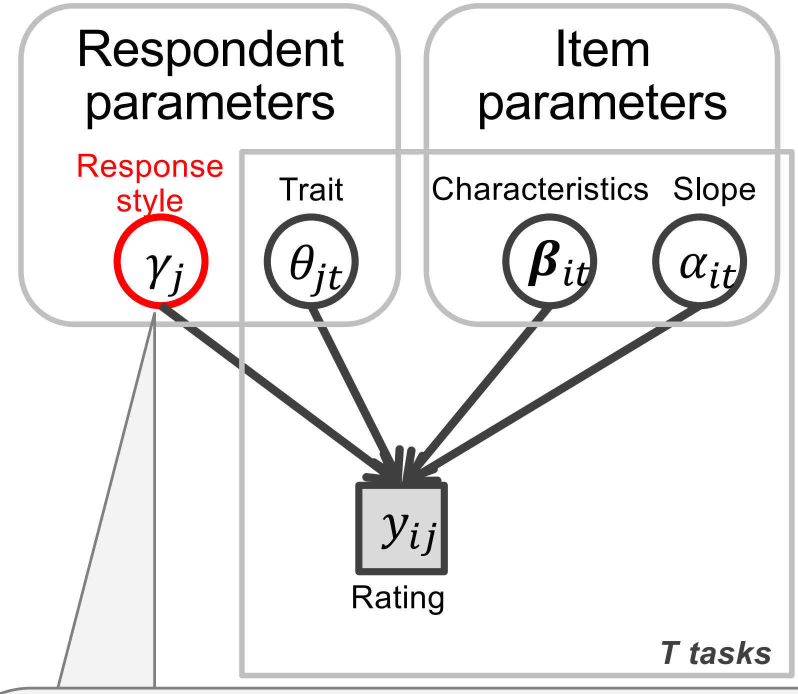
Modeling procedure



Response style (RS)
= Tendency to choose specific categories regardless of content
[Baumgartner & Steenkamp 2001]

Our multitask model (mtGPCMRS)

Extended version of [Tutz et al. 2018]



A family of ordinal regression

$$\log\left(\frac{P\left(y_{ij}=s \mid \Phi\right)}{P\left(y_{ij}=s-1 \mid \Phi\right)}\right)=\xi_{ijs}$$

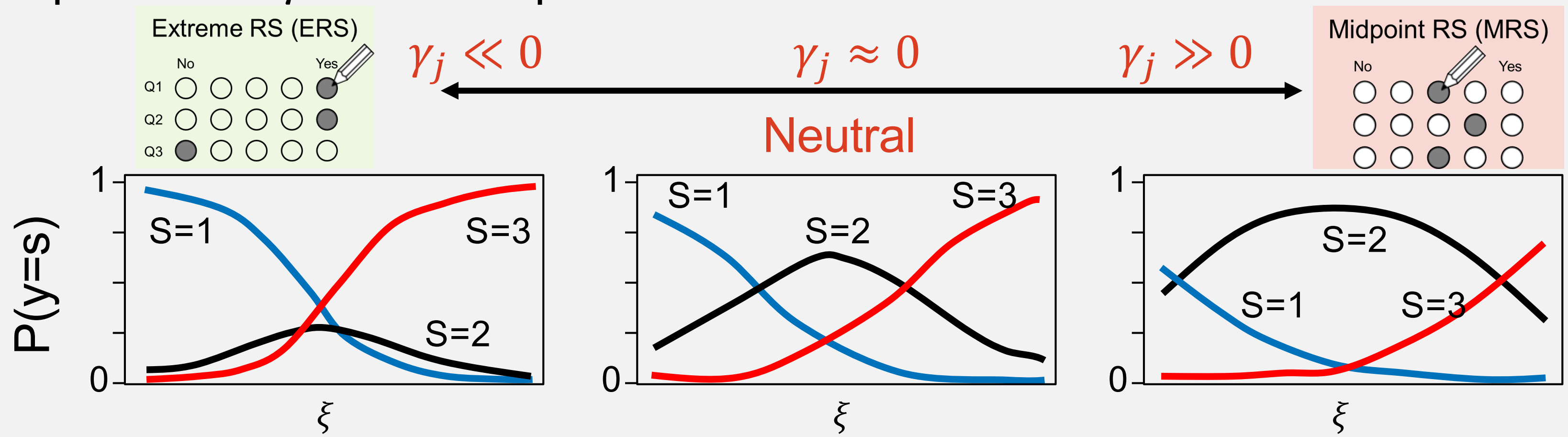
$$\xi_{ijst}=\underbrace{\alpha_{it}\theta_{jt}-\beta_{ist}}_{\text{Base terms}}+\underbrace{\left(m_{st}-s+0.5\right)\gamma_j}_{\text{Response style term}}$$

