

マウス軌跡パラデータを用いた 項目反応および回答時間予測の試み

分寺 杏介



加葉田 大志朗

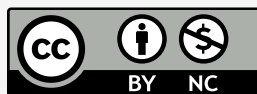
中田 匠哉

神戸大学



bunji@bear.kobe-u.ac.jp

日本テスト学会第24回大会
2026年9月15日



※ These slides are available under CC BY-NC 4.0.

1

問題と目的

- PC・スマホによるウェブ調査は今や当たり前になっている
- 回答時間 (RT) を利用する研究はすでに色々ある

【例】

van der Linden (2007):階層モデルによって回答内容と回答時間を同時に扱う

Meng et al. (2014): 期待反応確率のカテゴリ間分散が小さいほど期待RTが小さくなると仮定したモデル

Bunji & Okada (2020):ドリフト拡散モデルに当てはめて一対比較データのRTを分析するモデル

- RT以外のデータにも可能性はあるのでは？

特に,回答者に追加の負荷を与えない＝自動的に収集できるデータがあると嬉しい

(前提) 回答内容とRTは, 回答プロセスの「結果」の表出である

■ 回答の認知過程を観測する場合, 回答中のリアルタイムな行動記録が必要
e.g., 脳波 (EEG), アイトラッキング

■ 例えばアイトラッキングに関しては近年研究が増えている印象

Chauliac et al. (2020): 視線の動きと回答の一貫性の関係を調べた

Man et al. (2022): van der Linden の階層モデルに「注視回数」を追加したモデル拡張を提案した

Tu et al. (2026): 注視回数の確率分布のパラメータに, 期待反応確率のカテゴリ間分散を追加

プロセスデータには, 回答内容やRTとは別の情報が含まれているはず

ただ, 脳波もアイトラッキングも専門の機器が必要なので一般的には利用が難しく…

- 21世紀に入ったあたりから急速に増えてきた
理由の一つは,分析のためのソフトウェア (e.g., Freeman & Ambady, 2010)が開発されてきたこと
- 手の運動に「回答時の葛藤」などが表出するとされている (e.g., MacLeod, 1991; Longstaff & Heath, 2003)
 - ▶ カーソルの動きには意思決定の微細な認知プロセスが反映されるかも
- しかし現時点では,認知心理学の二肢選択課題での利用がほとんど
【例】 di Palma et al. (2022): IATでのカーソル軌跡を利用している

迷わないほどカーソルはまっすぐ
正解の選択肢へ進むと考えられる

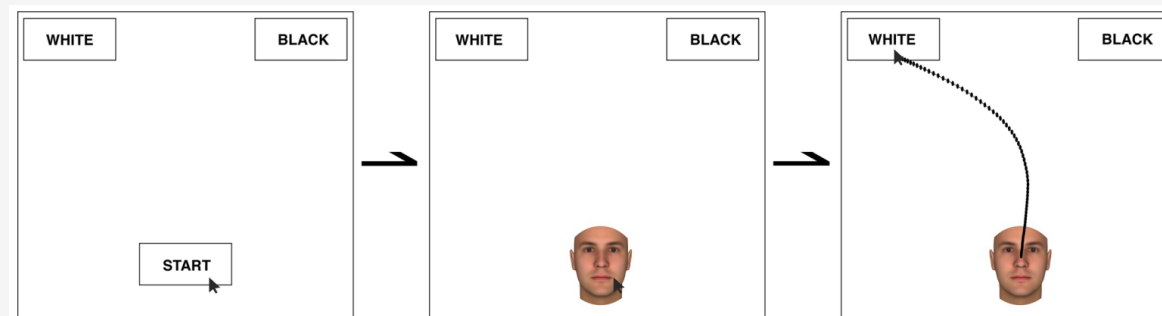


Fig 1. Representative illustration of a typical trial of MT-race-IAT. After clicking on 'Start' button at the bottom centre of the screen, the stimulus is presented in its position. The task consists in categorising the stimulus as fast as possible by moving and clicking over the chosen on the top-left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

- 特に評定尺度(リッカート尺度)における研究は数えるほどしか存在しない
- それらの研究も,基本的には補助的な利用に留まっている

【例】

Mok et al. (2017): 質の低いワーカー(回答者)の検出に利用する

Wang et al. (2020): マウスカーソル軌跡を用いて顔の信頼性判断の初期評価過程を検討,
これとは別にリッカート尺度によって最終的な判断結果を測定した

椎名 (2019): リッカート尺度における回答内容(選択肢)とRTとカーソル軌跡の関連を
探索的に調べた

リッカート尺度への回答について, カーソル軌跡は
回答内容とRTとは異なる情報を提供するのだろうか?

■ 回答内容とRTに加えて,カーソル軌跡を利用する価値を検証する

回答内容とRTでは捉えきれなかった情報をカーソル軌跡が提供してくれるのか?

あるいは「回答内容とRTがあればもう十分」なのか?

