Trial, Task, And Trait of Theory of Mind

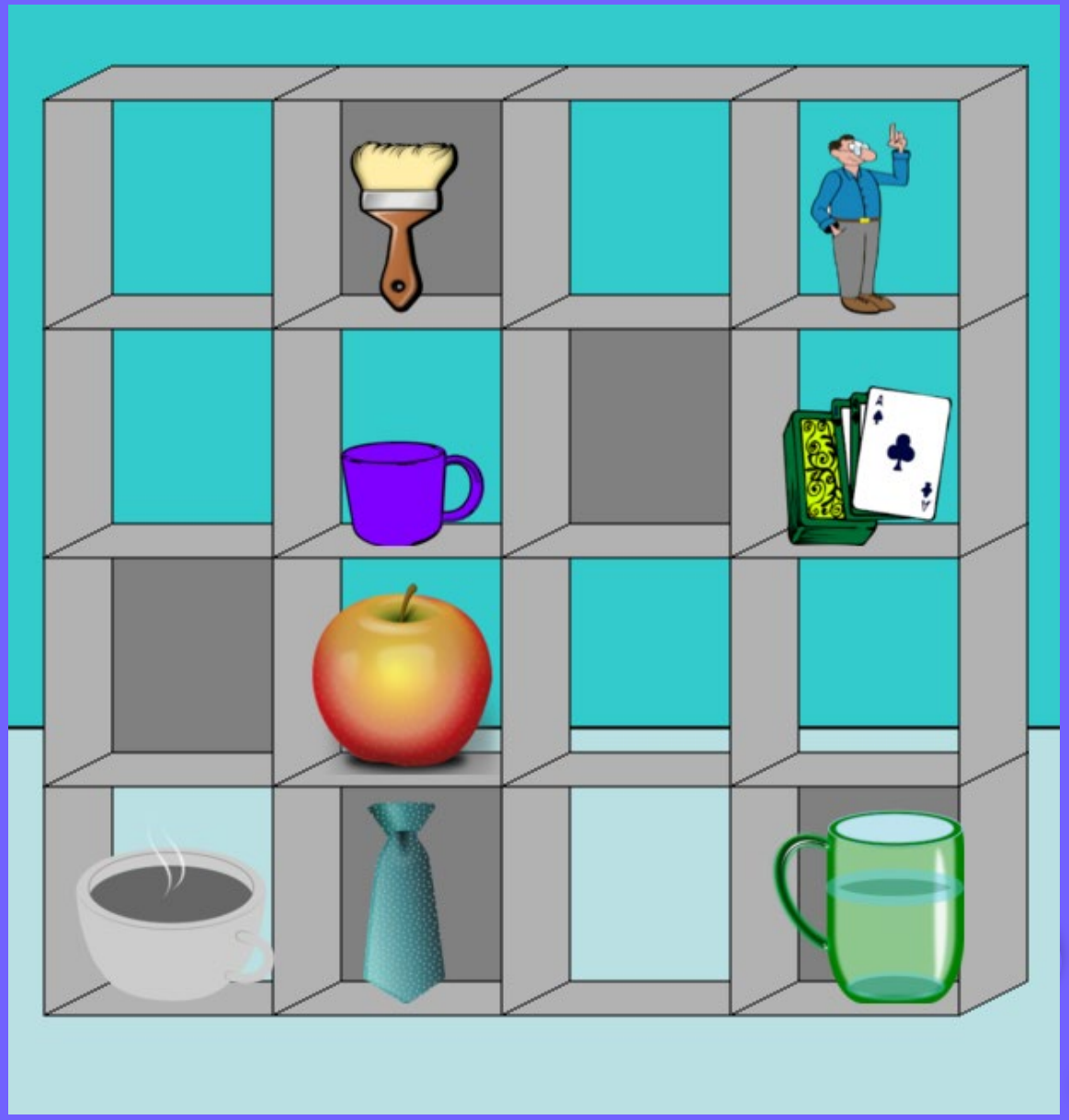
Reassessing the Director Task for Measuring Individual Differences in Theory of Mind

Han Hao, Ph.D.

Department of Psych Sciences

Tarleton State University

@SWPA 2026



Experimental effects and individual differences are not the same question

Experimental research asks whether a manipulation changes performance on average.

Individual-differences research asks whether a score reliably separates people with different standings on a trait/status.