Base terms i.e. GPCM

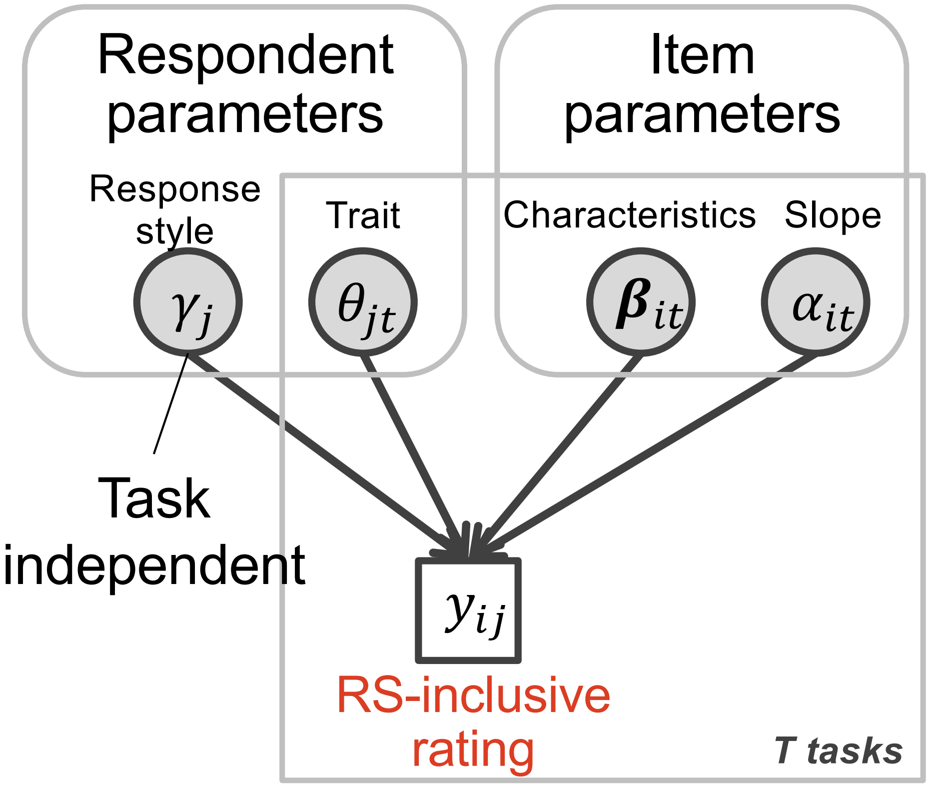
Response style term

y: Response
i : Item (stimulus)
j : Respondent
Φ: Parameter set
m_s: Middle response

RS parameter γ: Task independent factor



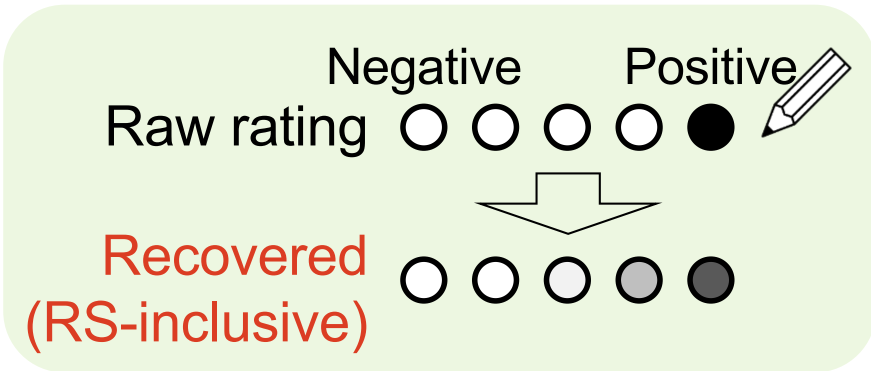
Rating re-generation process: RS-inclusive