■ 回答時間とRTの予測タスクとして実施

▲ もしカーソル軌跡に追加の情報があるならば,予測精度が向上するはず

■ 本発表で扱わないこと

回答者の特性値 θ の測定モデルにカーソル軌跡を組み込んだモデルの開発

階層モデルのような

▲ そもそも θ の構成概念としての定義から考える必要があるため(今後の課題)

カーソル軌跡を用いて推定した θ のほうが外的指標との相関が高くなるか,など

2

方法

クラウドソーシング (CrowdWorks) によって収集したデータを使用

■ 尺度: HEXACO日本語版 (Wakabayashi, 2014)

100項目中32項目 (AgrとeXtの各16項目)のみを使用 ◀ 回答者の負担を考慮して

■ 最終的なサンプルサイズ: 1,418 (収集した数は2,000)

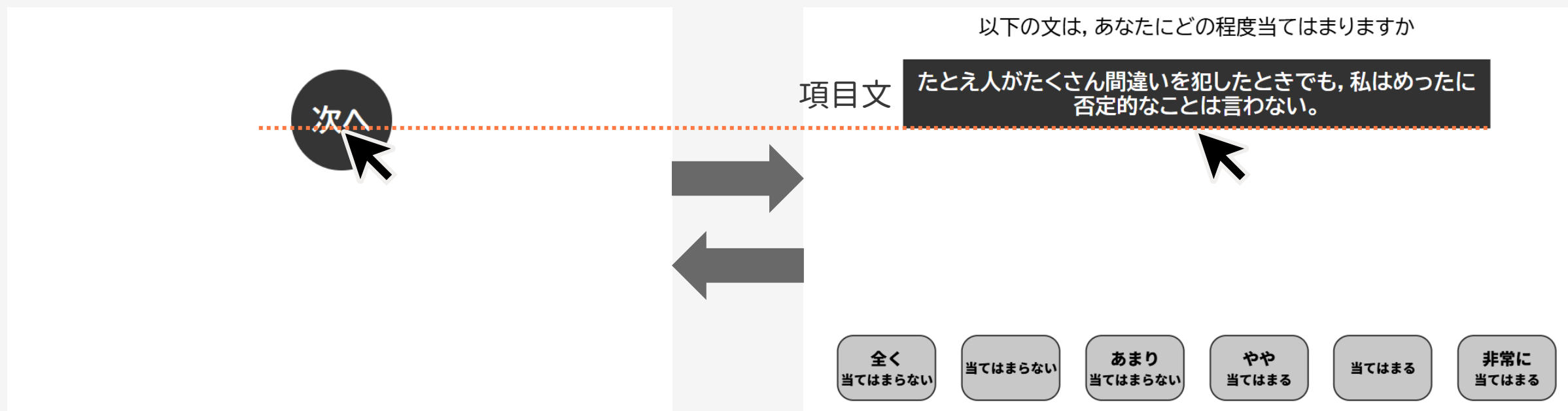
【除外基準】

- a. マウス以外のデバイス(トラックボール・トラックパッドなど)を使用した人
- b. RTが1秒未満または30秒以上の項目が1つでもあった人
- c. システムの問題でカーソル軌跡が正しく記録されなかった人
- d. 回答中にブラウザからのフォーカスが外れた(例:別のウィンドウを開いた)人
- e. 「回答エリア(後ほど説明)」からカーソルが大きく外れた人

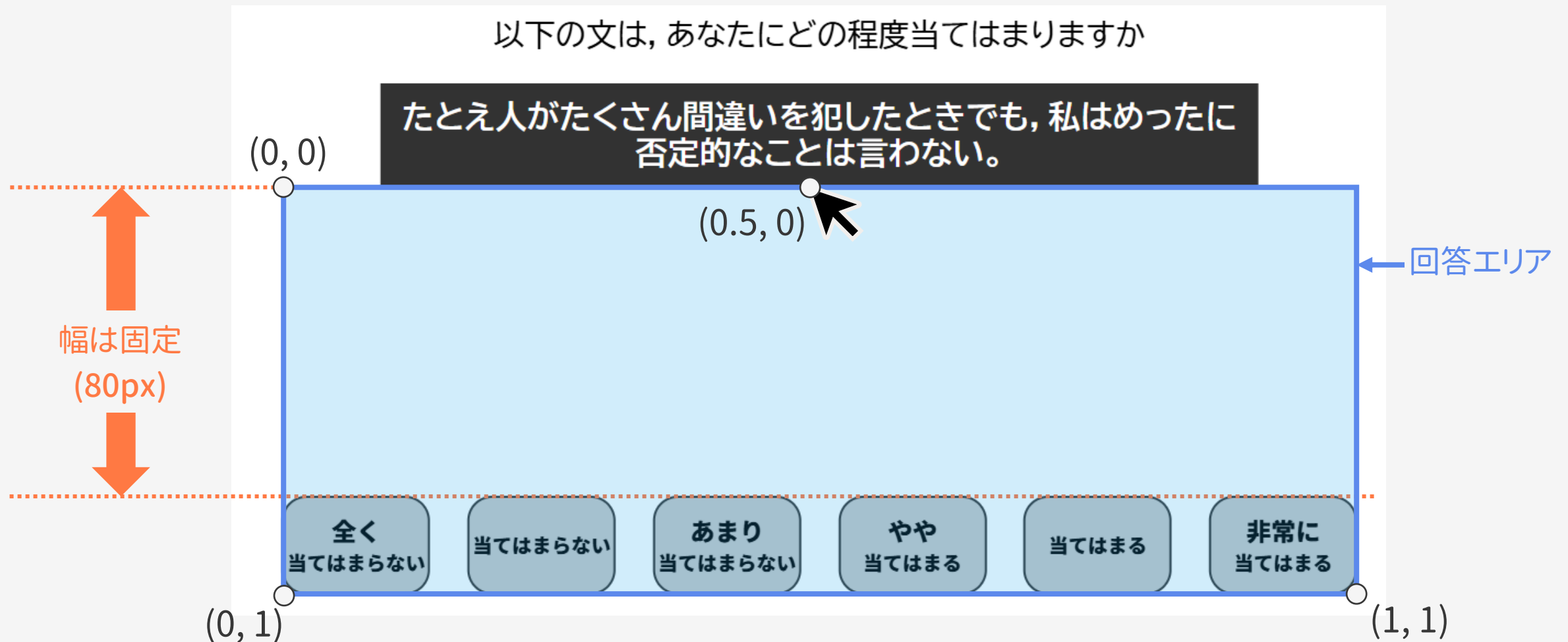
カーソルが円の中で停止すると、

(注視点のような役割)