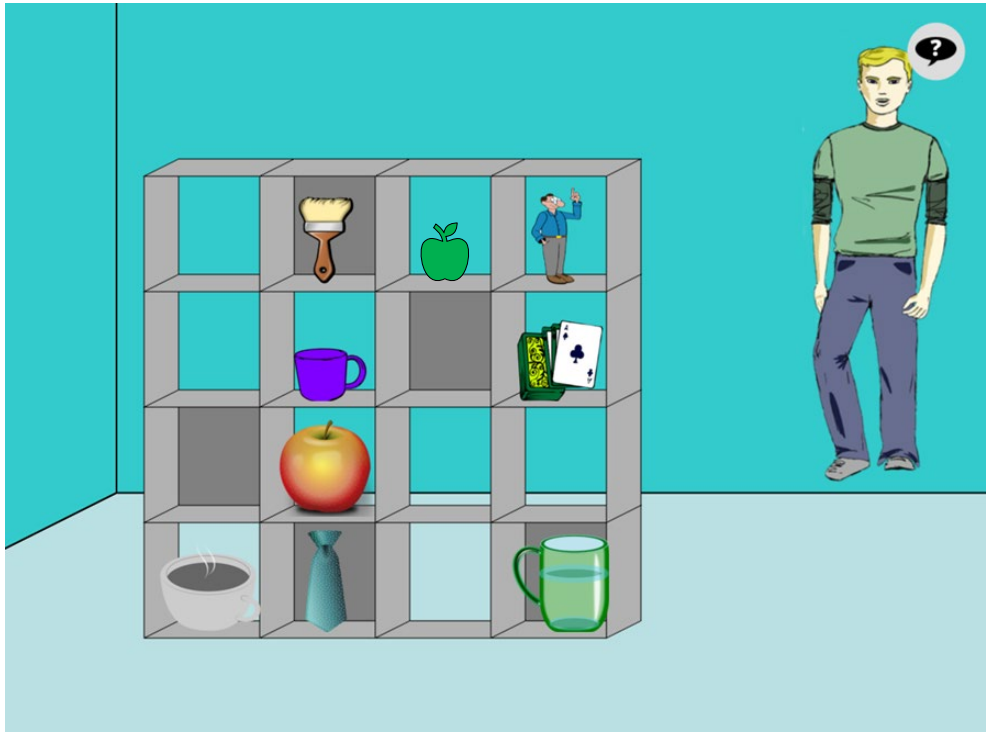
A task (and/or its measures) can succeed at the 1st goal and still fail at the 2nd.

This matters for paradigms like the **Director Task**, which was designed for experimental use and moved to person-level interpretation to investigate **Theory of Mind (ToM)**.

(Hedge et al., 2018; Goodhew & Edwards, 2019; Brysbaert, 2024)

The Director Task (DT) as a cognitive measure of ToM

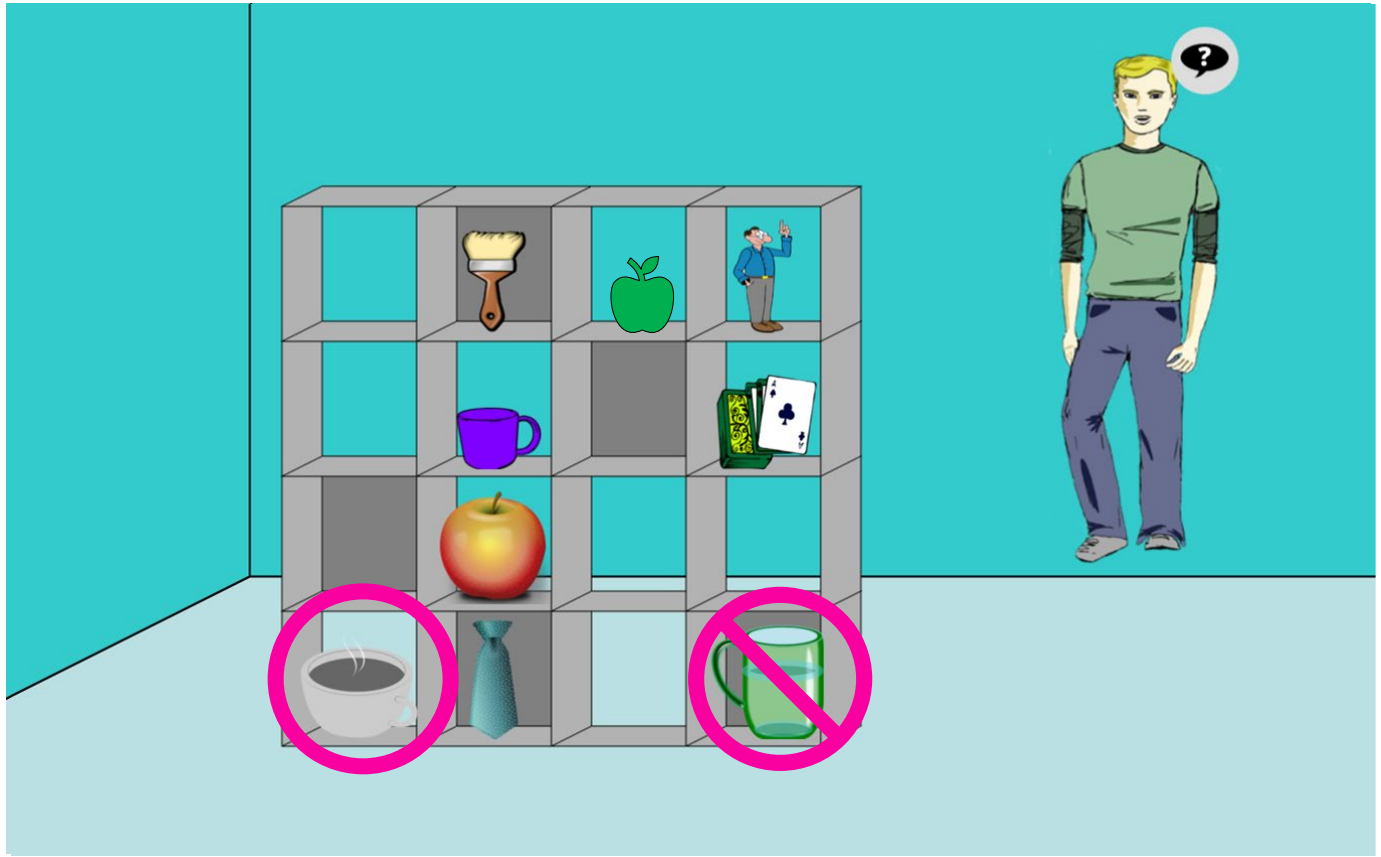
- The Director's task (DT) is a cognitive measure of ToM due to its visual perspective-taking demands
- Participants interpret instructions from the director's perspective
- Errors reveal interference from the participant's own privileged view