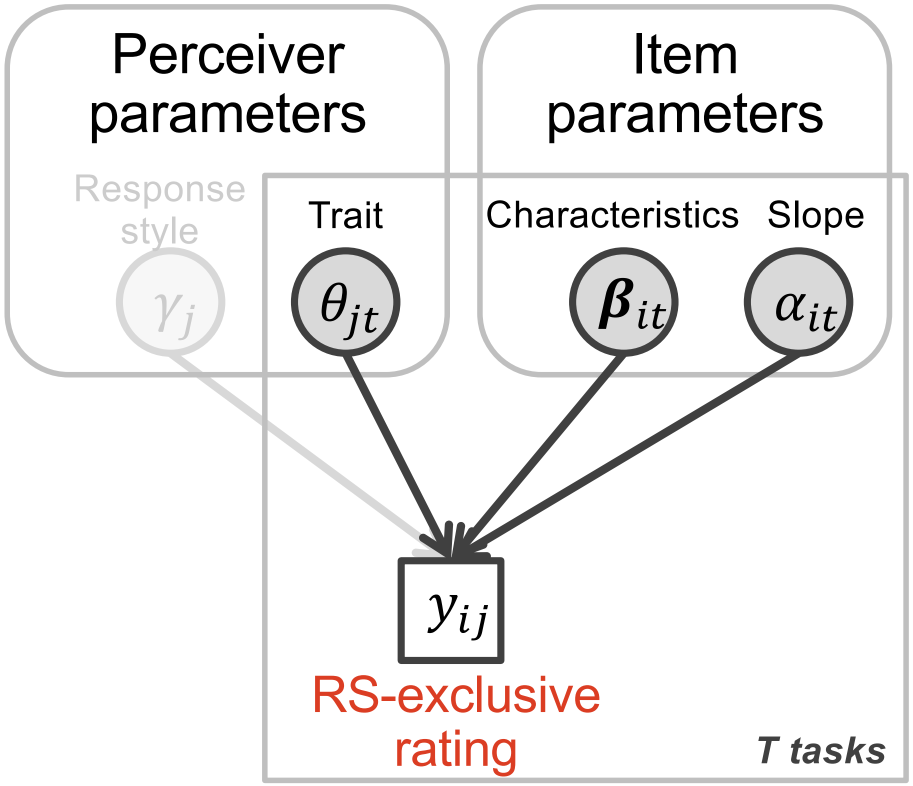
$$\xi_{ijst} = \alpha_{it}\theta_{jt} - \beta_{ist} + \left(m_{st} - s + 0.5\right)\gamma_j$$