項目と選択肢が表示される



■ 回答エリア内で 0-1 に標準化した



特徴量は出处をもとに「回答内容」「RT」「カーソル」の3タイプに分類

タイプ	名前	説明
回答内容	response	各項目への回答内容 (リッカート尺度: 0-5)
	centered response	選択肢の中央度 ($\{0, 5\} \rightarrow 0$; ($\{1, 4\} \rightarrow 1$; ($\{2, 3\} \rightarrow 2$)
	fscore	カテゴリカルCFA(1因子)での因子得点
	fscore2	因子得点の2乗
RT	rt	各項目に対するRT
	rt_sd	回答者内でのRTのSD(ばらつき)
カーソル	MAD	始点と終点を結んだ直線からの距離の最大値
	AUC	始点と終点を結んだ直線と実際の軌跡に囲まれた面積
	xpos_flips	X(左右)方向での方向転換の回数
	total_dist	カーソルの総移動距離
	vel_max	カーソルの最高速度
	acc_max	カーソルの最大加速度
	idle_time	カーソルがY(上下)方向に動き始めるまでの時間
	curvature	total_distと始点と終点を結んだ直線の比
	start_angle	最初にカーソルが動いた方向(角度)

赤字の特徴量は

- 回答者内で標準化した値(std)と生の値を両方使用
- 各バージョンにPCAを実施後第1-3主成分得点のみを保持

(補足)カーソル特徴量の定義について

- およそ 20-30 ミリ秒ごとに, JavaScriptの pointermove イベントを記録した

