The role of reinforcement learning in pragmatic reasoning tasks

Modeling individual differences in ACT-R

John Duff ❖ Alexandra Mayn ❖ Vera Demberg

Saarland University, Dept. of Language Science & Technology

UNIVERSITÄT DES SAARLANDES

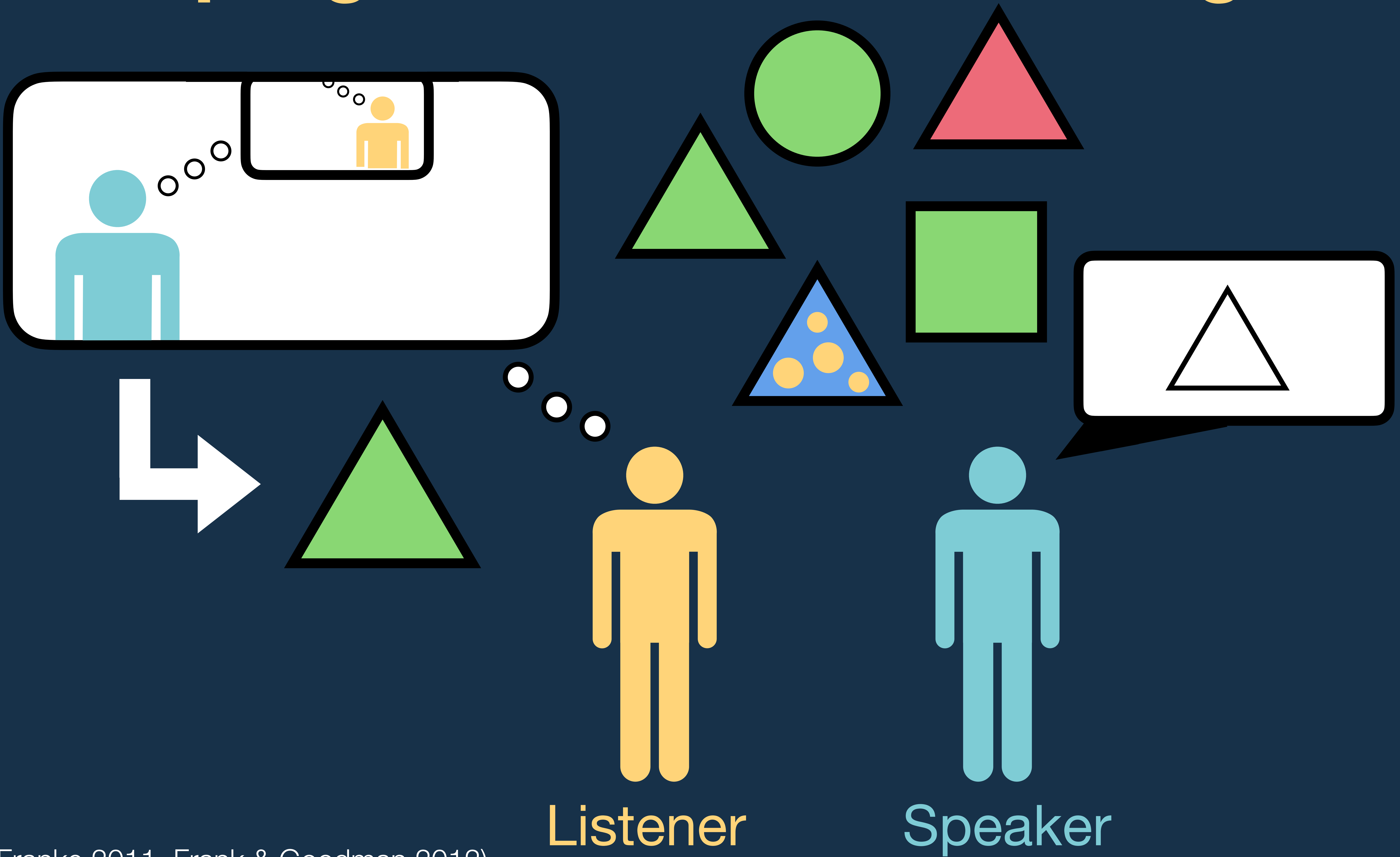
European Research Council
Established by the European Commission

XPRAG.it

27 September 2024

jduff@lst.uni-saarland.de

Gricean pragmatics in a reference game

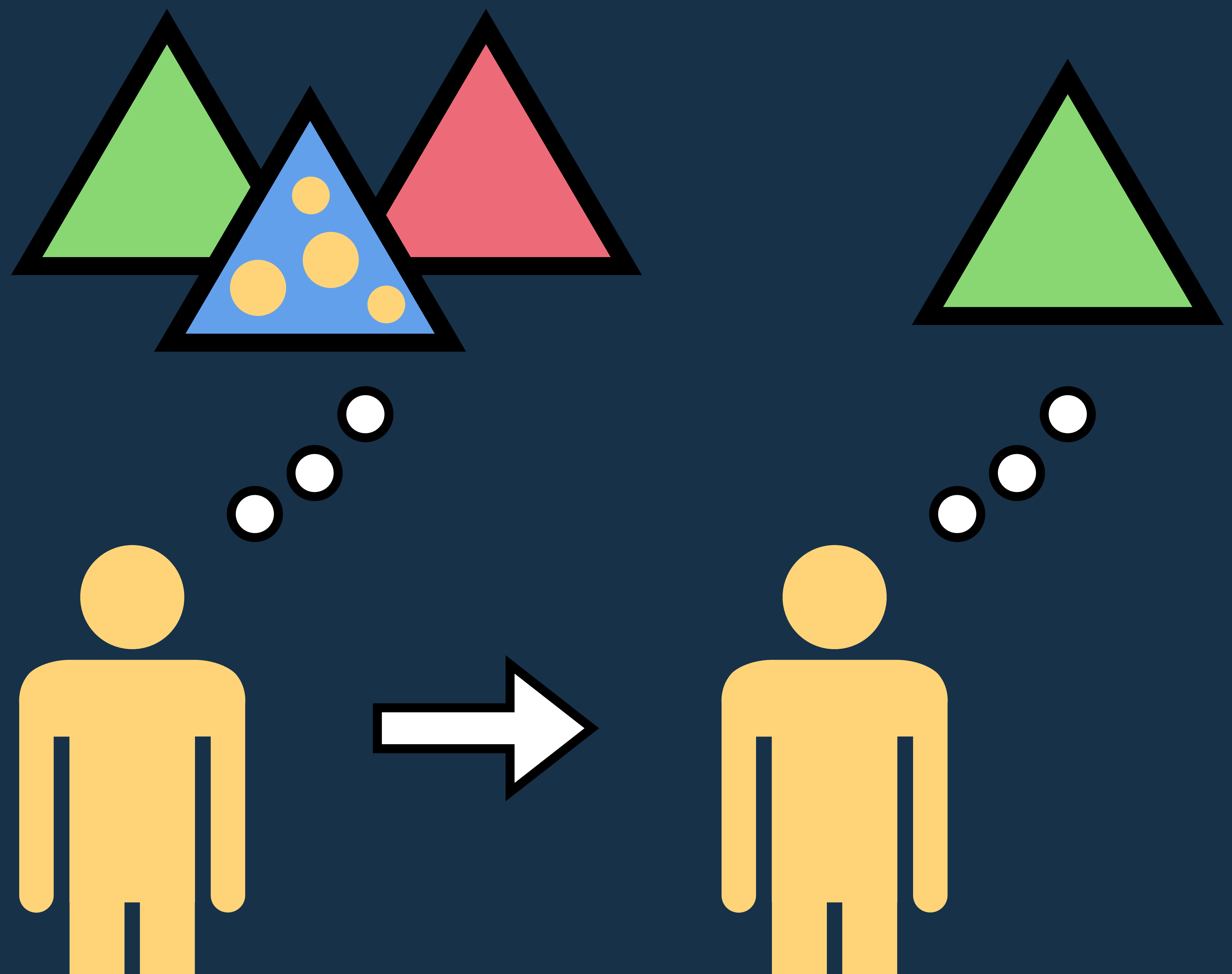


(Grice 1975, Franke 2011, Frank & Goodman 2012)

Two empirical complications

Pragmatic reasoning in games only emerges over time

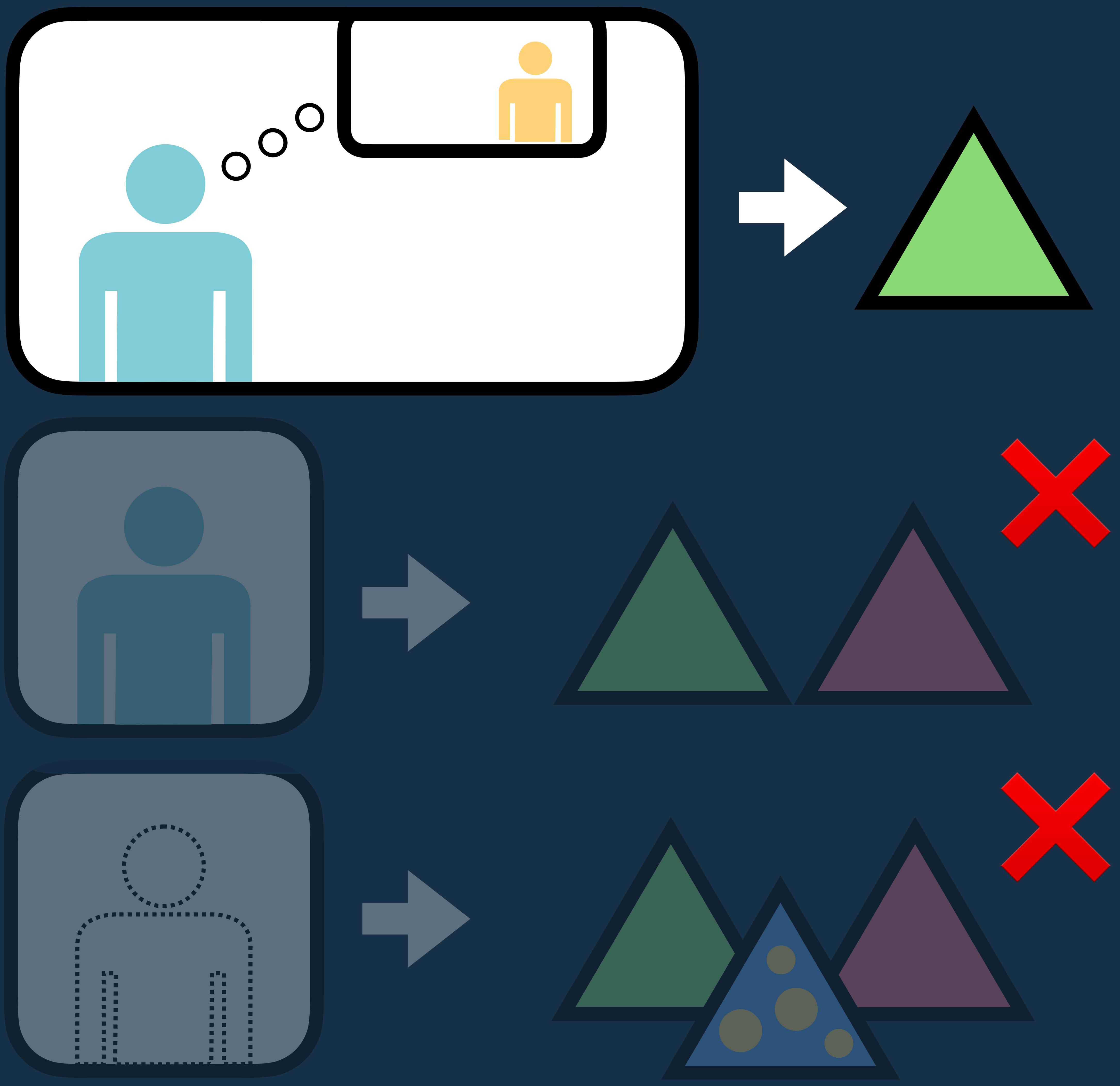
(Sikos et al. 2021)



Individuals vary in their depth of pragmatic reasoning

(Franke & Degen 2016, Mayn & Demberg 2023)

Modeling performance via reinforcement learning



Comprehenders find an optimal strategy through exploration and failure

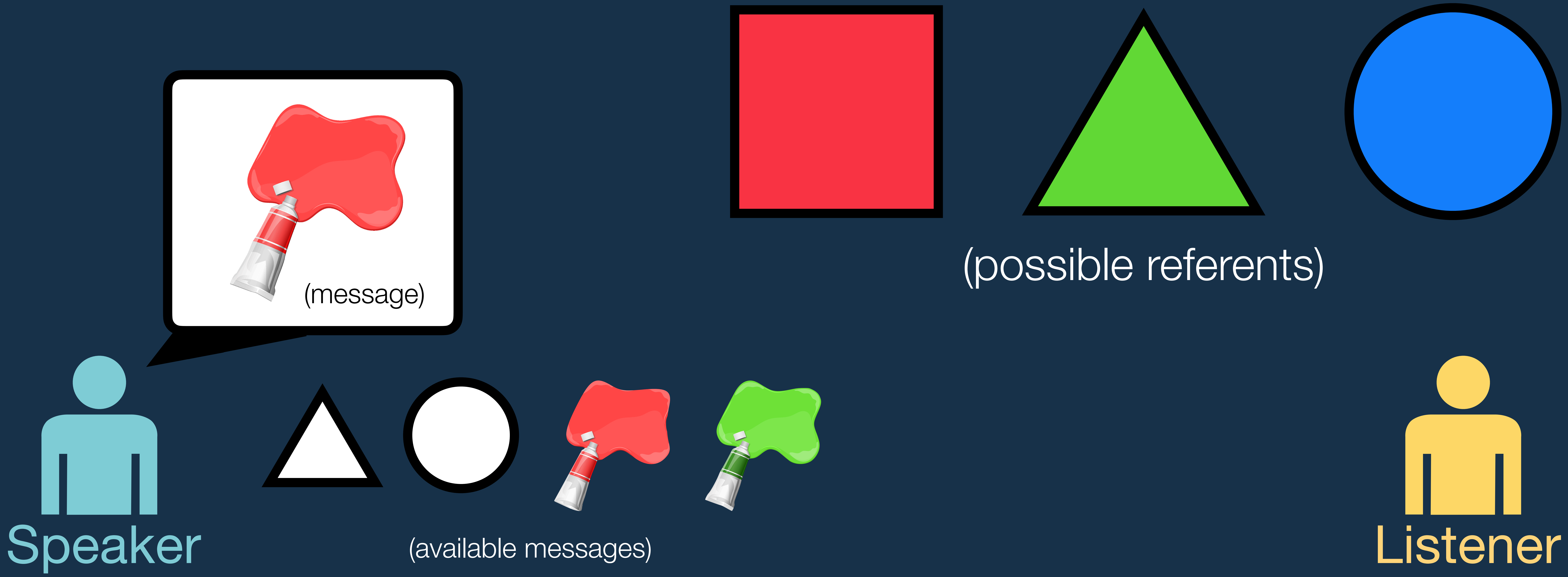
(cf. Stocco et al. 2021)

Roadmap

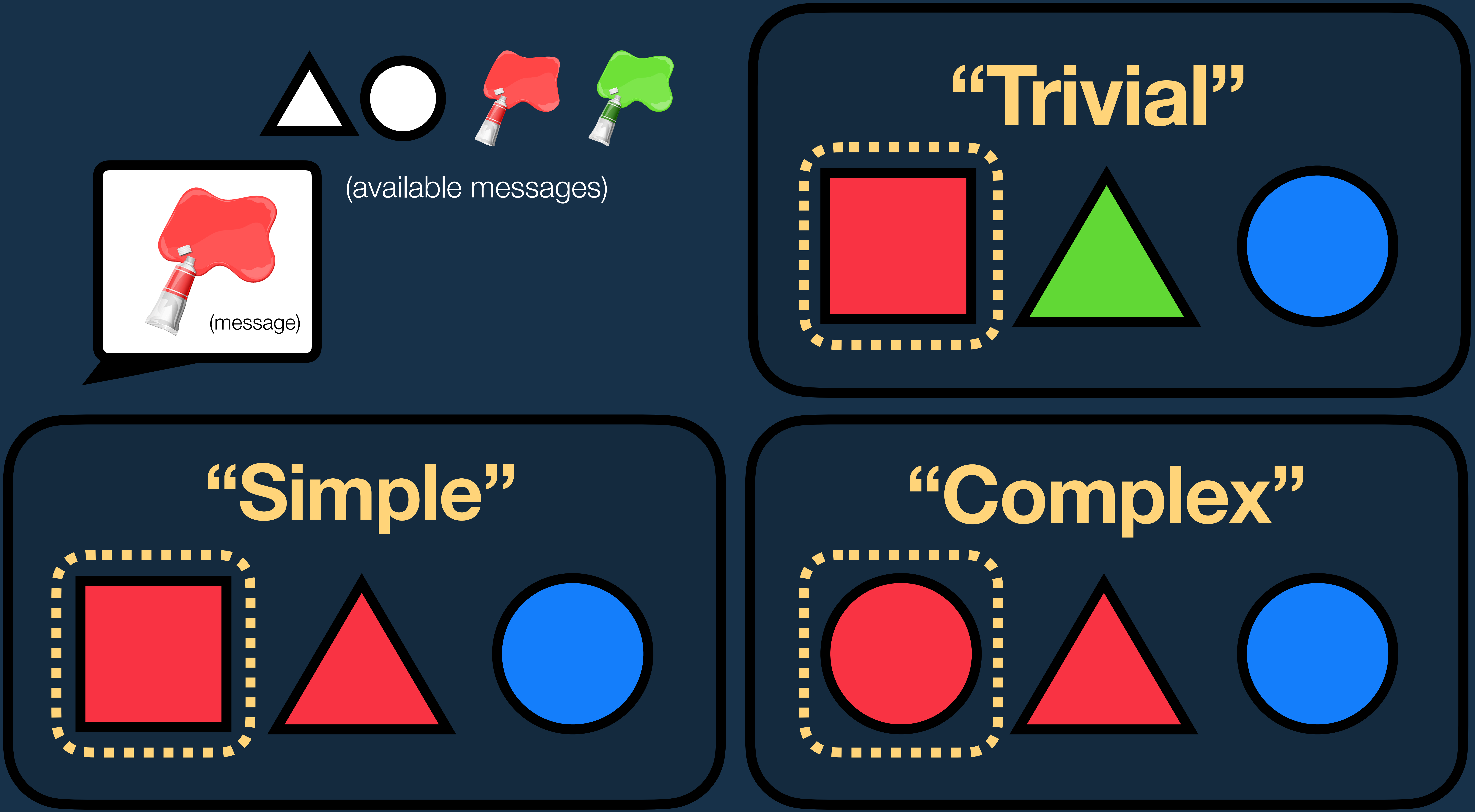
1. Background
2. Our ACT-R model
3. Validating the role of learning resources

The reference game task (RefGame)

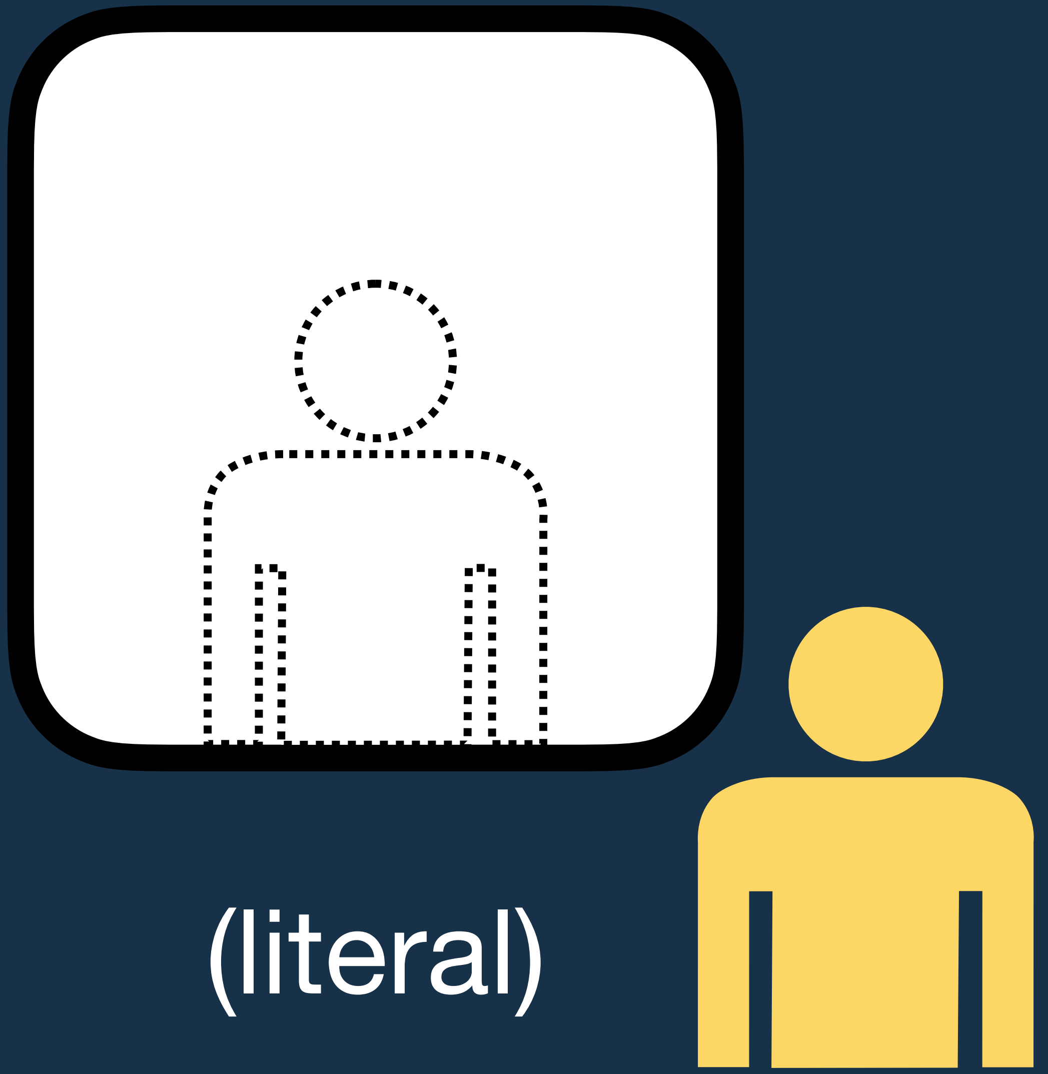
(Frank & Goodman 2012 and following; cf. Wittgenstein 1953)



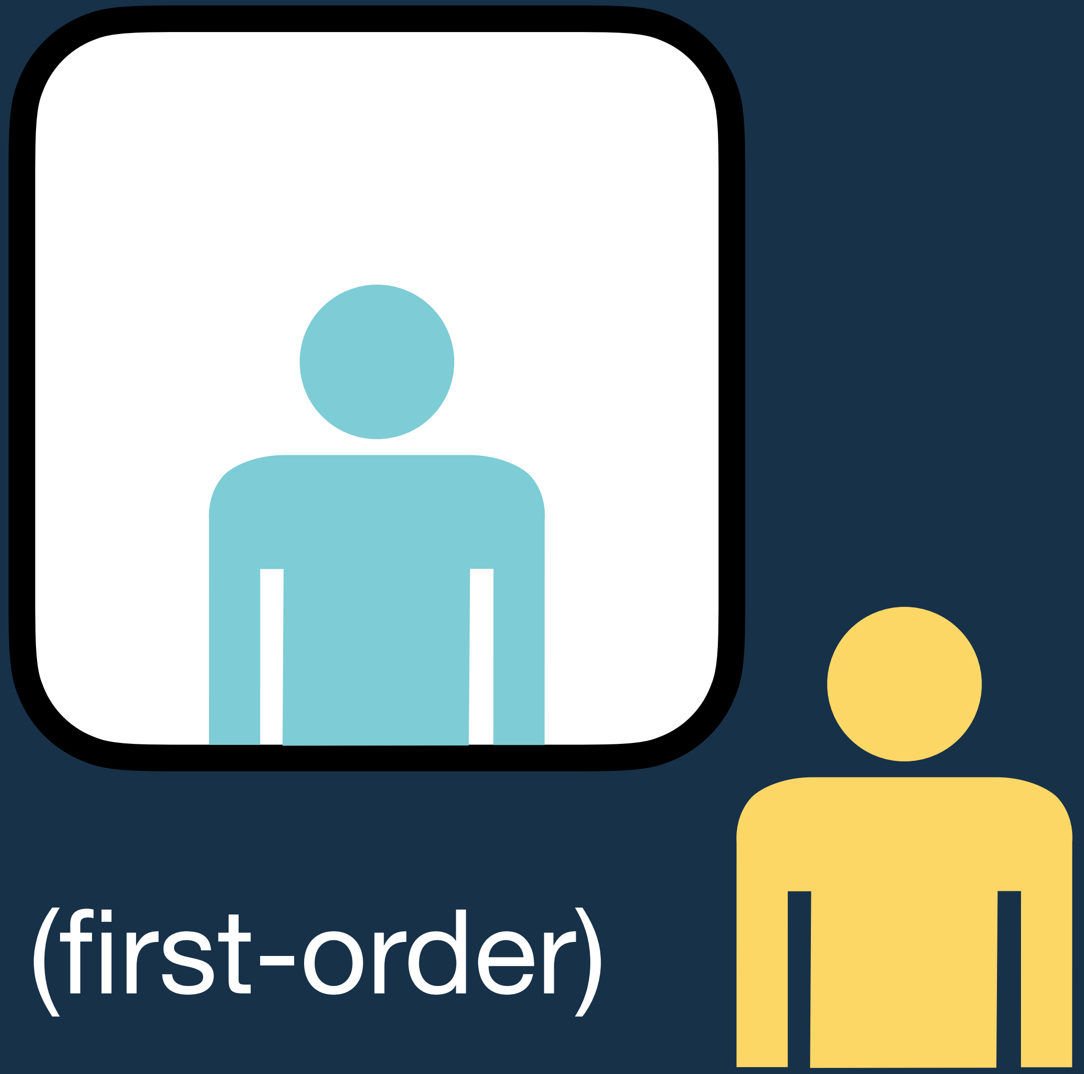
Three RefGame conditions (Franke & Degen 2016)



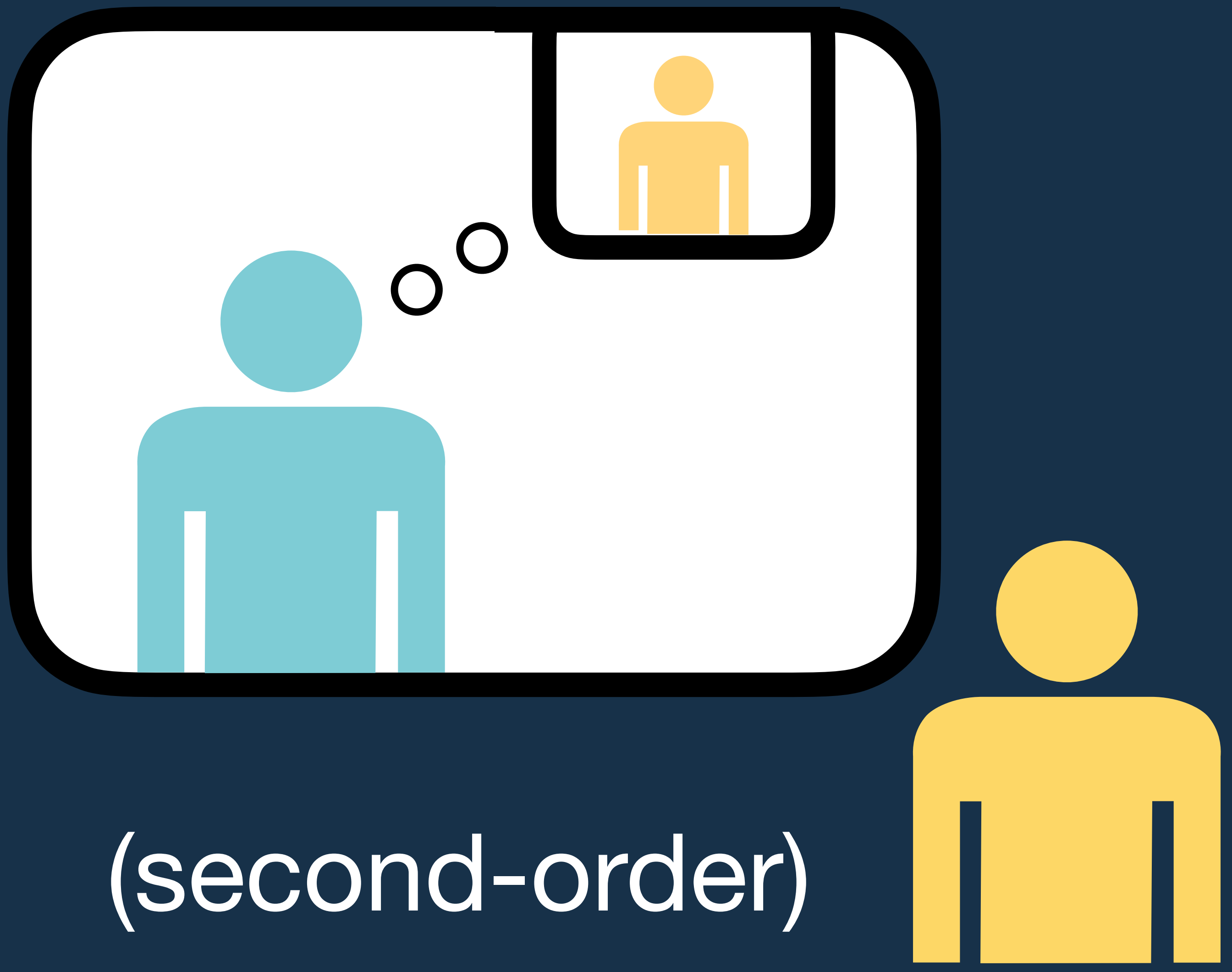
Expected success by strategy (Franke & Degen 2016)



(literal)

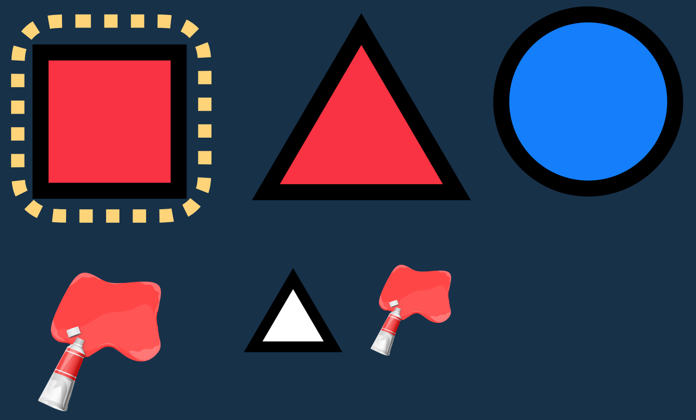


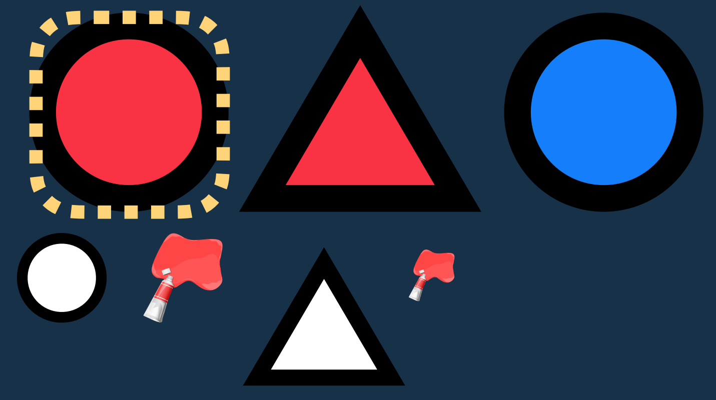
(first-order)



(second-order)

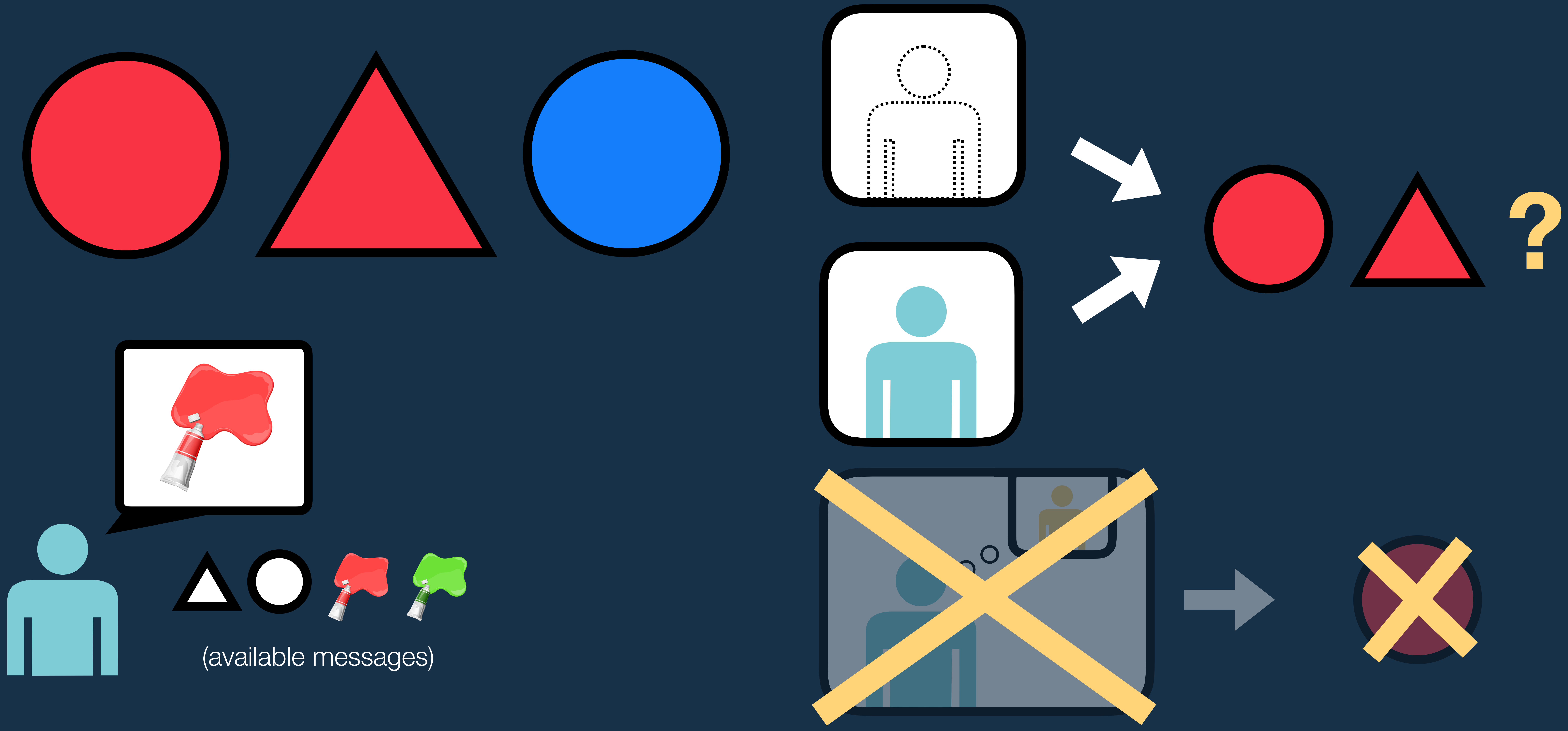
	(literal)	(first-order)	(second-order)
Picks:	Matching referents	Matching referents with fewest alternative messages	Matching referents with no more-informative messages
Trivial:	✅	✅	✅
Simple:	–	✅	✅
Complex:	–	–	✅



