Predicting Item Responses Using Mouse Trajectory Paradata and Machine Learning

Kyosuke Bunji

Kobe University

bunji@bear.kobe-u.ac.jp

IMPS2026
June 24, 2026

1

Introduction

Beyond Responses and Response Times

- **Online responses via PCs and smartphones have become standard today**
- **Numerous studies utilize response times (RT)**

[Examples]

van der Linden (2007): Models RT parallel to IRT, integrated within a hierarchical framework

Meng et al. (2014): A model where a smaller variance in the expected response probabilities across categories leads to a smaller expected RT

Bunji & Okada (2020): A model combining IRT and a diffusion model to handle the joint distribution of responses and RT

- **Are there other useful sources of data?**

Specifically, I want to utilize data **that can be collected automatically without additional respondent burden**

Psychology and Paradata (Freeman, 2018)

- Response choices and RTs only reflect the final "outcomes"
- To understand cognitive processes, we need data that capture behavior in real time

  e.g., EEG (electroencephalography), eye-tracking data
- Several studies have used eye-tracking data

  Chauliac et al. (2020): Investigated eye movement features and response consistency

  Man et al. (2022): Extended van der Linden's (2007) hierarchical framework to model visual fixation counts

  Tu et al. (2026): Used variance of response probability as a parameter for modeling visual fixation counts

Process data may contain information beyond responses and RTs.

Cursor Trajectories in Psychology

- **Research on mouse tracking has expanded rapidly in the 21st century**

  One major driver: the development of specialized mouse-tracking software

- **Hand movements and brain activity are closely linked**

  ➔ Cursor movements may reflect the microstructure of decision-making

- **However, trajectories has been used primarily in cognitive psychology tasks**

  [Example] Di Palma et al. (2022) used cursor trajectories to analyze the IAT

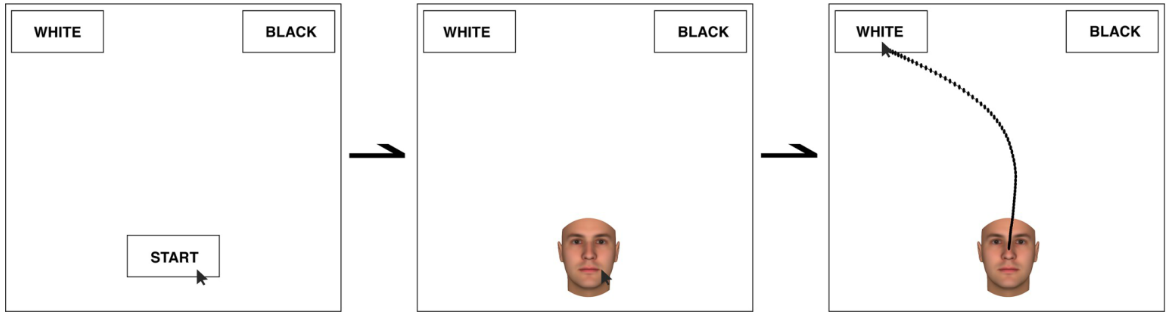


Fig 1. Representative illustration of a typical trial of MT-race-IAT. After clicking on 'Start' button at the bottom centre of the screen, the stimulus is presented in its position. The task consists in categorising the stimulus as fast as possible by moving and clicking over the chosen on the top-left or top-right target response alternatives with the computer mouse while the real-time mouse trajectory is recorded.

Cursor Trajectories and Psychometrics

- **Very few studies have applied mouse tracking to rating scales**
- **Existing studies have mainly used trajectories for auxiliary purposes**

[Examples]

Mok et al. (2017) used cursor trajectories **to detect low-quality workers**

Wang et al. (2020) collected cursor trajectories and Likert responses as separate data for the same task and treated them as independent **to investigate the initial decision-making process**

Shiina (2019, in Japanese) **examined the relationship between RT and cursor trajectories** for Likert-scale responses

Do mouse trajectories contain information beyond responses and RT?

Purpose of the Study

- **To examine the utility of using cursor trajectories in addition to response choices and RT**

  Are responses and RT sufficient, or do cursor trajectories contain hidden information not captured by them?

- **Predictive Task**

  Predicting response and RT

  If cursor trajectories contain relevant information, prediction accuracy should improve

- **Not the focus of this presentation**

  Building latent trait model to estimate θ using cursor trajectory data

  (Whether incorporating mouse trajectories into θ estimation models improves validity)

2

Methods

Dataset

Note: The dataset analyzed here differs from the one described in the submitted abstract.

Data were collected via a crowdsourcing platform in Japan (CrowdWorks).