以下の文は, あなたにどの程度当てはまりますか

たとえ人がたくさん間違いを犯したときでも, 私はめったに否定的なことは言わない。

idle time: Y座標が0.1を最初に超えた時間

start_angle: 1つ目の座標と3つ目の座標で作られる角度

AUC

MAD

xpos_flip = 2

全く
当てはまらない

当てはまらない

あまり
当てはまらない

やや
当てはまる

当てはまる

完全に
当てはまる

vel_max =

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

■ 単一の項目への反応を機械学習モデルによって予測

目的変数: (a) 回答内容, (b) 選択肢の中央度, (c) RT

モデル: XGBoost

Elastic net, Random forest, SVM も試しましたが, 基本的には同じ結果だったので省略

評価指標: RMSE

ハイパーパラメータチューニング: 5-fold CVを用いたベイズ最適化

■ 入力特徴量

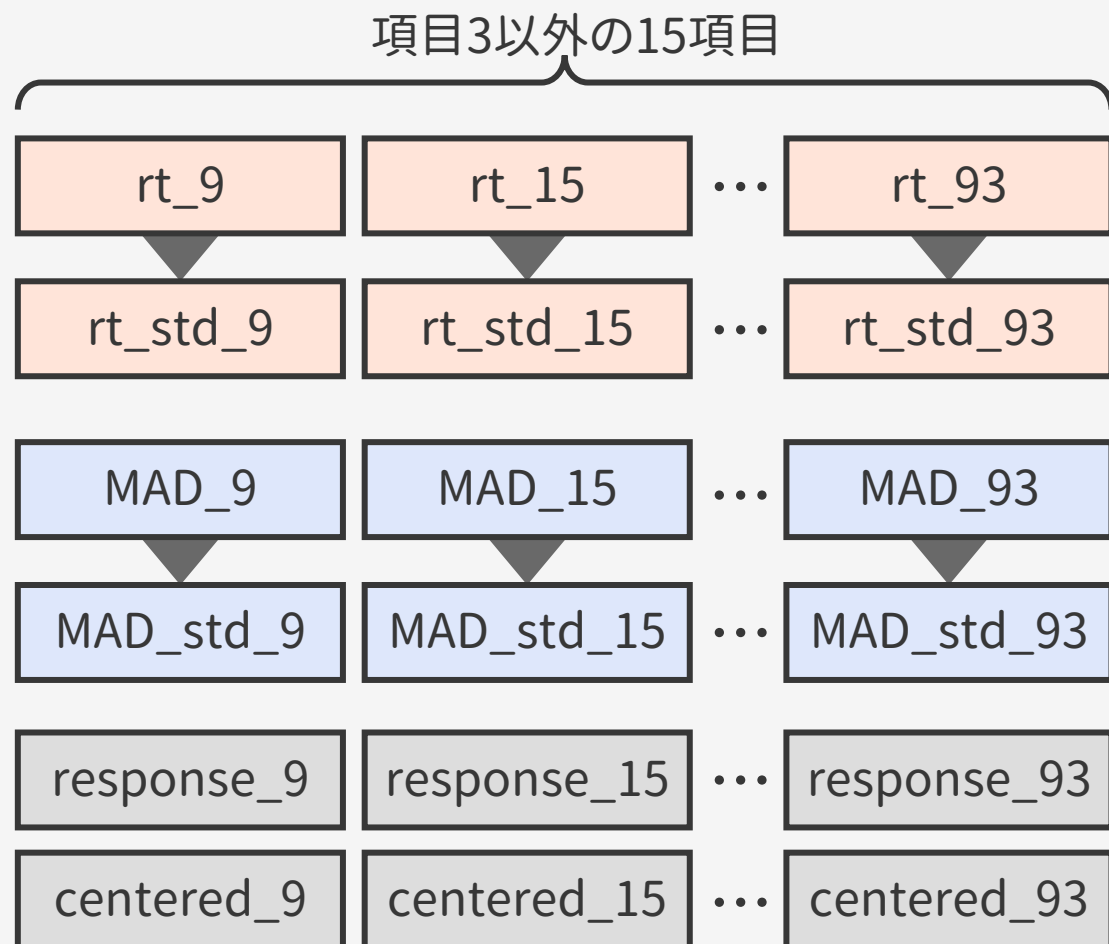
目的変数と同じ因子に属する15項目から得られた特徴量

使用する特徴量グループの全組み合わせ: 7 (+1 null) パターンで精度を比較

▲ カーソル特徴量の追加による精度の増分を評価するため

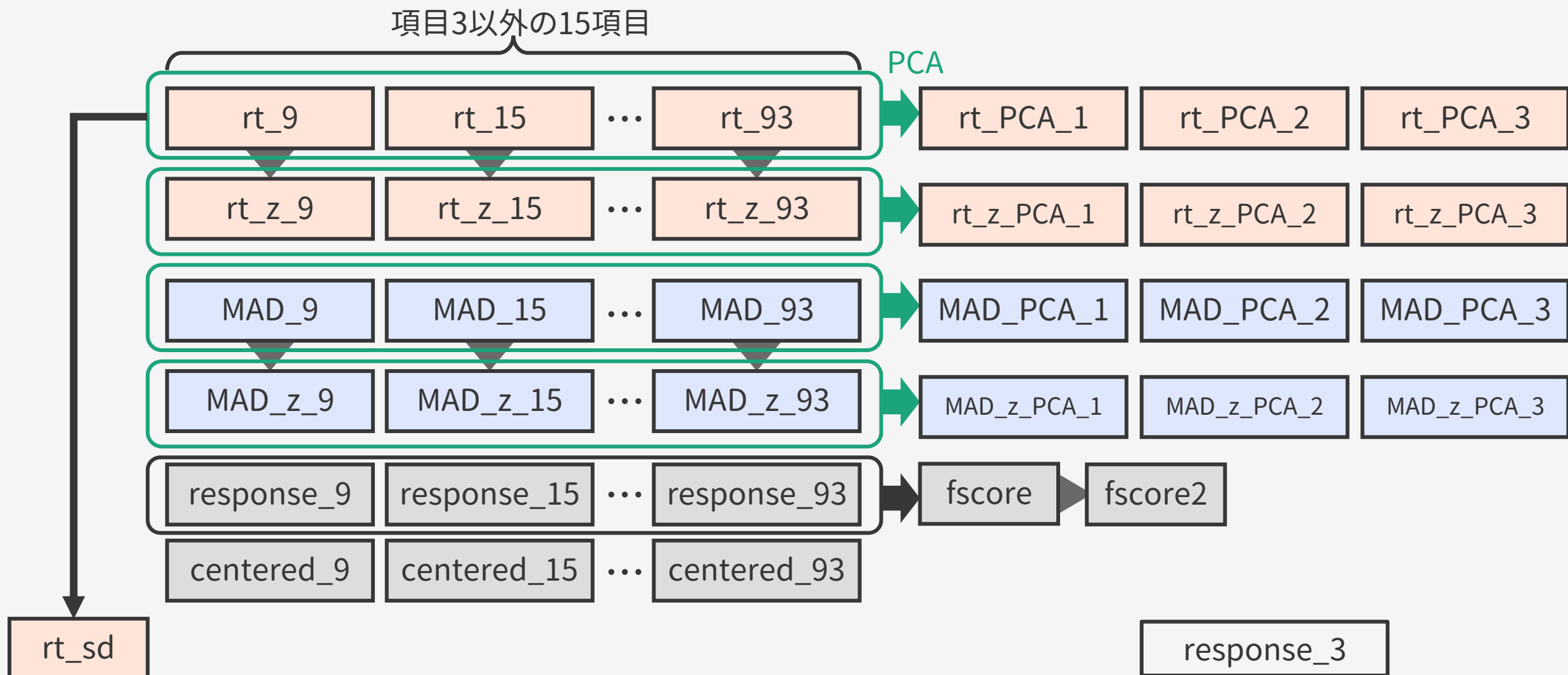
回答内容		✓		✓		✓		✓
RT			✓	✓			✓	✓
カーソル					✓	✓	✓	✓
(null)					(full)			