What does a DT trial look like

Trial Type

+ **Filler Trial**: “Move the person left”
+ **Control Trial**: “Move the small apple down”
+ **Exp Trial**: “Move the large cup up”



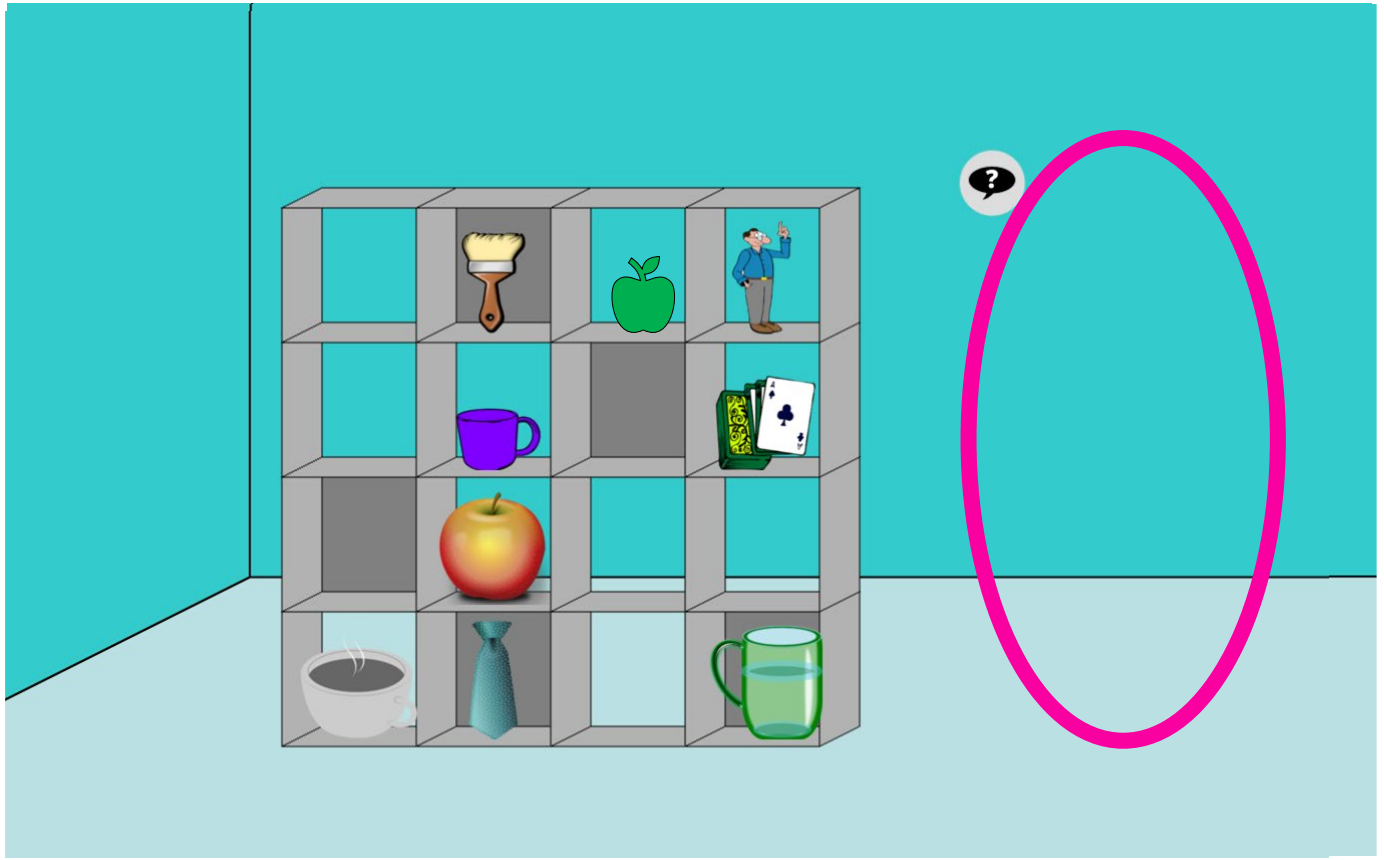
What does a DT trial look like

Trial Type

+ **Filler Trial**: “Move the person left”
+ **Control Trial**: “Move the small apple down”
+ **Exp Trial**: “Move the large cup up”

Director Type

+ Instructor vs. No Instructor



Conditions and corresponding demands

Trial type

Filler: no meaningful conflict

Control: some level of **object search** demand

Experimental: object search + **goal maintenance** demand

Director condition

No Instructor: explicit rules, no perspective-taking demand

Instructor: implicit rules (PT demand, **but only in Exp trials**)

Ins – F	**NoIns – F**
Ins – C	**NoIns – C**
Ins – E	**NoIns – E**

DT for multiple purposes

Does a paradigm designed for trial-level contrasts also work as a psychometric measure of cognitive individual differences?

