More Trials, Better Trials, or More Tasks?

Simulation Insights on Cognitive Task Design for Individual Differences

Han Hao

Tarleton State University

A question of measurement and design

Which trials, in which tasks, are sufficient to measure an intended cognitive construct?

Processes are inferred from performance.
Manipulations vary proposed processing demands, and differences in accuracy, RT, or derived indices are attributed to those demands.

Performance operationalize (?) cognitive abilities.
Scores and/or contrasts are used to estimate individual standings on corresponding abilities.

But scores also reflect how tasks are implemented.
Paradigm and stimulus demands contribute stable impacts, so any single task will be biased.

Score arithmetic ≠ process separation.
Stable composites may reflect complex mechanisms. Subtraction or division assumes that unwanted contributions cancel arithmetically.

Two task design choices motivated this simulation

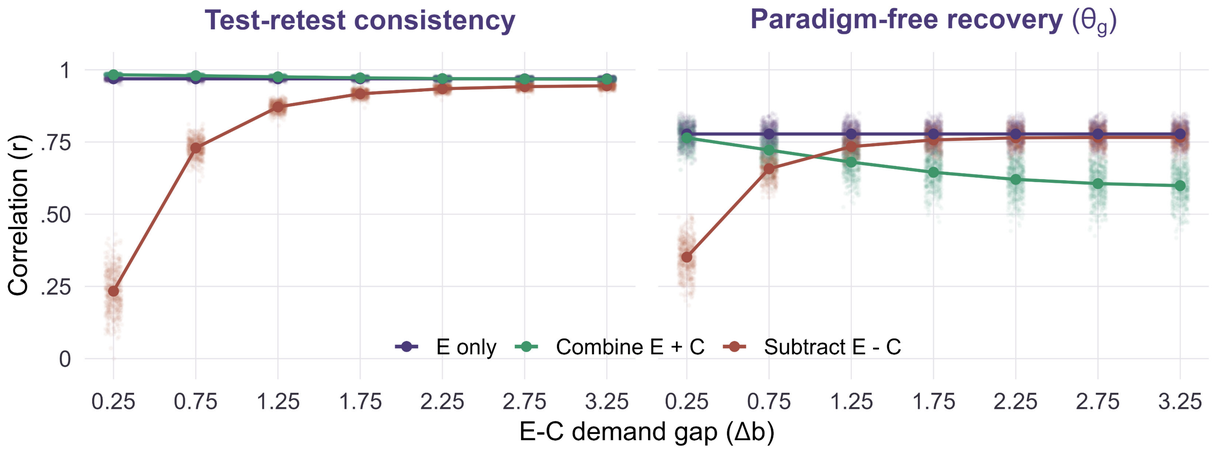
- **Difference scores vs. Total or condition scores**
- **Longer tasks vs. Broader paradigms and formats**

This study was inspired by WM, AC, and ToM research but aims to address broader questions about cognitive task measurement and design.

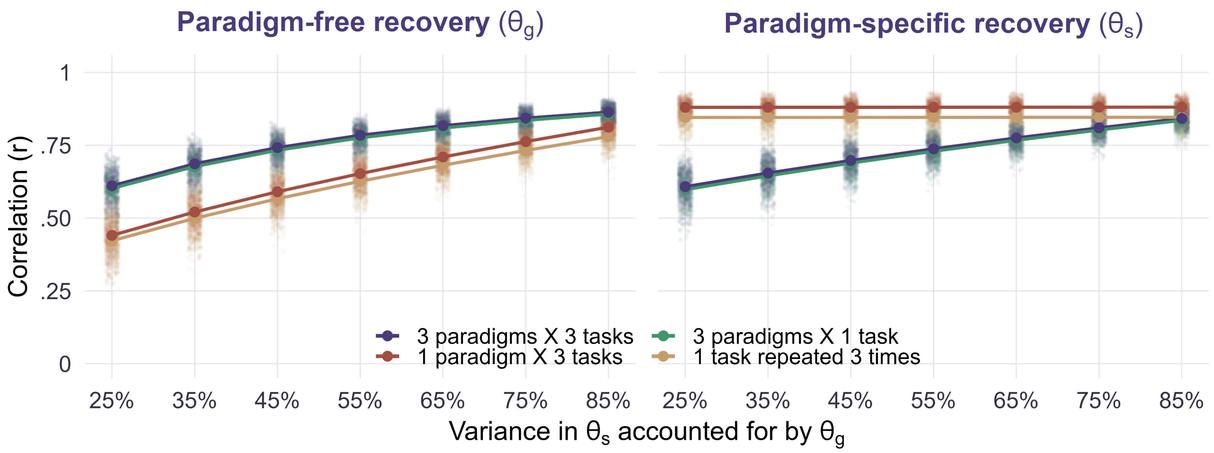
- How do reliability and process recovery vary across scoring methods as trial demands become more or less distinct across experimental conditions?
- How does the allocation of tasks and trials affect process recovery as the degree of shared trait variance across paradigms changes?
- Do more consistent scores also provide better recovery of the intended trait?