■ (例) HEXACOのAgrの項目3の回答内容を予測するタスクの場合



response_3

■ (例) HEXACOのAgrの項目3の回答内容を予測するタスクの場合



■ (例) HEXACOのAgrの項目3の回答内容を予測するタスクの場合

response_9 response_15 ... response_93
centered_9 centered_15 ... centered_93

rt_sd

rt_PCA_1

rt_PCA_2

rt_PCA_3

rt_z_PCA_1

rt_z_PCA_2

rt_z_PCA_3

MAD_PCA_1

MAD_PCA_2

MAD_PCA_3

MAD_z_PCA_1

MAD_z_PCA_2

MAD_z_PCA_3

fscore

fscore2



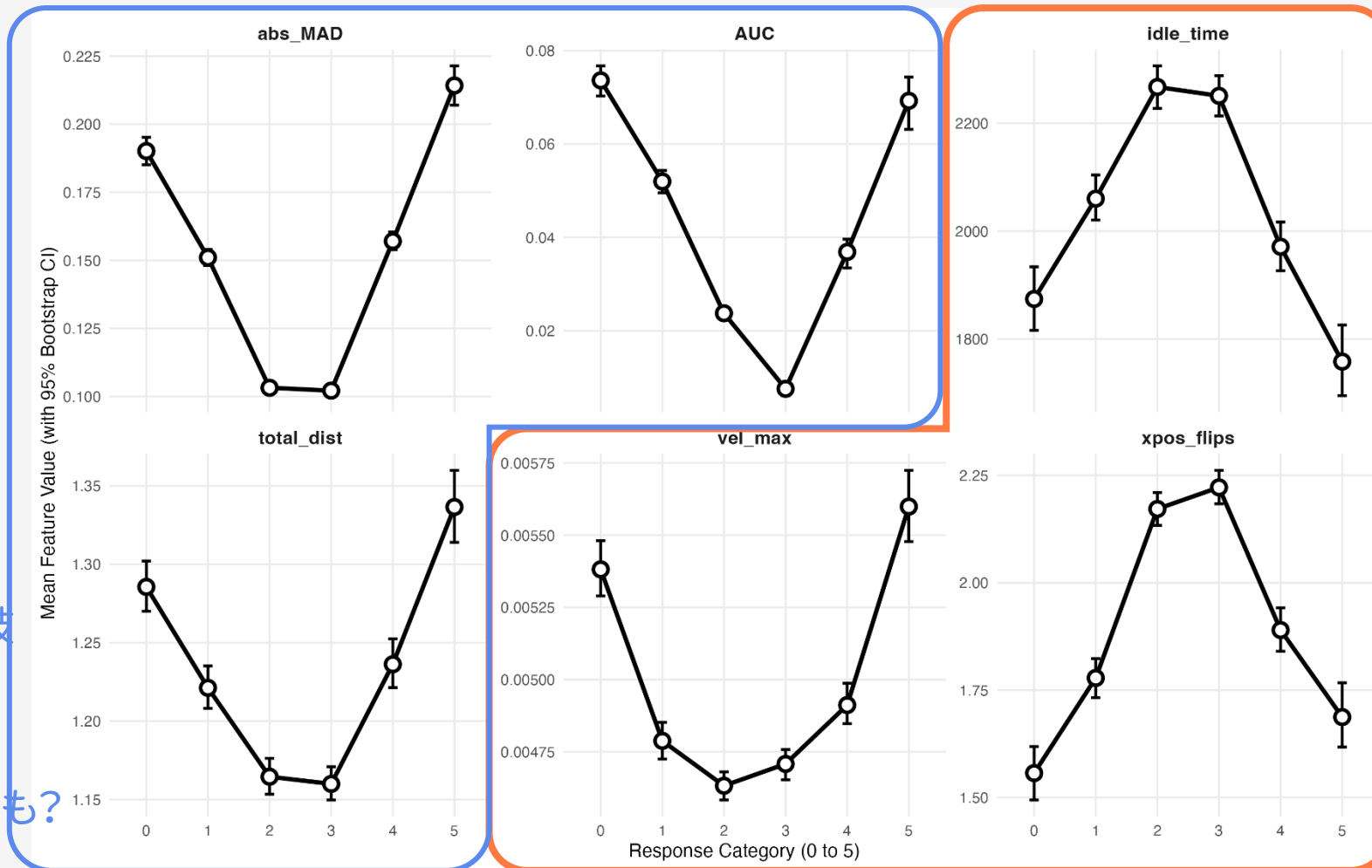
すべて投入

response_3

3

結果

実際に選択したカテゴリに応じて記述的な平均値は大きく変動している



RTで見られる現象(逆U字)とよく似た傾向が観測された

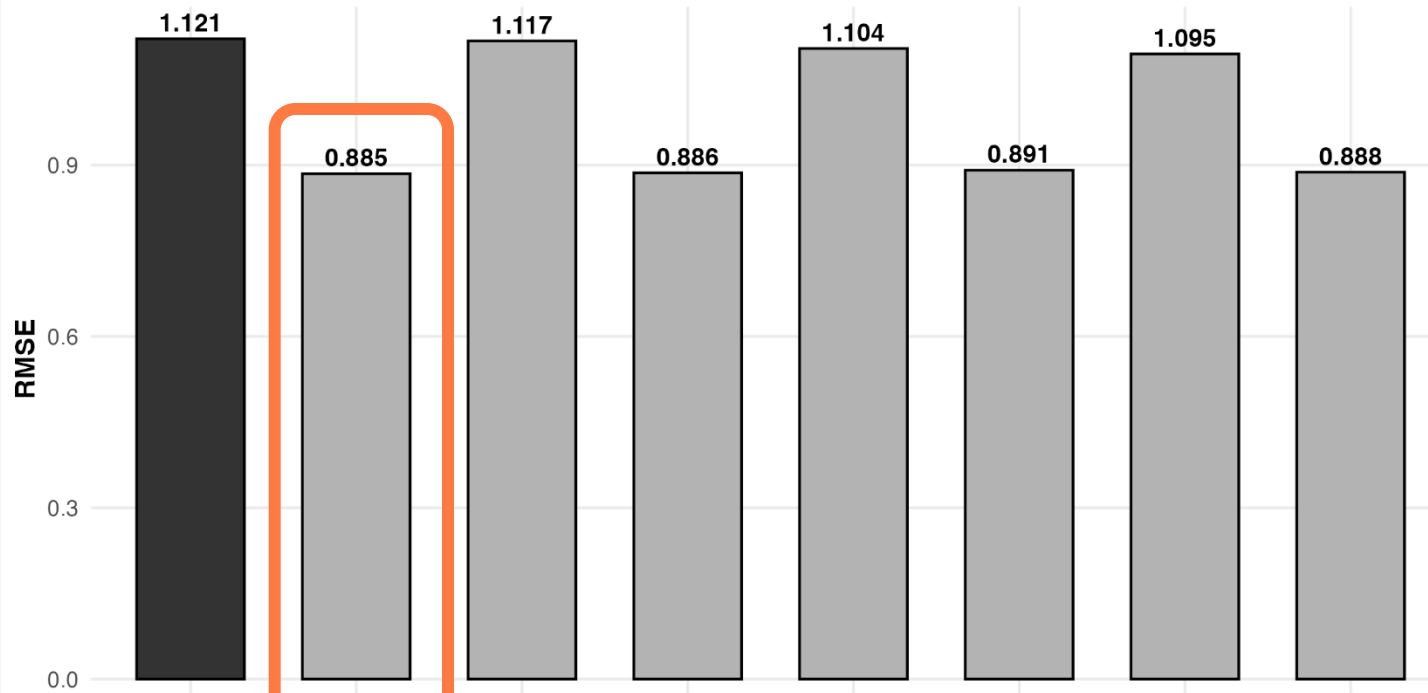


回答の「迷い」を反映している特徴量と言えるかも

これらは単純に始点から選択肢ボタンまでの距離の違いを反映しているかも?

RMSE across Feature Sets (Outcome: response)

Baseline is calculated against the Item-specific Null Model (Model: XGBoost)