- Measure: Japanese HEXACO inventory (Wakabayashi, 2014)

  Only 32 items were used (16 items each for Agr and eXt) to reduce respondent burden

- Final sample size: 1,418 (out of 2,000 collected)

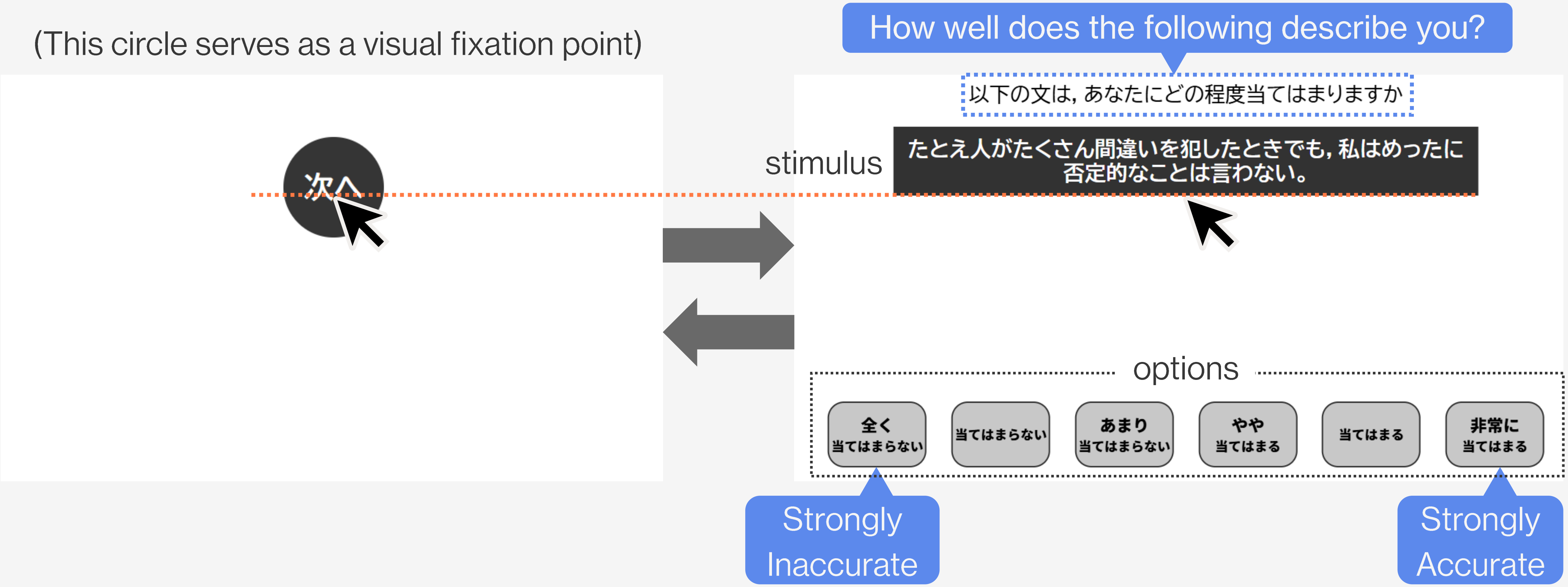
  [Exclusion criteria]

  a. Self-reported pointing device other than a mouse (e.g., trackball or trackpad)
  b. RT < 1 s or ≥ 30 s on at least one item
  c. Incomplete or failed cursor log recording
  d. Loss of browser focus during responding
  e. Significant cursor deviation from the response area

Experimental Interface

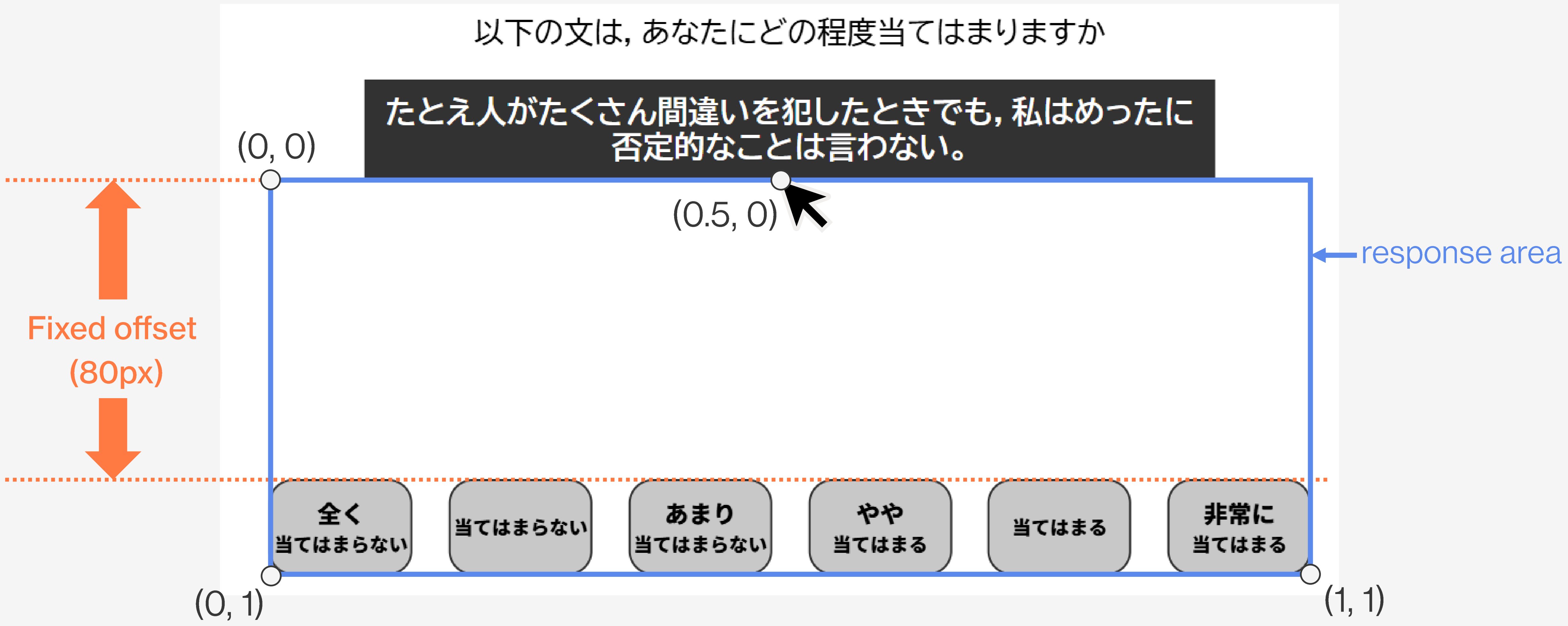
When the cursor enters and paused on the circle, the item and response options appear on the screen.