Perspective use in referential communication in **experimental studies** (condition differences)

Visual perspective-taking ToM paradigm for **cognitive measurement** (composite scores across certain trials)

Individual-differences claims in **correlational studies** (associations with other cognitive measures)

Sample and data structure

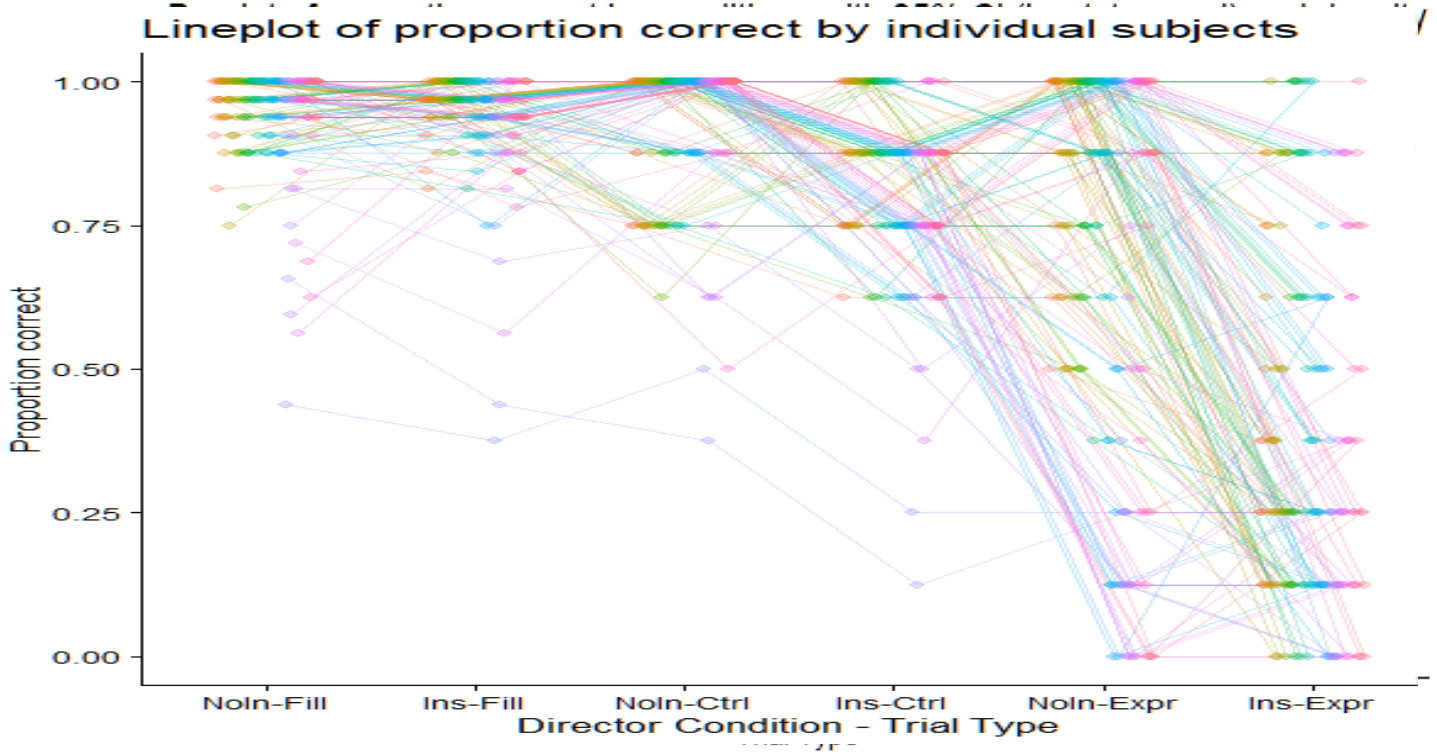
For trial-level investigations by condition and item analyses

+ Aggregated sample: $N_1 + N_2 = 155$
+ $(32 + 8 + 8) \times 2 = 96$ trials
+ Mainly focused on accuracy, **RT for complement due to exp program**

For predictive validity investigations

+ External-measure subset: $N_1 = 47$
+ **Fluid Intelligence (Gf):** RAPM/MARS
+ **Affective ToM:** RMET (Baron-Cohen et al., 2001; Oakley, 2016)
+ **Attention Control:** The squared tasks (Burgoyne et al., 2023)