Posterior predictive
RS-inclusive distribution

$$P(\tilde{y}|y) = \int \boxed{P(\tilde{y}|\Phi)} p(\Phi|y) d\Phi$$



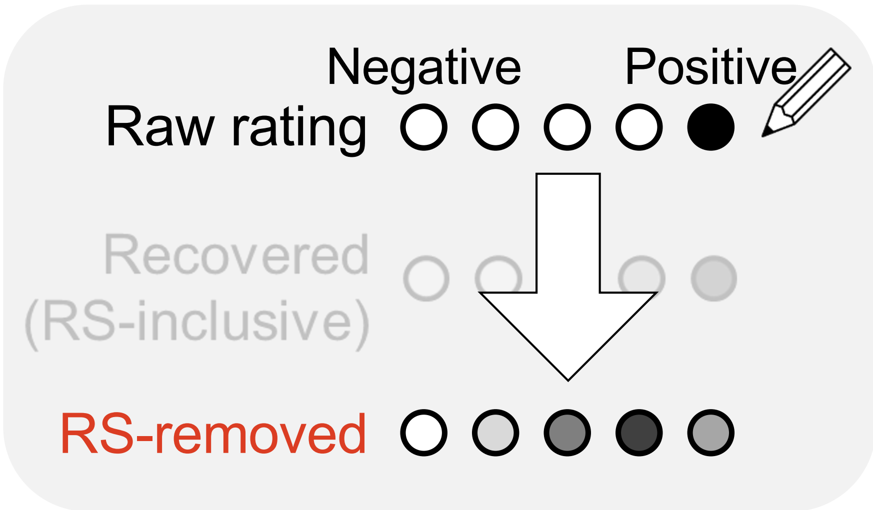
Rating re-generation process: RS-free



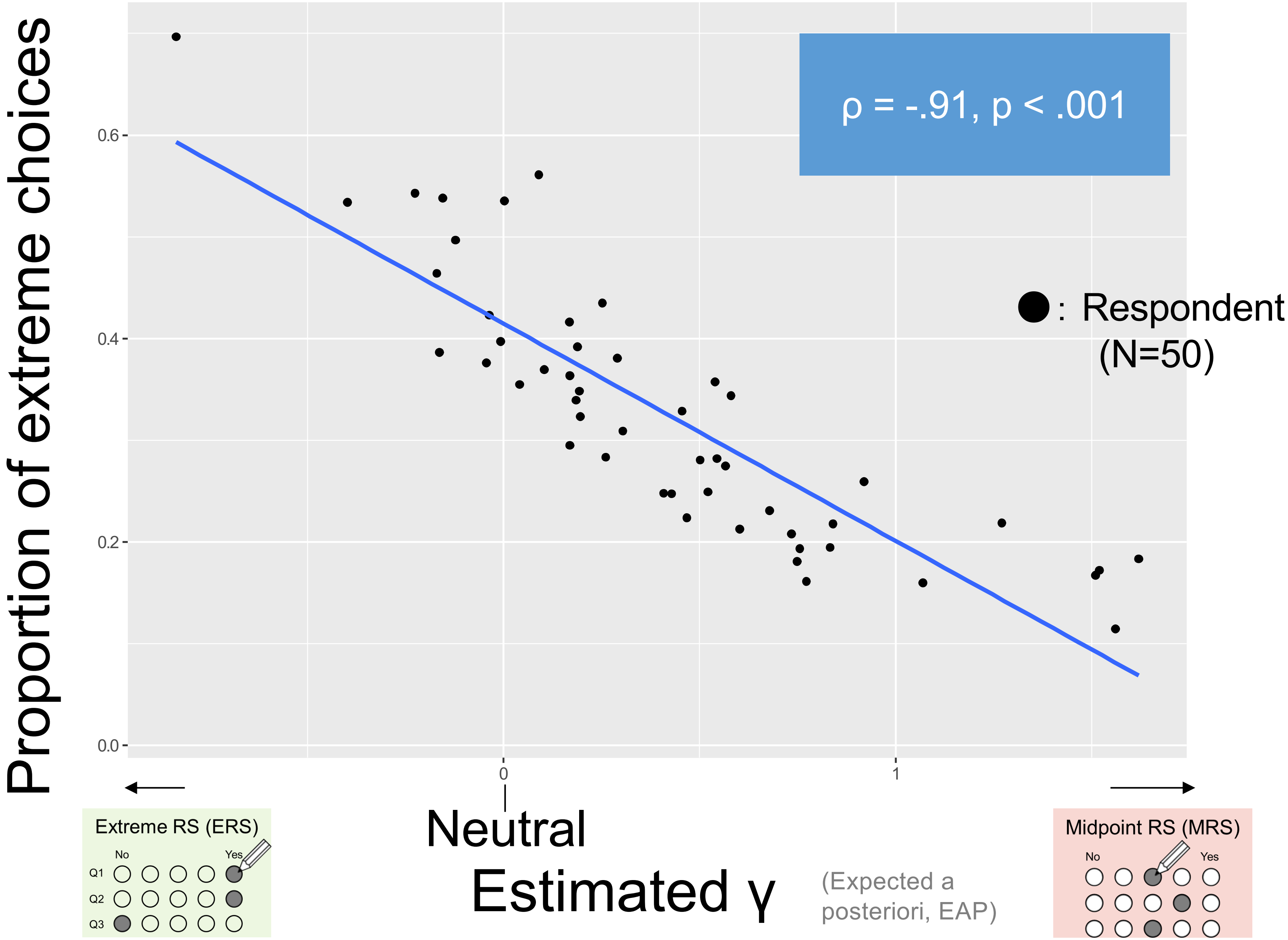
$$\xi_{ijst} = \alpha_{it}\theta_{jt} - \beta_{ist} + (m_{st} - s + 0.5)\gamma_j$$