(This circle serves as a visual fixation point)



Coordinate Normalization

- Cursor coordinates were normalized within the response area (0-1)



Input Feature Groups (mouse features were based on Kieslich, Henninger, et al., 2019; Pino et al., 2025)

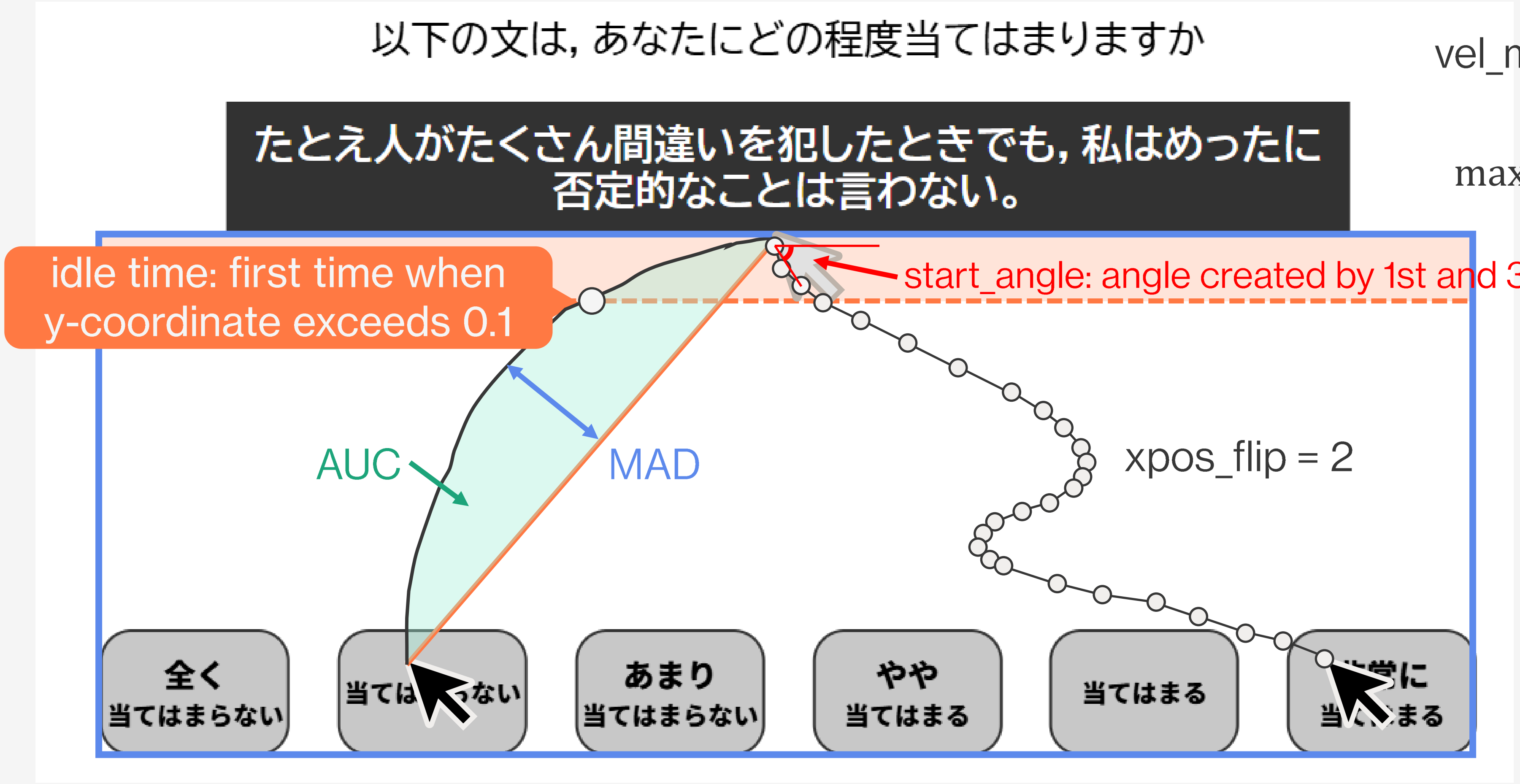
Features were grouped into response, RT, and mouse variables and were compared.

type	name	brief description
response	response	Response to each item (0-5 Likert scale)
	centered response	centered response ({0, 5} -> 0; ({1, 4} -> 1; ({2, 3} -> 2)
	fscore	Factor score from categorical CFA
	fscore2	Square of fscore
RT	rt	RT on each item
	rt_sd	Within-respondent SD of RT
mouse	MAD	Maximum deviation from the straight-line trajectory
	AUC	Area Under the Curve (AUC) relative to the straight-line trajectory
	xpos_flips	Directional changes along the x-axis
	total_dist	Total Euclidean distance traveled
	vel_max	Maximum velocity
	acc_max	Maximum acceleration
	idle_time	Onset time of cursor movement in the y-direction
	curvature	Ratio of total_dist to the length of the straight-line trajectory
	start_angle	Initial movement angle

For the features in red,
- both standardized (within-person) and non-standardized versions were used.
- PCA was conducted on each version, retaining the first three principal component scores.