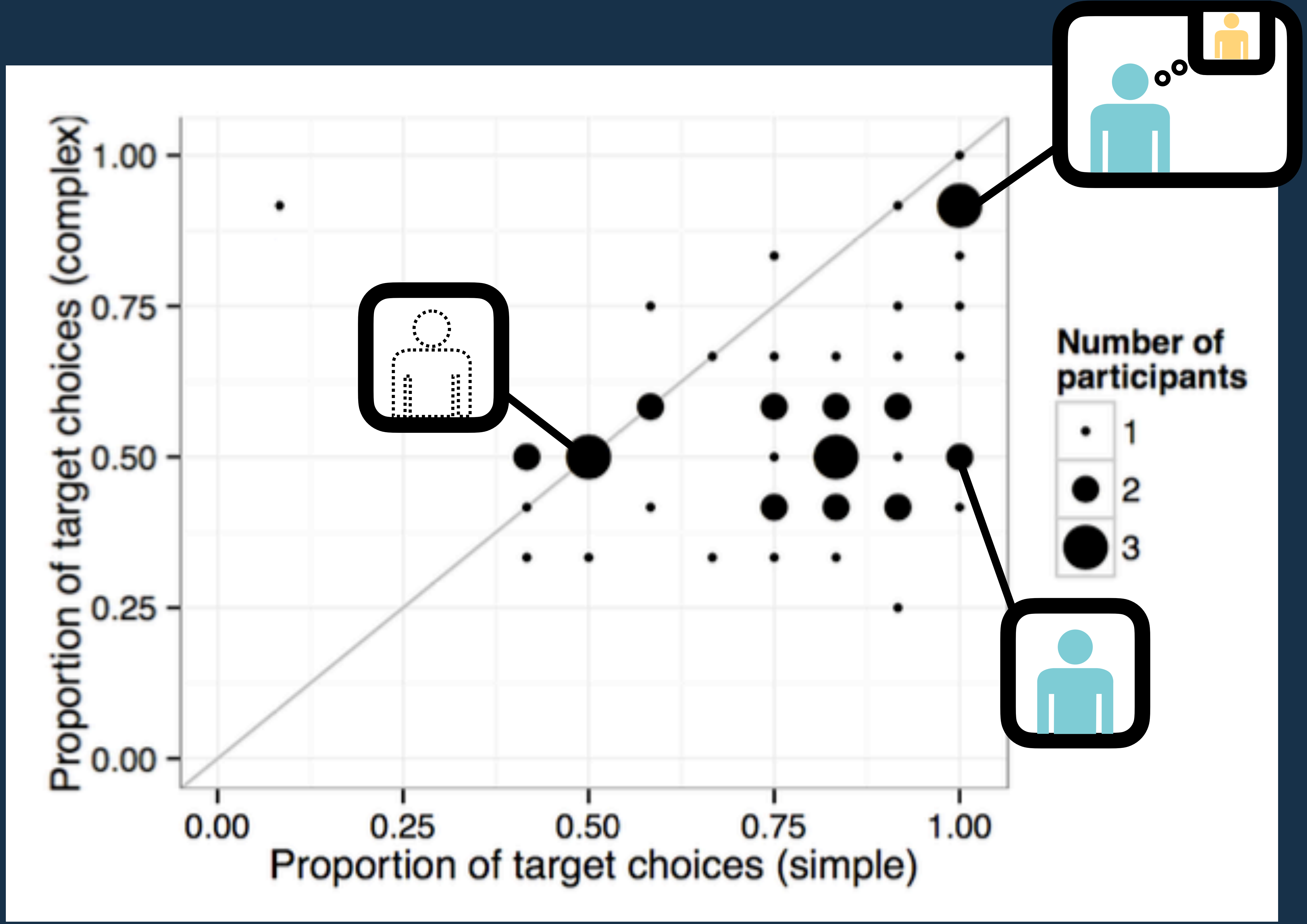


Observation #1: No second-order reasoning in one-shot experiments

Sikos et al. (2021)

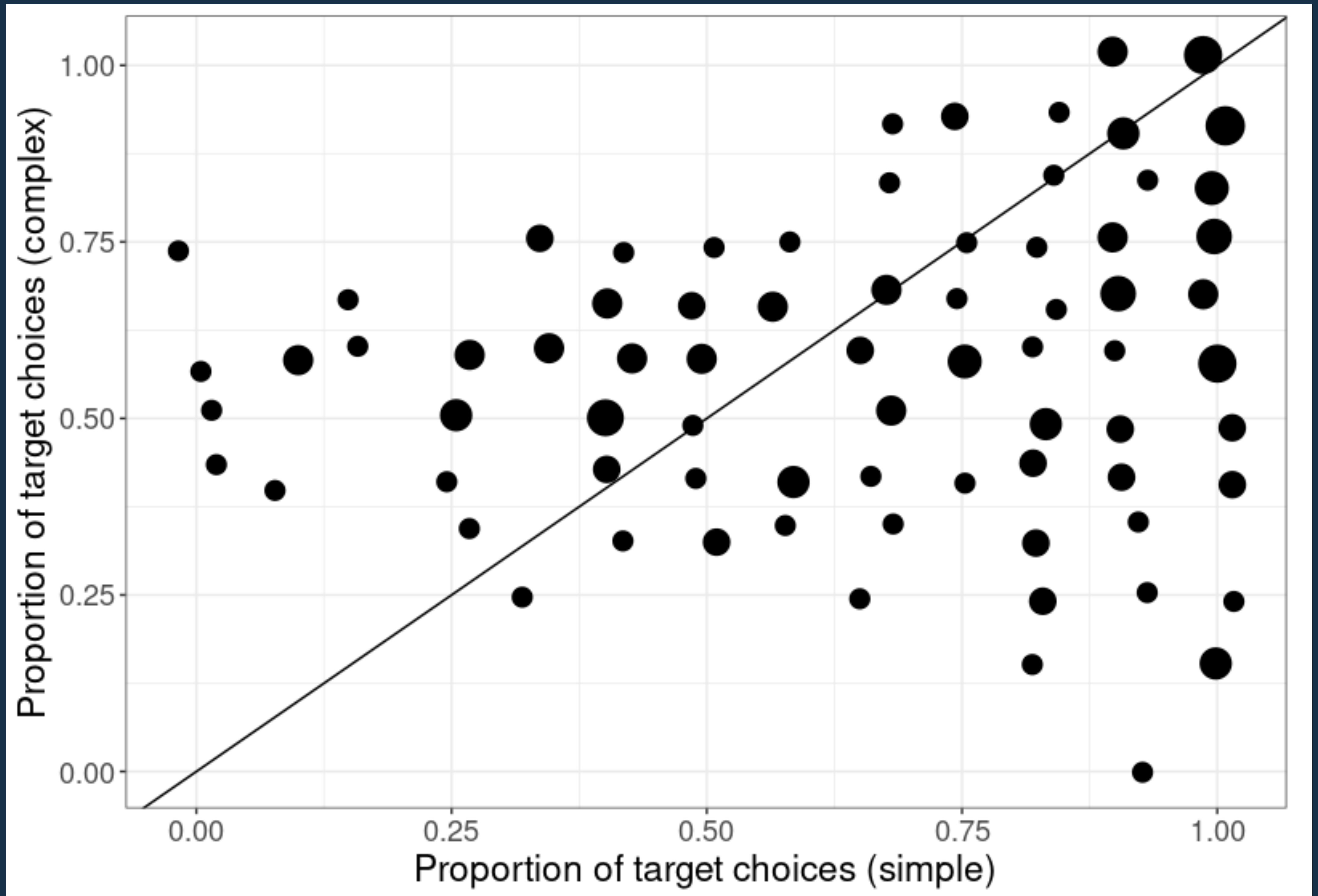


Observation #2: Individual differences in many-shot performance



Franke & Degen (2016)

(n = 60, 12 obs/condition)



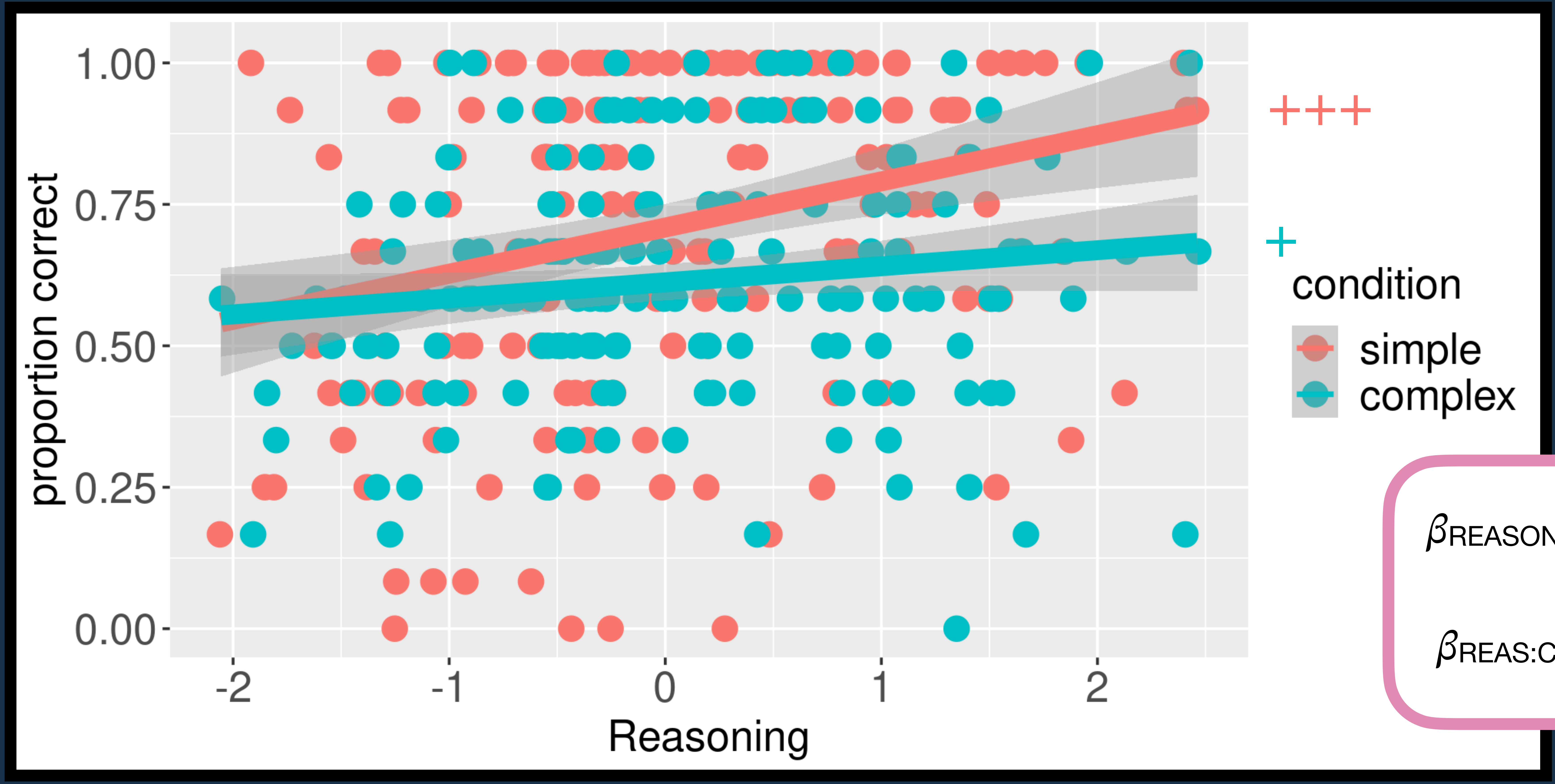
Mayn & Demberg (2023)

(n = 173, 12 obs/condition)

(debiased stimuli, cf. Mayn 2023)

Unexpected covariate: Reasoning performance

:= Raven’s Matrices + Cognitive Reflection Task



$\beta_{\text{REASONING}} = (0.00, 0.31)_{95\%}$

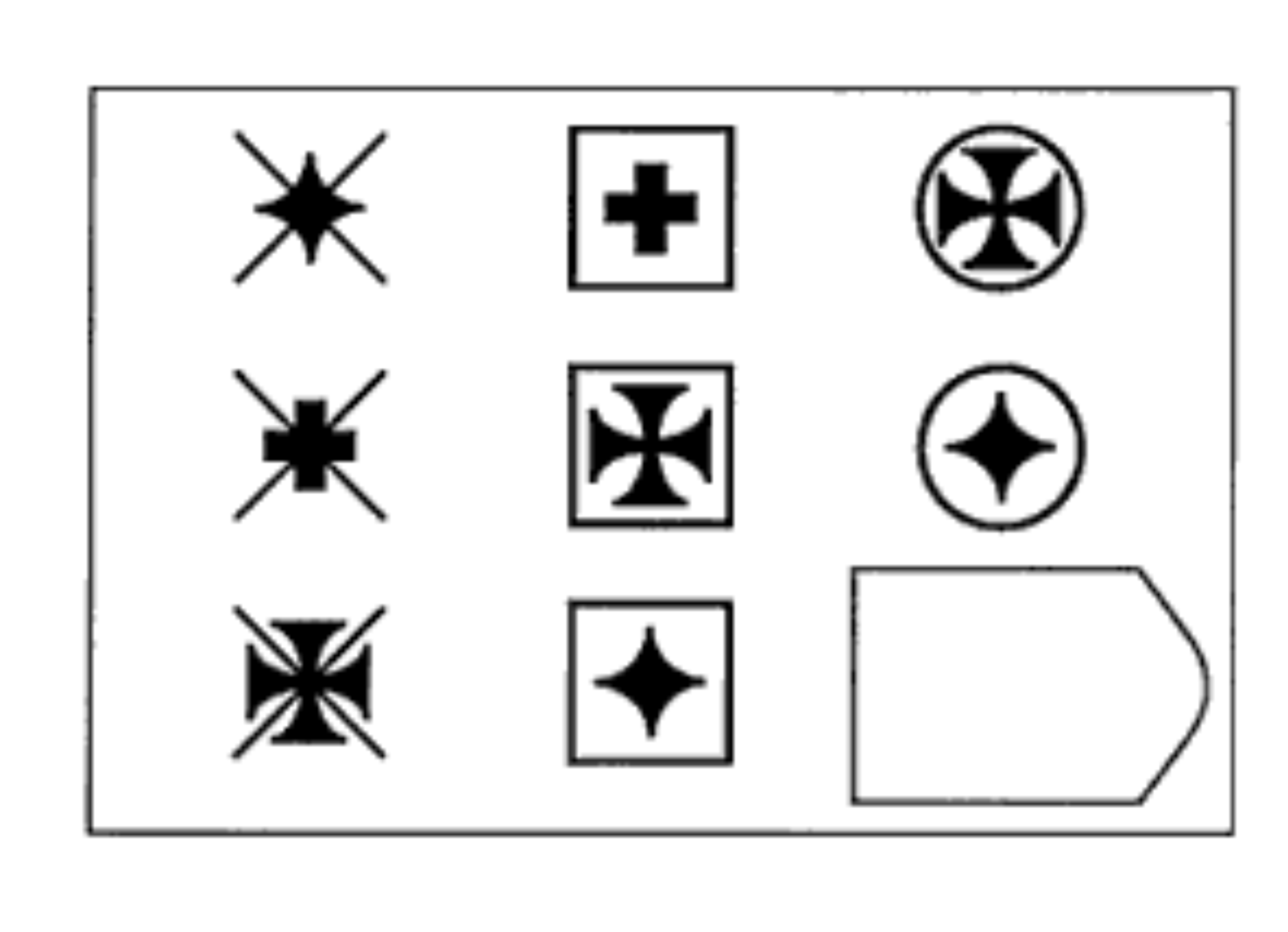
qualified by

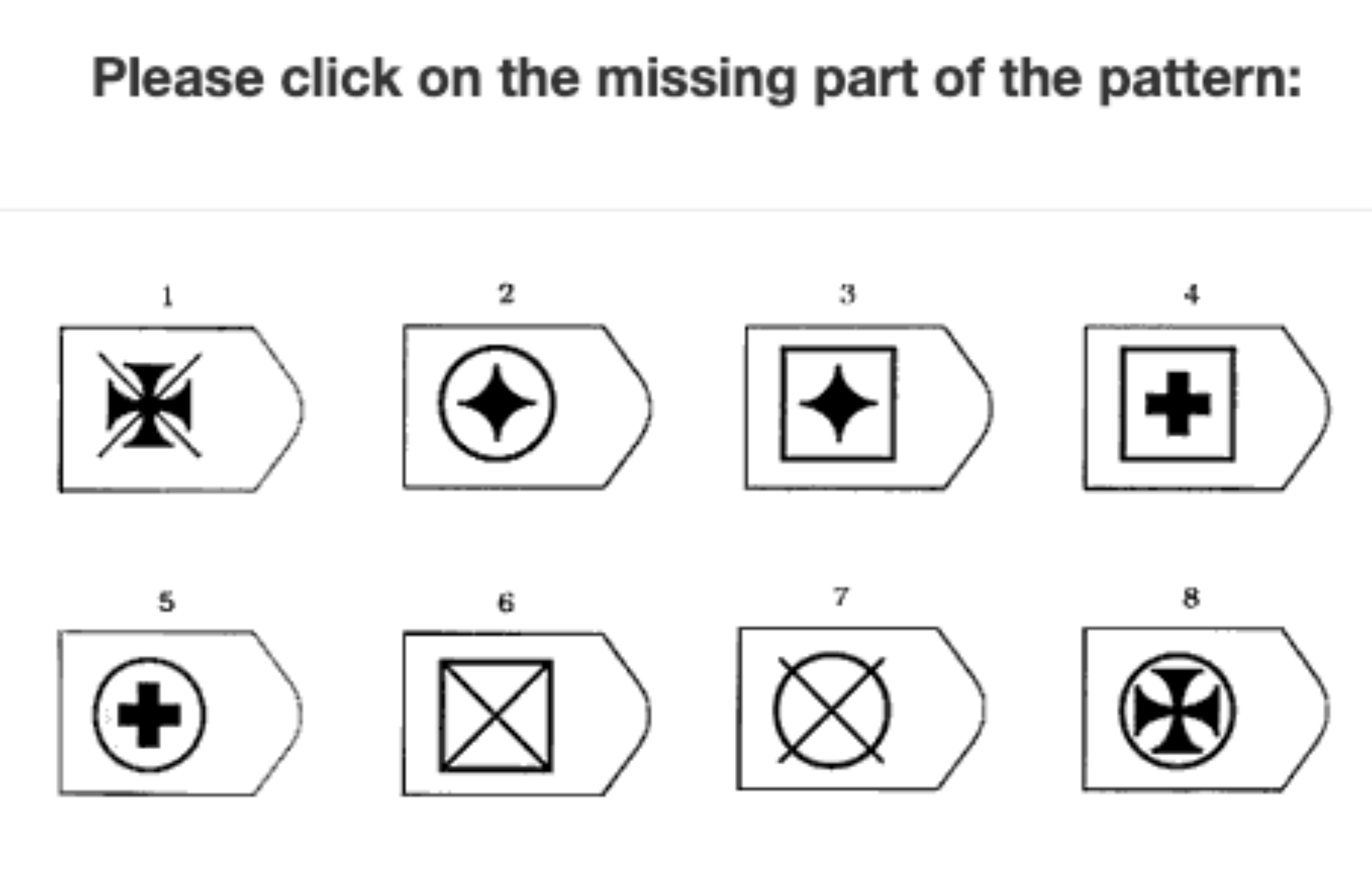
$\beta_{\text{REAS:COND}} = (0.21, 0.85)_{95\%}$

(Mayn & Demberg 2023)

(also Theory of Mind, but not Working Memory)

Raven’s Matrices





Success requires **efficient pattern induction** in a large hypothesis space.

(Carpenter et al. 1990, Gonthier & Thomassin 2015, Gonthier & Roulin 2020, Stocco et al. 2021)

Modeling individual differences in Raven's

ACT-R: Computational modeling framework for simulating real-time task performance given realistic memory, visual processing, and learning mechanisms.

(Anderson et al. 2004; see uses in Lewis & Vasishth 2005, Hendriks 2016, Brasoveanu & Dotlačil 2020)

Stocco et al. (2021):

ACT-R model for Raven's performance as rule induction via exploration and reinforcement learning

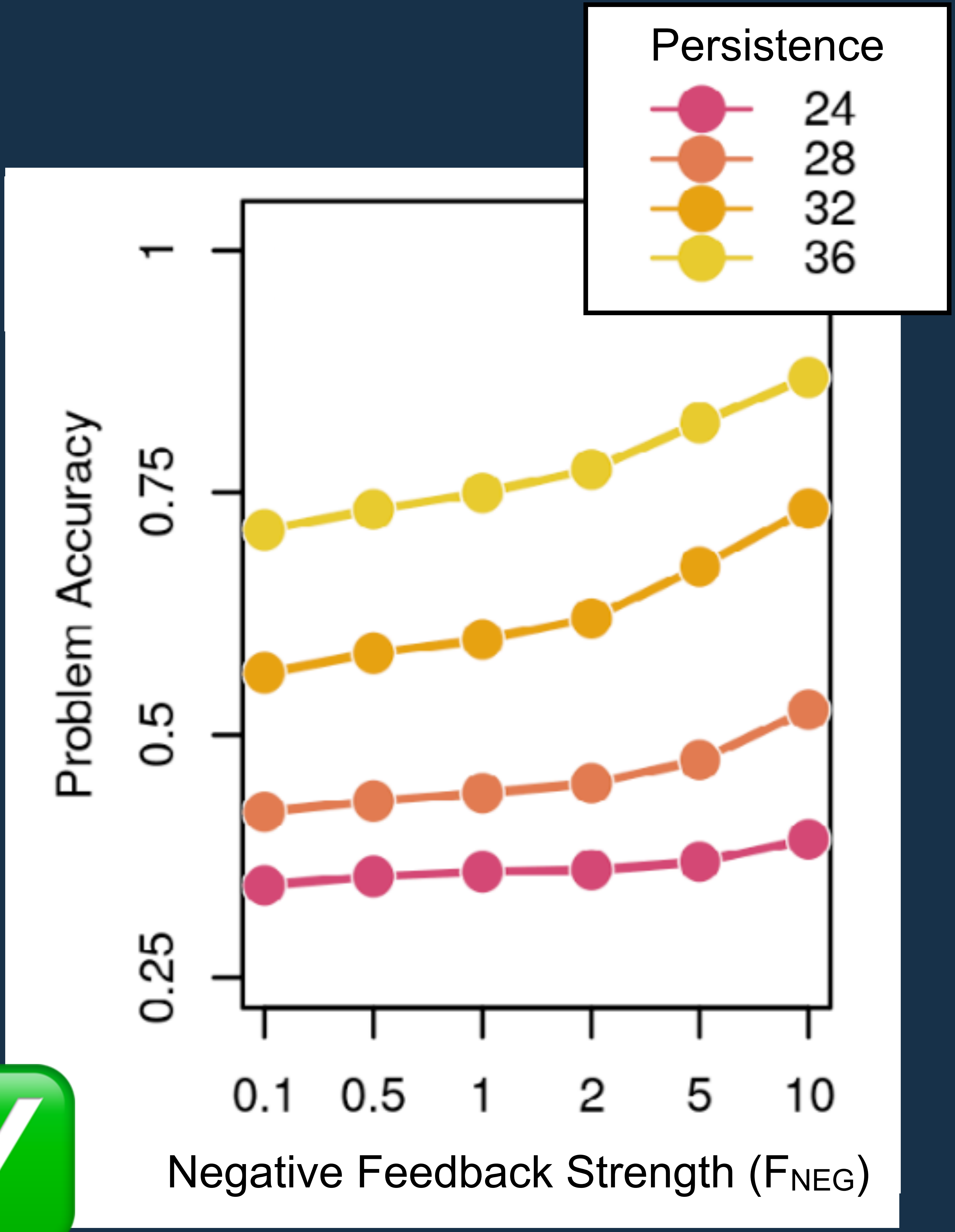
individually parameterized by:

persistence

(Eisenberger & Leonard 1980)

neg. feedback strength (F_{NEG})

(Frank et al. 2004)



Our contribution

Introduce an ACT-R model of RefGame as a problem of strategy exploration and learning

Successfully models learning effects and individual differences

Correctly predicts patterns of RTs and concrete differences in learning behavior

First step towards cognitively-realistic models of pragmatic performance

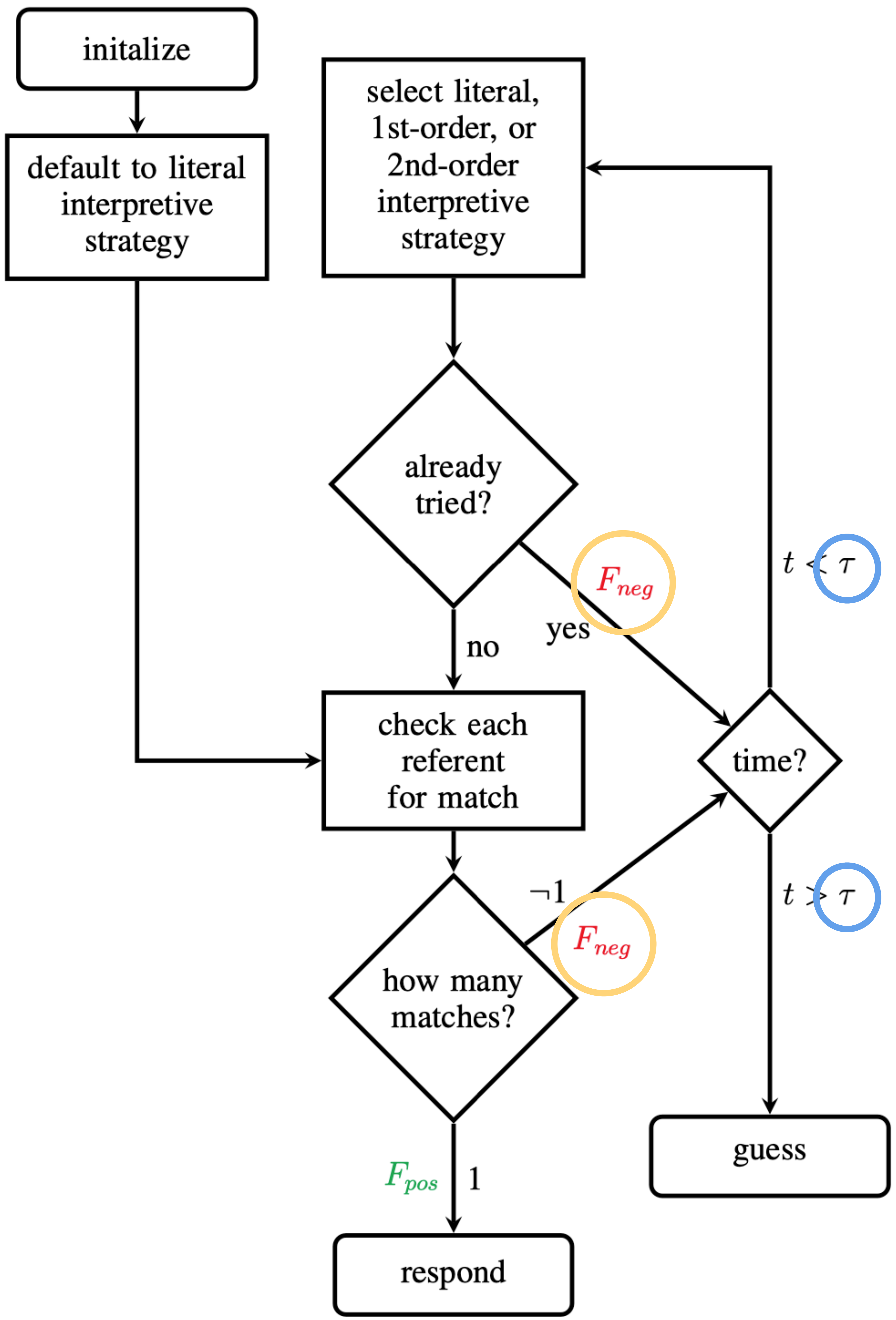
Roadmap

1. Background
2. Our ACT-R model
3. Validating the role of learning resources

RefGame as exploration

(implemented in pyactr: Brasoveanu & Dotlačil 2020)

- Attempt literal interpretation
  - Check informativity (number of matches)
  - If informative (1 match), select match
  - Else, penalize utility with FNEG
  - If time remains, return to…
- Select highest-utility strategy (with noise)
  - If already checked, penalize utility with FNEG
  - Else, evaluate; select or return again
- If time ever exceeds persistence (τ) guess