Posterior predictive RS-free distribution

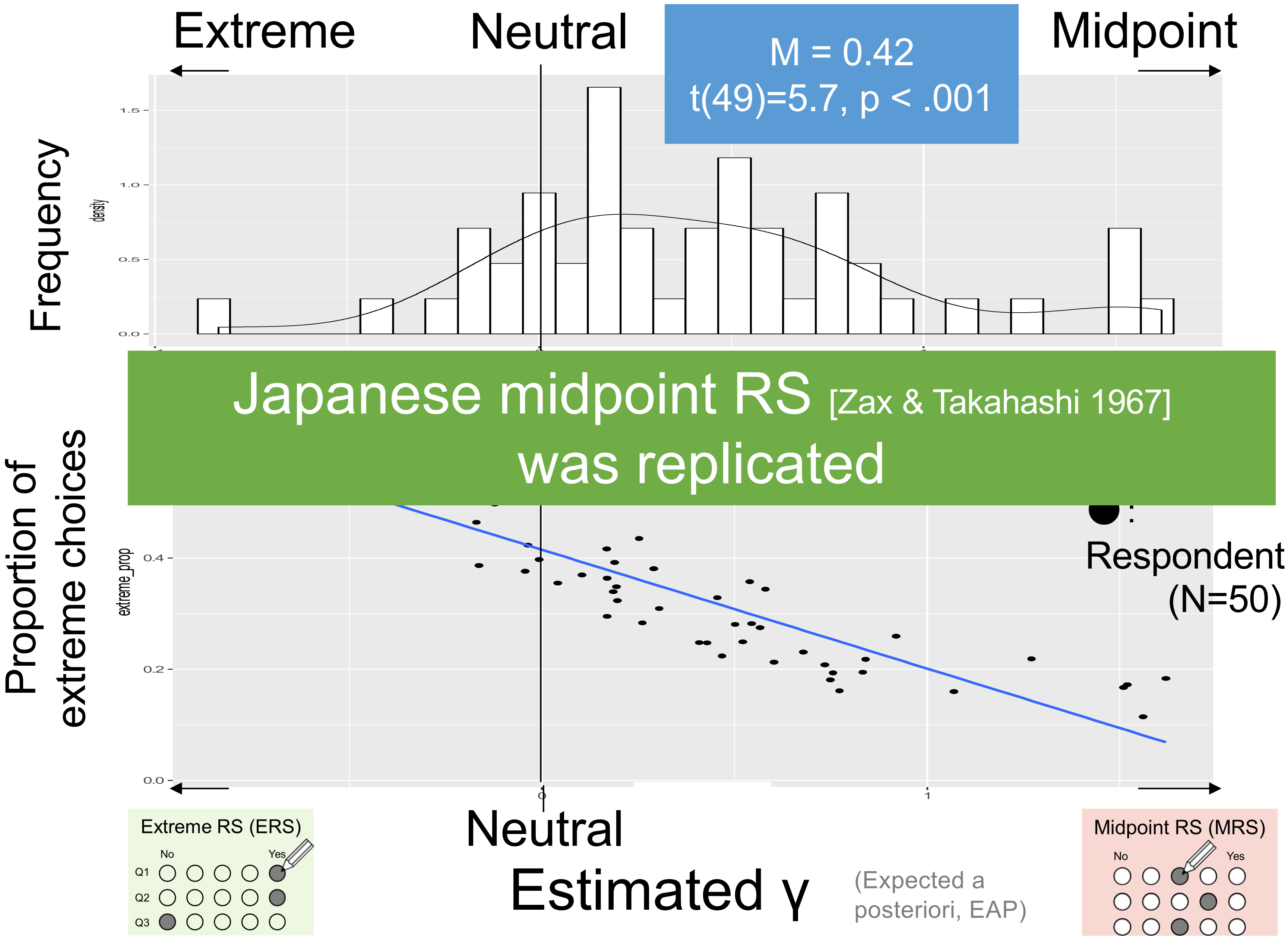
$$P(\tilde{y}|y) = \int \boxed{P(\tilde{y}|\Phi_{-\gamma})} p(\Phi|y)d\Phi$$



Our estimates vs traditional metric (Extreme RS)