Examples of Mouse-Trajectory Features

- Cursor coordinates are recorded approximately every 20–30 ms via the JavaScript `pointermove` event.



vel_max =

$$\max\left(\frac{\sqrt{(x_{k+1}-x_k)^2+(y_{k+1}+y_k)^2}}{t_{k+1}-t_k}\right)$$

Prediction Task

ML to predict response behavior on specific items

Targets: (a) raw response, (b) centered response, (c) response time

Model: XGBoost

Elastic net, Random forest and SVM were also tried but not reported here.

Evaluation criterion: RMSE

Hyperparameter Tuning: Bayesian optimization with 5-fold CV

Input features

Features from remaining 15 items (within the same factor)

7 (+1 null) models based on all feature-group combinations

To evaluate the incremental value of mouse features

	(null)							(full)
response		✓		✓		✓		✓
rt			✓	✓			✓	✓
mouse					✓	✓	✓	✓

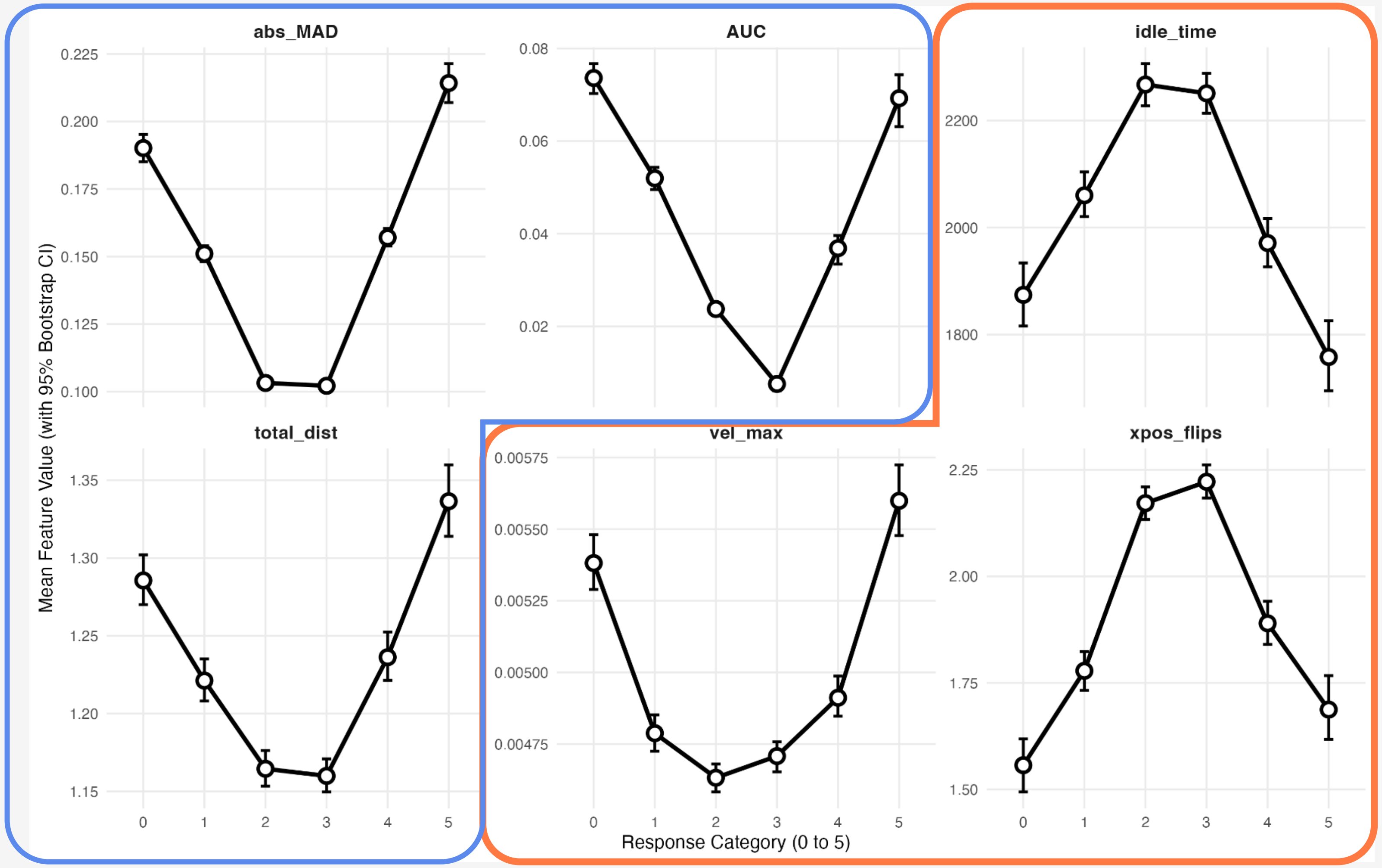
3

Results

Representative mouse features by response category

- Mouse features systematically vary across response categories.