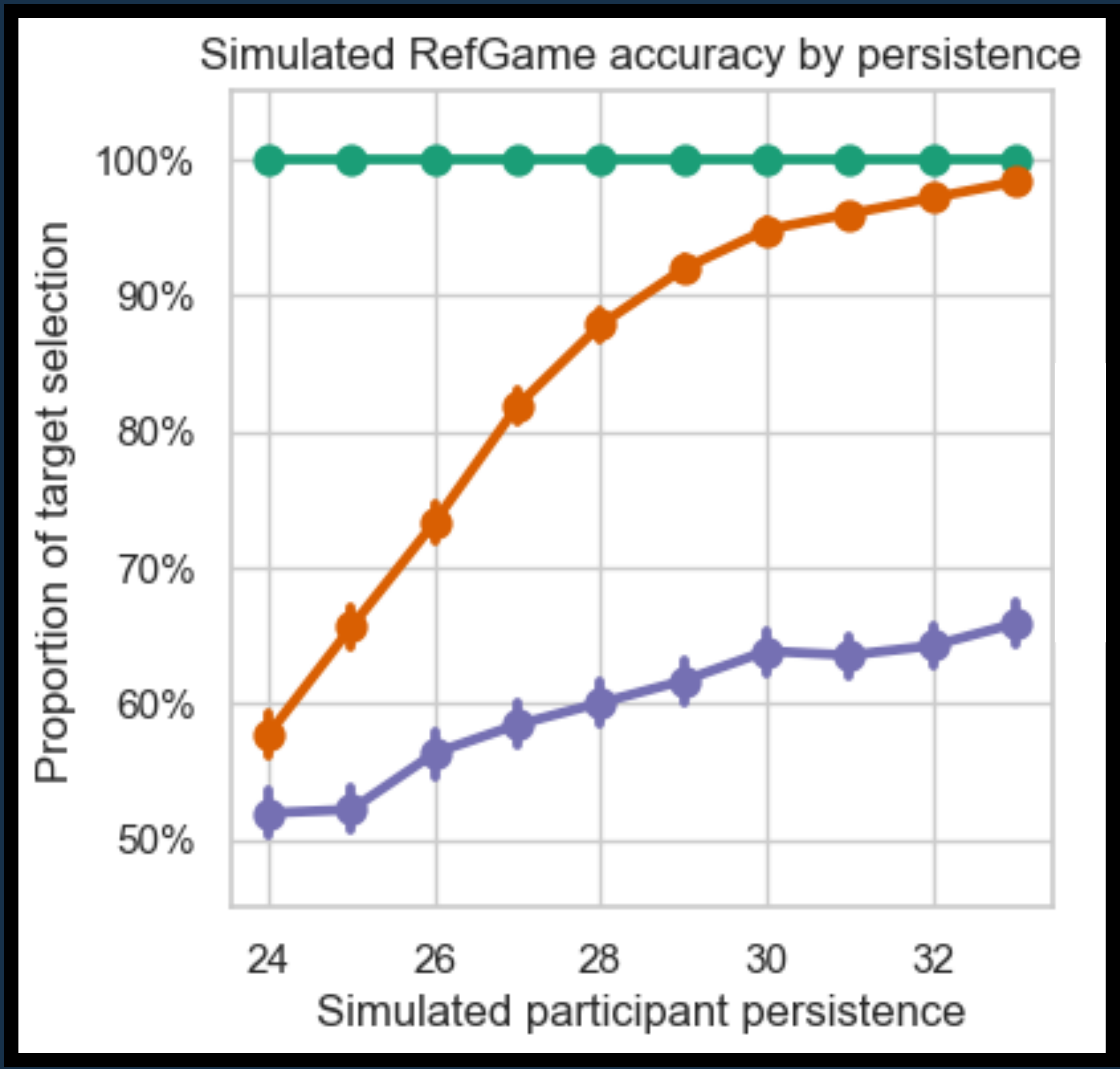


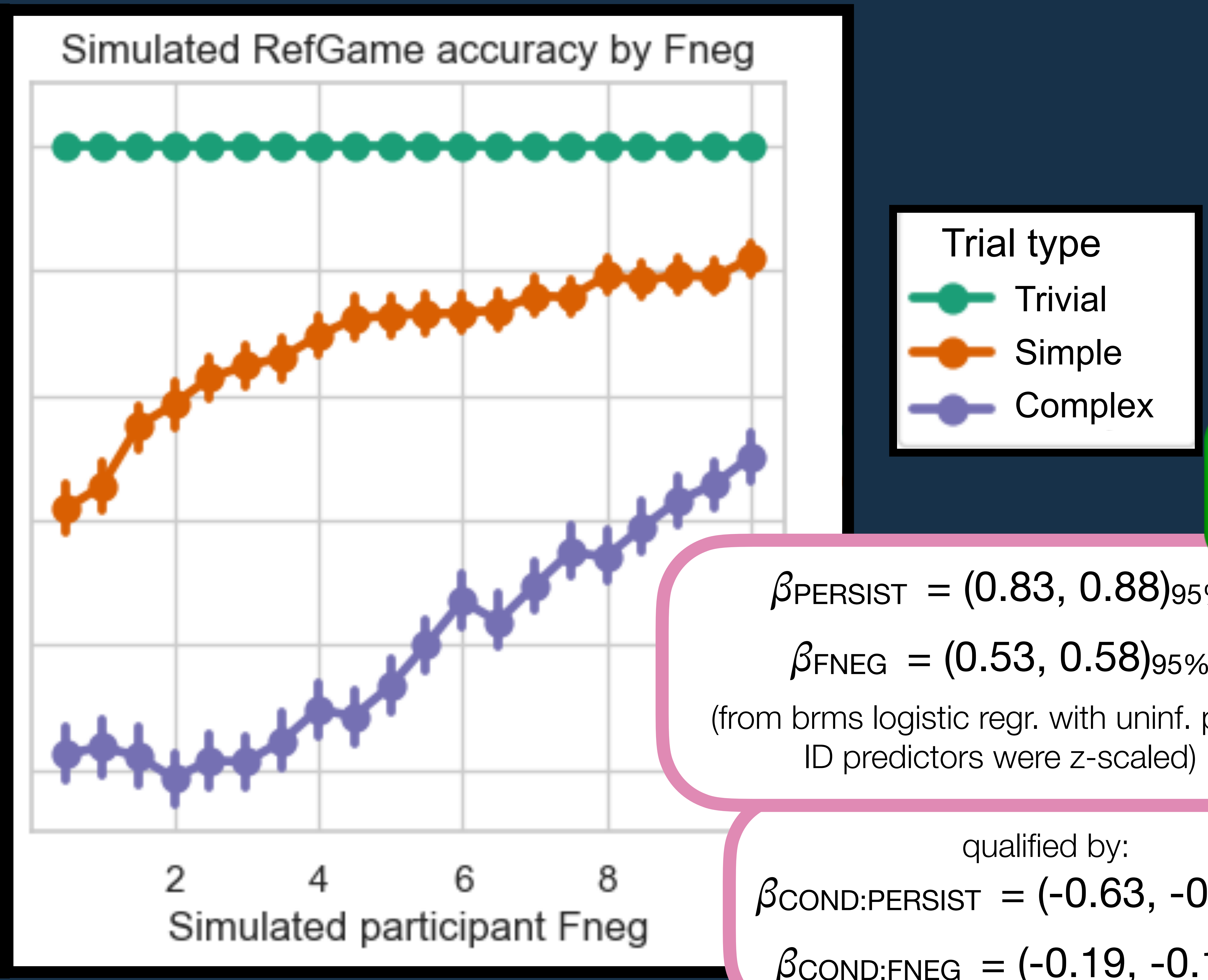
Model experiment

- Simulated task: Randomized 36-trial RefGame (16 trivial, 8 simple, 8 complex)
- Simulated participants: 10 persistence values x 20 F_{NEG} values, 25 per cell
- Critical strategy utilities begin as a fixed stair-step

Literal: 5 First-Order: -2.5 Second-Order: -5

Learning-related individual differences





β_{PERSIST} = (0.83, 0.88)$_{95\%}$

β_{FNEG} = (0.53, 0.58)$_{95\%}$

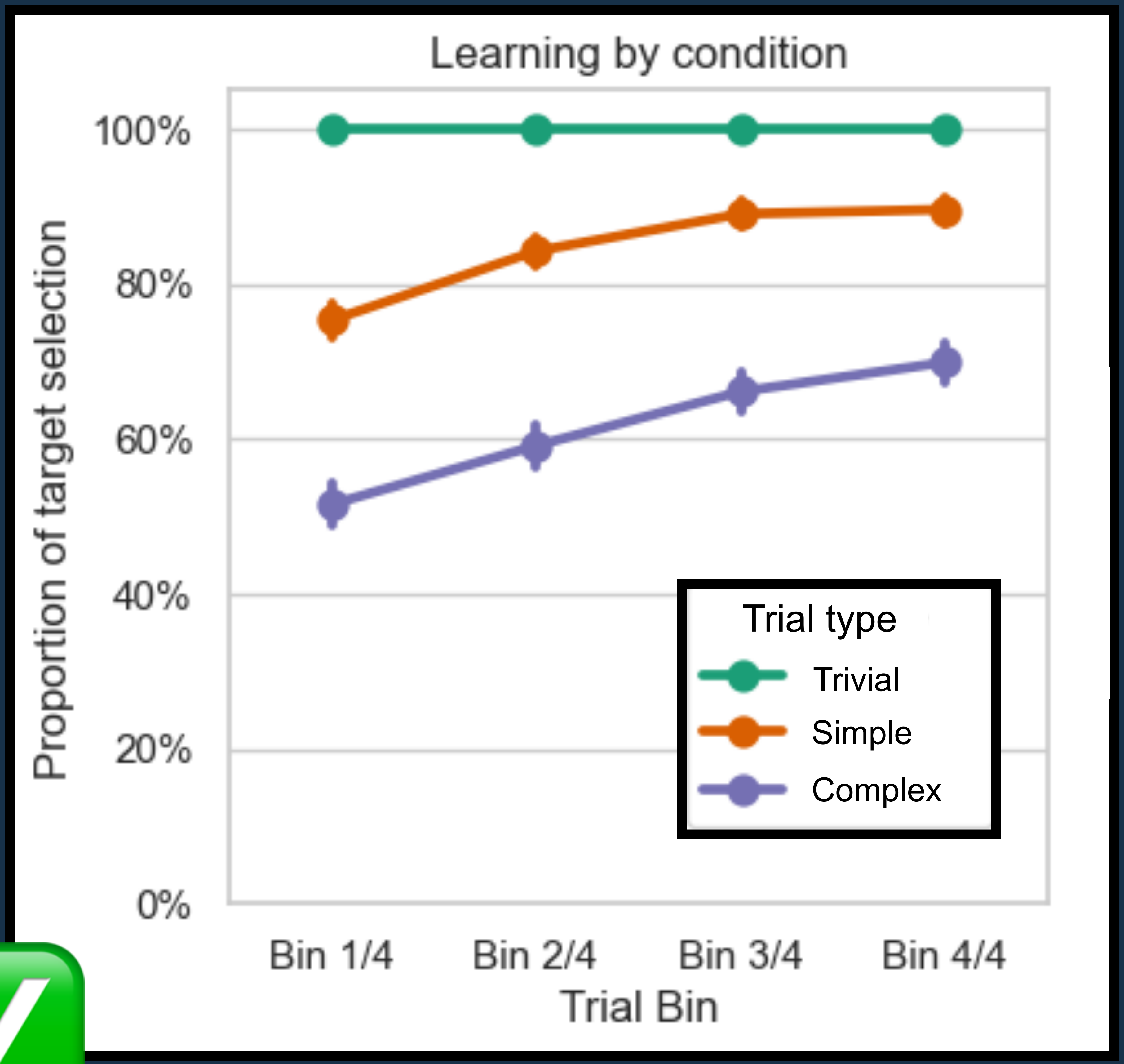
(from brms logistic regr. with uninf. priors, ID predictors were z-scaled)

qualified by:

$\beta_{\text{COND:PERSIST}}$ = (-0.63, -0.59)$_{95\%}$

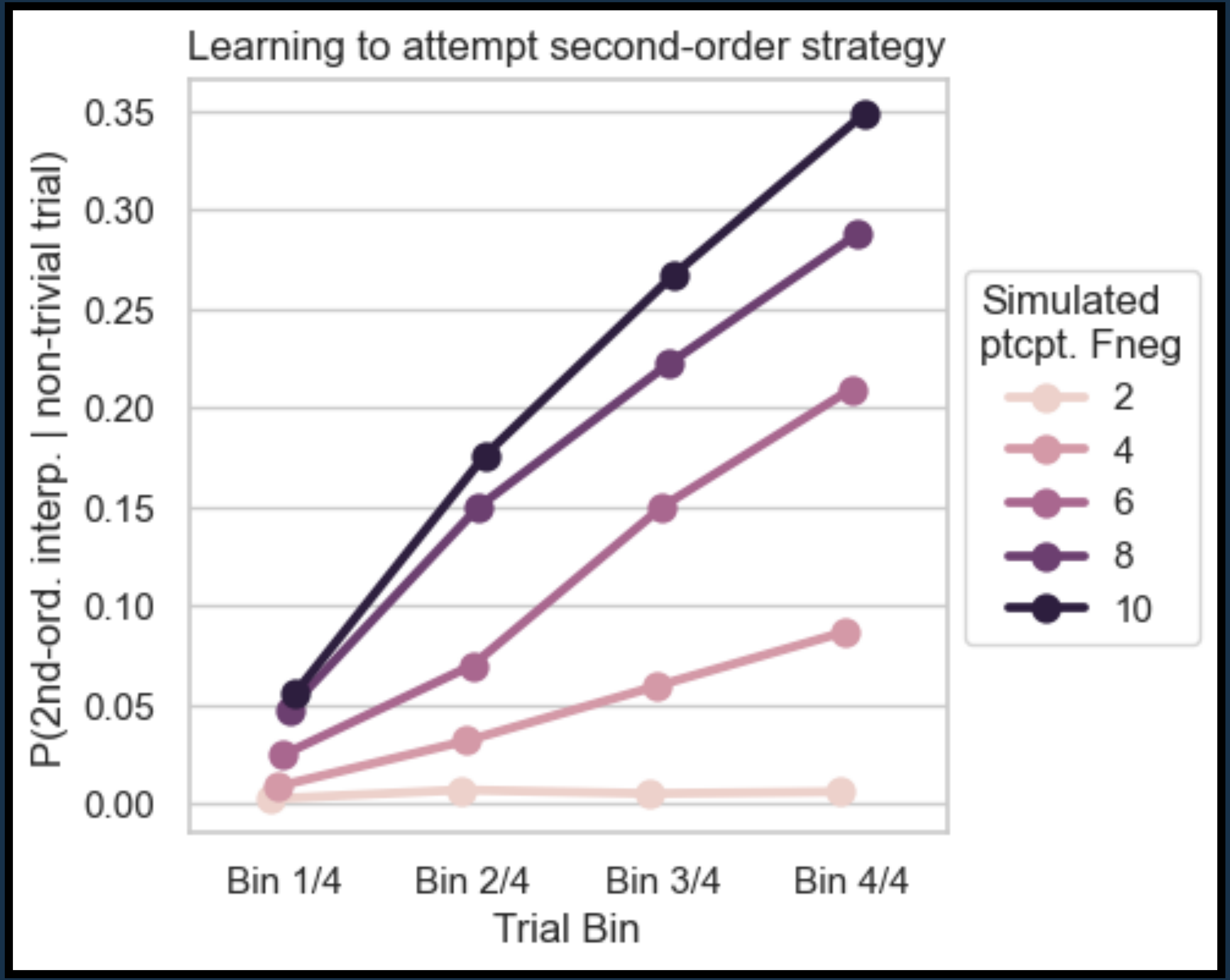
$\beta_{\text{COND:FNEG}}$ = (-0.19, -0.15)$_{95\%}$

Predicted learning behavior



$\beta_{\text{TRIAL}} = (0.05, 0.05)_{95\%}$

(from brms logistic regr. with uninf. priors, trial was centered and not scaled)



Roadmap

1. Background
2. Our ACT-R model
3. Validating the role of learning resources

New pre-registered experiment

- Randomized 36-trial RefGame (16 trivial, 8 simple, 8 complex), collecting RTs
- 150 participants from Prolific
- After RefGame, participants completed various individual difference tasks, including tasks measuring persistence and F_{NEG}

Predictions:

(A) Accuracy $\propto$ persistence, F_{NEG}

(B) Accuracy $\propto$ progress (a learning effect)

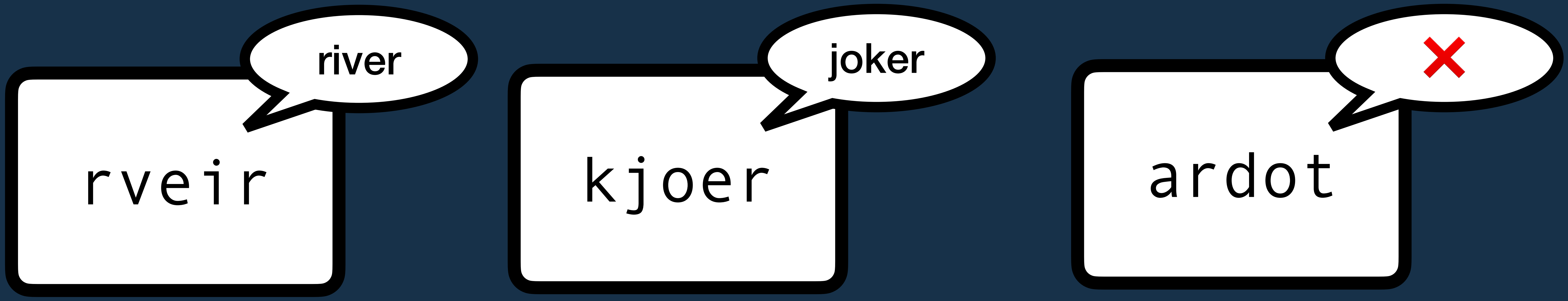
(C) RTs should vary by condition as the ACT-R model predicts

Measuring Persistence:

Impossible Anagrams

(Ventura & Shute 2013)

(see also Eisenberg & Leonard 1980; Dale et al. 2018)



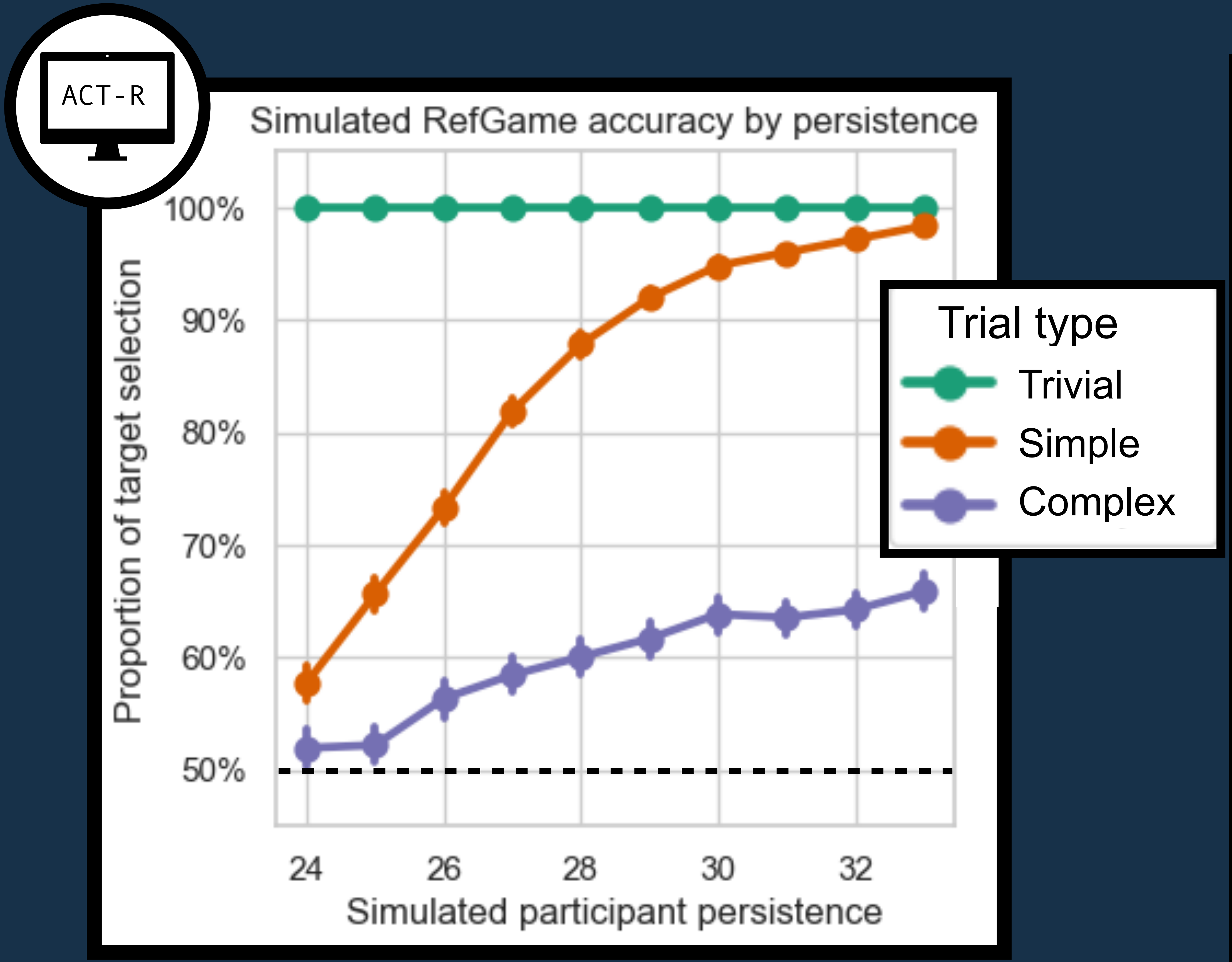
(easy, 10) (hard, 5) (impossible, 5)

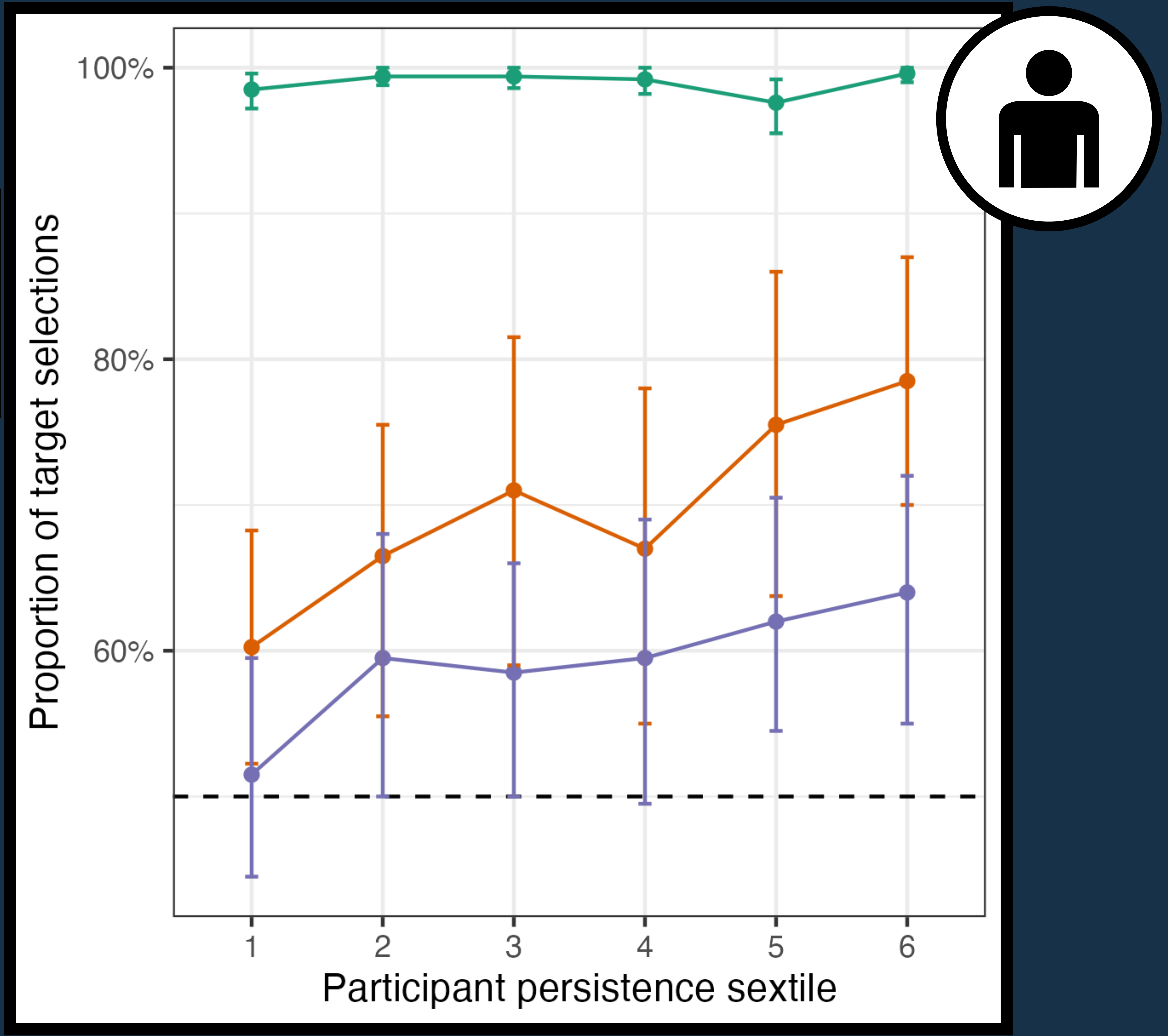
Anagram Persistence:

$$\frac{\text{SkipTime}_{\text{IMPOSS}}}{\text{Correct RT}_{\text{EASY}}}$$

- Initial validation: This measure correlated with…
  - Time spent on (task-final) impossible Raven's problem (Dale et al. 2018) — $R = 0.18$
  - "Grit" score derived from self-assessment (Duckworth & Quinn 2009) — $R = 0.20$

RefGame accuracy by measured anagram persistence





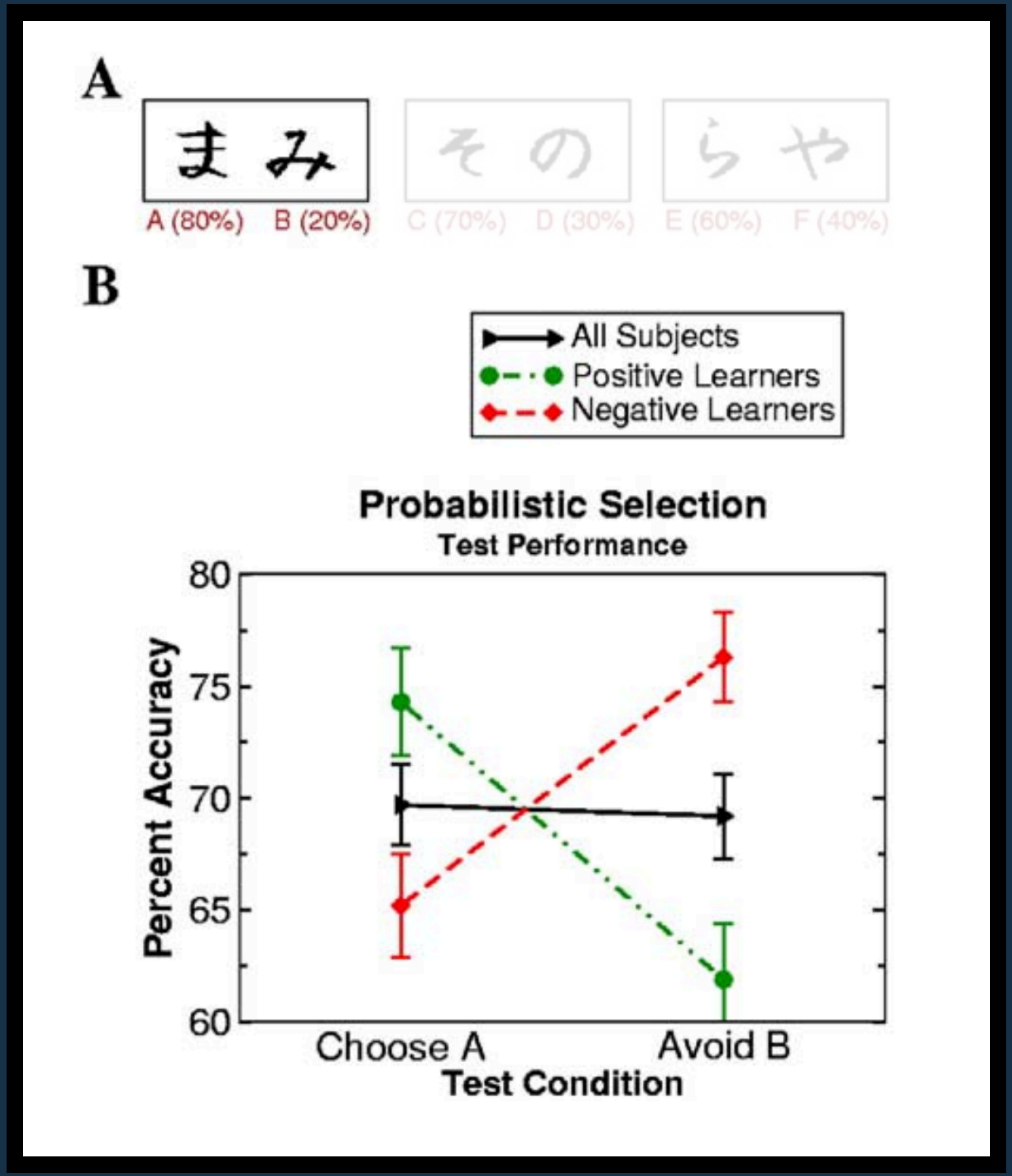
Model $\beta_{PERSIST}$ = $(0.83, 0.88)_{95\%}$

Human $\beta_{PERSIST}$ = $(0.08, 0.58)_{95\%}$

(from brms logistic regr. with uninf. priors, ID predictors were z-scaled)

Measuring F_{NEG}:

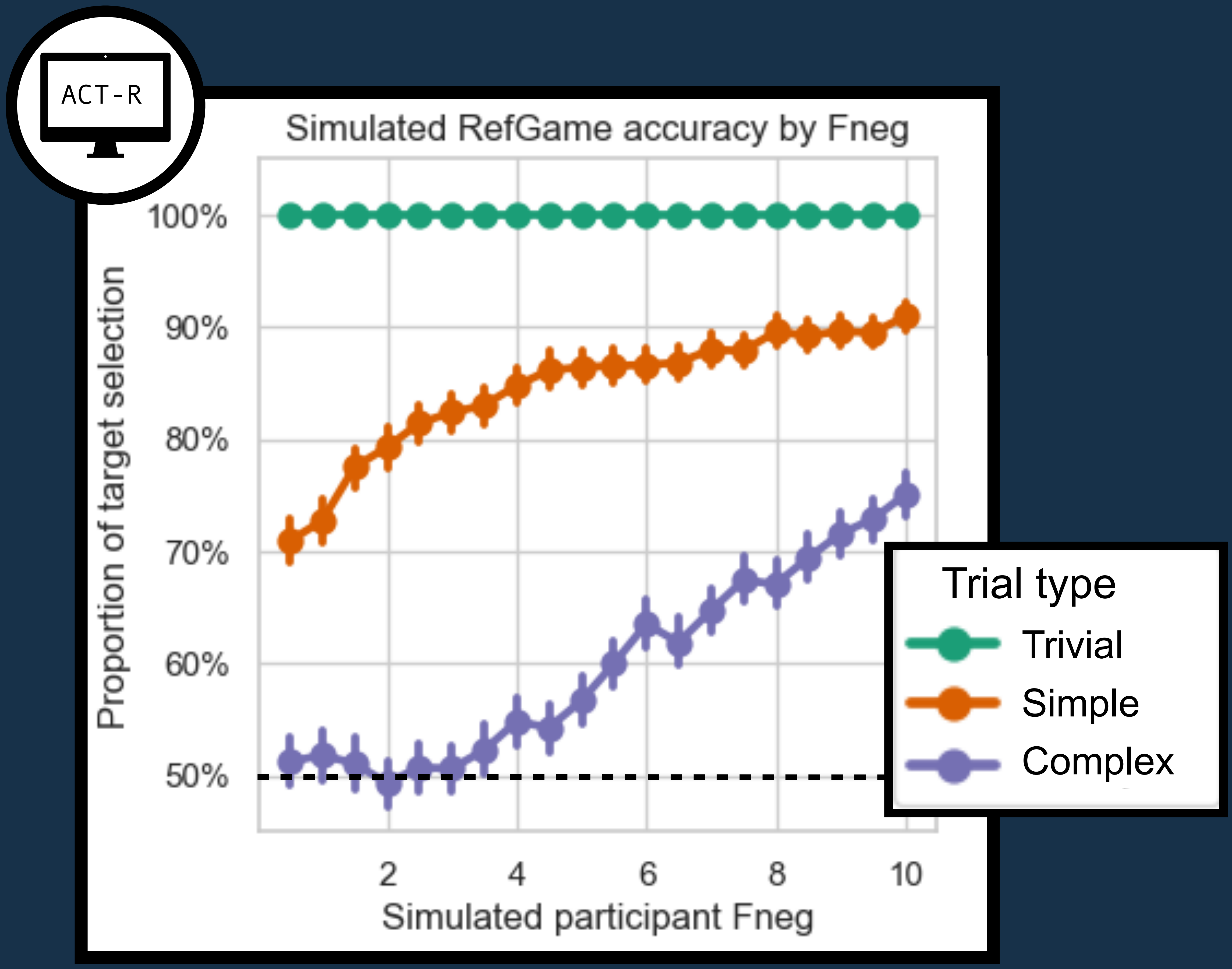
The Probabilistic Stimulus Selection task

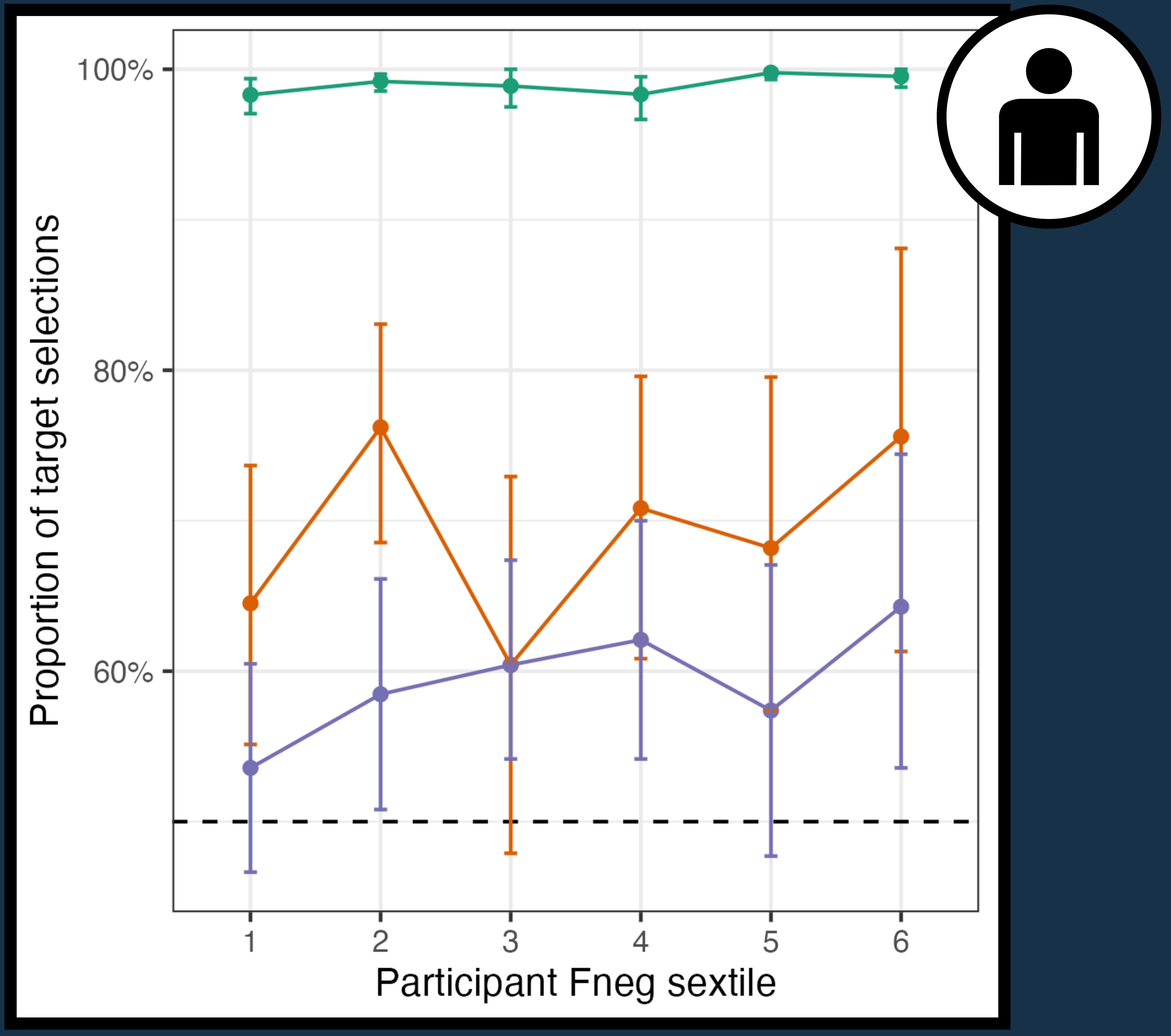


(Frank et al. 2004, 2005, 2007)

- A is a better choice than B, prompts two types of learned behavior:
  - Learn positive value of A (via F_{POS})
  - Learn negative value of B (via F_{NEG})
- Corresponds to individual differences in error-related negativity in ERPs and dopamine levels in basal ganglia.