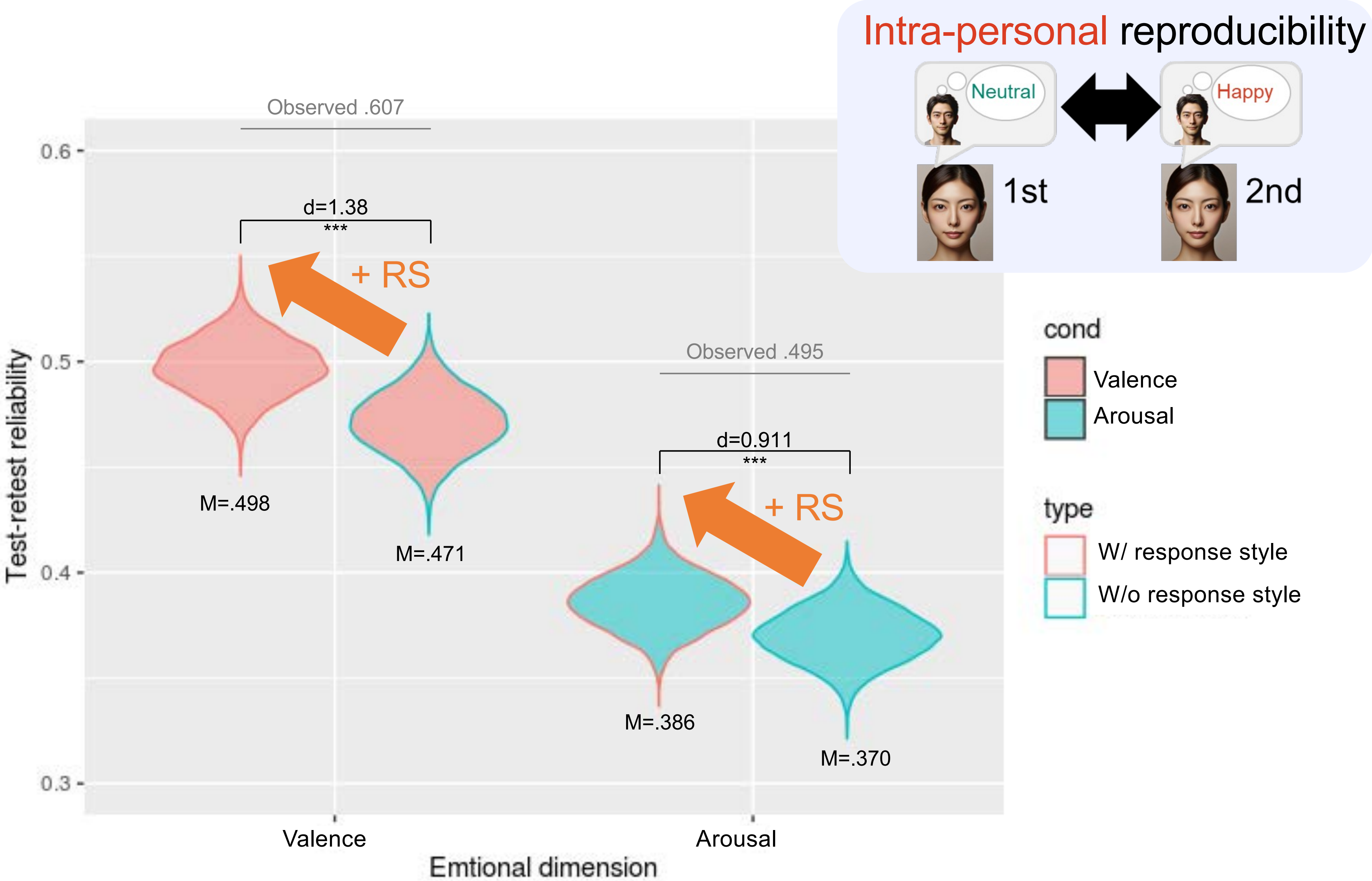


Our estimates vs traditional metric (Extreme RS)



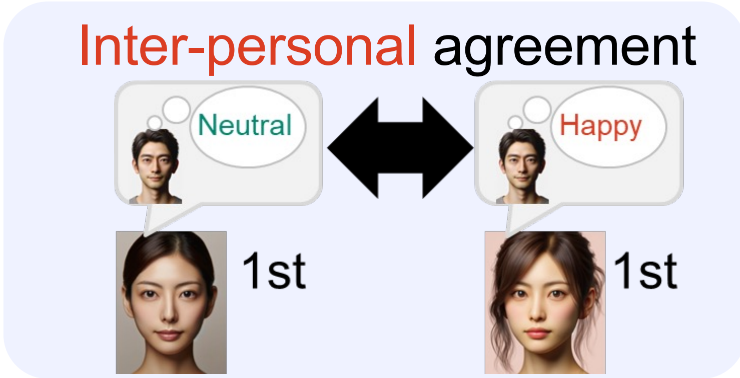
RS increased individual's test-retest reliability