Results: when scoring and allocation work

Sim 1. High consistency did not always mean better process recovery; control trials contamination could be a potential issue in certain situations



Sim 2. Broader coverage was favored especially when the target trait is "complex"; is the construct defining the paradigm or the other way around?



Simulation design & procedures

General Setup (500 repetitions; n = 200)

- **9 tasks: 3 paradigms and 3 stimulus formats:** e.g., verbal, numerical, and spatial
- **8 simulated processes:** 1 θ_g, 3 θ_{pd}, 3 θ_{ft}, & 1 θ_{bg}
- **θ_g:** the bias-free target process.
- **θ_s:** the same process but biased by 1 θ_{pd} (fixed biases across repetitions in Sim1 and varied in Sim2).
- **Trials in a 2PL setup:** varying discrimination (small, unstructured) and difficulty (structured).
- **Each trial response** requires 1 θ_s, 1 θ_{ft}, & the θ_{bg}.
- **Non-compensatory bottleneck:** Trial accuracy depended on the weakest process (a bottleneck).

Sim1. Compare scoring methods

- **540 trials:** 9 tasks × (30 Experimental + 30 Control)
- **E-only:** average E responses; **E+C:** average E and C accuracy; **E−C:** subtract C from E accuracy

E-C demand gaps: difficulty parameter difference from 0.25 to 3.25. Larger gaps eased C's requirement for the target process; E trials remained unchanged across gap settings (but still vary across trials, M = −0.8, SD = 0.45).

Sim2. Compare trial allocations

E-only scoring; 270 E trials throughout the 4 plans

- **3 paradigms with 3 tasks in each:** 3 × 3 × 30 = 270
- **3 paradigms with 1 task in each:** 3 × 1 × 90 = 270
- **1 paradigm with 3 tasks in each:** 1 × 3 × 90 = 270
- **1 task repeatedly tested 3 times:** 1 × 1 × 90 × 3 = 270

Shared trait variance: θ_g was set to account for 25% to 85% of each θ_s variance, in 10-percentage step-breaks (λ^2 =.25 to .85). Larger shares made θ_ss more similar (aka stronger higher-order loadings).

Implications for task design

Under the current non-compensatory bottleneck assumptions, E−C recovery improved as experimental and control target demands separated. **Difference scores could be reliable and/or valid when E and C trials have very significant contrasts in cognitive demands.**

E+C scores remained highly consistent even as their recovery of θ_g weakened. **High reliability doesn't necessarily mean good recovery of target process.**

At a fixed trial budget, broader paradigm coverage favored θ_g, while focused task sampling favored the recovery of θ_s. **Task and trial allocations should reflect whether the intended construct is shared across paradigms or specific to one.**

The θ_g-recovery advantage of broader paradigm coverage narrowed as paradigm abilities shared more commmon variance. **Additional paradigms can be especially valuable for measuring a broadly defined trait where paradigms are less strongly related.**

Scope for next steps

Process assumptions: compare additive and different bottleneck mechanisms.

Task realism: use empirically calibrated 2PL parameters and condition/trial setups.

Score estimation: Accuracy-RT and latent models.

AI disclosure: The author designed the study and the simulations; AI assisted code development, visualization, and drafting. The code and results are not yet publicly available but will be made public on the linked page by the QR code.