Observed relation to measured F_{NEG}





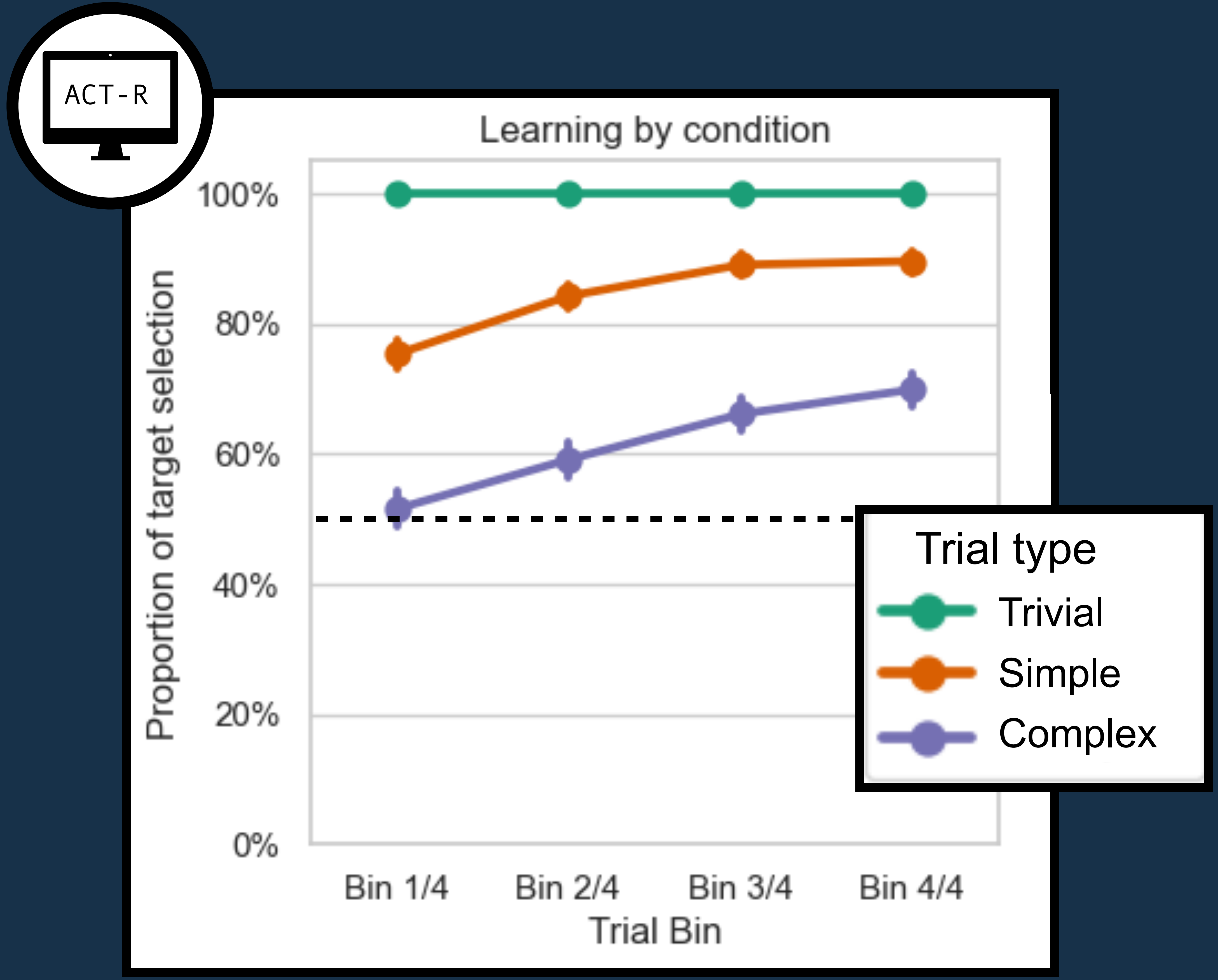
✅

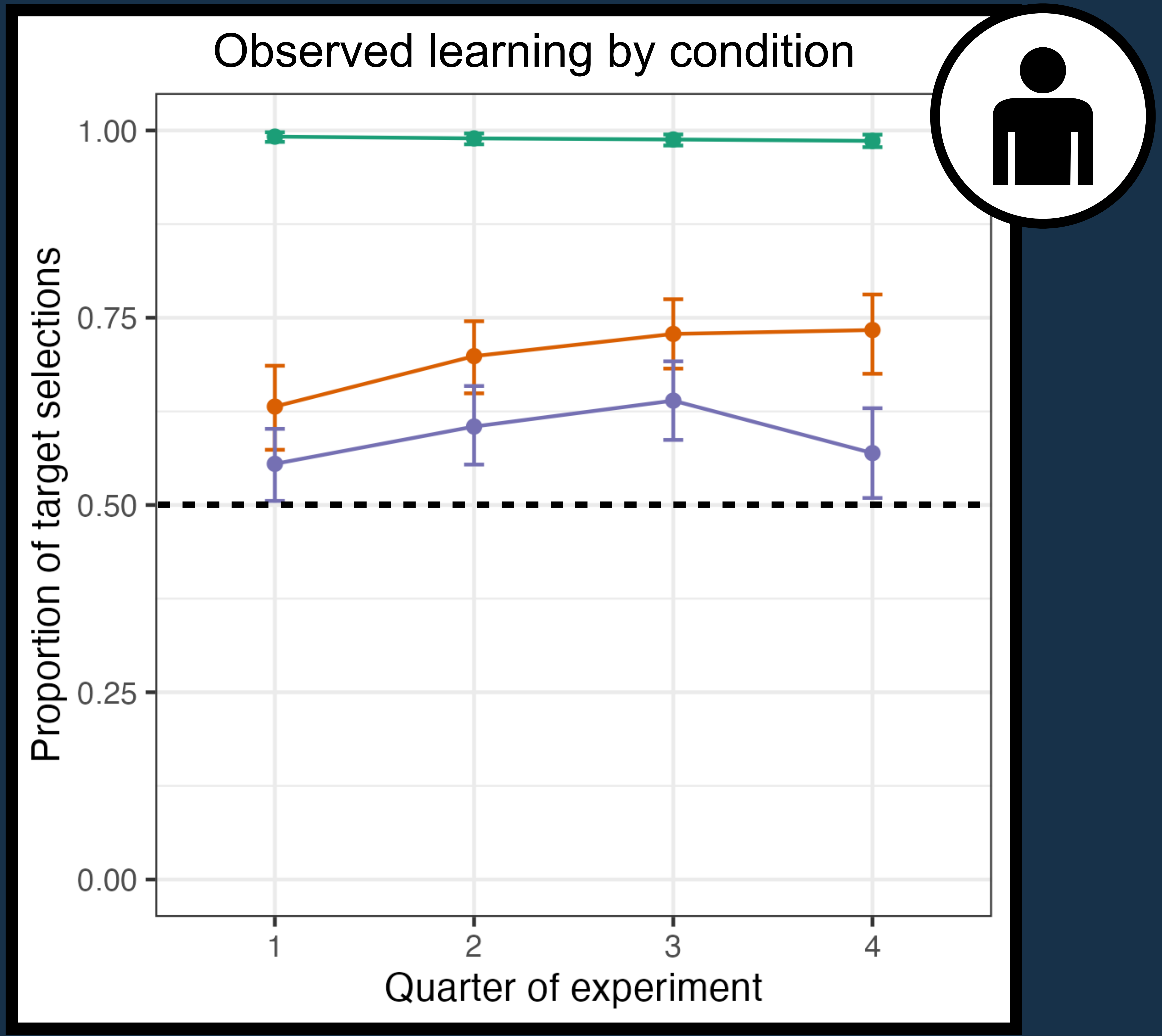
Model $\beta_{FNEG} = (0.53, 0.58)_{95\%}$

Human $\beta_{FNEG} = (-0.05, 0.40)_{95\%}$

(from brms logistic regr. with uninf. priors, ID predictors were z-scaled)

Further evidence for learning



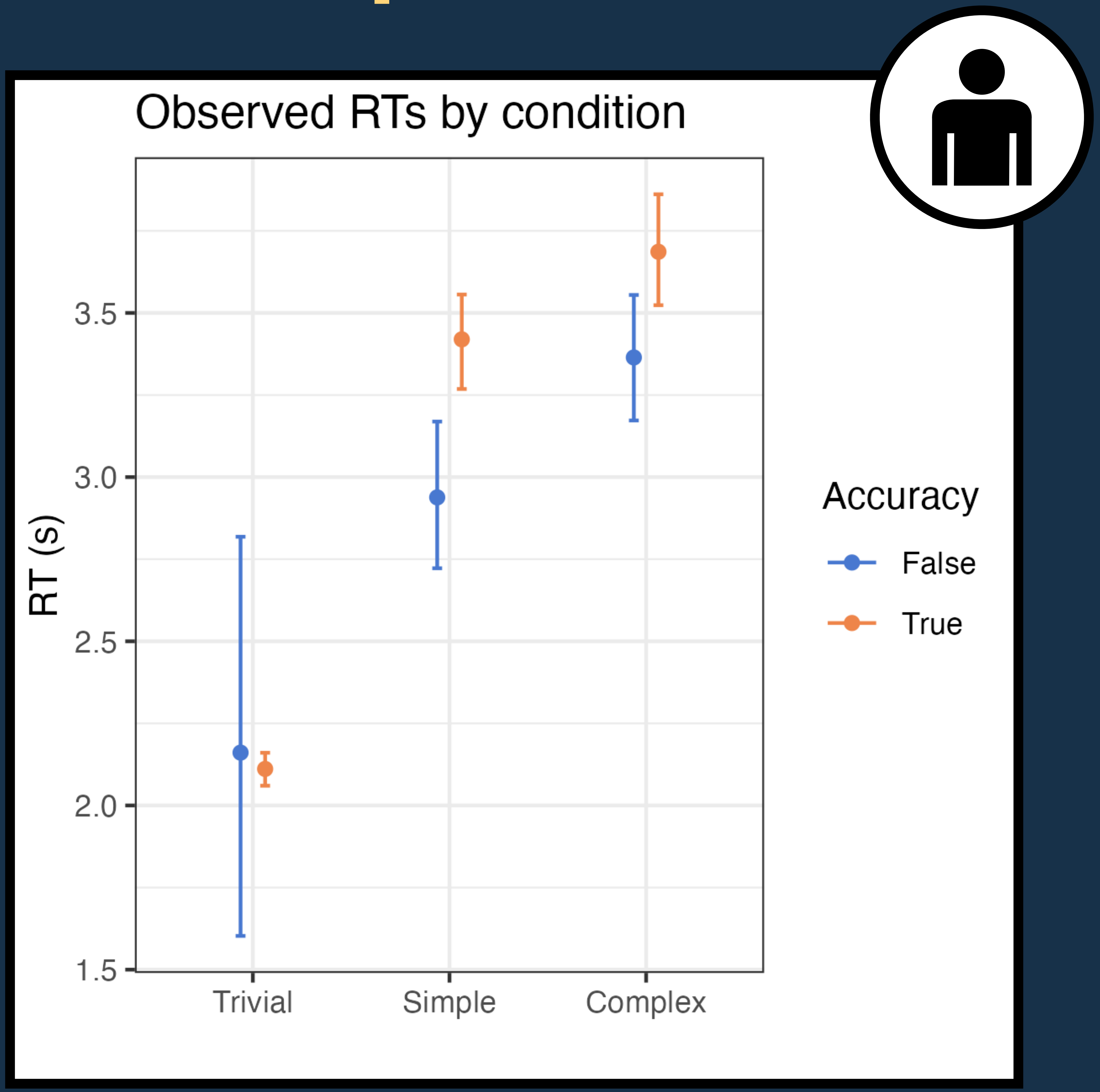


Model β_{FNEG} = $(0.05, 0.05)_{95\%}$

Human β_{FNEG} = $(0.01, 0.03)_{95\%}$

(from brms logistic regr. with uninf. priors, trial was centered and not scaled)

Comparison of response time patterns



Correct Trivial $<_{RT}$ Correct Critical ($P > 0.99$)

Incorrect Critical $<_{RT}$ Correct Critical ($P = 0.90, 0.95$)

(from brms logistic regr. with uninf. priors)

✅

Introduce an ACT-R model of RefGame as a problem of strategy exploration and learning

Successfully models learning effects and individual differences

Correctly predicts patterns of RTs and concrete differences in learning behavior

First step towards cognitively-realistic models of pragmatic performance

In support of algorithmic-level models

- Probabilistic models of pragmatic **competence** (e.g. Frank & Goodman's Rational Speech Act model) have been extremely influential, but they are not models of **processing**

- Processing models are needed to explain a host of more complex facts:

- On-task learning behavior
- Evidence for inference-specific cognitive load

(De Neys & Schaeken 2007, Marty & Chemla 2013, van Tiel et al. 2017)

- Effects of general cognitive differences
- Heuristics/failures of probabilistic reasoning

(Mayn, Duff, Bila & Demberg 2024, cf. Fox et al. 2004)

Independent from a core hypothesis of Gricean competence!

Beyond the game setting

- Current model is specific to a highly controlled, novel game.
- Still, core may be plausible for ad-hoc inferences in natural comprehension:
  - Rational preference to avoid effort
  - Search for alternative meanings triggered by low informativity/relevance
  - Experience-based tuning of reasoning depth for a given interaction
- Indeed, Raven's scores also correlate with ad-hoc atypicality inferences.

(Ryzhova, Mayn & Demberg 2023)

- We aim to extend our model in this direction.

Alexandra Mayn

Vera Demberg

UNIVERSITÄT DES SAARLANDES

European Research Council
Established by the European Commission

ERC Grant #948878 to V. Demberg, “Individualized interactions in discourse”

Thanks also to Sebastian Schuster, Michael Franke, Niels Taatgen, and audiences at MathPsych 2024 for suggestions and feedback.

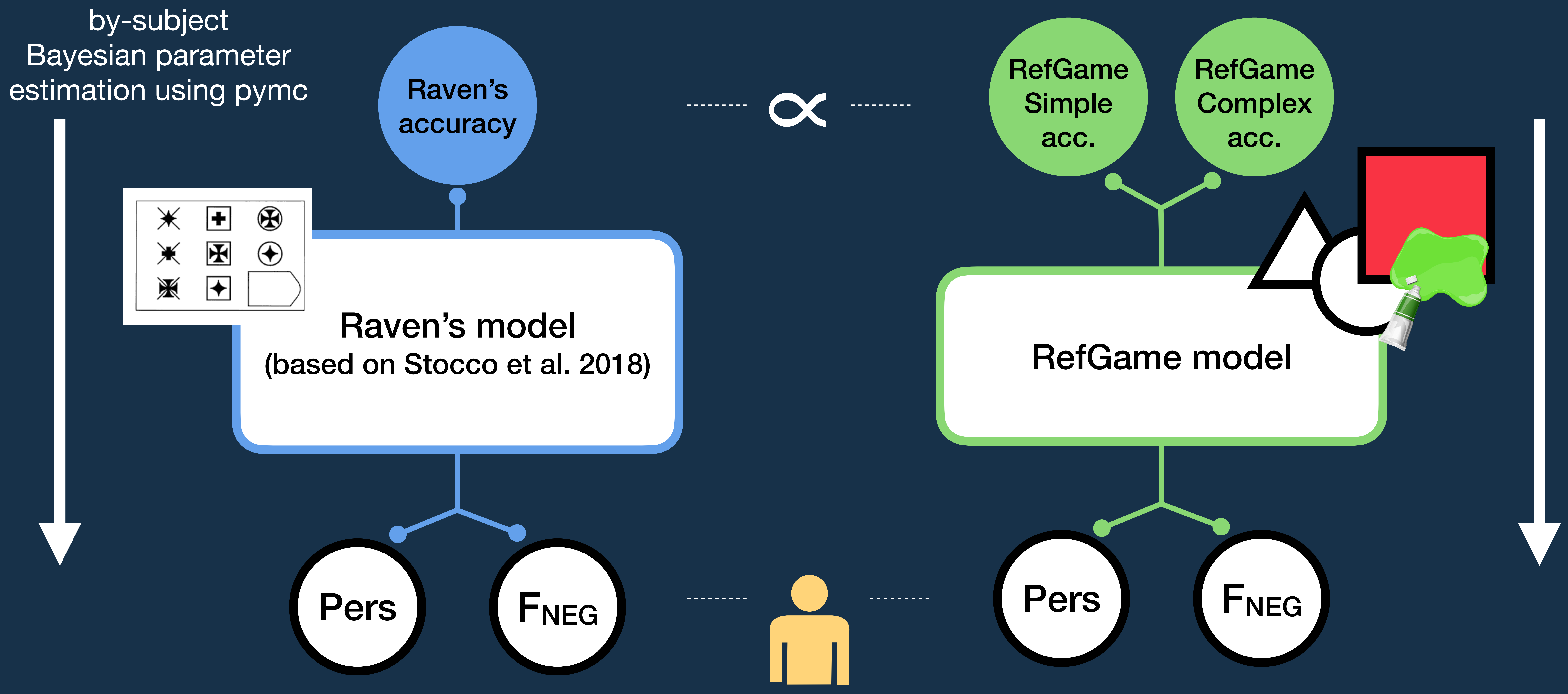
Thanks!

Ask us about:

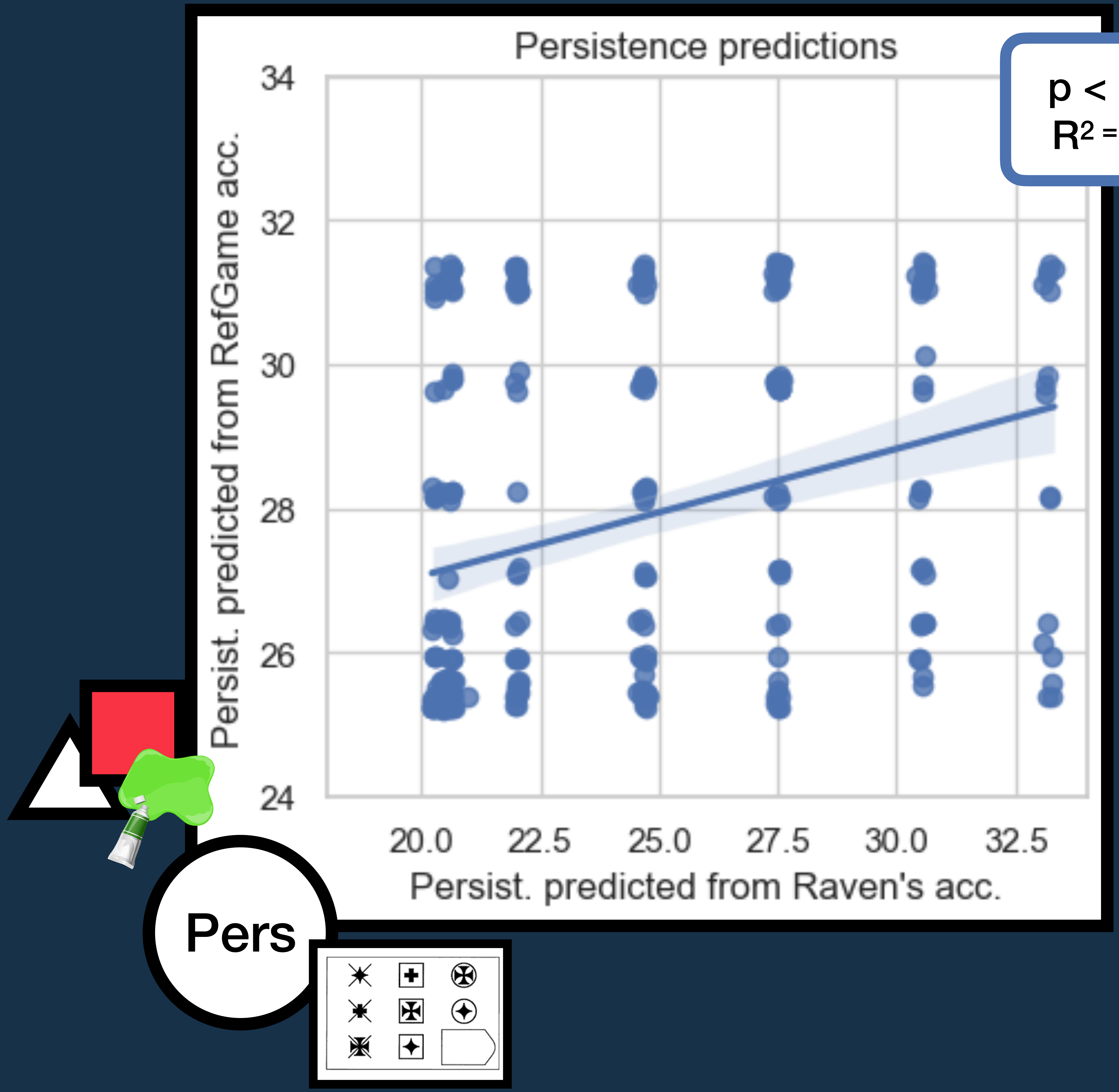
- A parameter estimation analysis assessing the connection between Raven’s and RefGame
- Our related poster on probability fallacies in first-order reasoning
- Finer details of model simulations and experimental data
- A more complex model accounting for individual differences in tendency towards pragmatic reasoning