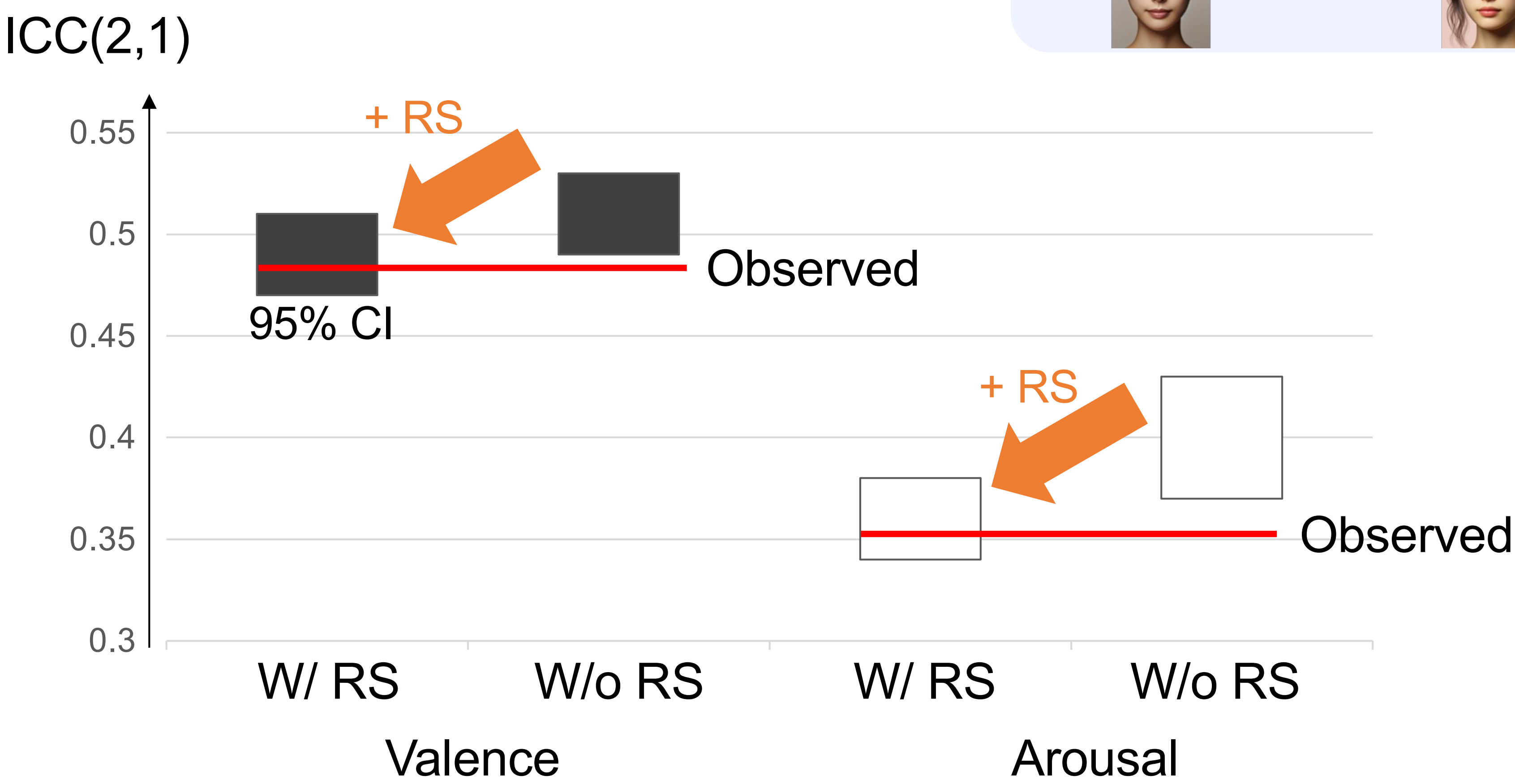
NTT [in prep.]



RS decreased inter-personal agreement

[Kumano & Nomura ACII 2019]





Summary: Approaches to 3 issues in personalized subjective affect estimation

1. Combined deep learning and item-response theory for balancing performance and explainability
2. Developed an absolute metric to measure how close models are to true model under aleatoric uncertainty
3. Proposed a model to remove response styles (ERS/MRS) from subjective ratings