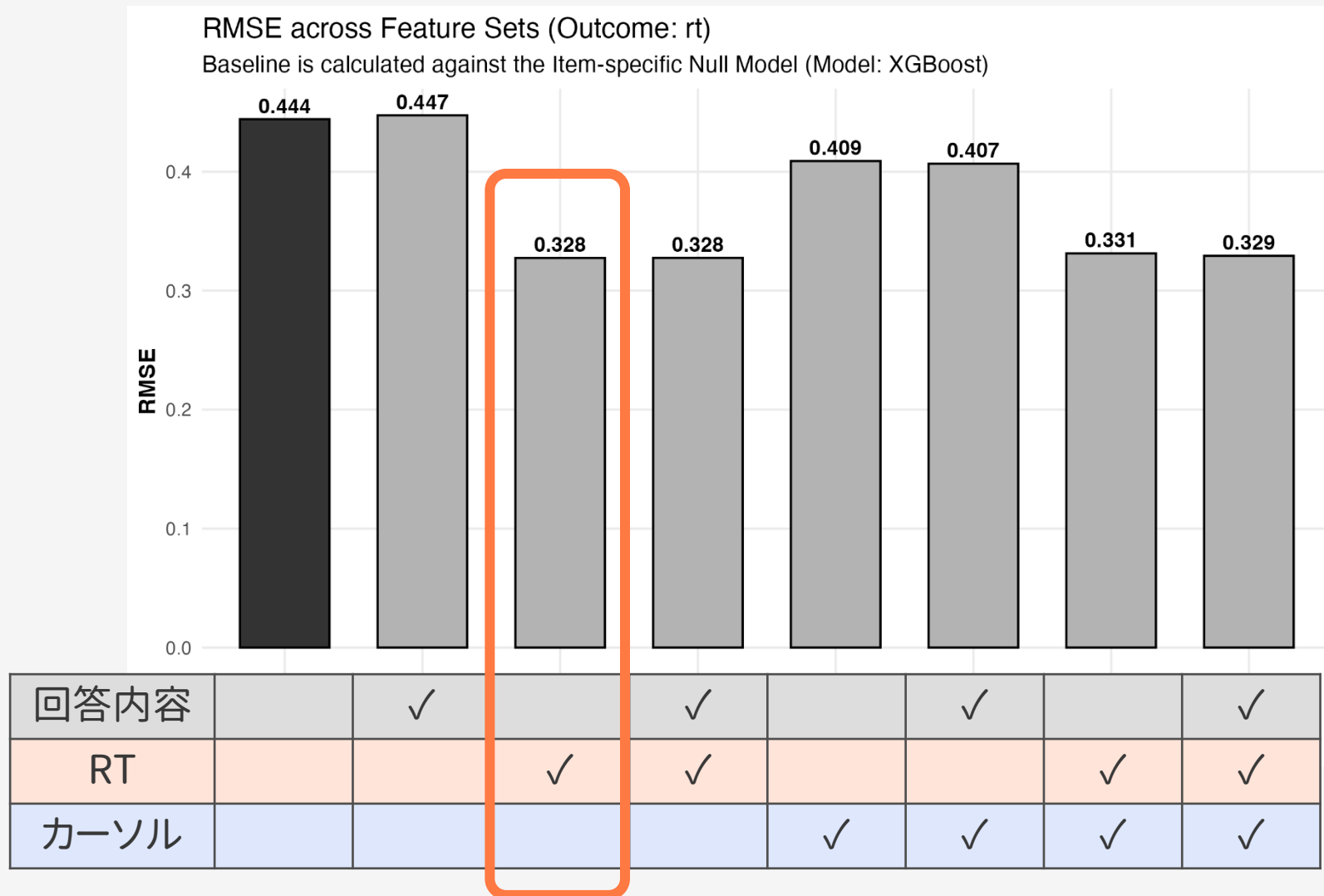


回答内容		✓		✓		✓		✓
RT			✓	✓			✓	✓
カーソル					✓	✓	✓	✓

- 一番は「**回答内容のみ**」モデル
- RTやカーソル特徴量を追加しても改善には繋がらなかった
- 回答内容の特徴量が無い場合はRTよりもカーソル特徴量のほうがわずかに精度を向上させる

目的変数が回答の中心度の場合も
ほぼ同じ結果でした

予測という意味では回答内容の特徴量があれば十分と言えそう



- 一番は「RTのみ」のモデル
- ここでもカーソル特徴量は精度の改善に寄与しなかった
- RT特徴量が無い場合は回答内容よりもカーソル特徴量のほうが精度を向上させる

予測という意味ではRT特徴量があれば十分と言えそう

■ 各特徴量の寄与は投入順によって大きく変わってしまう

▶ シャープレイ値を算出することで, 順序によらないフェアな比較を試みる

	目的変数		
グループ	回答内容	回答の中心度	RT
回答内容	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%)
カーソル	0.0152 (4.08%)	0.0109 (4.21%)	0.0779 (17.23%)

■ 回答内容を予測する場合

回答内容の特徴量が寄与の大半 (over 93%) を占めていた

RT特徴量よりはカーソル特徴量のほうが寄与していた

■ RTを予測する場合

やはりRT特徴量が大半を締めているが, カーソル特徴量も一定 (~17%) の貢献

グループレベルのシャープレイ値の計算方法

【例】カーソル特徴量のシャープレイ値を求める場合

	モデルへの投入順			
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	$v(\text{resp}, \text{RT}, \text{mouse}) - v(\text{RT}, \text{mouse})$

▲ この6つの値の平均値がシャープレイ値

$$\begin{aligned} \text{Shaplay}(\text{resp}) &= \frac{1}{3} [v(\text{resp}) - v(\phi)] + \frac{1}{3} [v(\text{resp}, \text{RT}, \text{mouse}) - v(\text{RT}, \text{mouse})] \\ &\quad + \frac{1}{6} [v(\text{resp}, \text{mouse}) - v(\text{mouse})] + \frac{1}{6} [v(\text{resp}, \text{RT}) - v(\text{RT})] \end{aligned}$$

(↑Lundberg and Lee, 2017)

全32項目でのSHAP値の平均値トップ10

回答内容	✓		✓		✓		✓
RT		✓	✓			✓	✓
カーソル				✓	✓	✓	✓
1	fscore	rt_z_pca1	fscore	AUC_sd	fscore	AUC_sd	fscore
2	response_4	rt_z_pca2	response_4	MAD_sd	fscore2	MAD_sd	response_25
3	centered_22	rt_pca2	centered_22	AUC_pca2	centered_22	MAD_pca1	fs2
4	centered_15	rt_pca1	response_25	MAD_pca1	response_4	AUC_pca2	response_4
5	response_25	rt_z_pca3	centered_20	AUC_z_pca3	response_9	rt_z_pca1	response_9
6	response_9	rt_sd	fscore2	start_angle_z_pca1	response_25	AUC_z_pca3	centered_22
7	fsore2		centered_1	AUC_pca1	centered_20	start_angle_z_pca1	centered_25
8	centered_1		centered_15	MAD_z_pca1	centered_2	rt_z_pca2	centered_20
9	centered_2		response_9	AUC_z_pca2	centered_25	AUC_pca1	response_1
10	centered_20		response_1	AUC_pca3	centered_1	MAD_z_pca1	centered_15