Model experiments linking the tasks

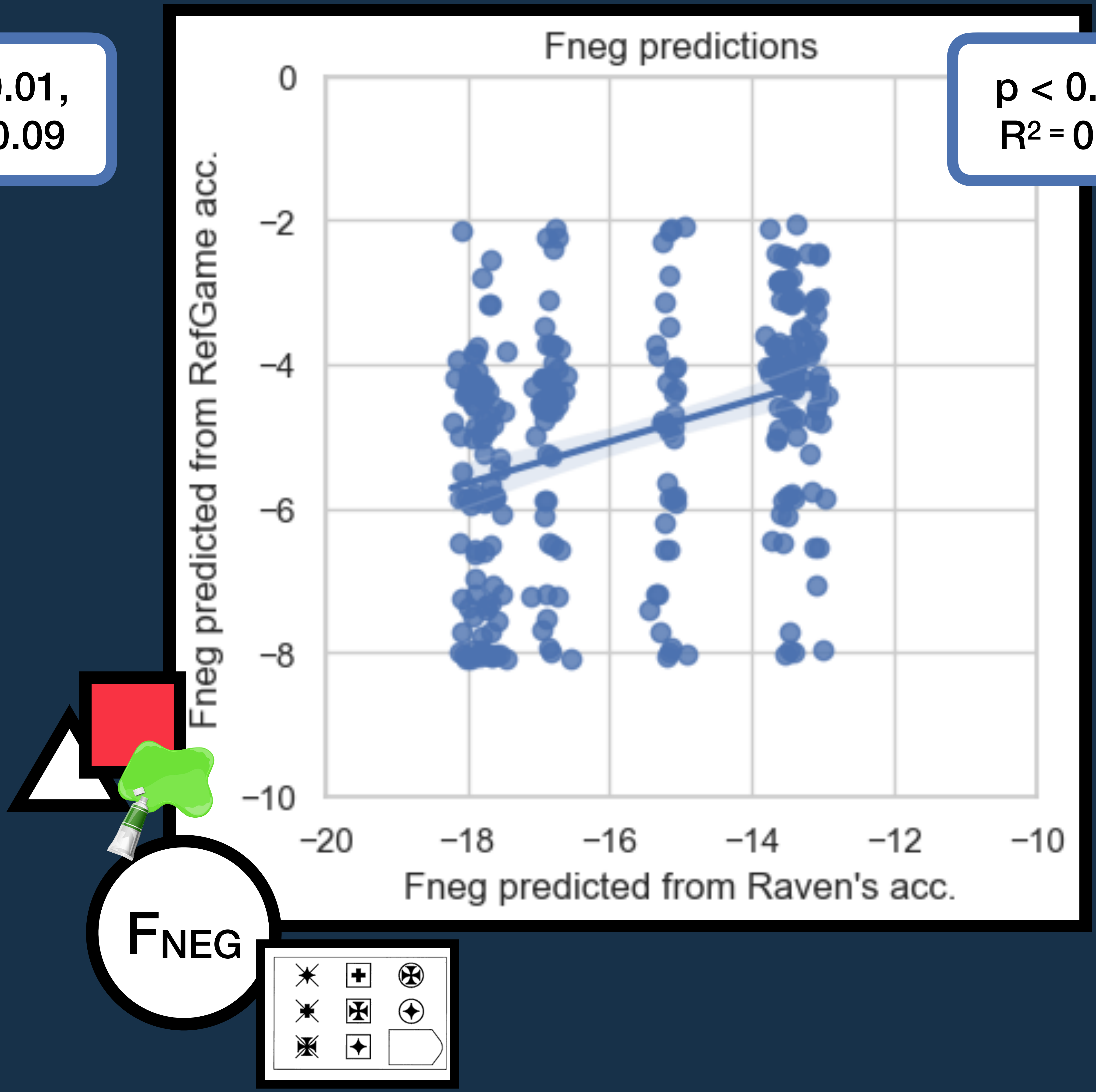
Jointly modeling Raven's and RefGame



Comparing parameters across tasks



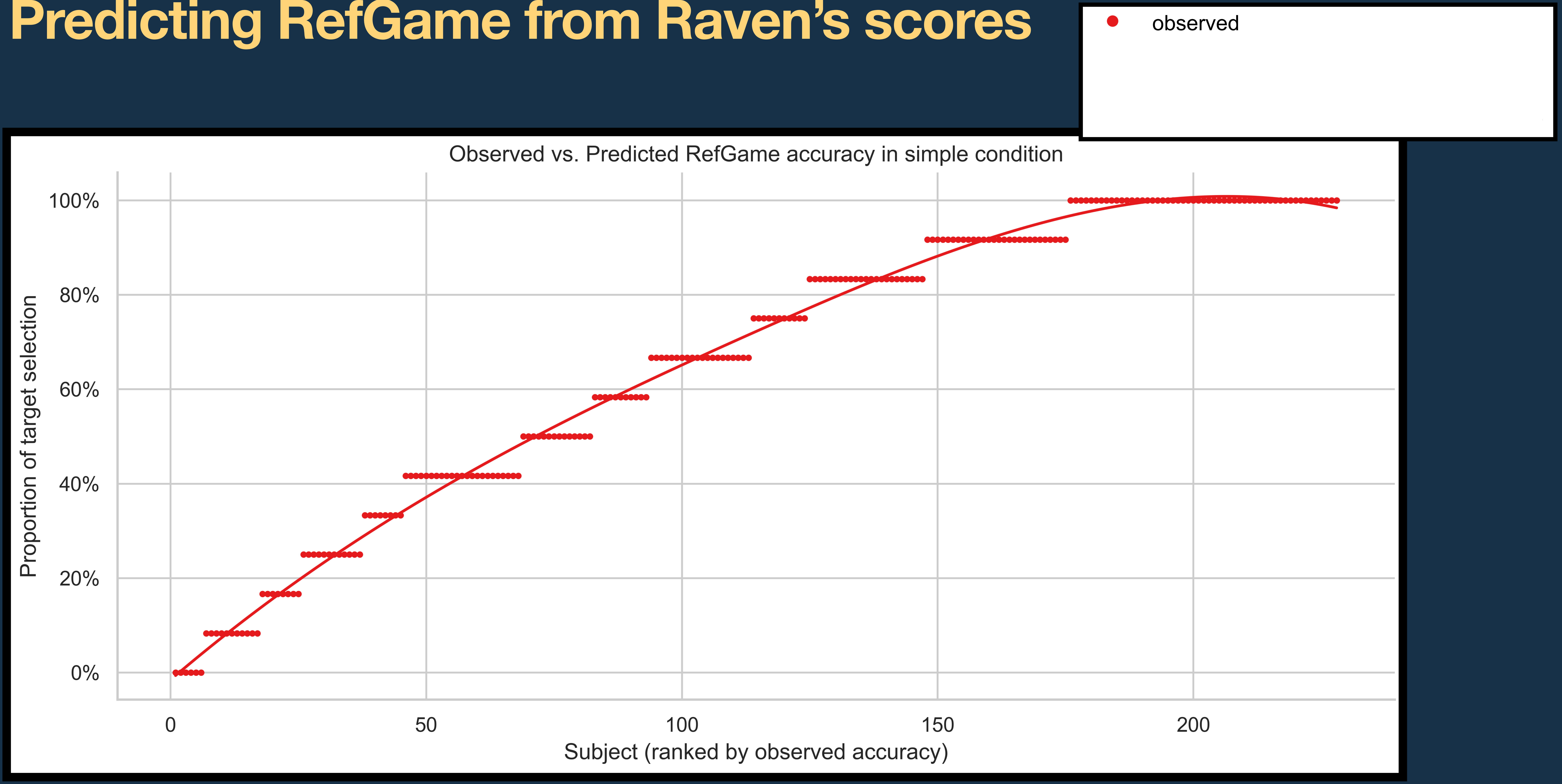
$p < 0.01$,
$R^2 = 0.09$



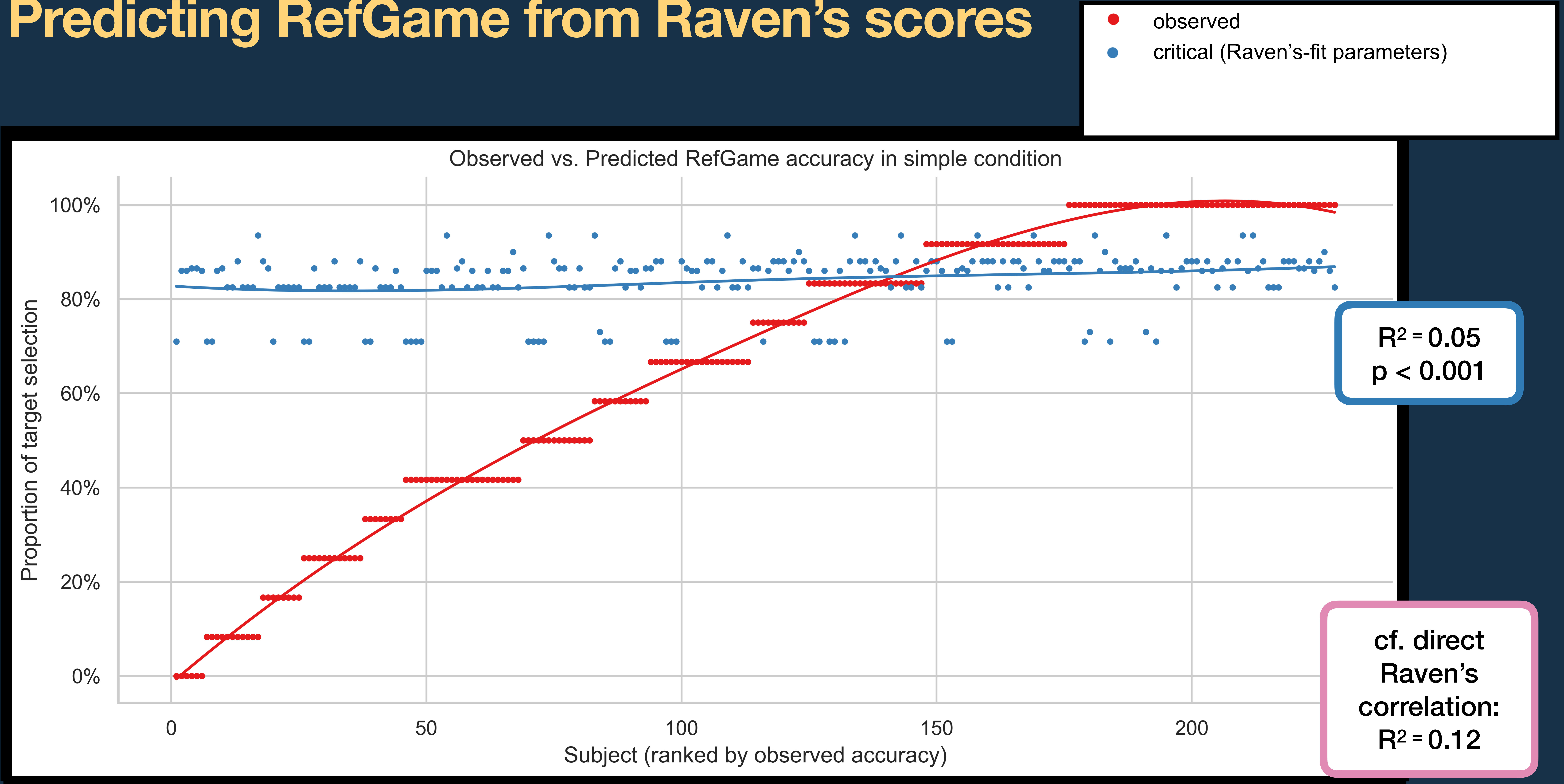
$p < 0.01$,
$R^2 = 0.11$

Predicting RefGame from Raven's scores

Predicting RefGame from Raven's scores



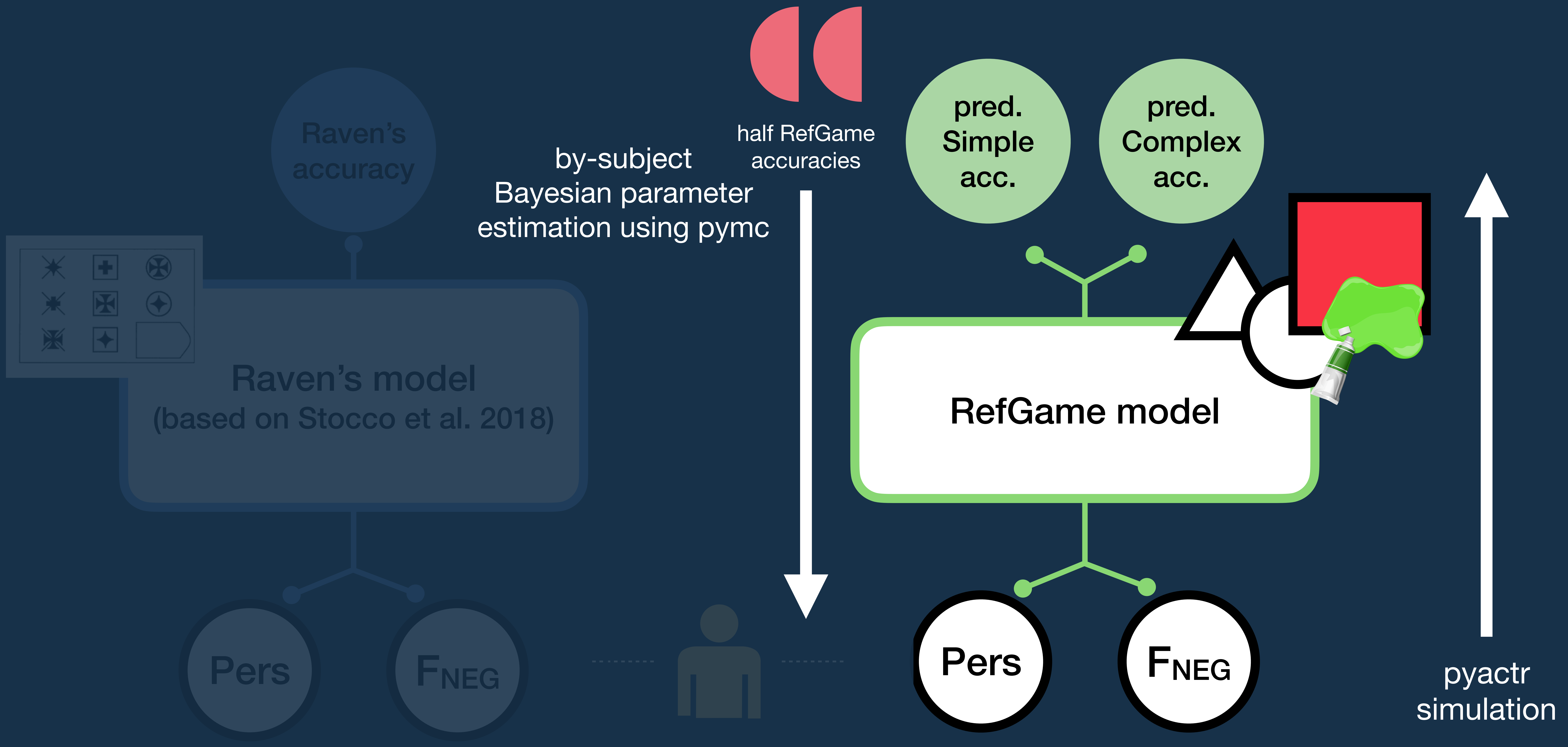
Predicting RefGame from Raven's scores



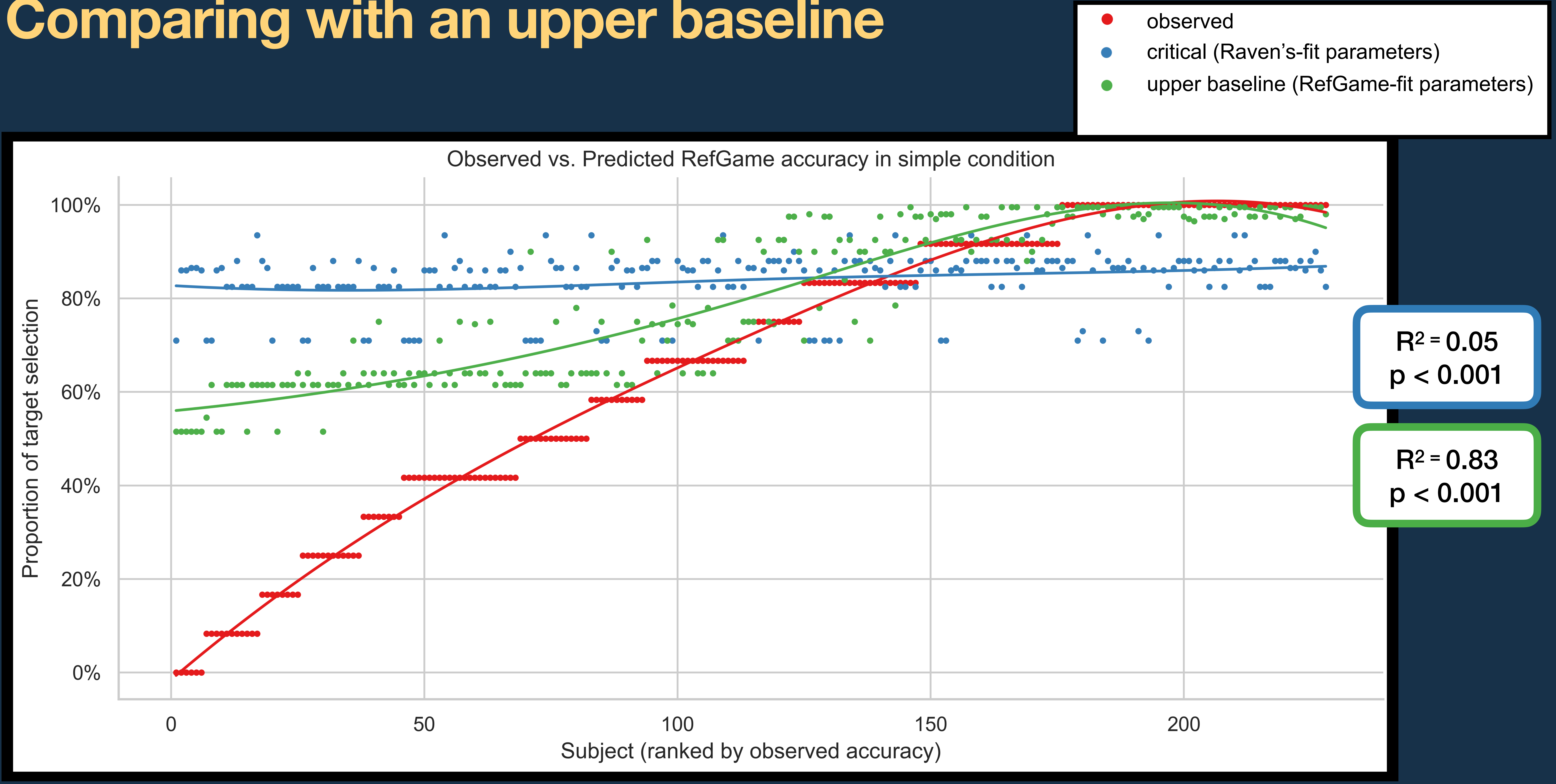
$R^2 = 0.05$
$p < 0.001$

cf. direct Raven's correlation: $R^2 = 0.12$

Deriving an upper baseline



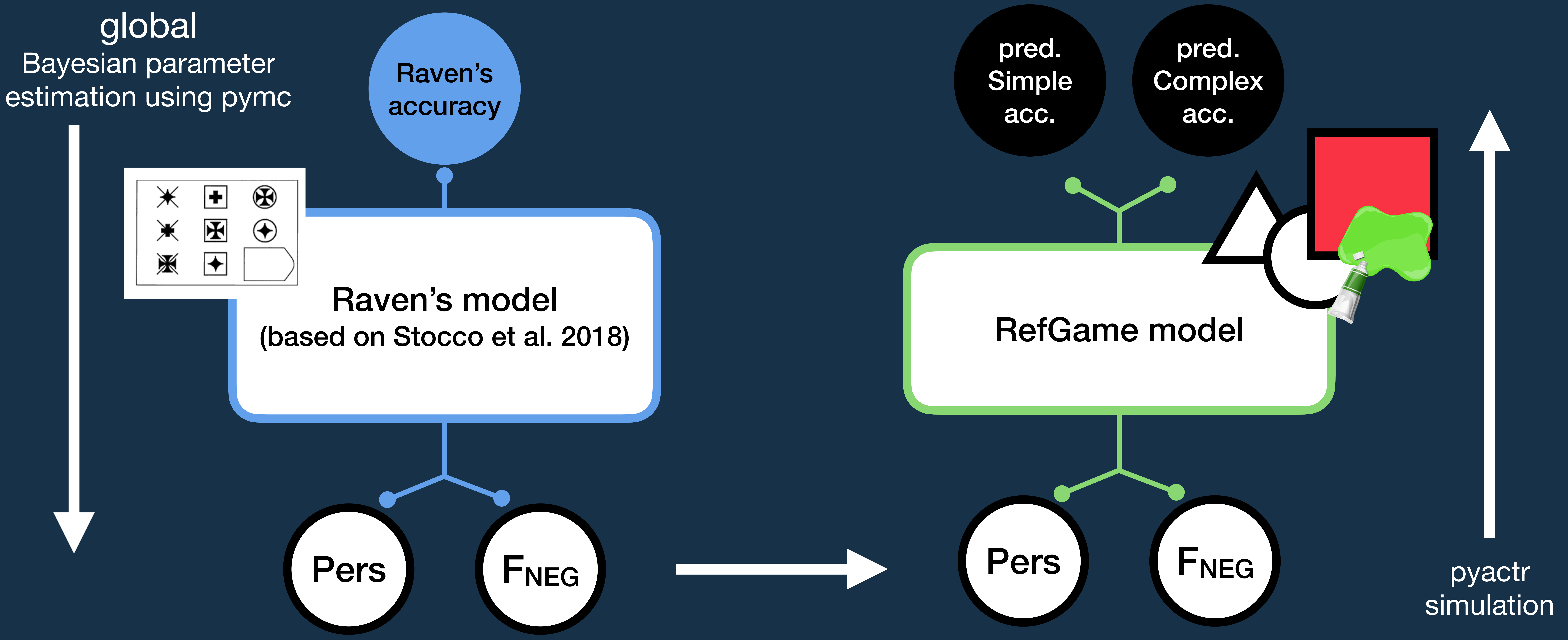
Comparing with an upper baseline



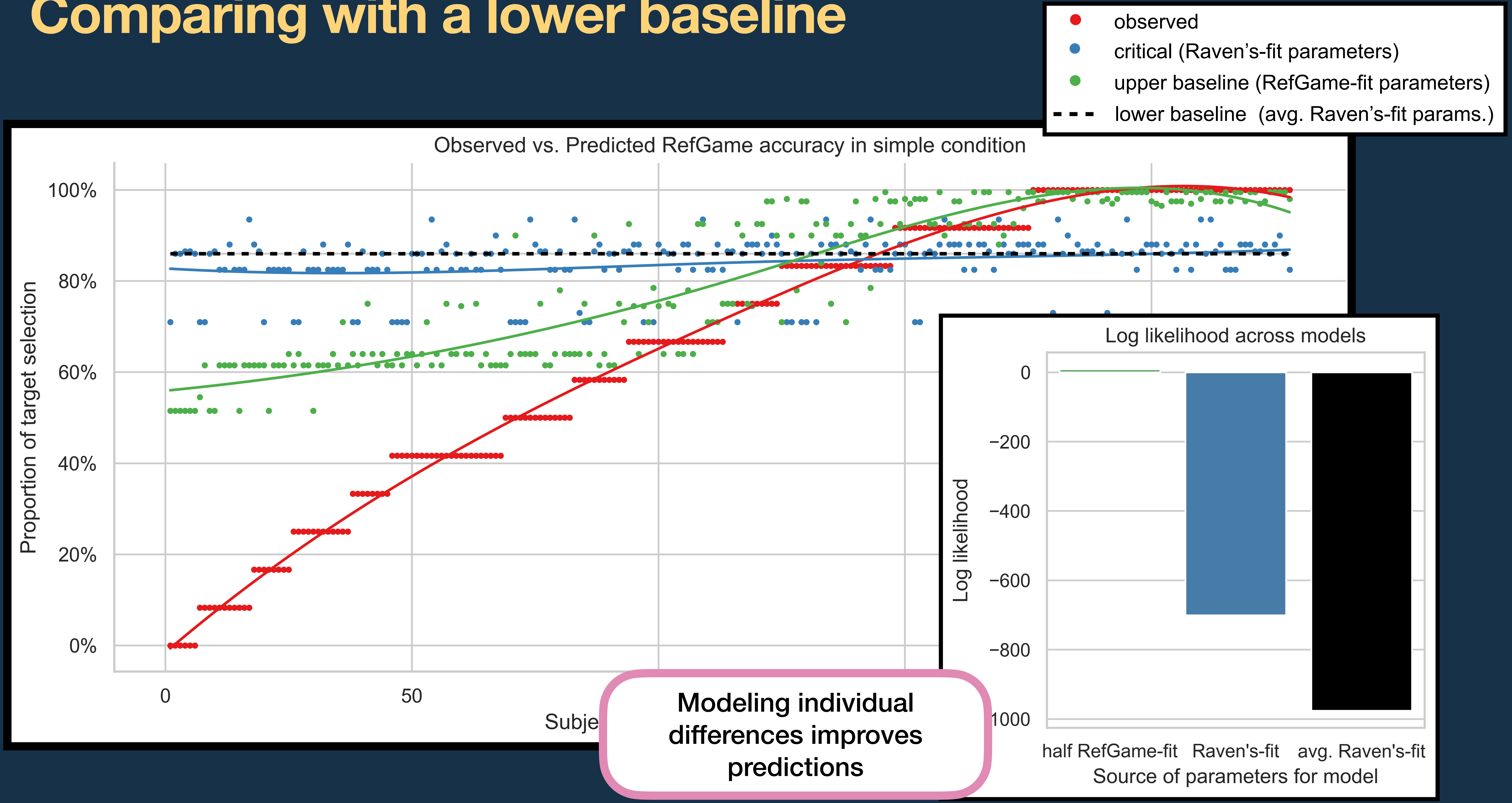
$R^2 = 0.05$
$p < 0.001$

$R^2 = 0.83$
$p < 0.001$

Deriving a lower baseline



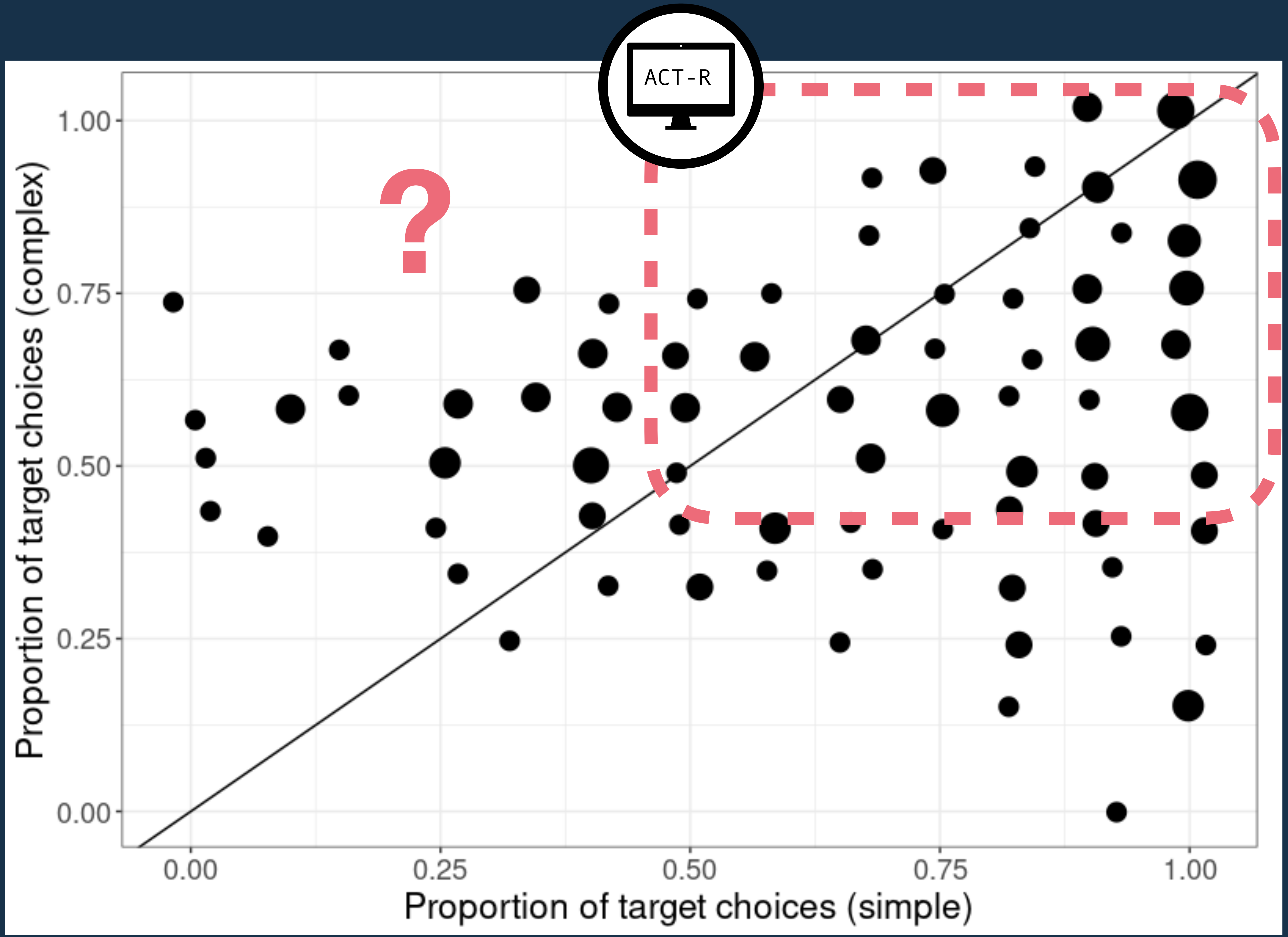
Comparing with a lower baseline



Modeling individual differences improves predictions

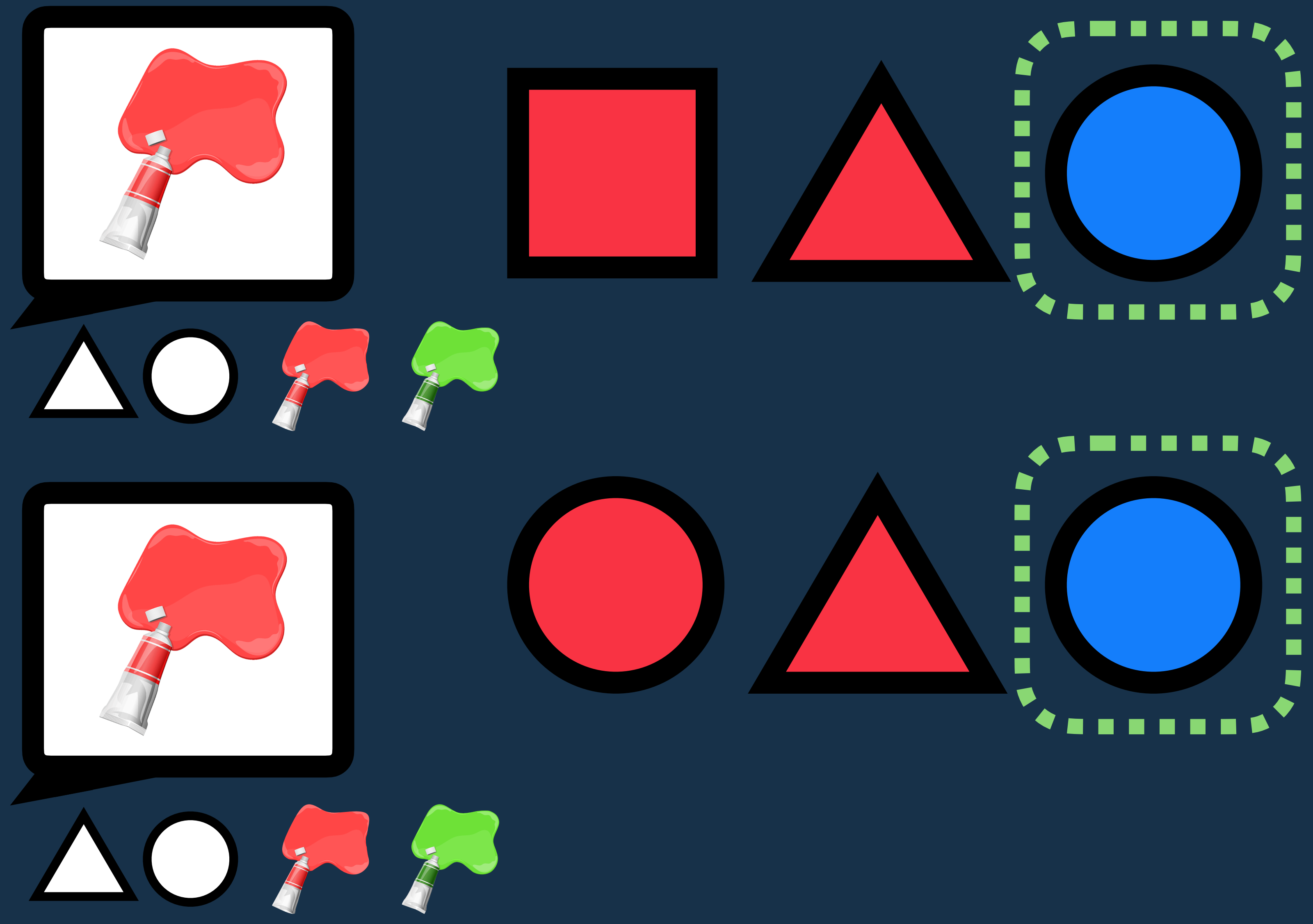
Modeling variable utility and “Odd One Out”

One unmodeled aspect of behavior



Mayn & Demberg (2023)

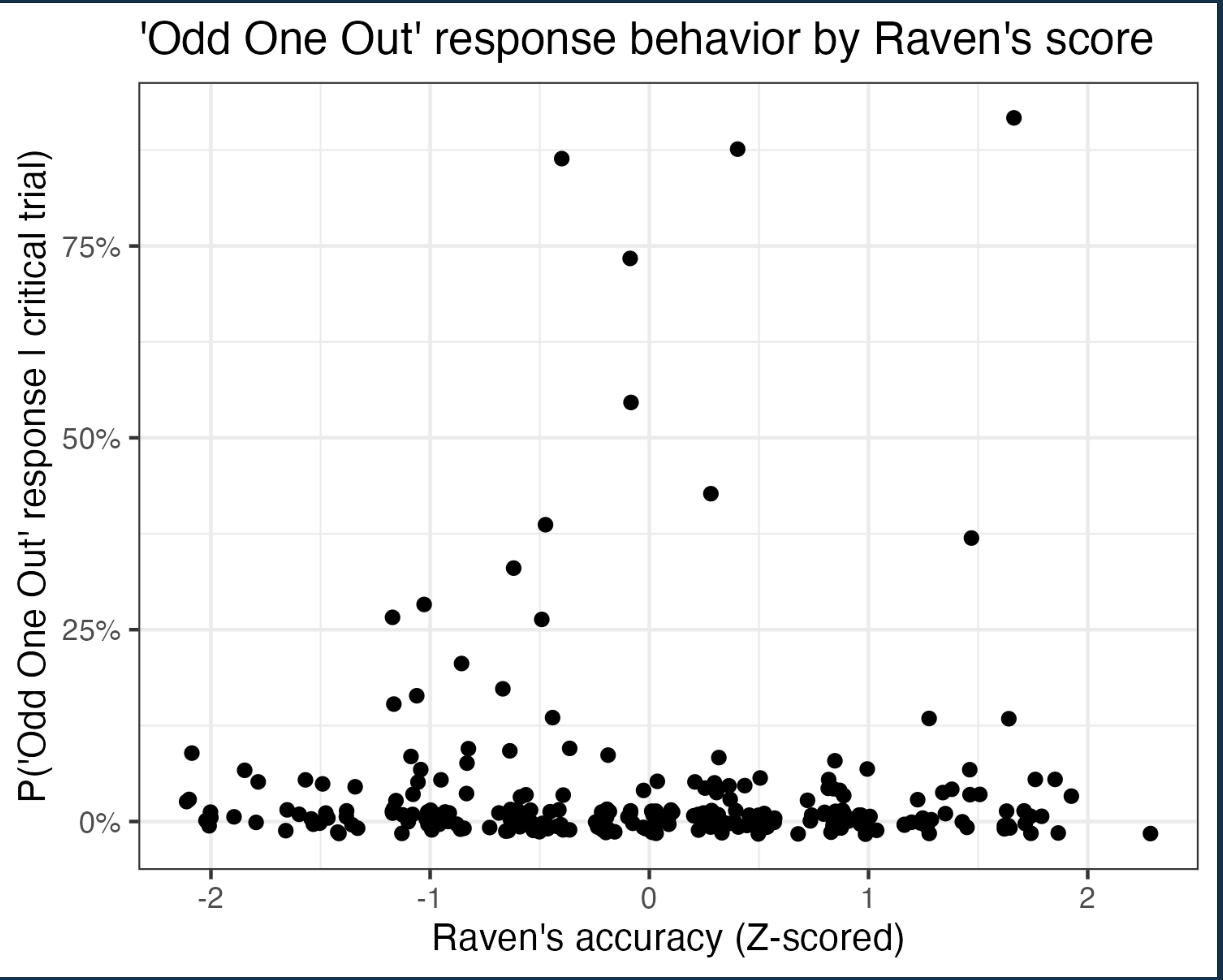
(n = 173, 12 obs/condition)



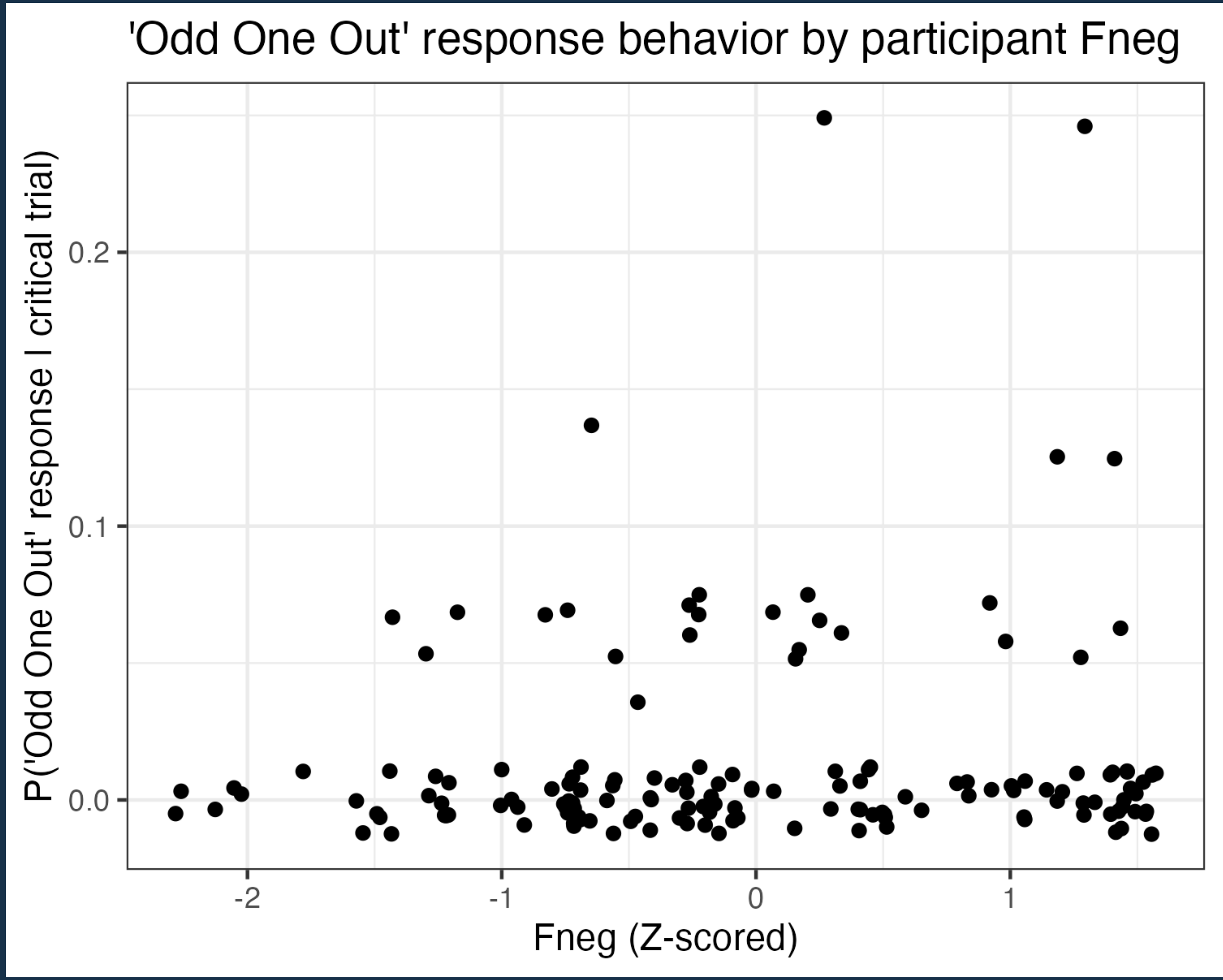
Mayn (2023): Some participants report an "Odd One Out" strategy.

Characteristics of OOO-responders

Some evidence that rapid learners are more likely to be unconventional.