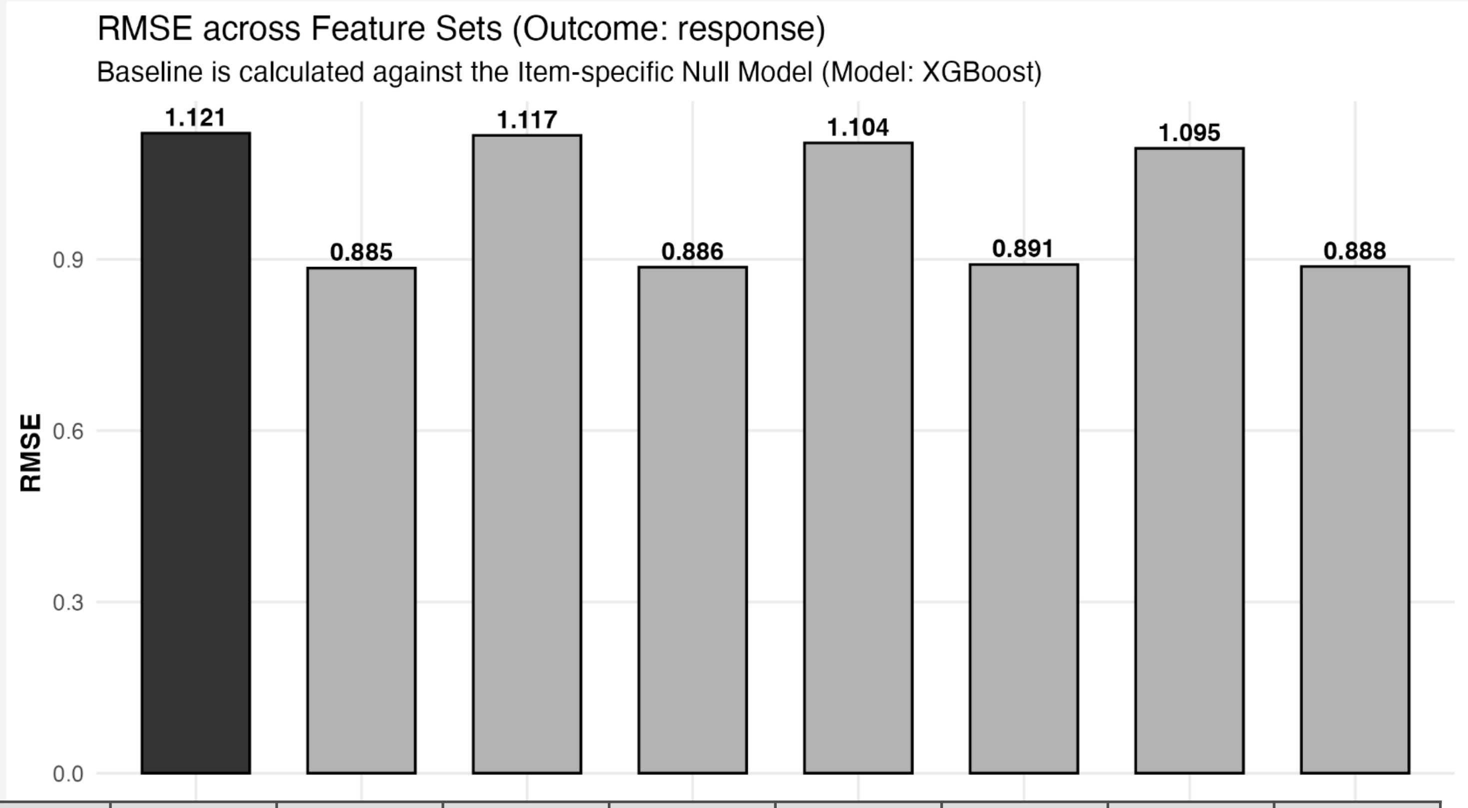
These patterns may be driven by geometric bias. (i.e., differing positions of option buttons)



Similar pattern was observed with RT (i.e., inverted-U relationship)

These features may reflect response "hesitation."

Model Performance (Target: raw response)