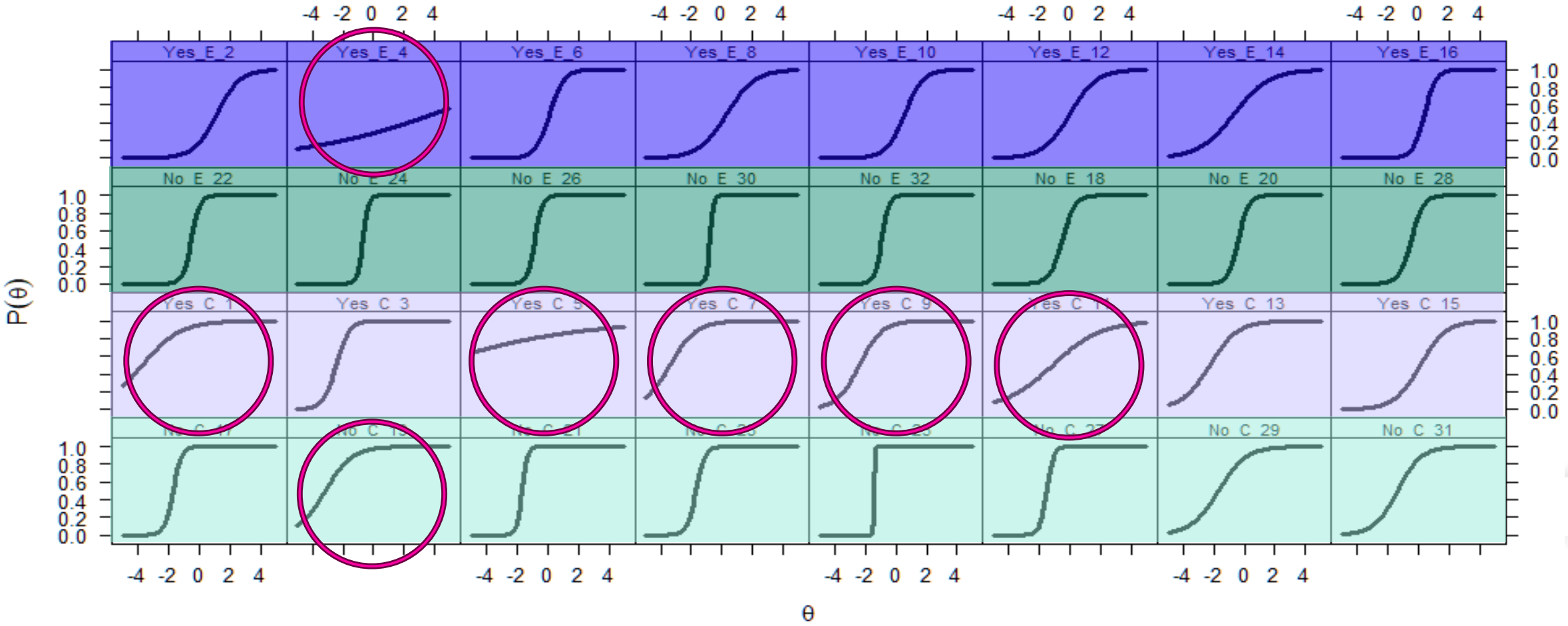
Study 1: Confirming experimental effects but …