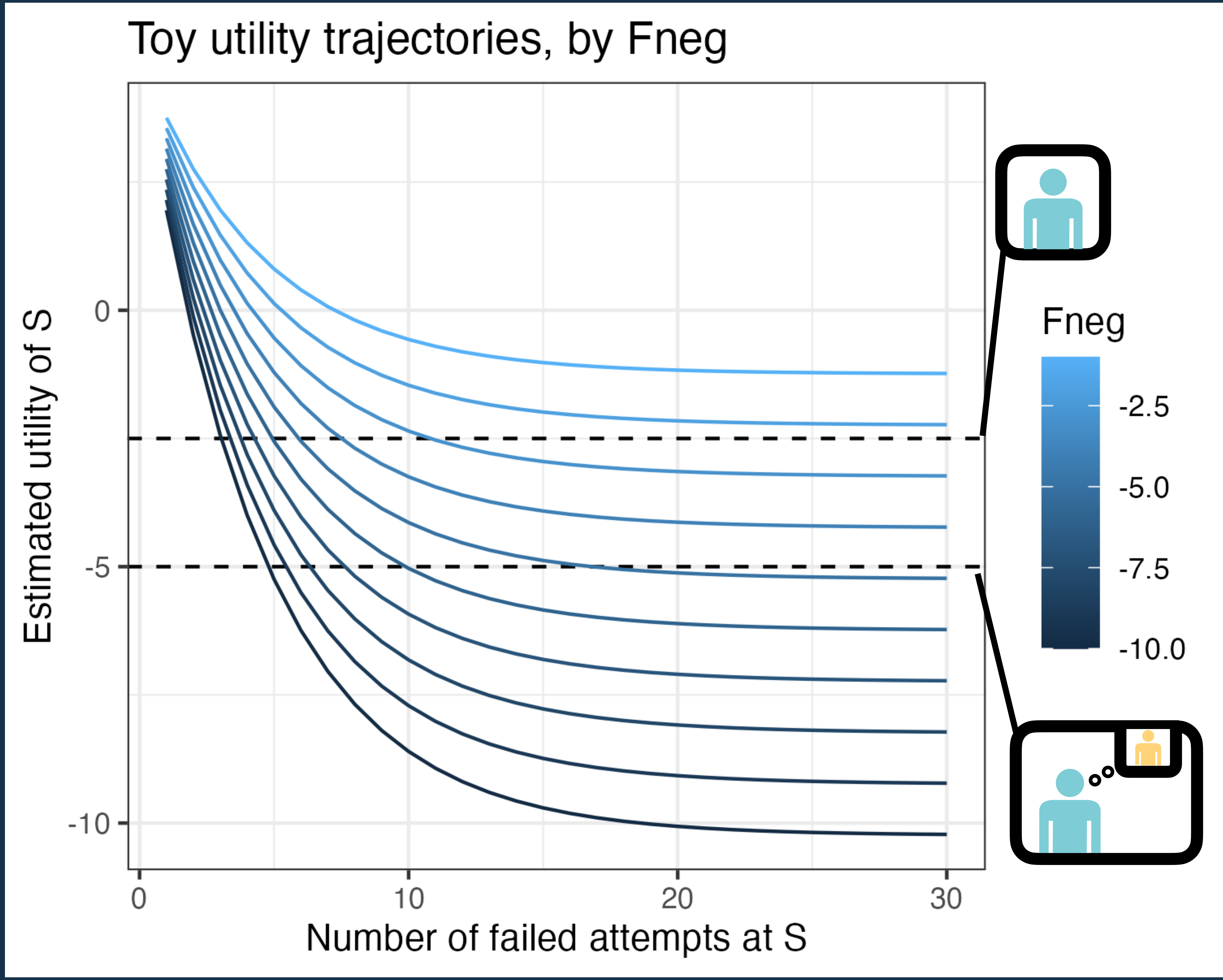


Mayn & Demberg (2023)



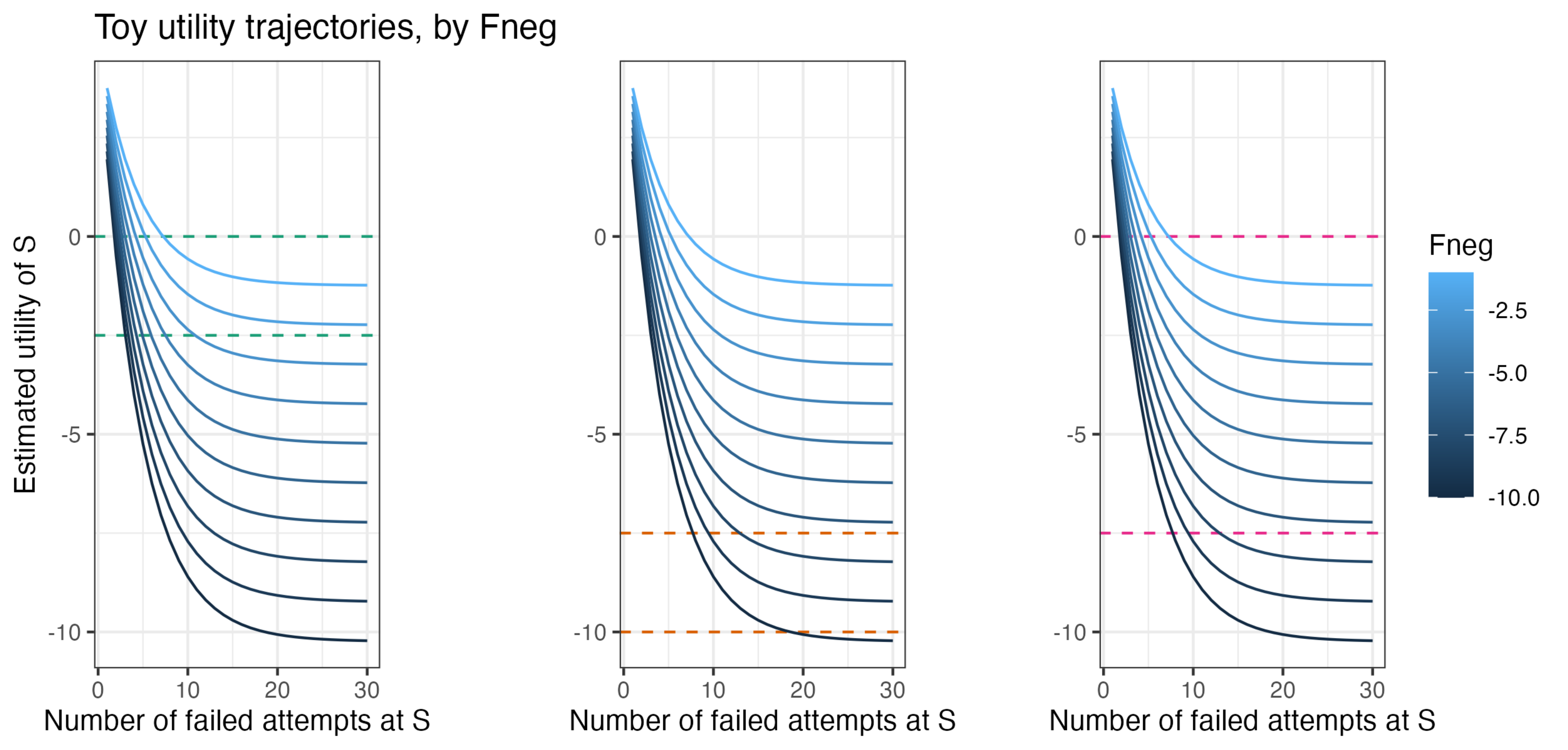
current study

How procedural learning works in ACT-R

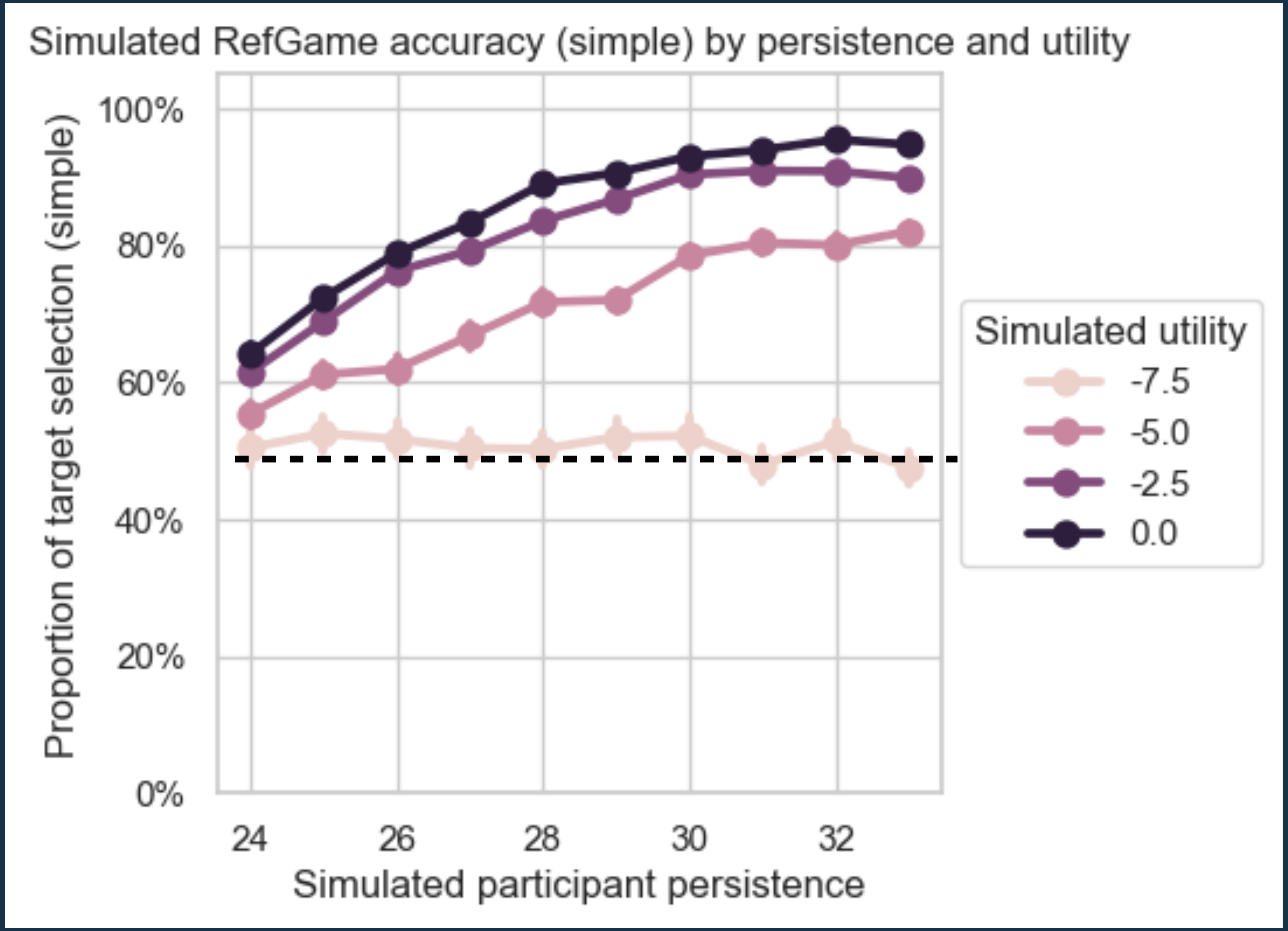


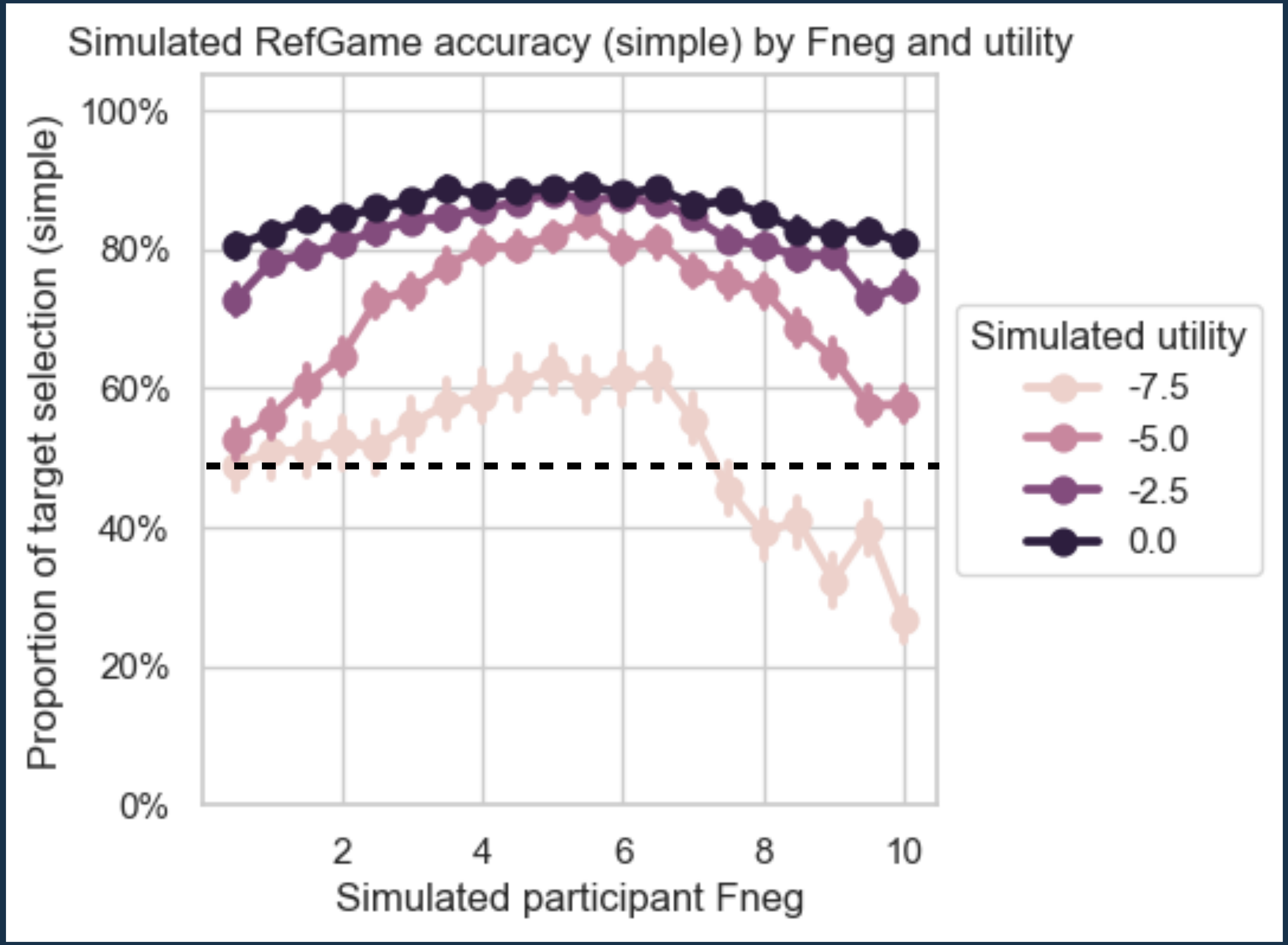
- ACT-R uses temporal difference learning, gradually updating estimated utilities towards their actual rewards
- Fneg determines that reward, therefore determines the floor for failed actions
- Actions which start with negative utilities can only be explored and adjusted in value if initial strategies can be penalized enough

Varying utility

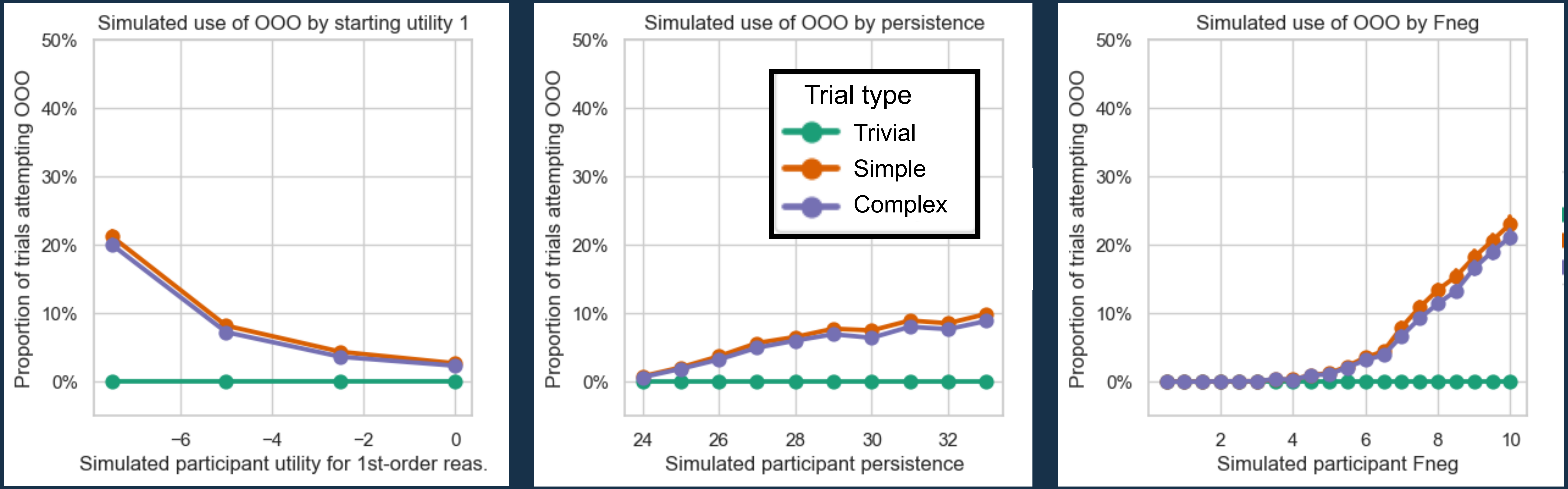


The effect of starting utility, and new exploration penalties

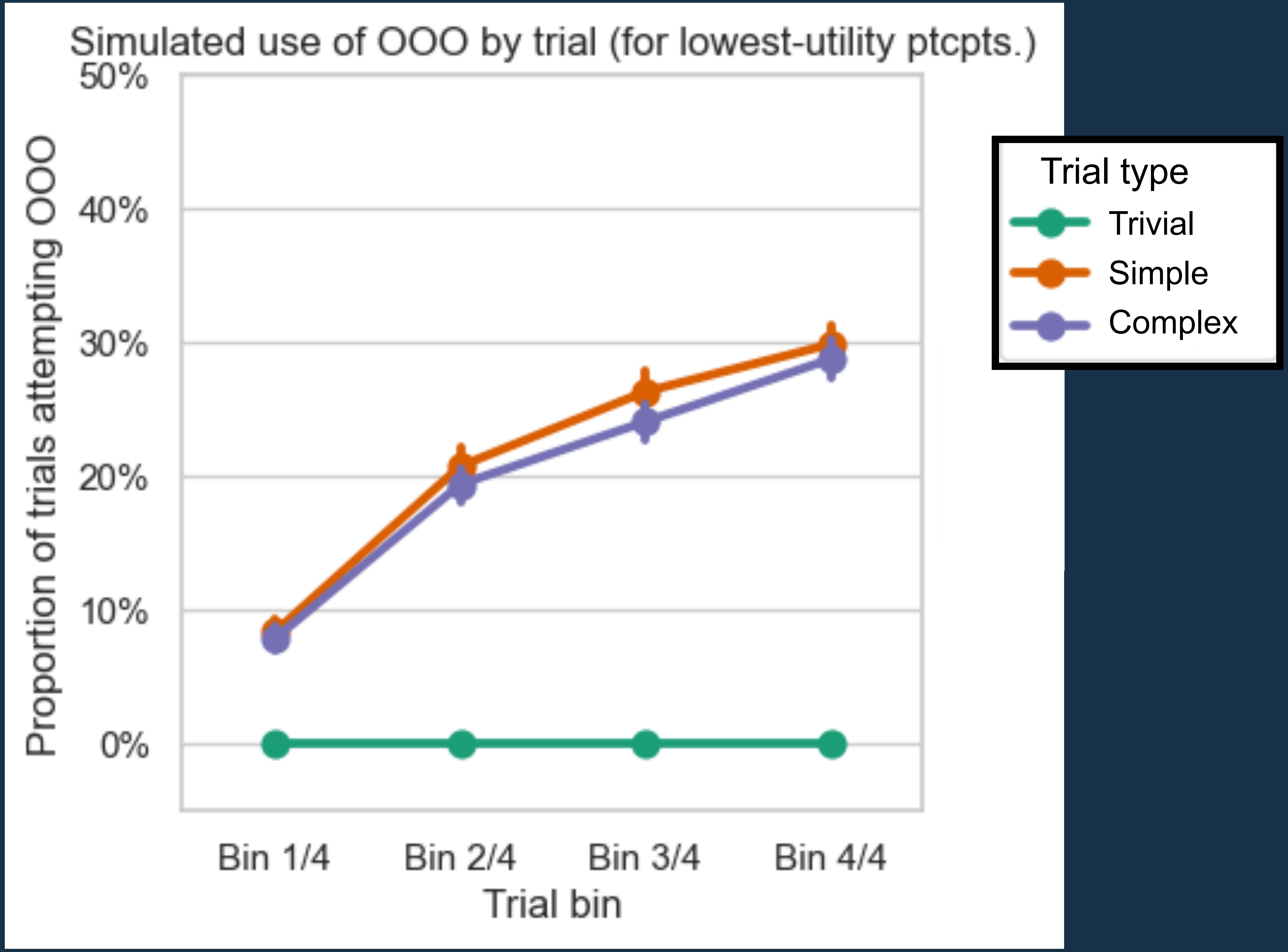




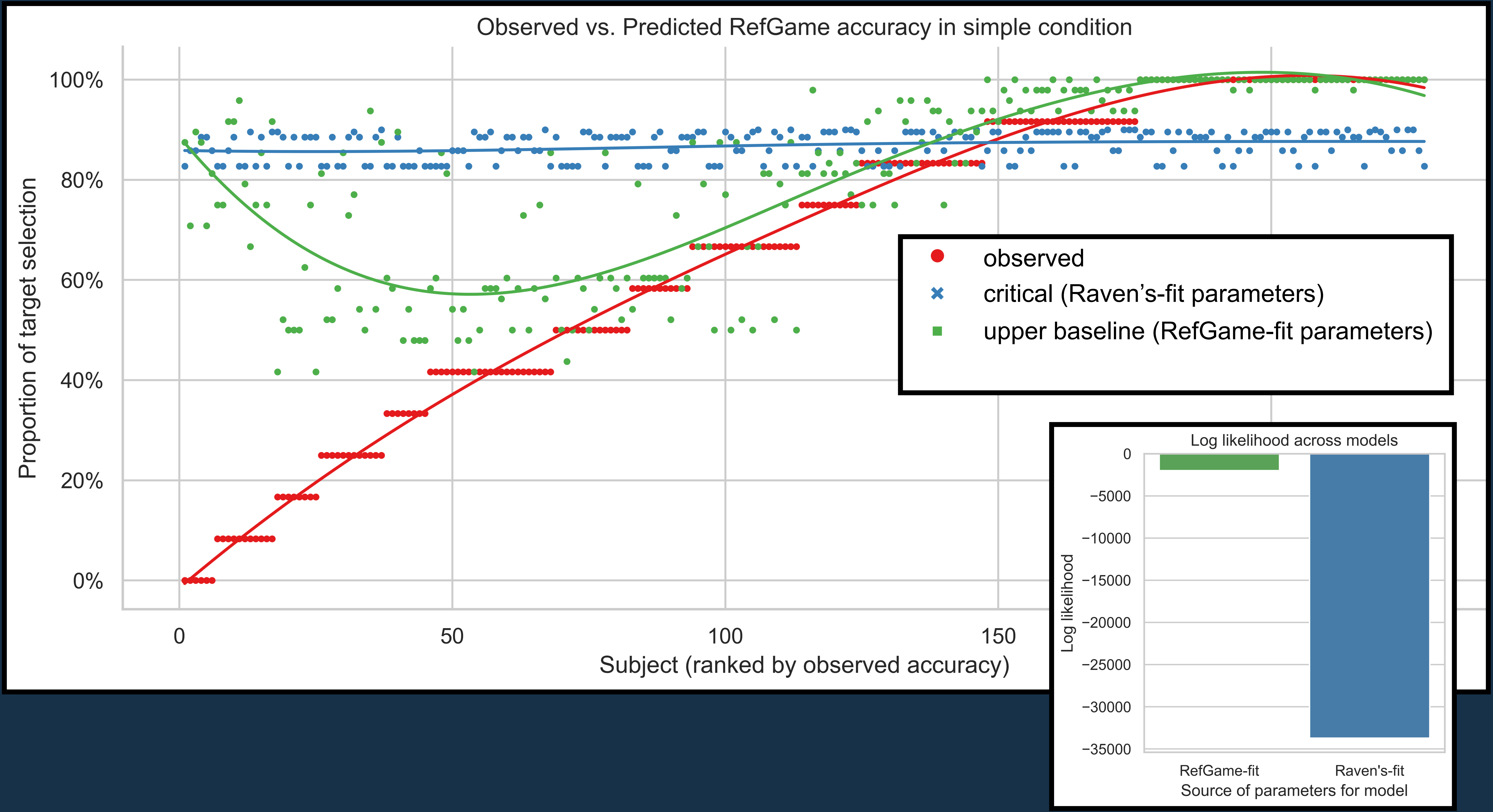
Variability in the discovery of OOO reasoning



Timecourse of discovery of OOO reasoning



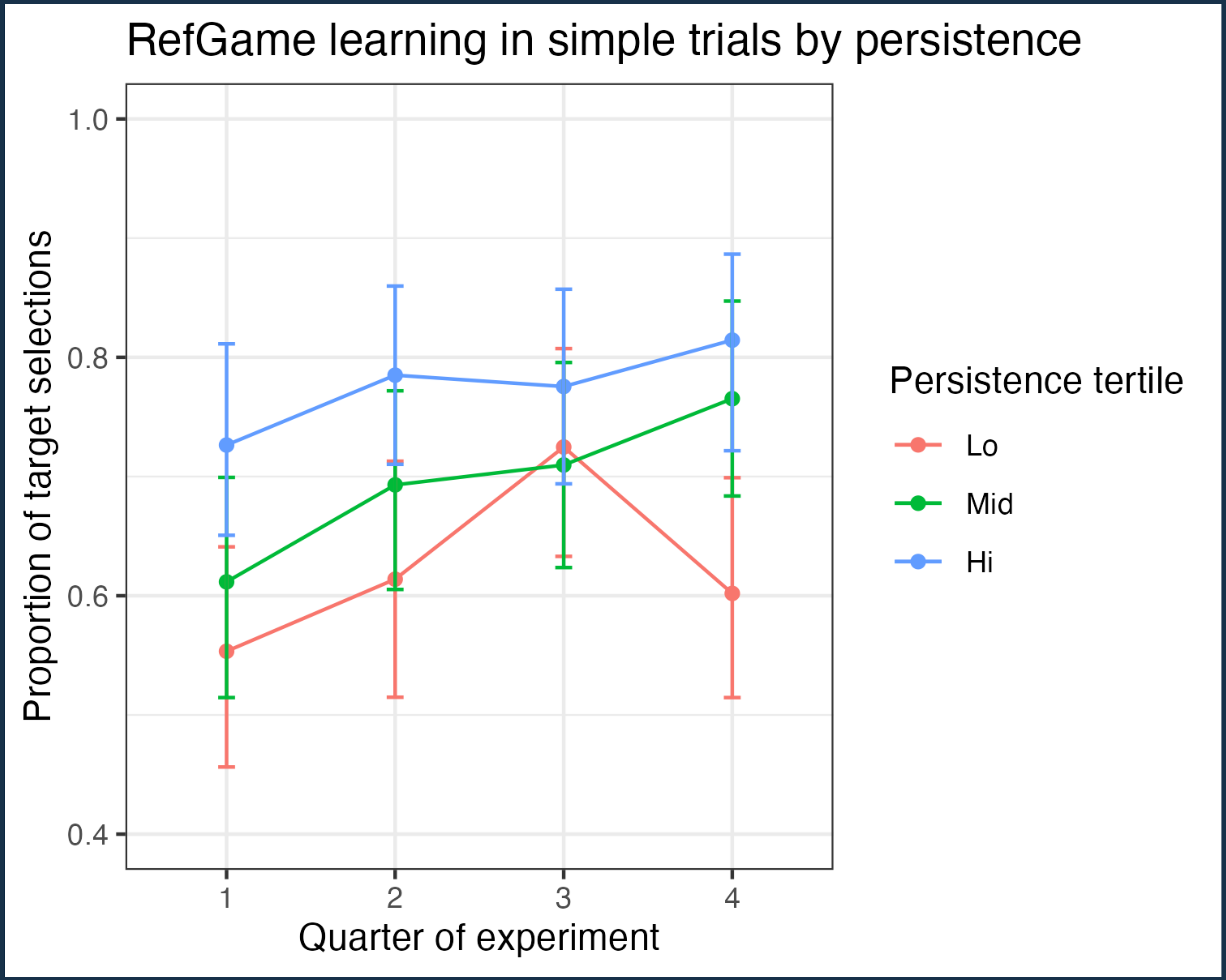
Predictions based on this model

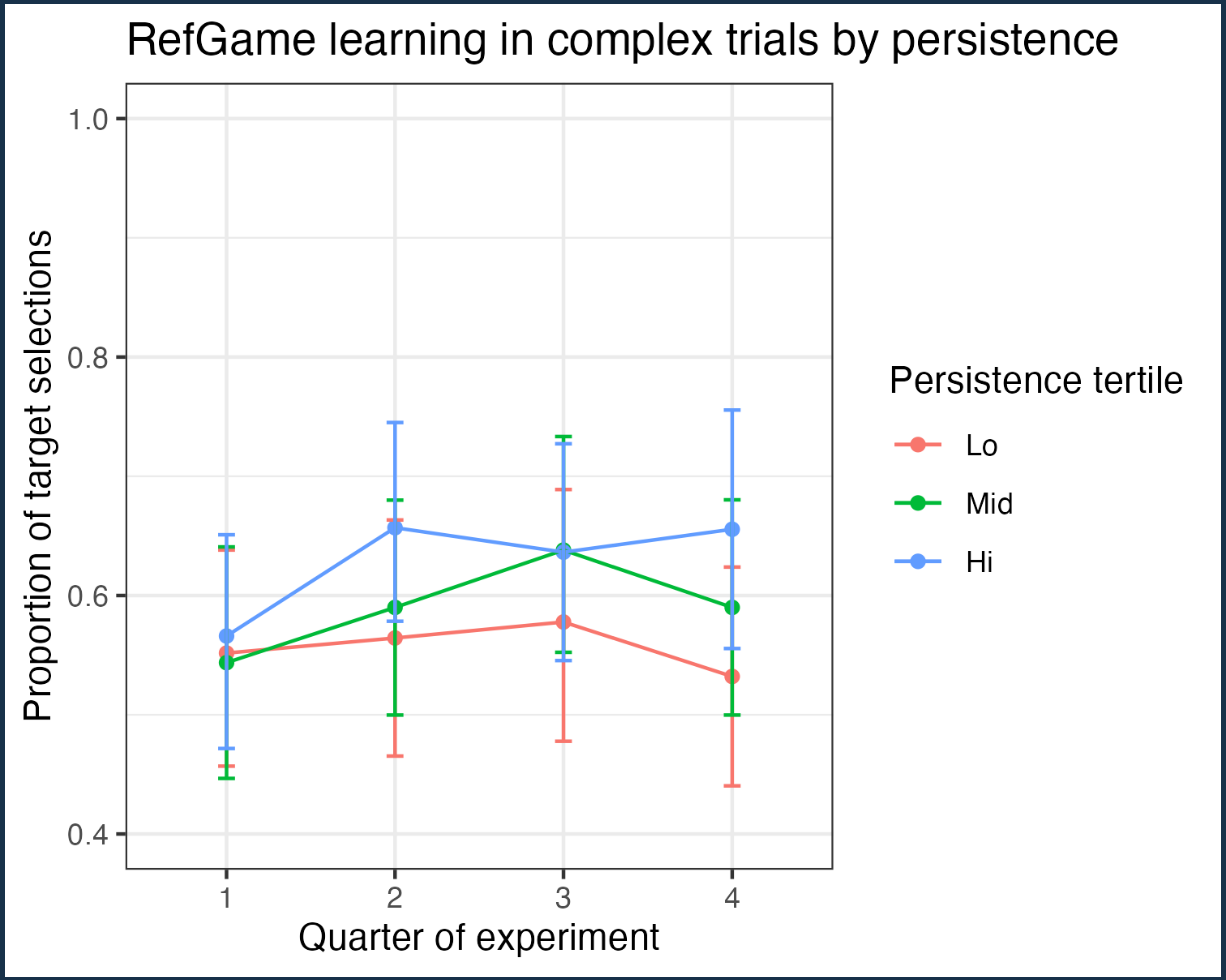


- No way to estimate initial utilities from Raven's, worse fits due to new uncertainty
- Self-fit is rather good now, except for the worst participants

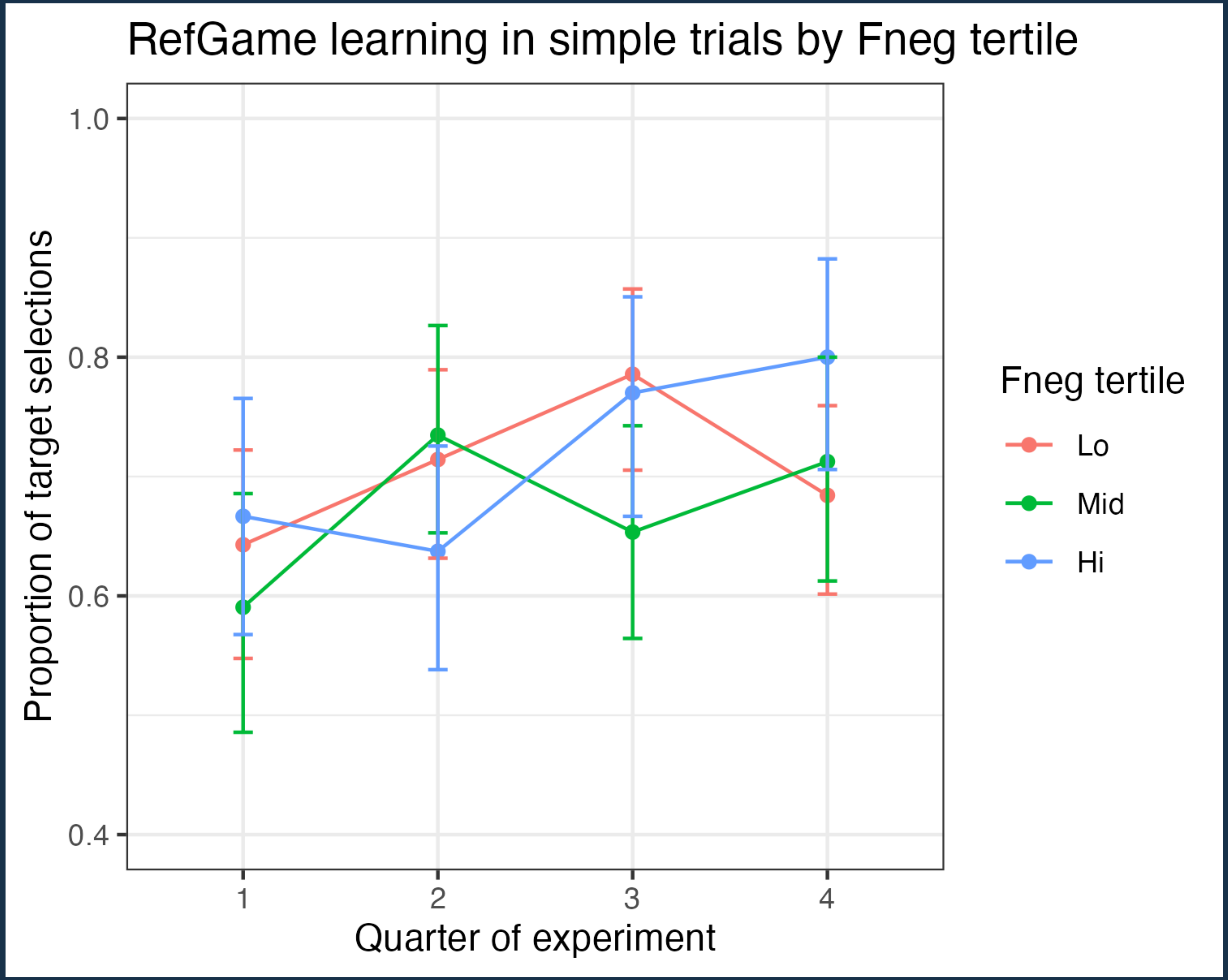
ID effects on learning trajectories

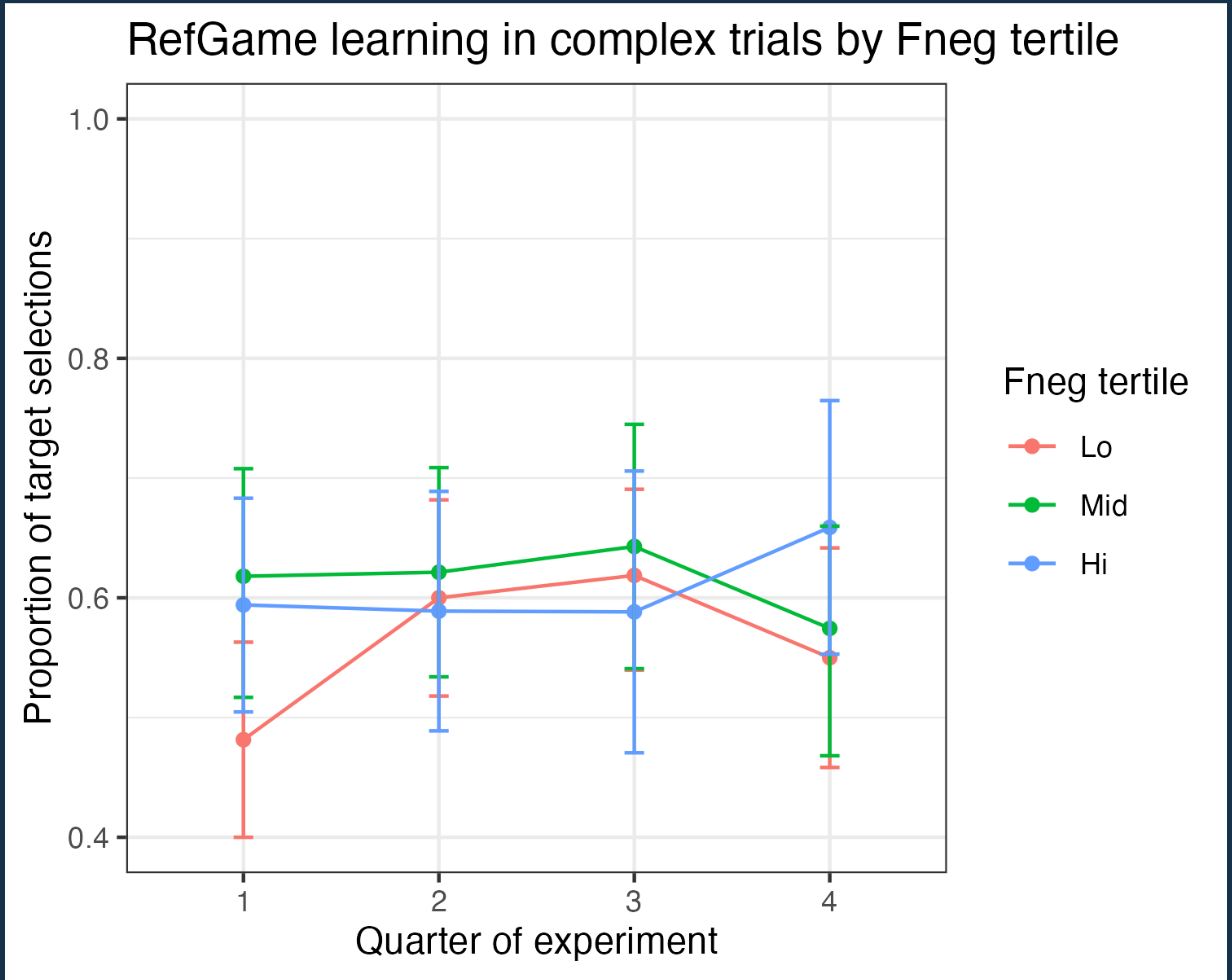
Persistence may indeed modulate learning