response		✓		✓		✓		✓
RT			✓	✓			✓	✓
mouse					✓	✓	✓	✓

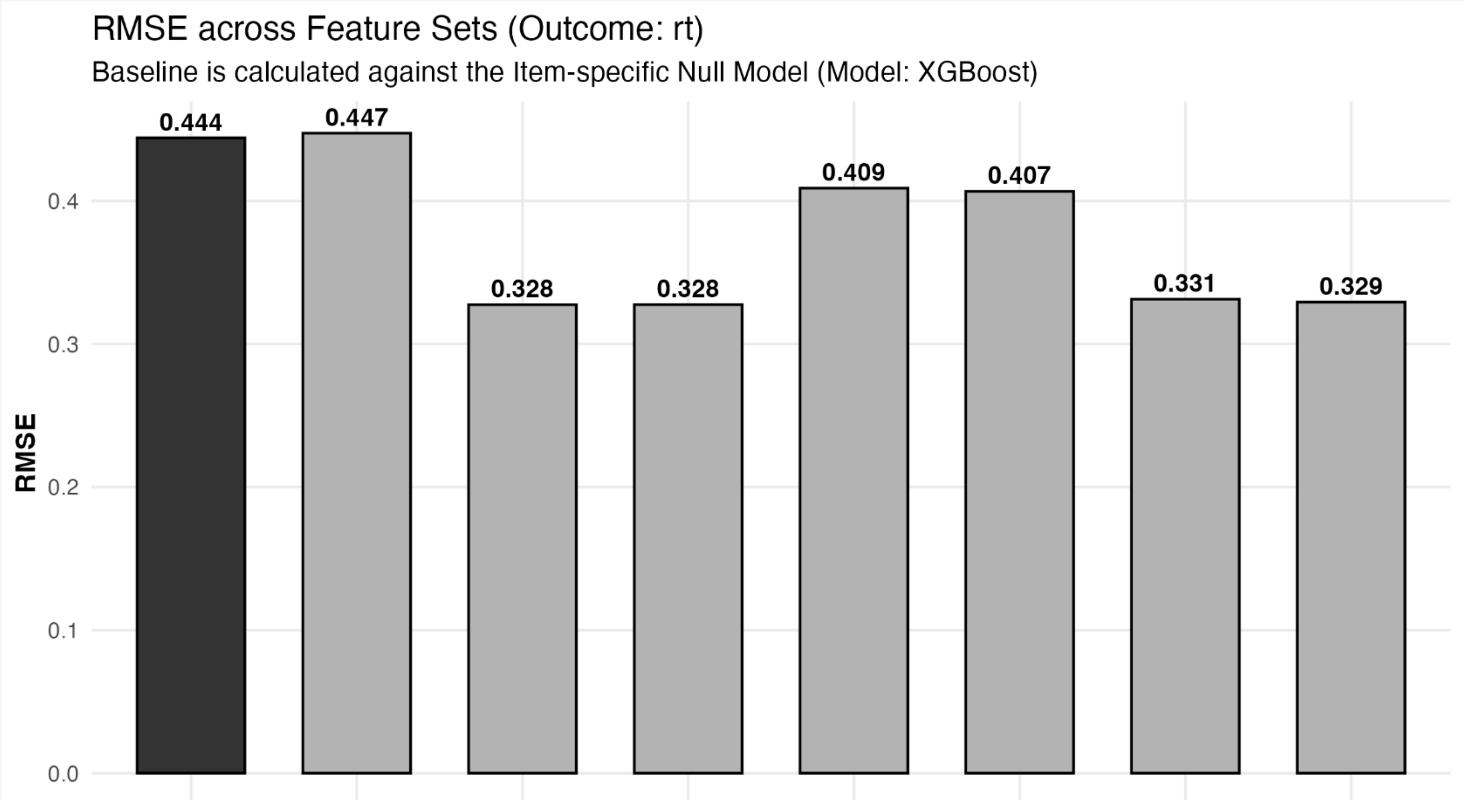
- The response-only model performed best
- Adding RT or mouse features did not improve performance

- When response features are not used, mouse features outperformed RT features

A similar tendency was observed when target is centered response.

Most predictive information is already captured by response features.

Model Performance (Target: RT)



response		✓		✓		✓		✓
RT			✓	✓			✓	✓
mouse					✓	✓	✓	✓

- The RT-only model performs best
- Mouse features did not improve performance in this task

- When RT features are not used, mouse features outperformed response features

Most predictive information is already captured by RT features.

Macro-Shapley value (Shapley, 1953)

The contribution of each feature group depends on the inclusion order

➔ The Shapley value fairly distributes each feature's contribution

	target		
feature	raw response	centered response	rt
response	0.3522 (93.33%)	0.2465 (95.45%)	−0.0012 (−0.27%)
RT	0.0064 (1.71%)	0.0009 (0.34%)	0.3754 (83.04%)
mouse	0.0152 (4.08%)	0.0109 (4.21%)	0.0779 (17.23%)

When predicting response

Response features dominate the contribution (over 93%)

Mouse features contribute more than RT features

When predicting RT

While RT features show a strong effect, mouse features also make a moderate contribution (~17%)

Definition of macro-Shapley value

[Example] mean contribution of mouse-related features

0	1st	2nd	3rd	Contribution
Null	response	RT	mouse	$v(\text{resp}) - v(\phi)$
Null	response	mouse	RT	$v(\text{resp}) - v(\phi)$
Null	RT	response	mouse	$v(\text{resp}, \text{RT}) - v(\text{RT})$
Null	RT	mouse	response	$v(\text{resp}, \text{RT}, \text{mouse}) - v(\text{RT}, \text{mouse})$
Null	mouse	response	RT	$v(\text{resp}, \text{mouse}) - v(\text{mouse})$
Null	mouse	RT	response	$v(\text{resp}, \text{RT}, \text{mouse}) - v(\text{RT}, \text{mouse})$