Saturday
3:15 to 4:00 PM
Frisco 6 Poster 27

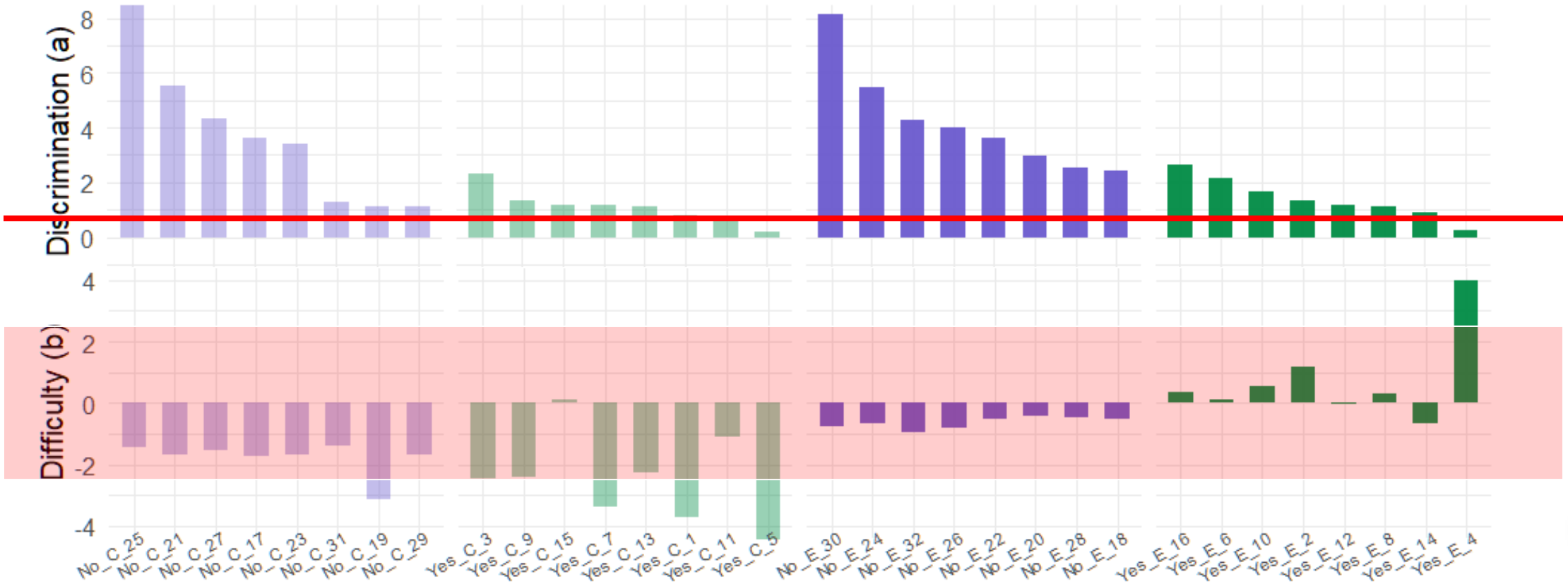
Study 2: Uni-IRT Results

Item trace plots



Study 2: Uni-IRT Results

Item parameters



DT for Experimental Effects of ToM, sure! DT for Individual differences of ToM, …?

Conventionally, we have been using composite scores of all or experimental trials as the cognitive ToM measure, but:

- If we take a composite score across all trials, we assume
  - The trials are unidimensional
- If we use composite scores only within certain (e.g., experimental) conditions, we are:
  - Performance in this condition represents a “cognitive ToM/Perspective-Taking” trait
  - Also losing information potentially useful from other conditions

Study 3: Exploring difference scoring for “process-pure” estimates

Candidate scores (other than conventional composite scores):

OS = object-search cost (avg. accuracy of Filler trials – avg. accuracy of Control trials)

GM = goal-maintenance cost (avg. accuracy of Control trials – avg. accuracy of Experimental trials)

PT = perspective-taking cost (GM_{no} - GM_{yes})

PT_2 = perspective-taking cost adjusted for order/practice

Study 3: Not all scores are reliable

Resampled Split-half reliability corrected by the Spearman-Brown method (k = 10000)

composite_all / composite_exp_only ≈ .80

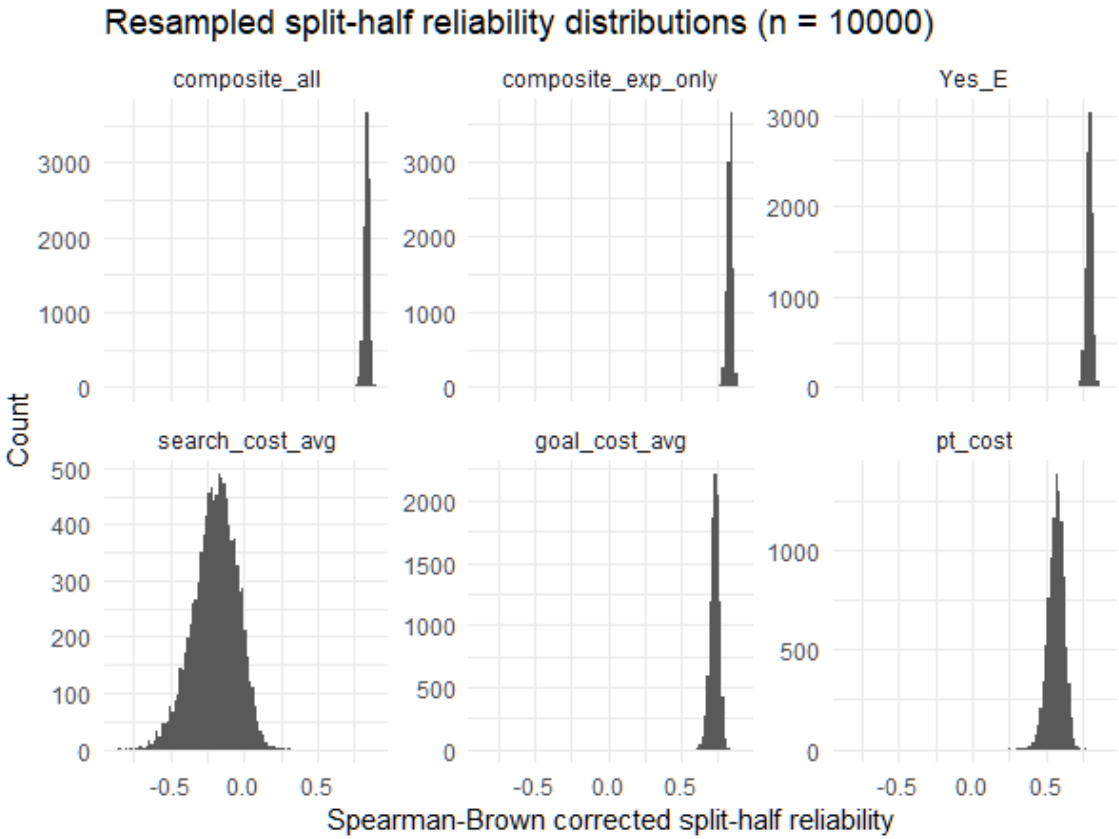
OS ≈ - 0.18; 95%CI[- 0.51, 0.07]