FNEG variation seems too noisy to tell

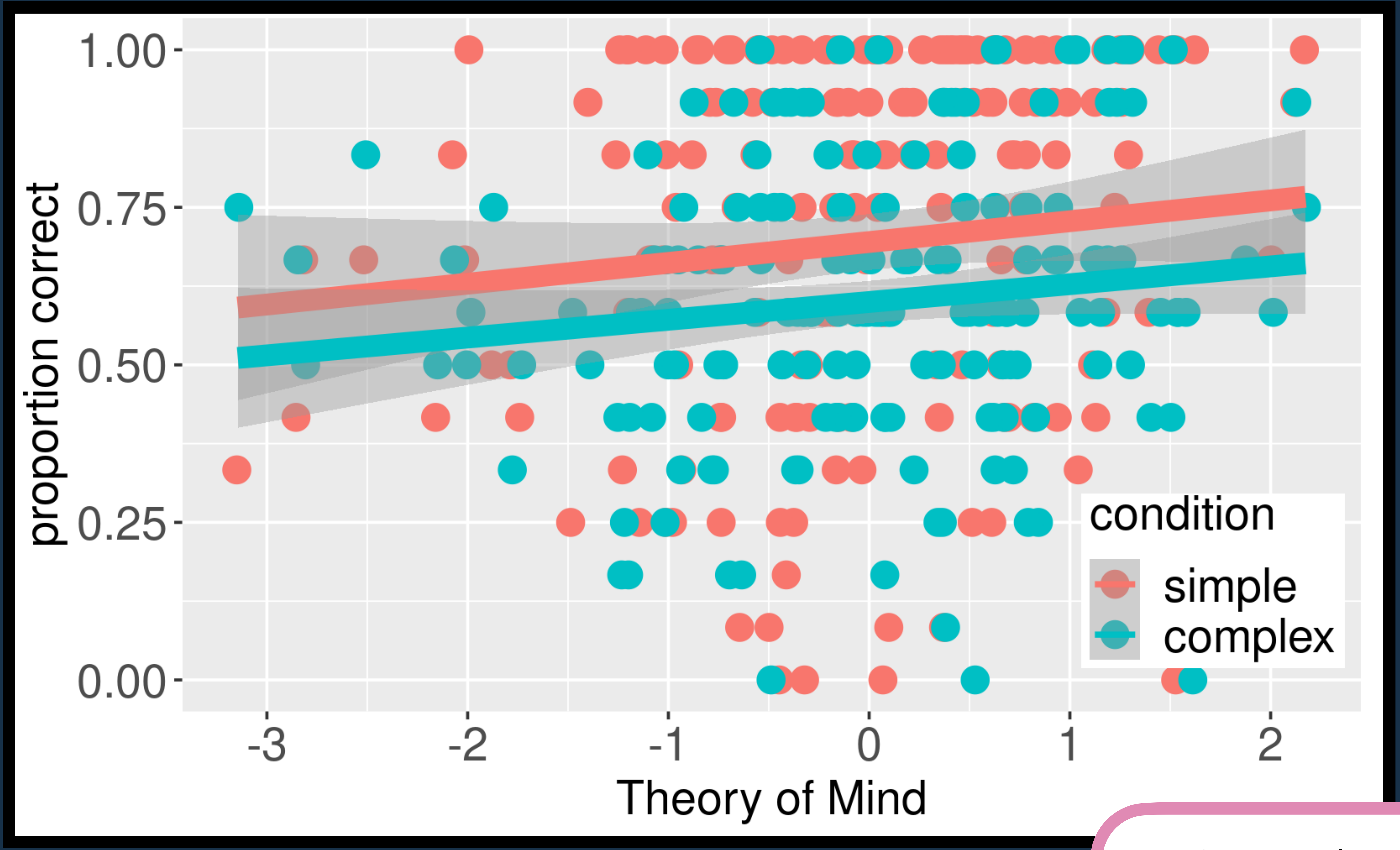




The role of Theory of Mind

Correlations with Theory of Mind ability

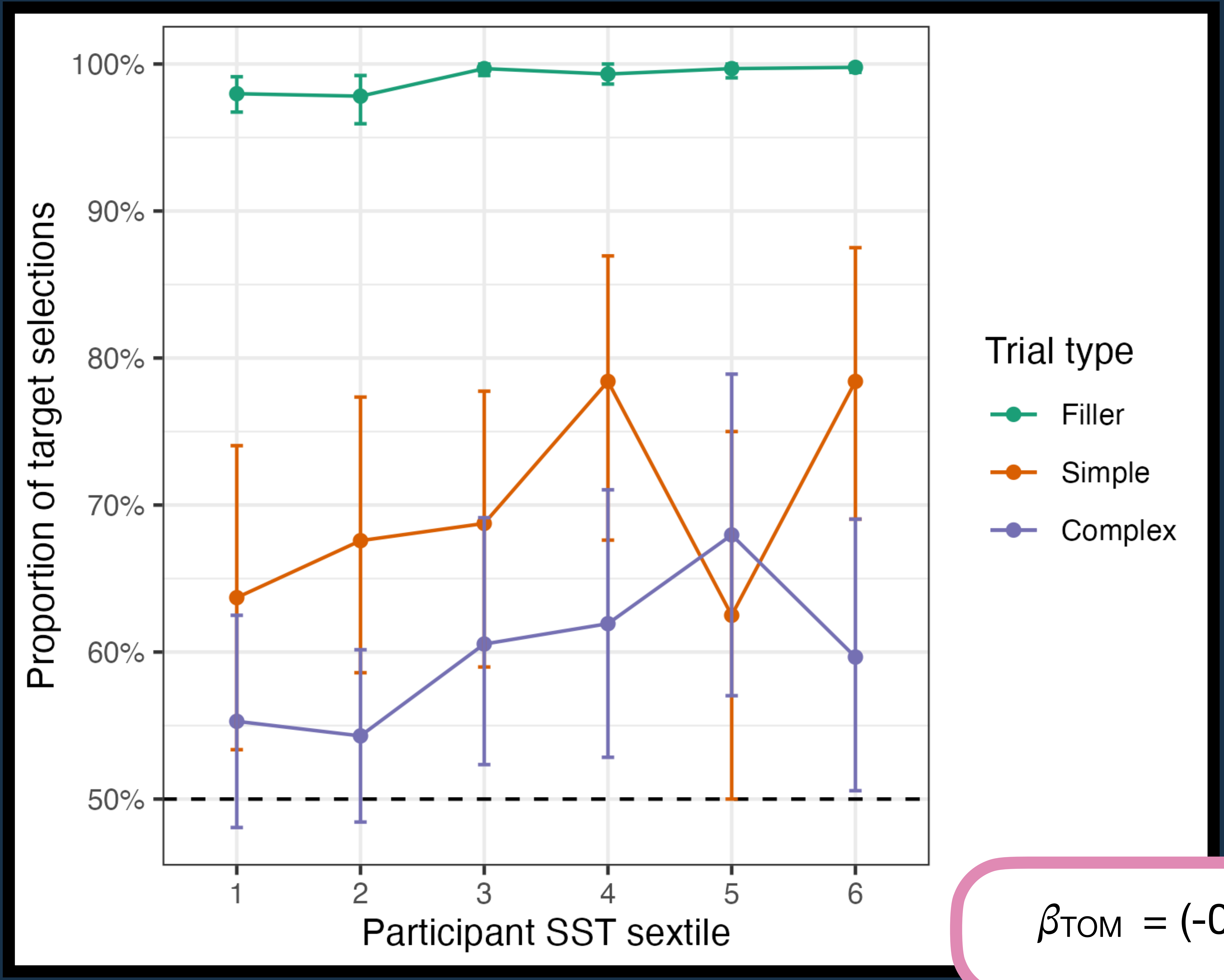
:= Reading the Mind in the Eyes + Short Story Task



$\beta_{TOM} = (0.01, 0.19)_{95\%}$

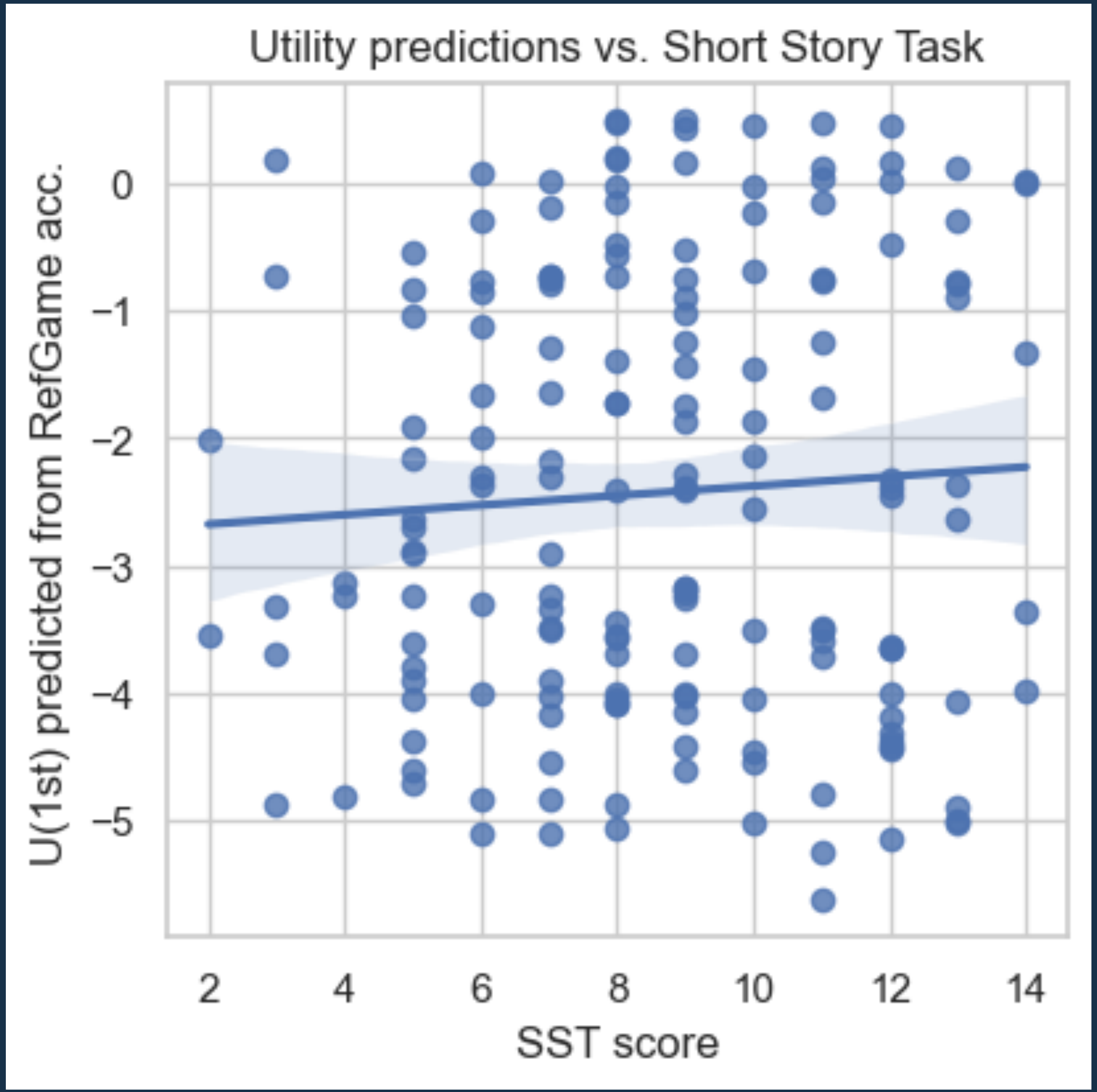
Mayn & Demberg (2023)

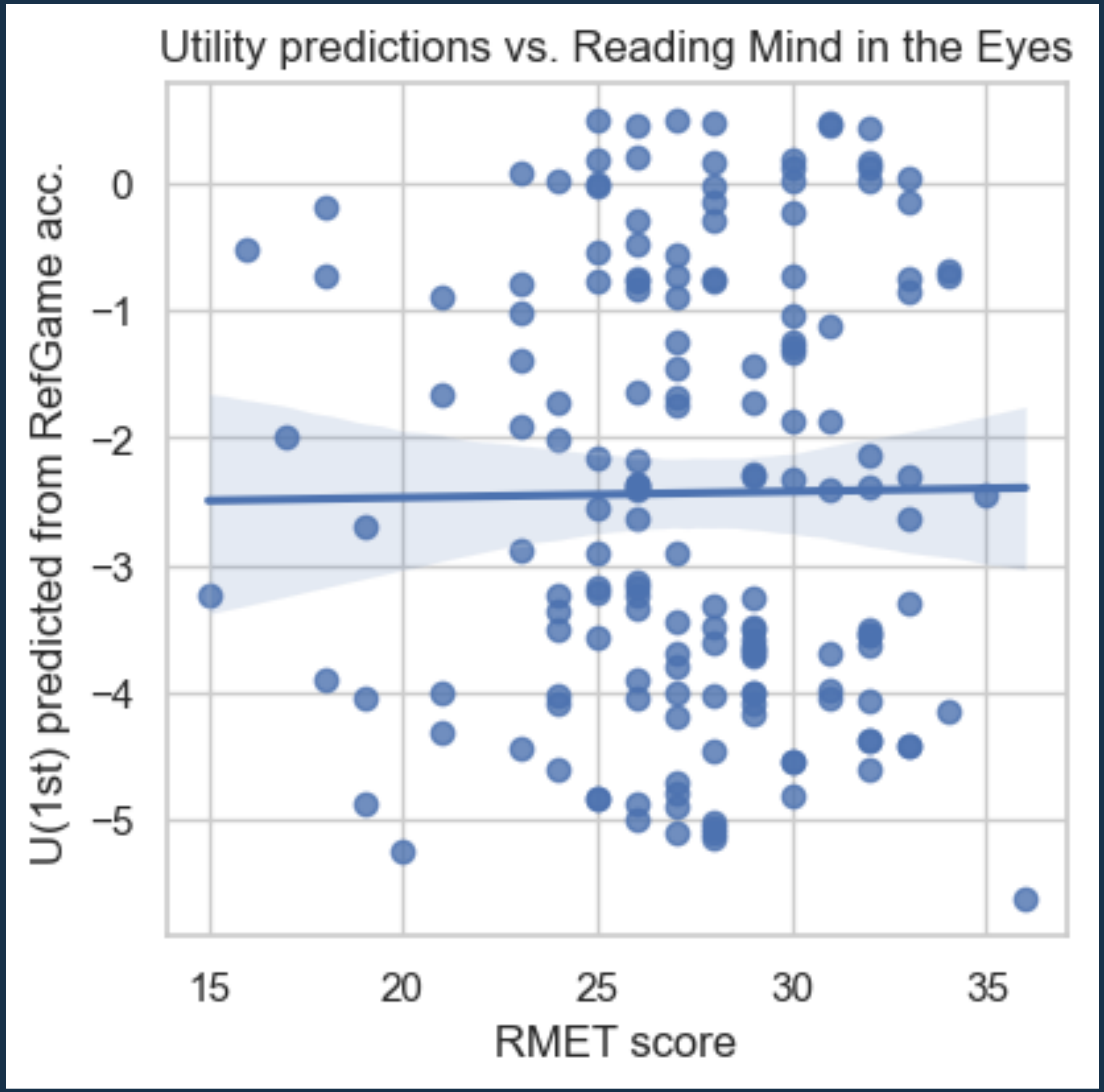
Replicated here merely as a trend



β_{TOM} = (-0.06, 0.42)$_{95\%}$

Theory of Mind tasks don’t track ACT-R estimated utilities

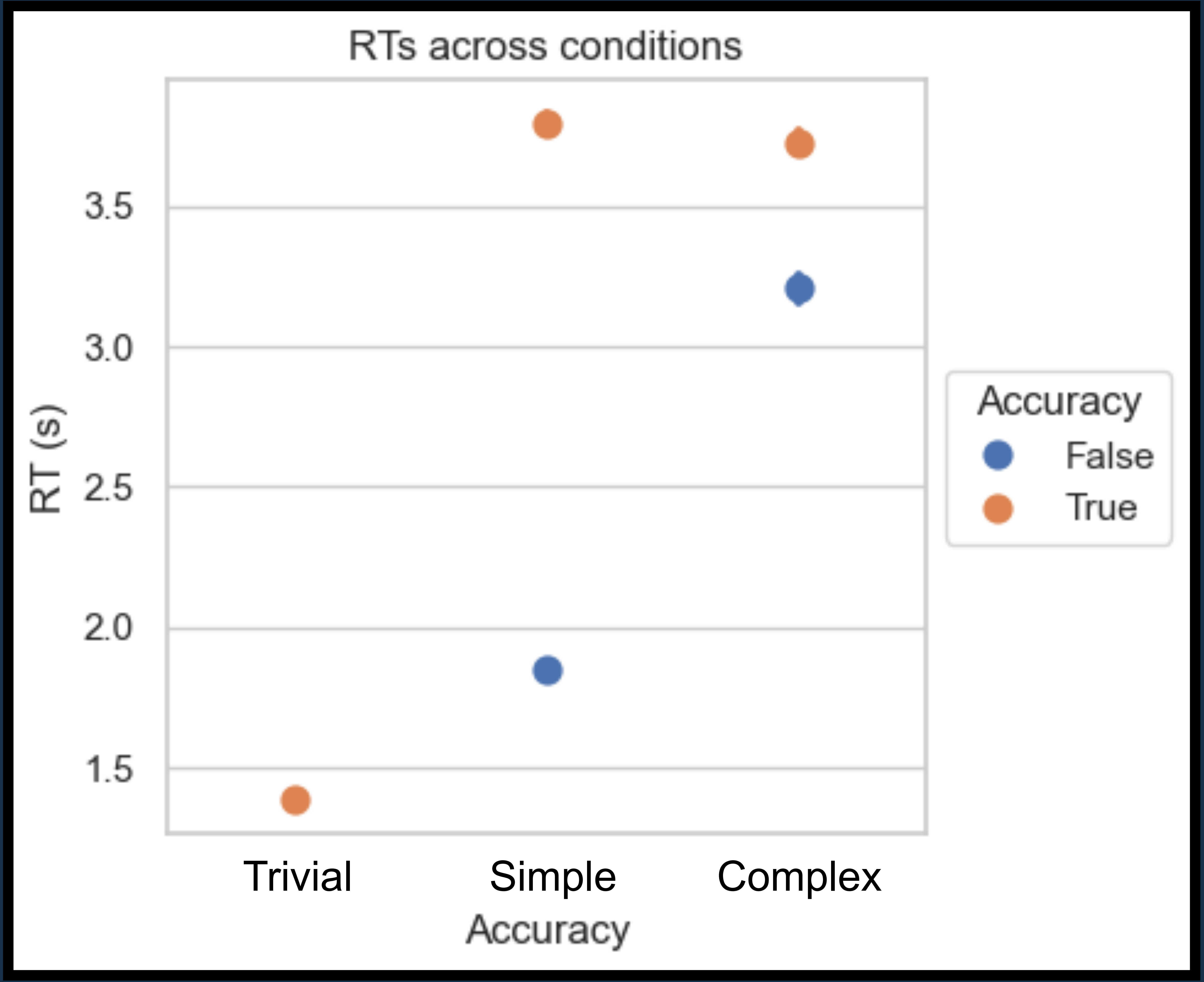




(using data from Mayn & Demberg, 2023)

More details on other tasks

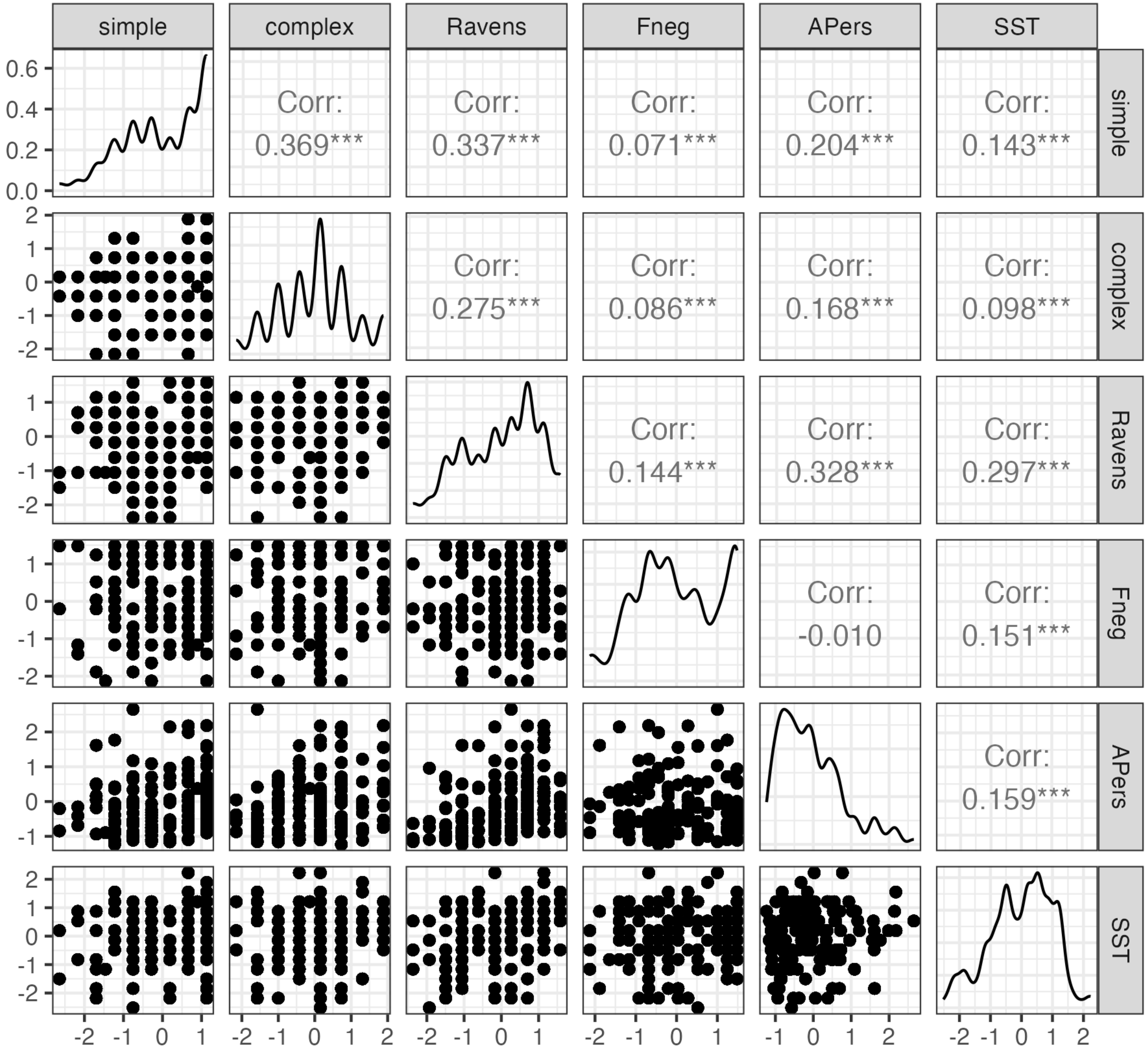
Further behavioral prediction: Variation in RTs



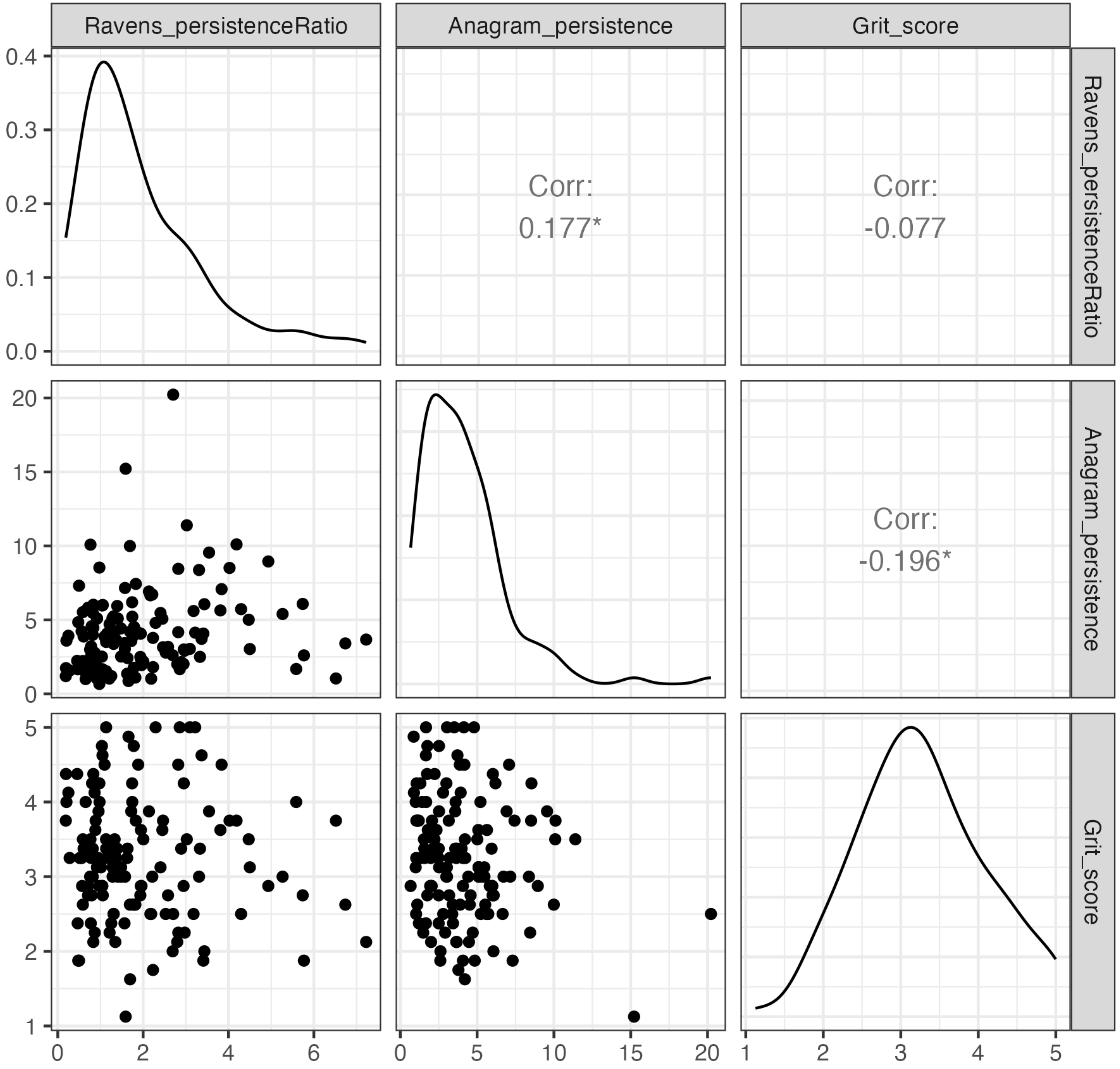
- Slower responses in more complex trials.
  - More complex reasoning, and more rounds of rejecting easier strategies.
- Trials with correct answers should be slower than incorrect.
  - Incorrect answers come from low-persistence participants.