↑ Mean of contributions is the macro-Shapley value

$$Shaplay(resp) = \frac{1}{3}[v(\text{resp}) - v(\phi)] + \frac{1}{3}[v(\text{resp}, \text{RT}, \text{mouse}) - v(\text{RT}, \text{mouse})\ ]$$

$$+\frac{1}{6}[v(\text{resp}, \text{mouse}) - v(\text{mouse})] + \frac{1}{6}[v(\text{resp}, \text{RT}) - v(\text{RT})\ ]$$

Top-ranked features (based on SHAP, Target: response)

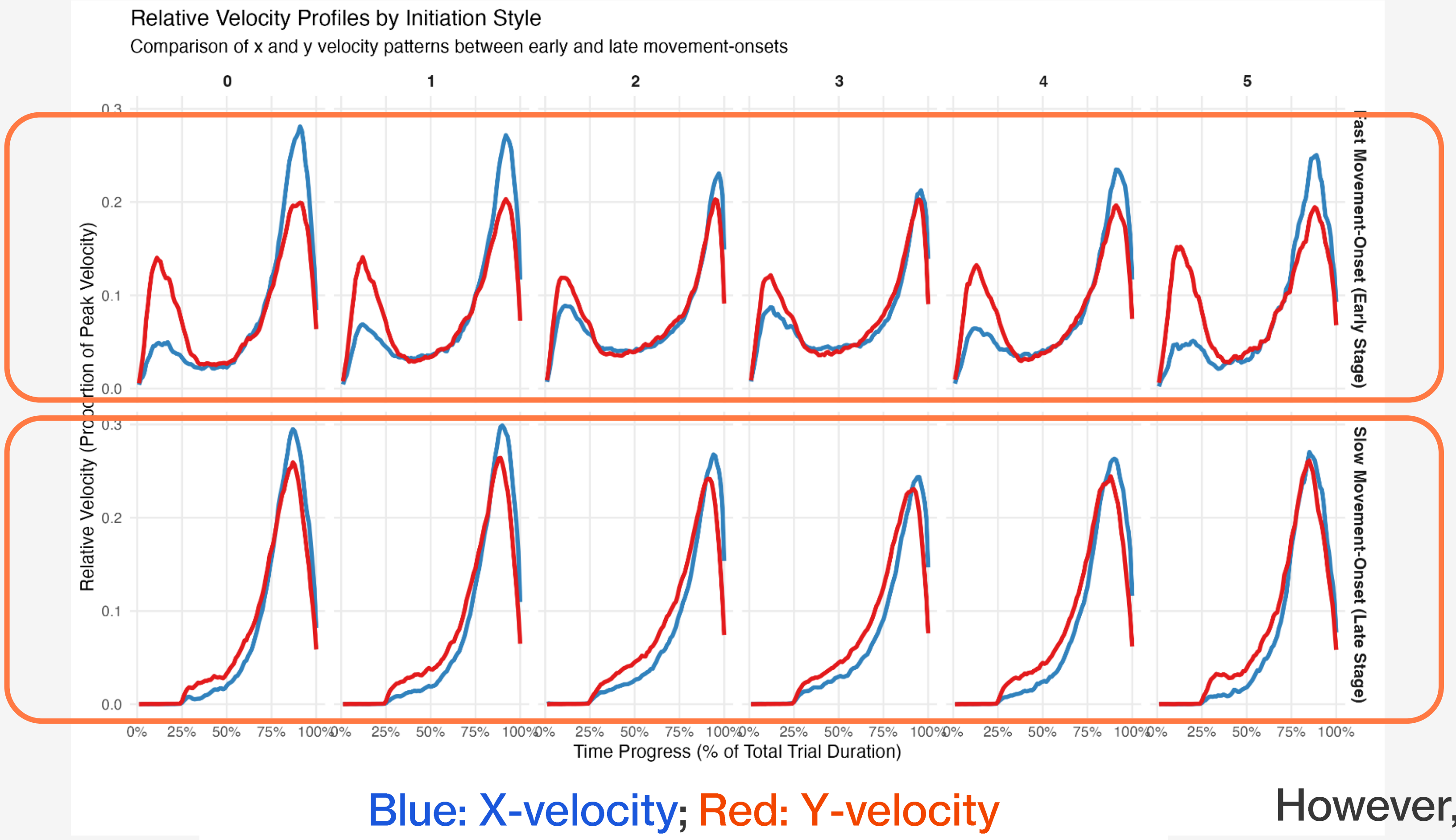
(↑Lundberg and Lee, 2017)

Average ranking across 32 target items

response	✓		✓		✓		✓
rt		✓	✓			✓	✓
mouse				✓	✓	✓	✓
1	fscore	rt_z_pca1	fscore	AUC_sd	fscore	AUC_sd	fscore
2	i4	rt_z_pca2	i4	MAD_sd	fs2	MAD_sd	i25
3	c_i22	rt_pca2	c_i22	AUC_pca2	c_i22	MAD_pca1	fs2
4	c_i15	rt_pca1	i25	MAD_pca1	i4	AUC_pca2	i4
5	i25	rt_z_pca3	c_i20	AUC_z_pca3	i9	rt_z_pca1	i9
6	i9	rt_sd	fs2	start_angle_z_pca1	i25	AUC_z_pca3	c_i22
7	fs2		c_i1	AUC_pca1	c_i20	start_angle_z_pca1	c_i25
8	c_i1		c_i15	MAD_z_pca1	c_i2	rt_z_pca2	c_i20
9	c_i2		i9	AUC_z_pca2	c_i25	AUC_pca1	i1
10	c_i20		i1	AUC_pca3	c_i1	MAD_z_pca1	c_i15

Exploratory Analysis

Different cursor-movement strategies were observed



The "Fast" group starts moving before the response is fully decided, then adjusts the trajectory.