GM ≈ 0.73; 95%CI[0.66, 0.78]

PT ≈ 0.58; 95%CI[0.46, 0.67]

The search score failed because it subtracts highly correlated conditions

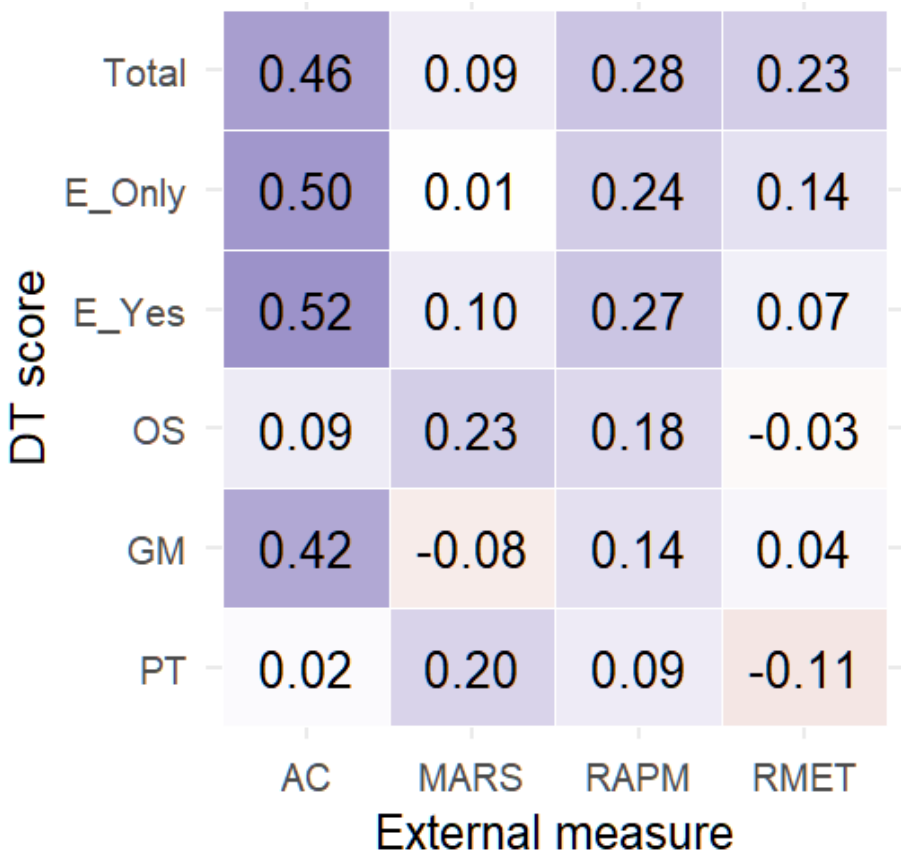
The PT score is not disastrous, but clearly weaker than the composite and weaker than GM



Study 3: Predictive validity

Subset N = 47 (non-robust pattern caution!)

- Good external relations with AC for conventional composites and GM
- Weak correlations with Gf and other ToM measures for all metrics
- External correlations of the PT scores are essentially flat (is it necessarily “bad” that PT and AC are not correlated?)



A potential Process overlap theory (Kovacs & Conway, 2022; Hao et al., 2025) interpretation?

What does the conventional score get right?

The common focus on aggregated performance in critical conditions may be **locally sensible**

- high reliability and predictive validity to some external measures

But this only tells us **where the strongest signal lives**

- It does not by itself tell us whether the signal **describes individual differences in a clean cognitive-ToM trait**

Task usefulness vs construct purity?

Implications for the Director Task re-design

Design changes for individual differences research purposes

- Increase the number and quality of Experimental items
- Remove filler items and/or revise pathological Control items
- Set up more precise RT administration

Scoring/modeling changes

- Consider Experimental-focused forms
- Evaluate multidimensional or individual-sensitive models

Implications for the ToM research

Be clear about the research goal:

+ Demonstration of a perspective-taking effect
+ Or measurement of a stable individual difference of a certain cognitive process or mechanism
+ If so, is the measure really a reliable and valid reflection of such a claimed process/mechanism (perspective-taking, emotion detection, etc.)?

Thank You!

Adrian D. Landry
Ph.D. Student

Tarleton State University

Ester D. Navarro
Assistant Professor

St. John's University

Manali Pathare
Ph.D. Student

New Mexico State University

+ Han Hao
+ hhao@tarleton.edu
+ https://hanhao23.github.io

Study Workflow and Methods

Study 1 Condition Comparisons

+ Full-within factorial **ANOVA** by conditions
+ Trial-level **binary logistic mixed model**
+ **Correlations** of condition accuracy and **ICCs** (exploratory)

Study 2 Unidimensional-IRT Analysis

+ Unidimensional **2PL IRT model**
+ 32 Control and Exp trials
+ Item parameters comparison by conditions (descriptive and exploratory)

Study 3 Alternative scoring analyses

+ Condition differences in accuracy as alternative ToM “process” scores
+ Monte-Carlo **split-half reliabilities**
+ **Criterion/predictive validity** of the ToM scores

Study 1: Condition level explorations for person-level info

Trial, task, & trait are not the same claim

Trial level: Does a manipulation create a perspective-related cost?