Correlations among critical individual difference measures

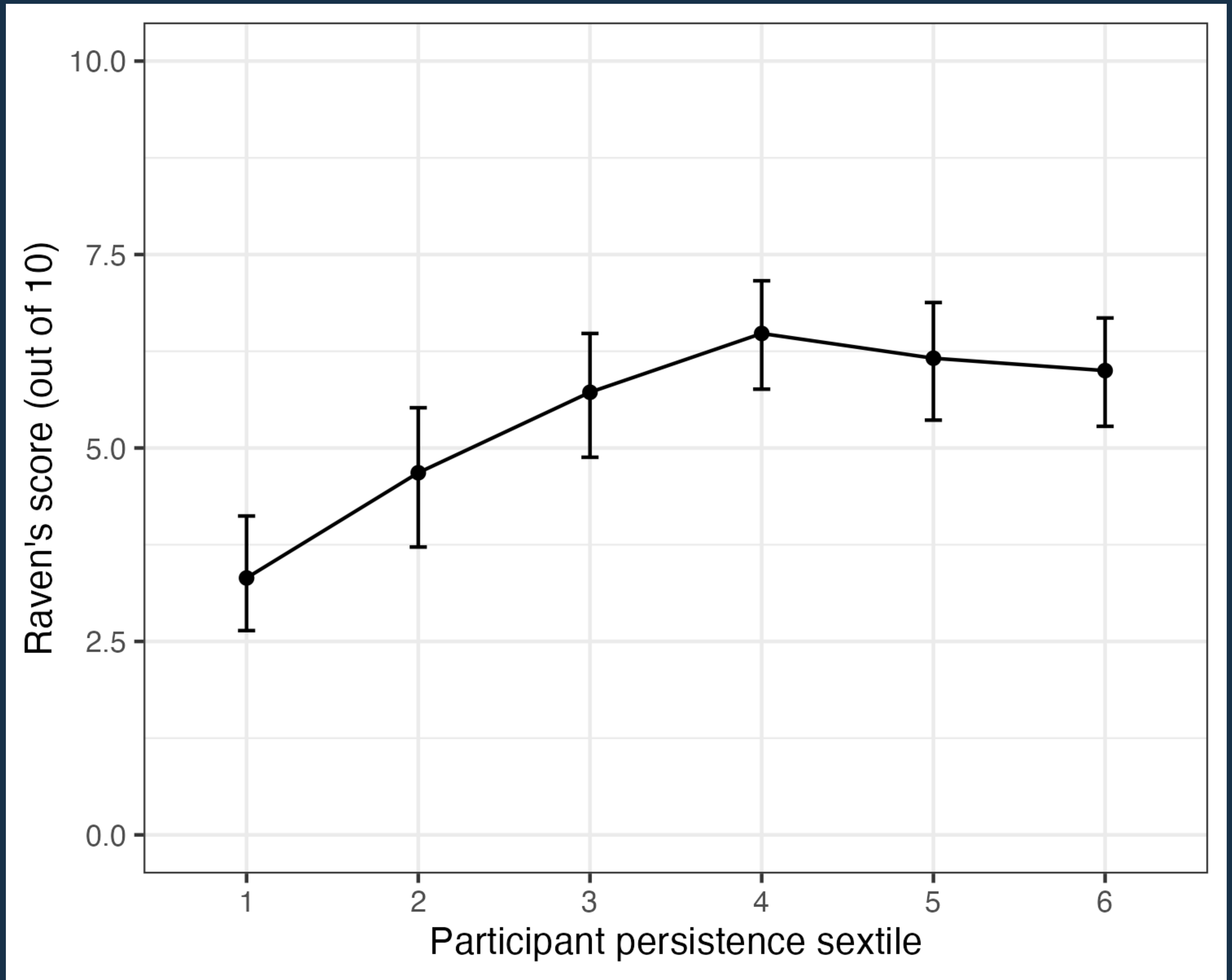
(z-scored and trimmed)

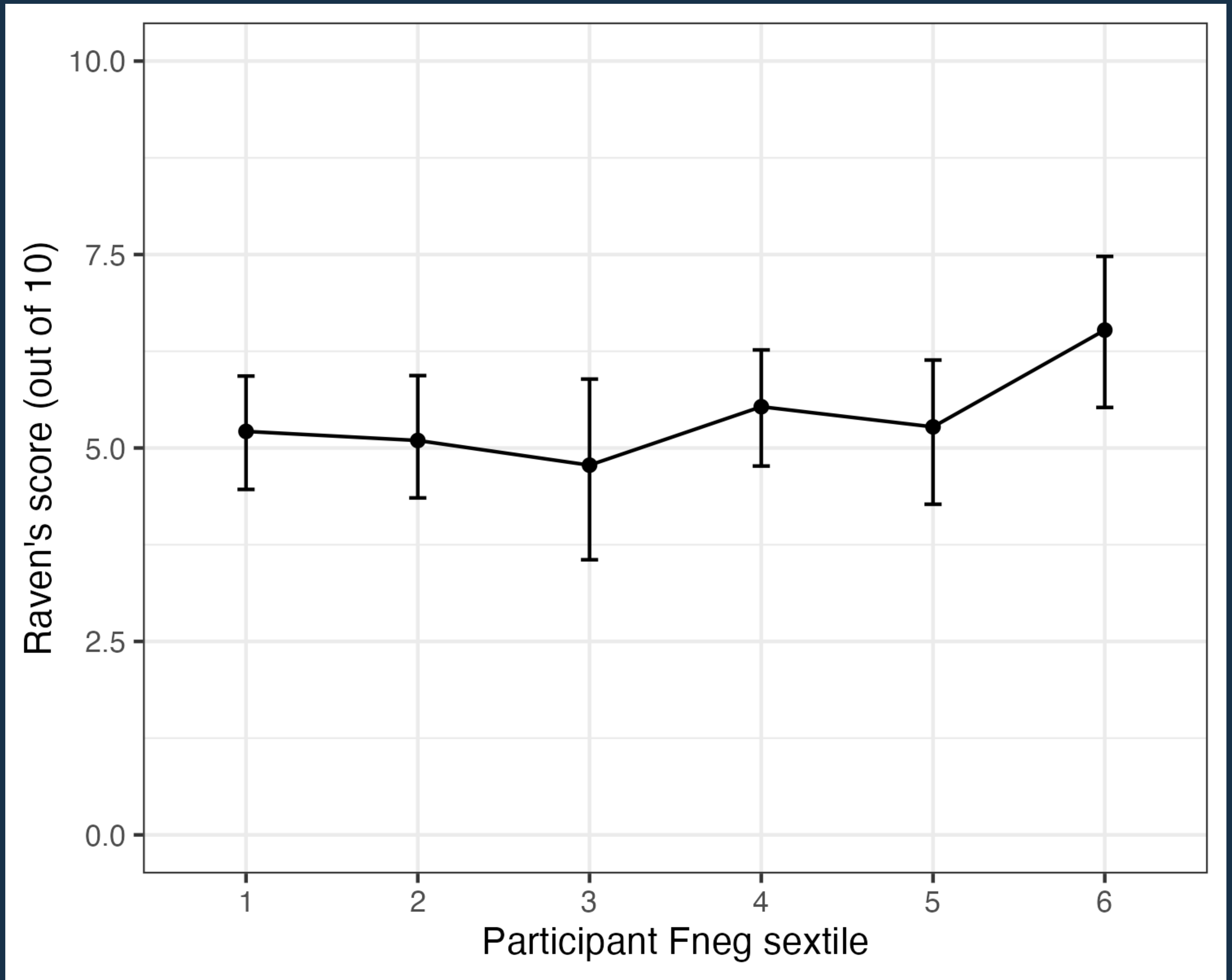


Measures of persistence



IDs in Raven's performance

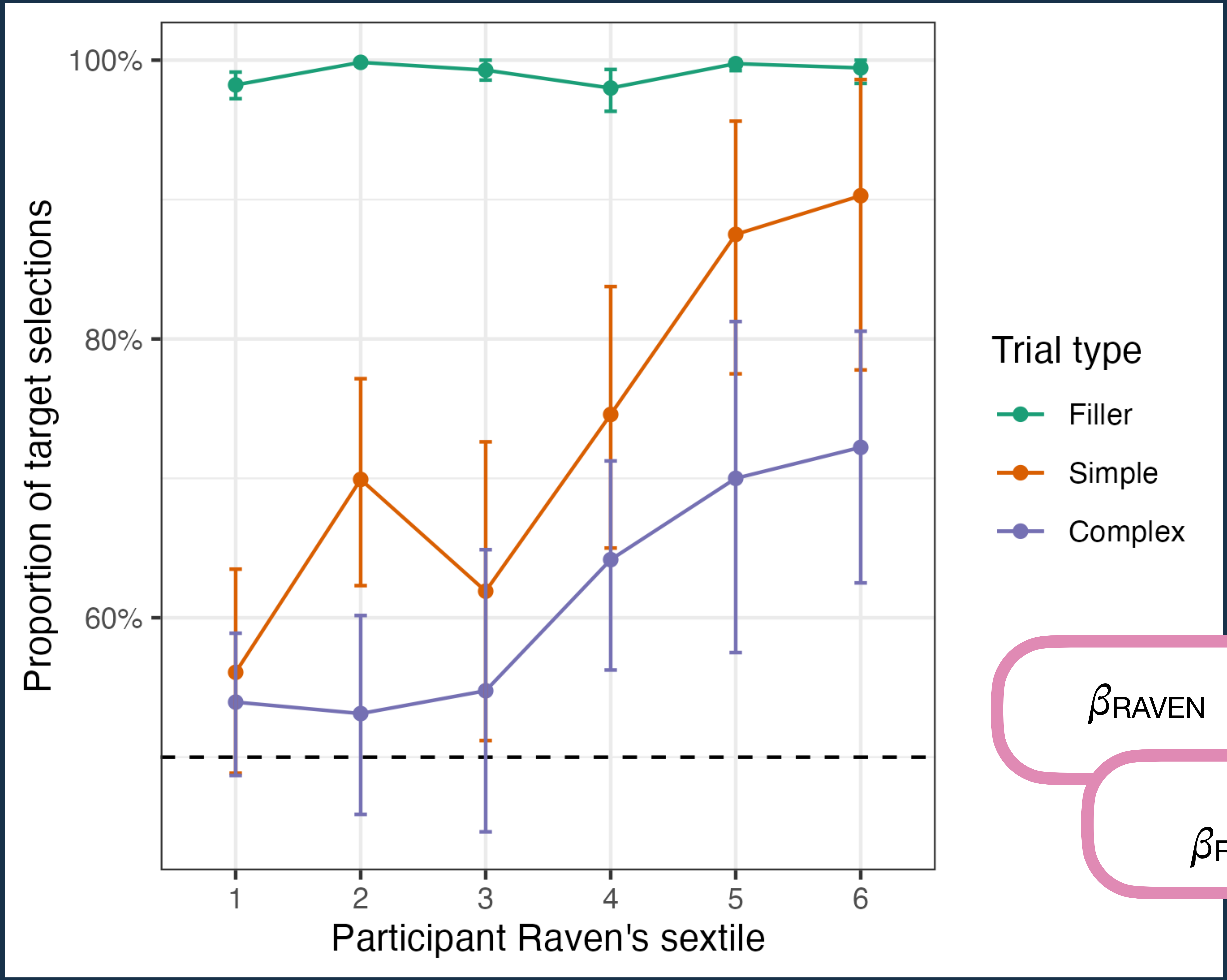




$\beta_{PERS} = (0.22, 0.45)_{95\%}$

$\beta_{FNEG} = (0.03, 0.25)_{95\%}$

Replicating the Raven's correlation



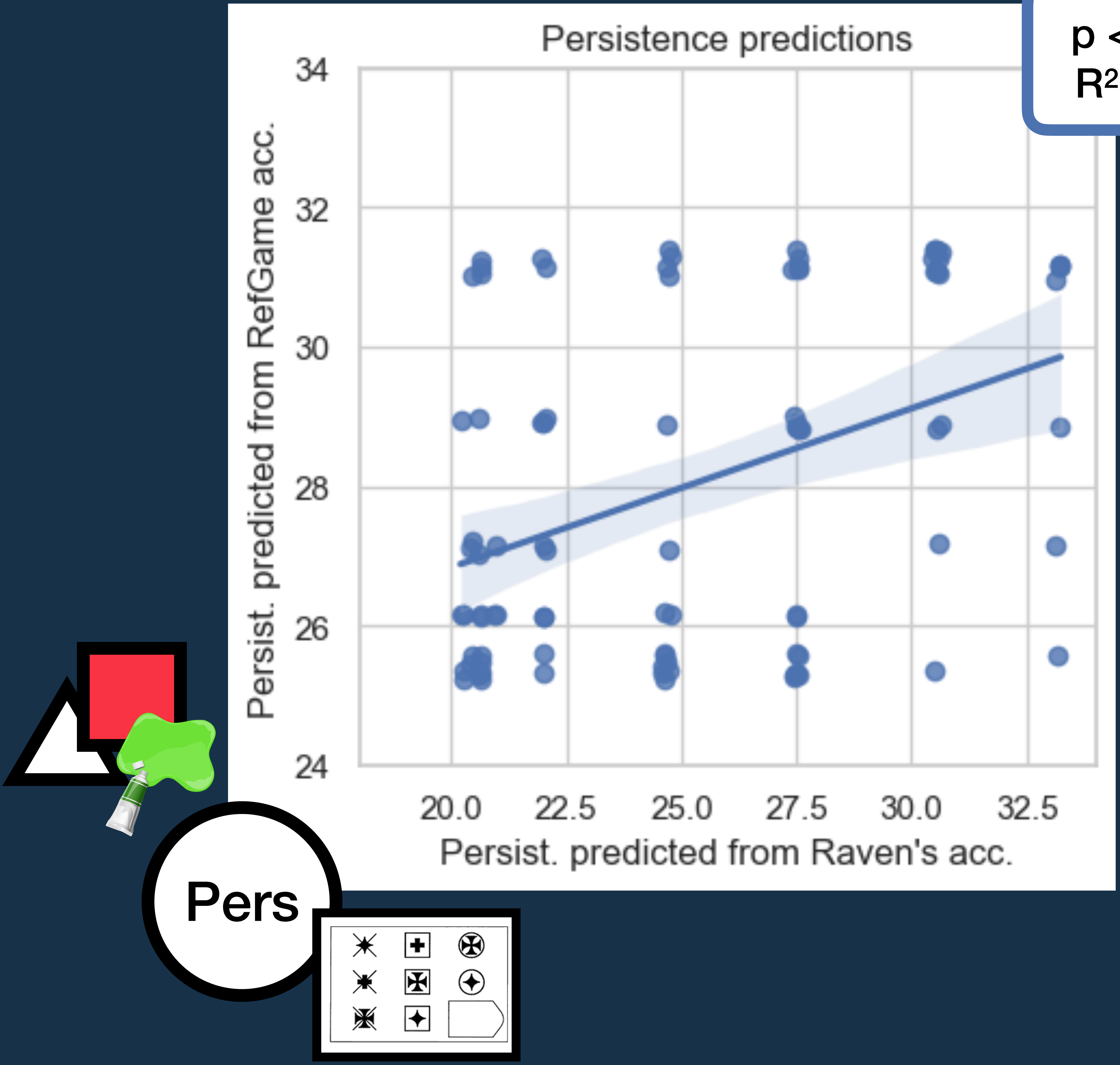
$\beta_{RAVEN} = (0.31, 0.71)_{95\%}$

even with other IDs:

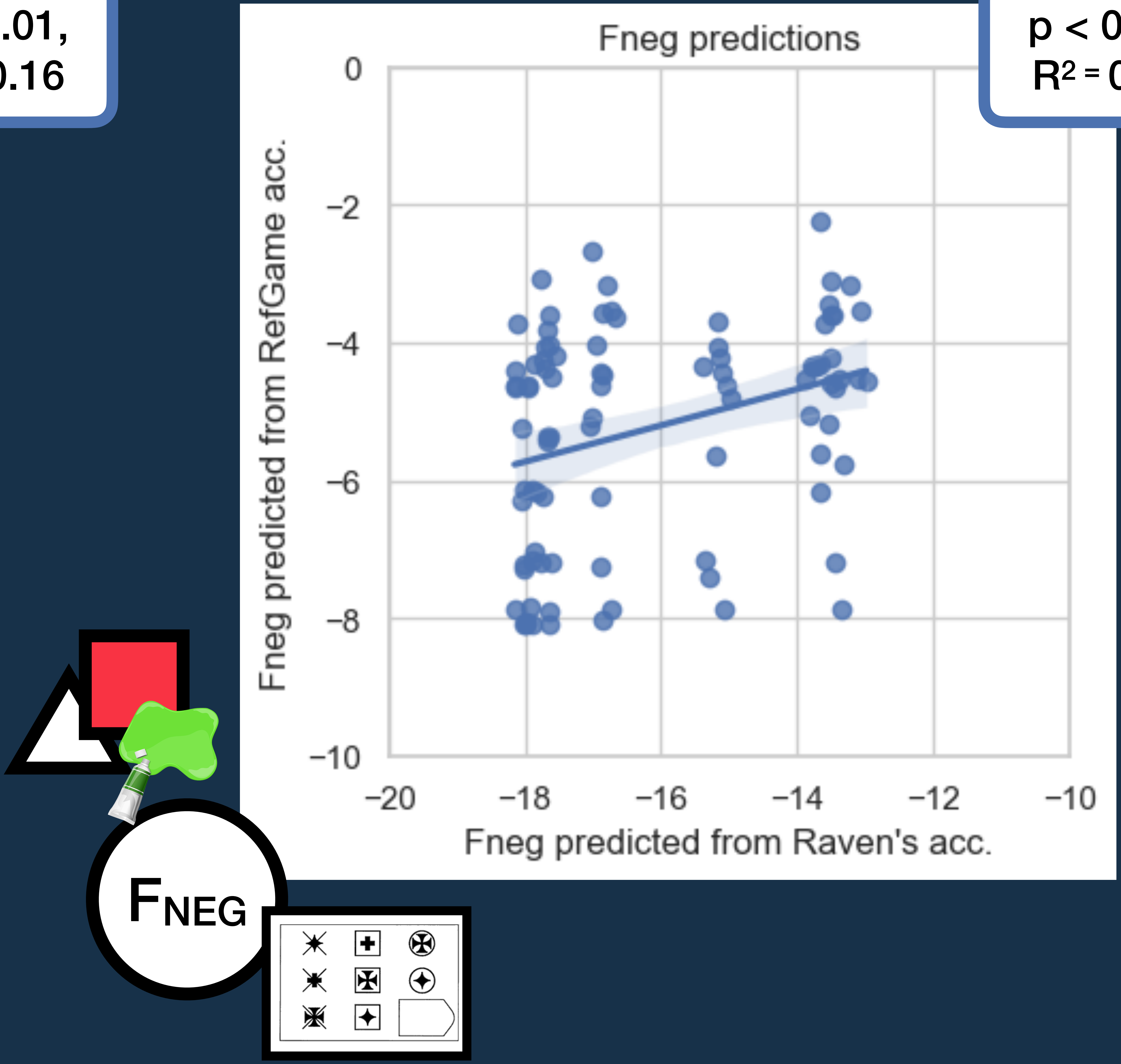
$\beta_{RAVEN} = (0.29, 0.79)_{95\%}$

Paradoxical relationships between parameter estimates and task measures

Parameter estimates again correlate across tasks

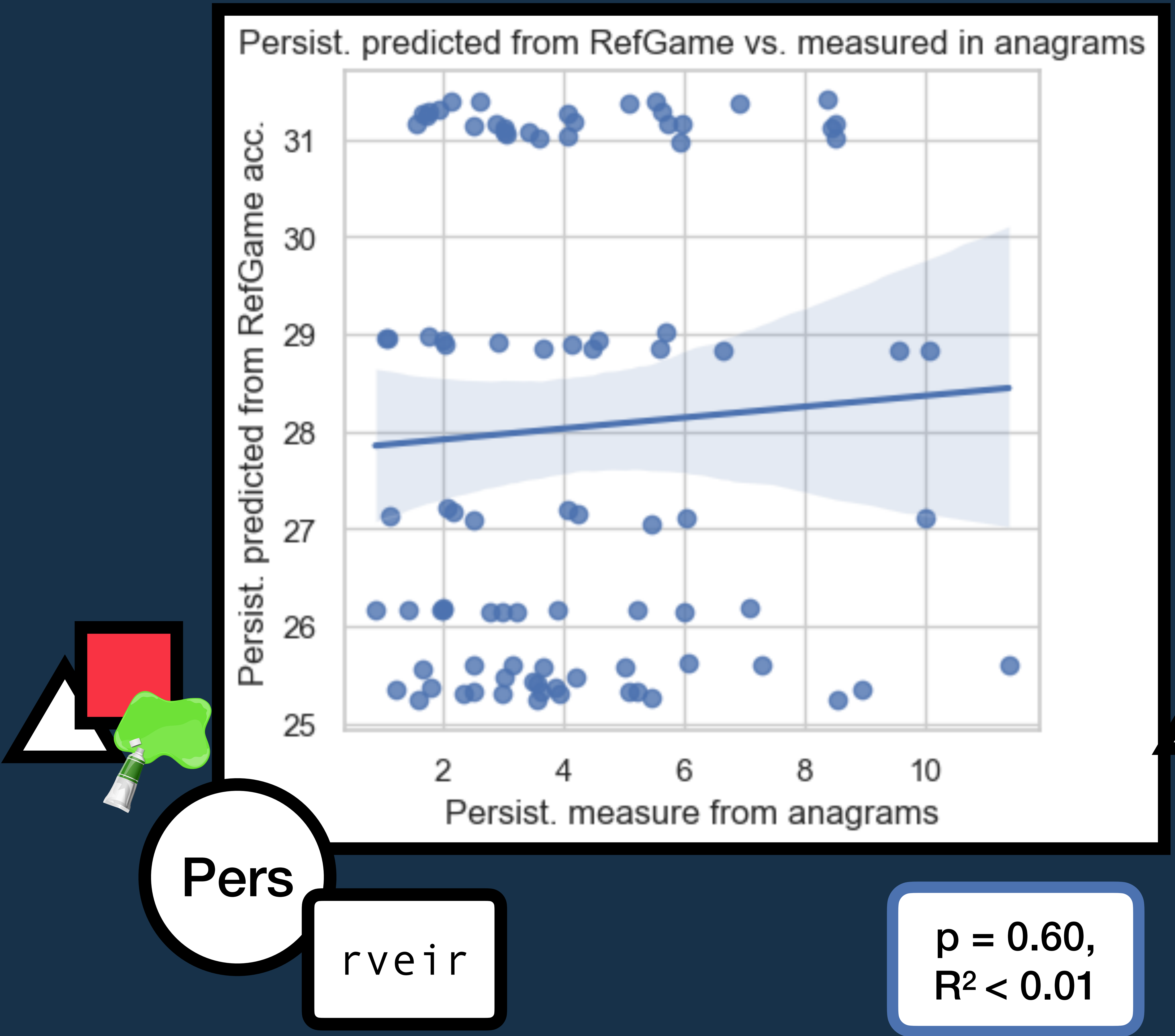


p < 0.01,
$R^2 = 0.16$

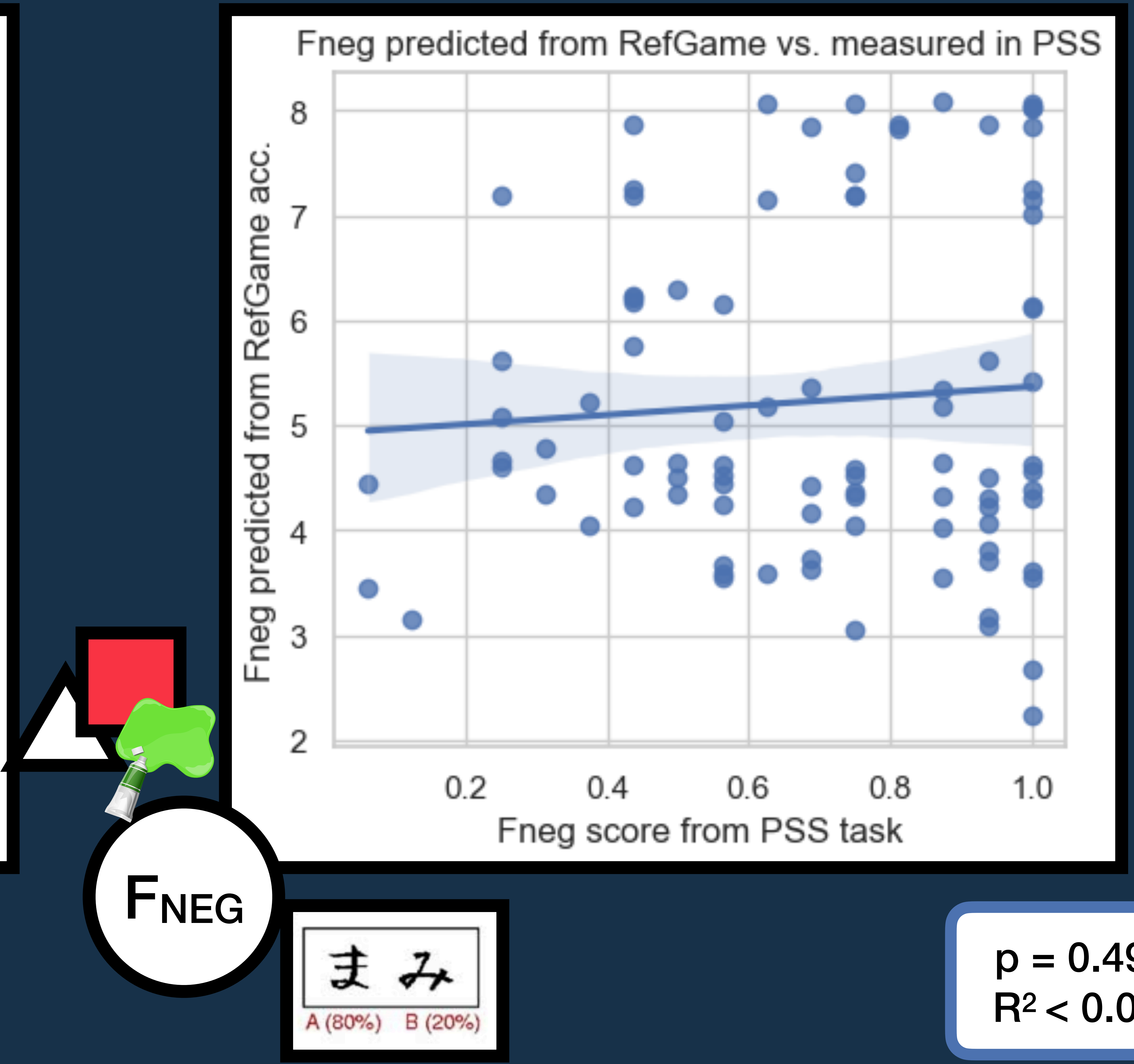


p < 0.01,
$R^2 = 0.10$

RefGame param. estimates do not correlate with new task measures

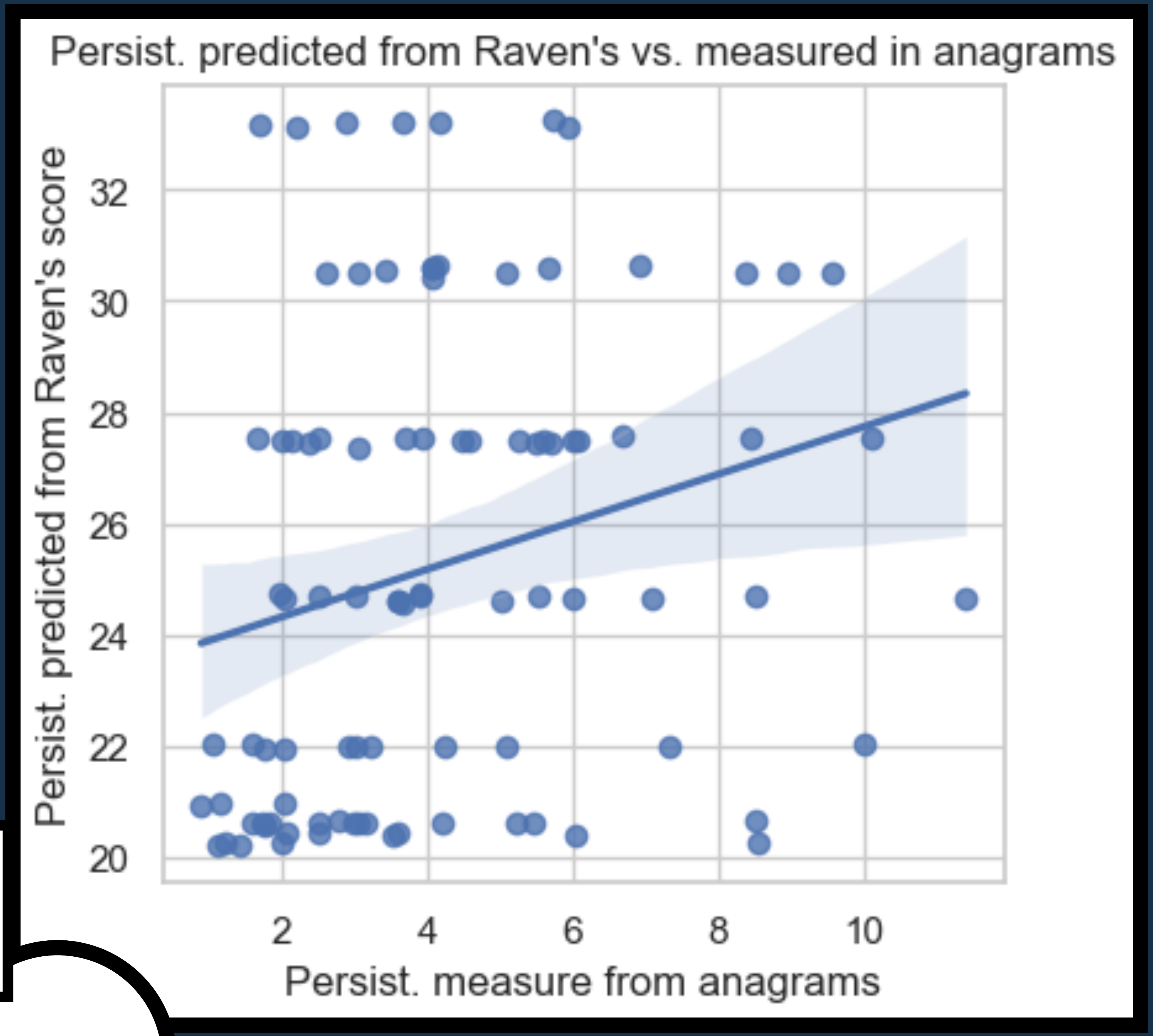


p = 0.60,
$R^2 < 0.01$



p = 0.49,
$R^2 < 0.01$

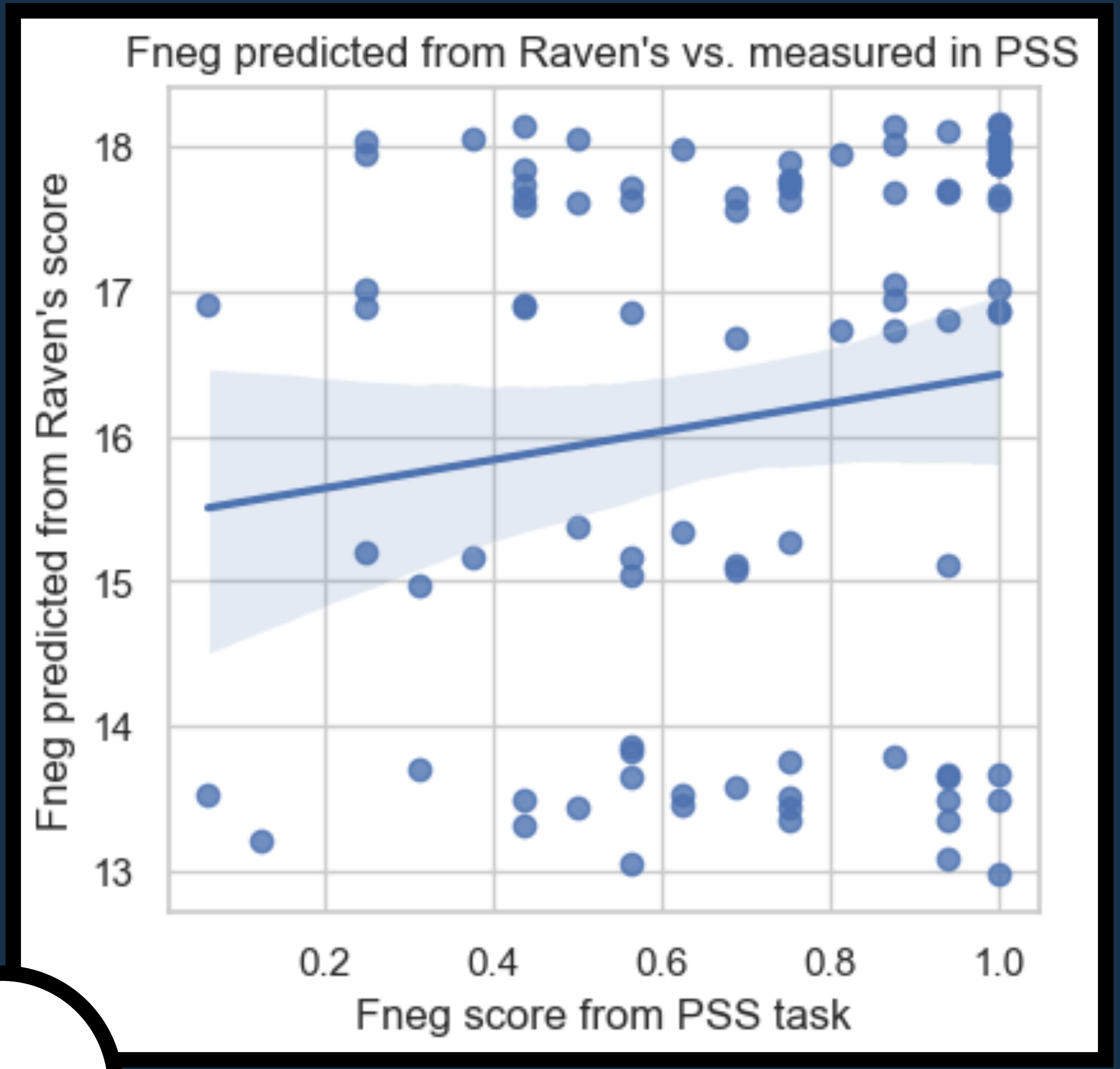
Raven’s param. estimates barely correlate with new task measures



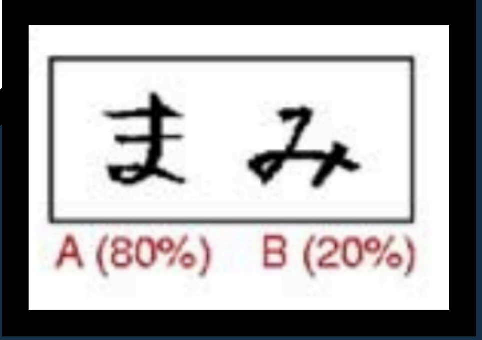
Pers

rveir

p = 0.02, $R^2 = 0.06$



FNEG



p = 0.20, $R^2 = 0.02$

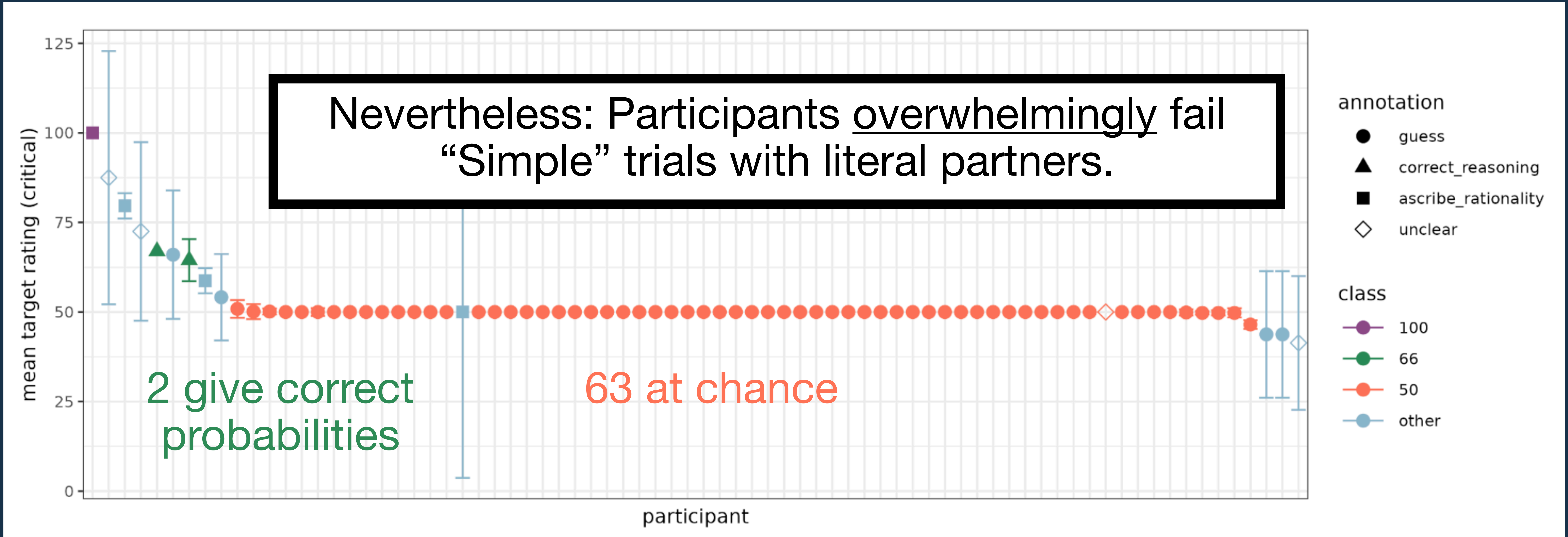
Probability fallacies in 1st-order reasoning

(Mayn, Duff, Bila & Demberg 2024)

- 1st-order pragmatic reasoning can solve "Simple" trials even with an **actual** literal (e.g. computer) speaker.

- Either 1st-order reasoning is never used, or participants apply it poorly.

(cf. Fox et al. 2004; Starns et al. 2019)



Atypicality inferences

(Ryzhova, Mayn & Demberg 2023)

`Mary went to a restaurant. She ate there!`

Mary must typically not eat when she goes to a restaurant.

?

- Participants with higher Raven's scores generated these inferences more often.
- Perhaps again, faster disengagement is supporting successful identification of a plausible candidate inference.