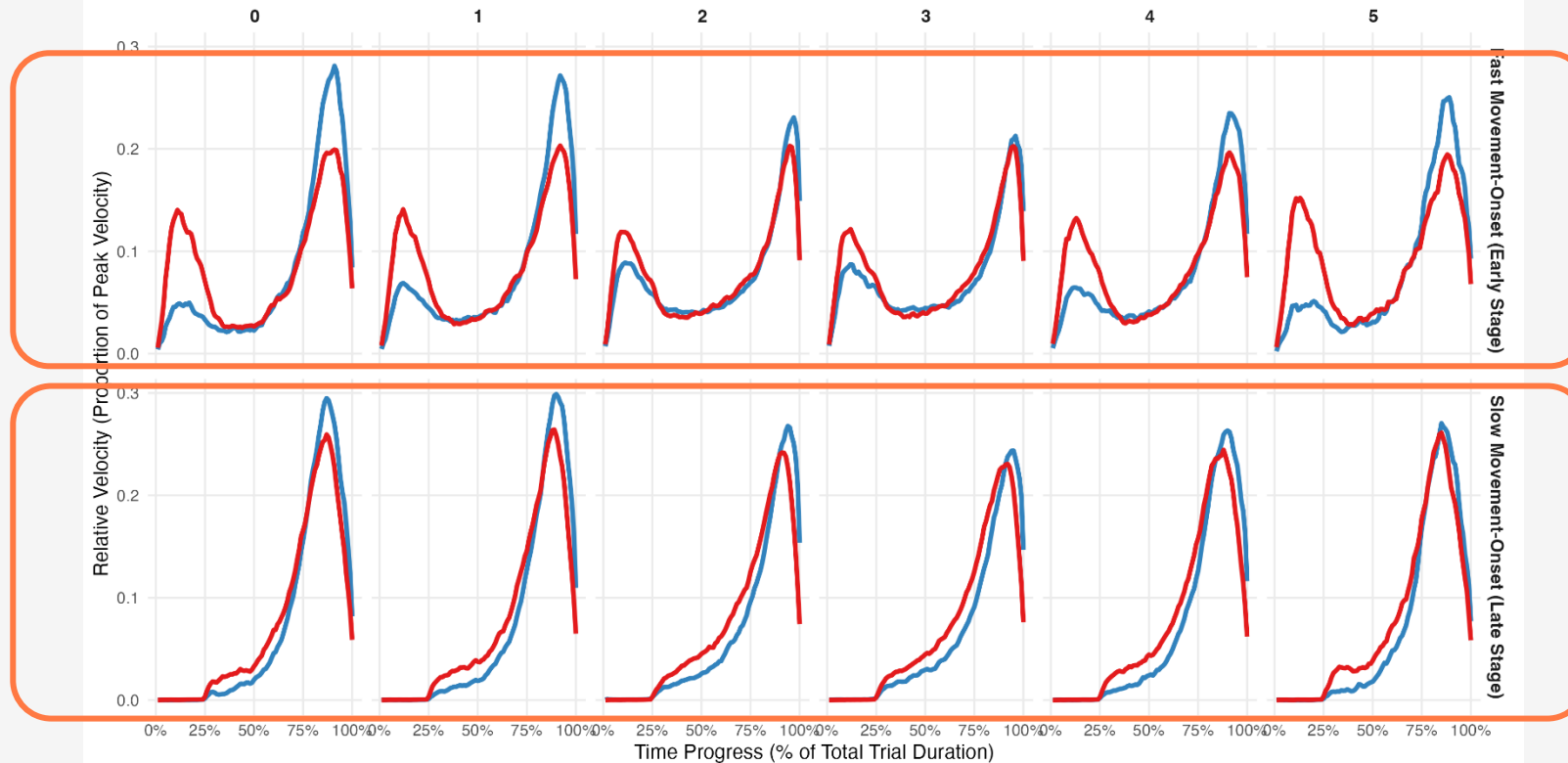
RT+カーソル特徴量のモデルでは
カーソル特徴量のほうが強く効いていた

カーソル操作の方略にパターンのようなものが見られるかもしれない

idle_timeの中央値で2グループに分けて,速度の変化を図示してみたところ…

Relative Velocity Profiles by Initiation Style

Comparison of x and y velocity patterns between early and late movement-onsets



青: X方向の速度; 赤: Y方向の速度

"Fast"グループでは,回答を決める前にとりあえず動かす
▶ 一旦止まってから選択肢に向かう

"Slow"グループでは基本的に大きな一回の移動で選択肢に直接向かっているように見える

※平均RTは"Fast"グループのほうが長かった

■ 回答およびRTの行動予測としてはカーソル軌跡はほぼ意味がない？

基本的には回答内容あるいはRT特徴量があればそれ以上の寄与は望めない
ただ、これらとは異なる情報を持っている可能性も高い(シャープレイ値より)

予測対象が回答やRTの場合には無意味, というだけの話

■ カーソル操作に異なる方略が見られた

IATなどの単純なタスクよりリッカート尺度では「思考」のプロセスが複雑(e.g., Tourangeau et al. 2000)

▶ 分析モデルにおいてこれを考慮する必要があるそう

■ 考えられる適用可能性

目的変数がより(認知プロセスに近い)複雑なものになると効果的かも？

あるいは特定の臨床的な診断には有効な可能性も高い (e.g., ADHD児診断: Pino et al., 2025)

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>
- Freeman, J. B., & Ambady, N. (2010). MouseTracker: Software for studying real-time mental processing using a computer mouse-tracking method. *Behavior Research Methods*, 42(1), 226–241. <https://doi.org/10.3758/BRM.42.1.22>
- 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.
- Longstaff, M. G., & Heath, R. A. (2003). The influence of motor system degradation on the control of handwriting movements: A dynamical systems analysis. *Human Movement Science*, 22, 91–110. [https://doi.org/10.1016/s0167-9457\(03\)00002-2](https://doi.org/10.1016/s0167-9457(03)00002-2)
- 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.
- MacLeod, C. M. (1991). Half a century of research on the stroop effect: An integrative review. *Psychological Bulletin*, 109(2), 163–203. <https://doi.org/10.1037/0033-2909.109.2.163>
- 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>
- 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>
- 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.
- Tourangeau, R., Rips, L. J., & Rasinski, K. A. (2000). *The psychology of survey response*. Cambridge University Press.
- 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>