Task level: Can the trials be summarized by one or more stable scores?

Trait level: Does that score index an individual difference in cognitive ToM?

The strongest trait claim requires support from all three levels, not just a successful experimental contrast (i.e., accuracy in experimental trials is lower than that in control)

In many correlational uses, we focus on critical experimental-trial performance

But it can be locally sensible at the task level without being process-pure at the trait level

(Goodhew & Edwards, 2019; Rouder & Haaf, 2019; Brysbaert, 2024)

A Process Overlap Theory interpretation

Process Overlap Theory:

A broad positive manifold of cognitive task measures emerges from overlapping mixtures of processes

Task complexity is associated with higher cognitive processes demand in certain trials and therefore conveys stronger correlations among measures

Applied here:

Experimental DT performance may carry shared executive-control variance

Subtracted PT scoring may remove much of that shared variance

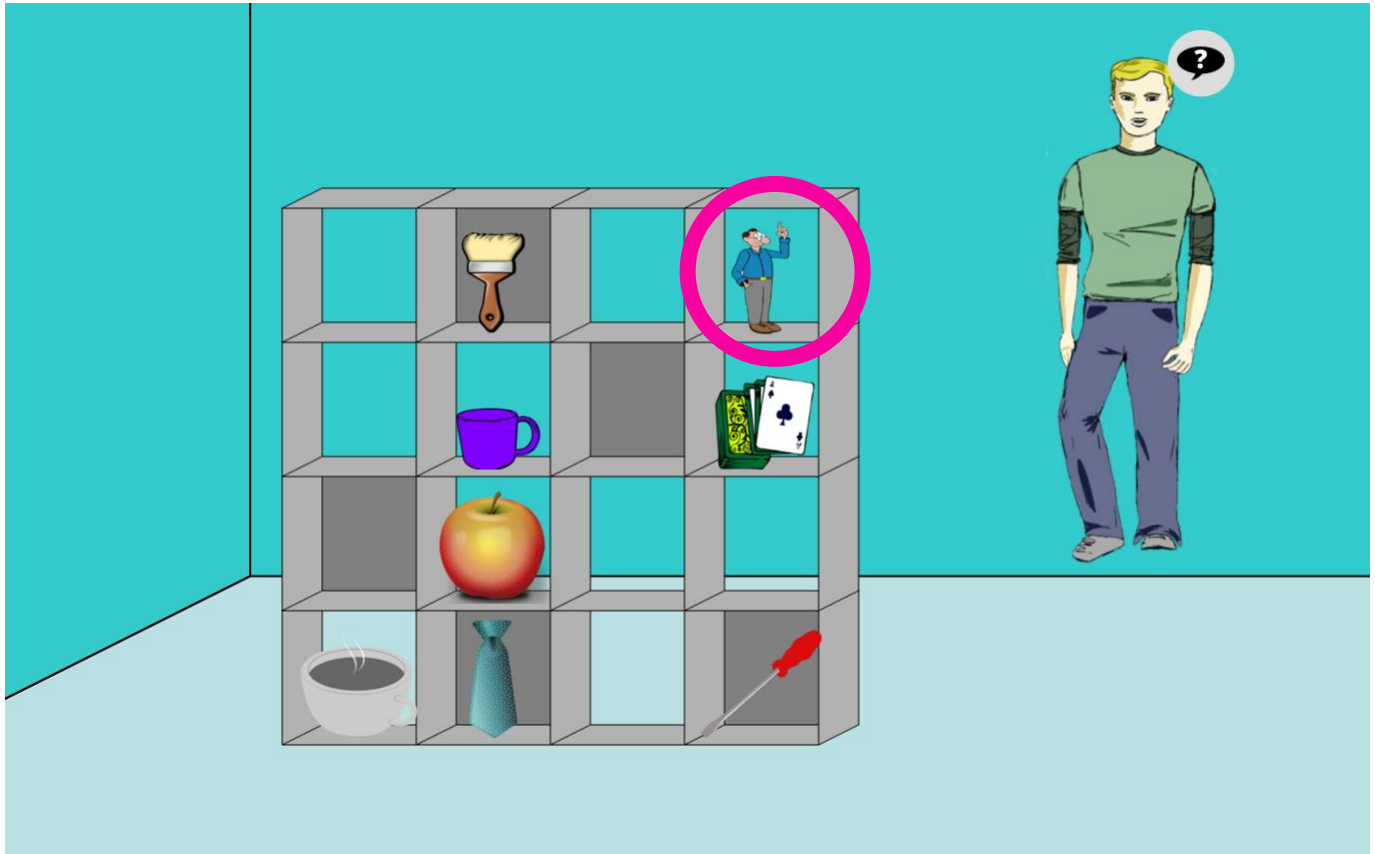
The PT score may be more process-pure, but also less psychometrically proven

What does a DT trial look like

Trial Type

+ **Filler Trial**: “Move the person left”

Director Type



What does a DT trial look like

Trial Type

+ **Filler Trial**: “Move the person left”
+ **Control Trial**: “Move the large cup up”

Director Type