The "Slow" group starts moving the cursor only after the response option is decided.

However, RT was longer in the "Fast" group.

Summary

No Incremental Utility for Response/RT Prediction

Cursor trajectories provide little additional value for predicting responses or response times (noise-prone due to response/RT dominance).

However, they may contain information not fully captured by responses or RTs (from macro-Shapley values).

Distinct Behavioral Profiles

Early-movement onset (deciding during movement) correlates with longer RTs, showing no strong relation to final response categories.

Limitations & Future Directions

May be promising for other complex prediction tasks or clinical assessments (e.g., detecting psychiatric disorders).

Thank you for your attention!

Predicting Item Responses Using Mouse Trajectory Paradata and Machine Learning

Kyosuke Bunji

Kobe University, Japan

bunji@bear.kobe-u.ac.jp

IMPS2026
June 24, 2026

References

Bunji, K., & Okada, K. (2020). Joint modeling of the two-alternative multidimensional forced-choice personality measurement and its response time by a Thurstonian D-diffusion item response model. *Behavior Research Methods*, *52*(3), 1091–1107. https://doi.org/10.3758/s13428-019-01302-5

Chauliac, M., Catrysse, L., Gijbels, D., & Donche, V. (2020). It is all in the surv-eye: Can eye tracking data shed light on the internal consistency in self-report questionnaires on cognitive processing strategies? *Frontline Learning Research*, *8*(3), 26–39. https://doi.org/10.14786/flr.v8i3.489

Di Palma, M., Carioti, D., Arcangeli, E., Rosazza, C., Ambrogini, P., Cuppini, R., Minelli, A., & Berlingeri, M. (2022). The biased hand. Mouse-tracking metrics to examine the conflict processing in a race-implicit association test. *PLOS ONE*, *17*(7), e0271748. https://doi.org/10.1371/journal.pone.0271748

Freeman, J. B. (2018). Doing psychological science by hand. *Current Directions in Psychological Science*, *27*(5), 315–323. https://doi.org/10.1177/0963721417746793

Kieslich, P. J., Henninger, F., Wulff, D. U., Haslbeck, J. M. B., & Schulte-Mecklenbeck, M. (2019). Mouse-tracking: A practical guide to implementation and analysis. In M. Schulte-Mecklenbeck, A. Kühberger, & J. G. Johnson, A Handbook of Process Tracing Methods (pp. 111–130). Routledge.

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17*, 4768–4777.

Man, K., Harring, J. R., & Zhan, P. (2022). Bridging models of biometric and psychometric assessment: A three-way joint modeling approach of item responses, response times, and gaze fixation counts. *Applied Psychological Measurement*, *46*(5), 361–381. https://doi.org/10.1177/01466216221089344

Meng, X.-B., Tao, J., & Shi, N.-Z. (2014). An item response model for Likert-type data that incorporates response time in personality measurements. *Journal of Statistical Computation and Simulation*, *84*(1), 1–21. https://doi.org/10.1080/00949655.2012.692368

References (cont.)

Mok, R. K. P., Chang, R. K. C., & Li, W. (2017). Detecting low-quality workers in QoE crowdtesting: A worker behavior-based approach. *IEEE Transactions on Multimedia*, *19*(3), 530–543. https://doi.org/10.1109/TMM.2016.2619901

Shapley, L. S. (1953). A value for n-person games. In Contributions to the theory of games II (pp. 307–317). Princeton University Press.

Shiina, K. (2019). Cursor trajectories can predict judgment times (written in Japanese). Academic studies and scientific research. Cultural science and social science, *67*, 97-111.

Tu, D., Zhang, F., Peng, S., Wang, D., & Cai, Y. (2026). A multimodal item response modeling for personality assessment in organizational research. *Organizational Research Methods*, 10944281261457337. https://doi.org/10.1177/10944281261457337

Pino, A., Papatheodorou, N., Kouroupetroglou, G., Giannopoulos, P.-A., Makris, G., & Papageorgiou, C. (2025). Hand dexterity evaluation grounded on cursor trajectory investigation in children with ADHD using a mouse and a joystick. *Technologies*, *13*(3), Article 3. https://doi.org/10.3390/technologies13030099

van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. Psychometrika, 72(3), 287–308. https://doi.org/10.1007/s11336-006-1478-z

Wakabayashi, A. (2014). A sixth personality domain that is independent of the Big Five domains: The psychometric properties of the HEXACO Personality Inventory in a Japanese sample. Japanese Psychological Research, 56(3), 211–223. https://doi.org/10.1111/jpr.12045

Wang, J., Lin, W., Fang, X., & Mo, L. (2020). The influence of emotional visual context on the judgment of face trustworthiness. *Psychology Research and Behavior Management*, *Volume 13*, 963–976. https://doi.org/10.2147/PRBM.S269543