Study Design for Robust and Meaningful Research in Computational Psychiatry

Saren H. Seeley, Ph.D.

Berner Lab, Center for Computational Psychiatry

saren.seeley@mssm.edu

New York Computational Psychiatry Workshop

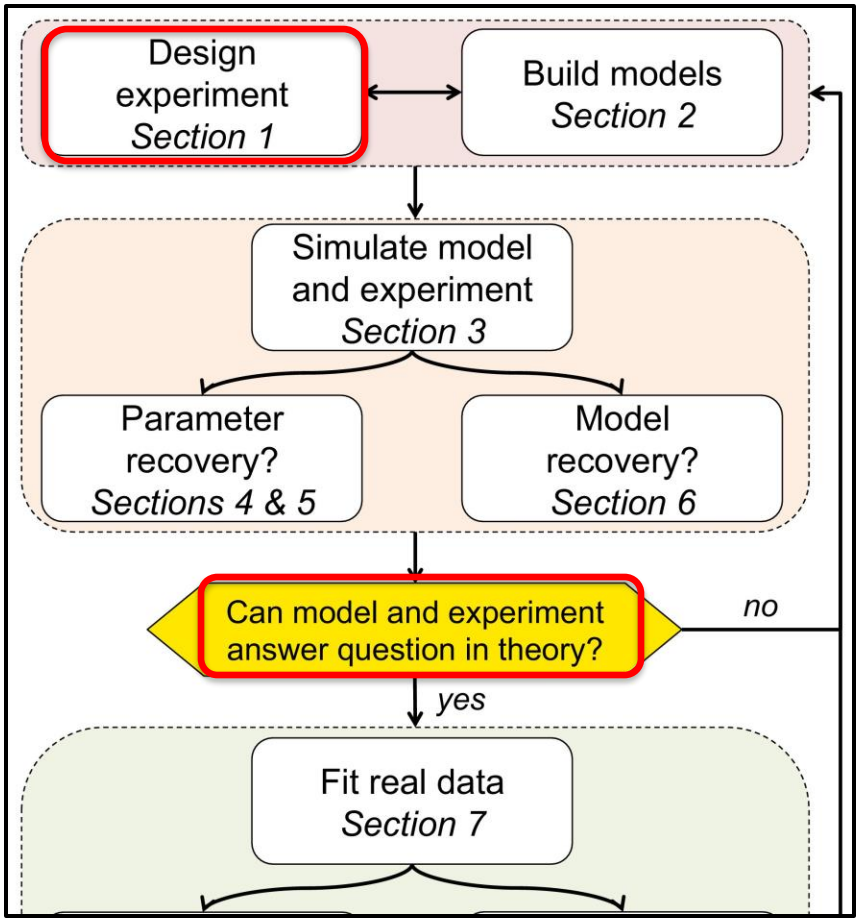
November 12, 2025

Icahn School of Medicine at Mount Sinai

This will not be a Research Methods lecture.

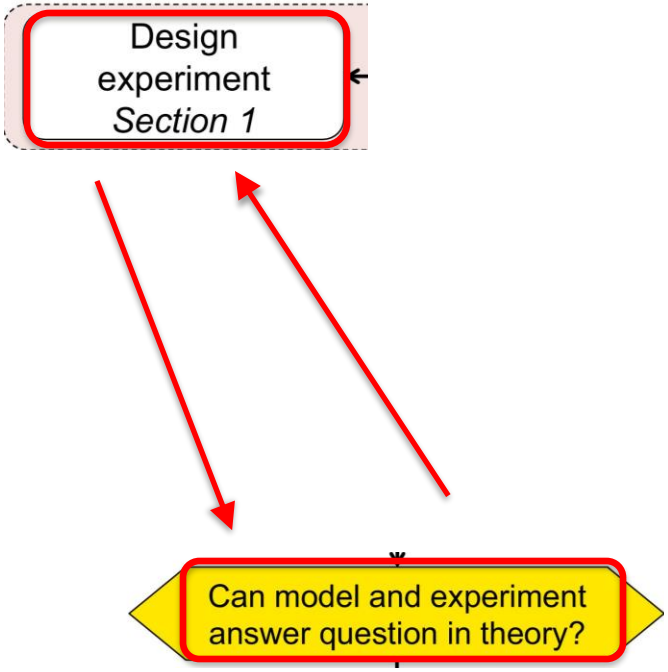
Ten simple rules for the computational modeling of behavioral data

Robert C Wilson[1,2†*], Anne GE Collins[3,4†*]



This will not be a Research Methods lecture.

We are here:



It does not matter how sophisticated or powerful your modeling approach is, if your experimental design cannot answer the question you are asking.

Optimal study design and reproducibility/replicability are not just about technical practices but how we think.

Computational psychiatry has advantages and unique considerations in both domains.

Good theories don't just predict direction—they predict *how much*, *when*, and *under what conditions*.

Well-specified theories contribute to reliable, replicable, & useful empirical results.

From: *Test broad directional hypotheses*

→

To: *Build cumulative knowledge through constrained, falsifiable predictions*

We often care about latent constructs.

The challenge in psych/neuro research is to create testable theories of things (that cannot be directly observed) that can...

- Establish latent constructs' existence
- Explain how they interact
- Generate more precise predictions
- Refine explanations

How do we theorize about things we cannot directly observe?

Example: Panic disorder

Clinical Phenomenon:

- Panic attacks are triggered by misinterpretation of body sensations.
  - Observational learning +/- past experience (e.g., illness in parent/self as a child) increases propensity for catastrophic misinterpretation.

Theory:

- Reduced/absent inhibitory learning (i.e., that feared outcome does not occur) maintains catastrophic misinterpretation.

Hypothesis:

- Higher initial Q values for threat (Rescorla-Wagner model)
  - **Alternative H:** Lower α (slower updating)

Example: Panic disorder in Rescorla-Wagner terms

Q(state, interpretation):

- Expected value given state:
  Q(heart racing, "heart attack") = high threat

Outcome(*t*):

- Actual [observed] outcome at time *t* (e.g., *"did not have heart attack*")

Prediction Error: δ = Outcome(t) - Q(t)

Updated belief for next time:

Q(t+1) = Q(t) + α × δ

- Unhelpful: *"The only reason I was okay last time was because I went to the ED, who knows what would have happened if I didn't?"*

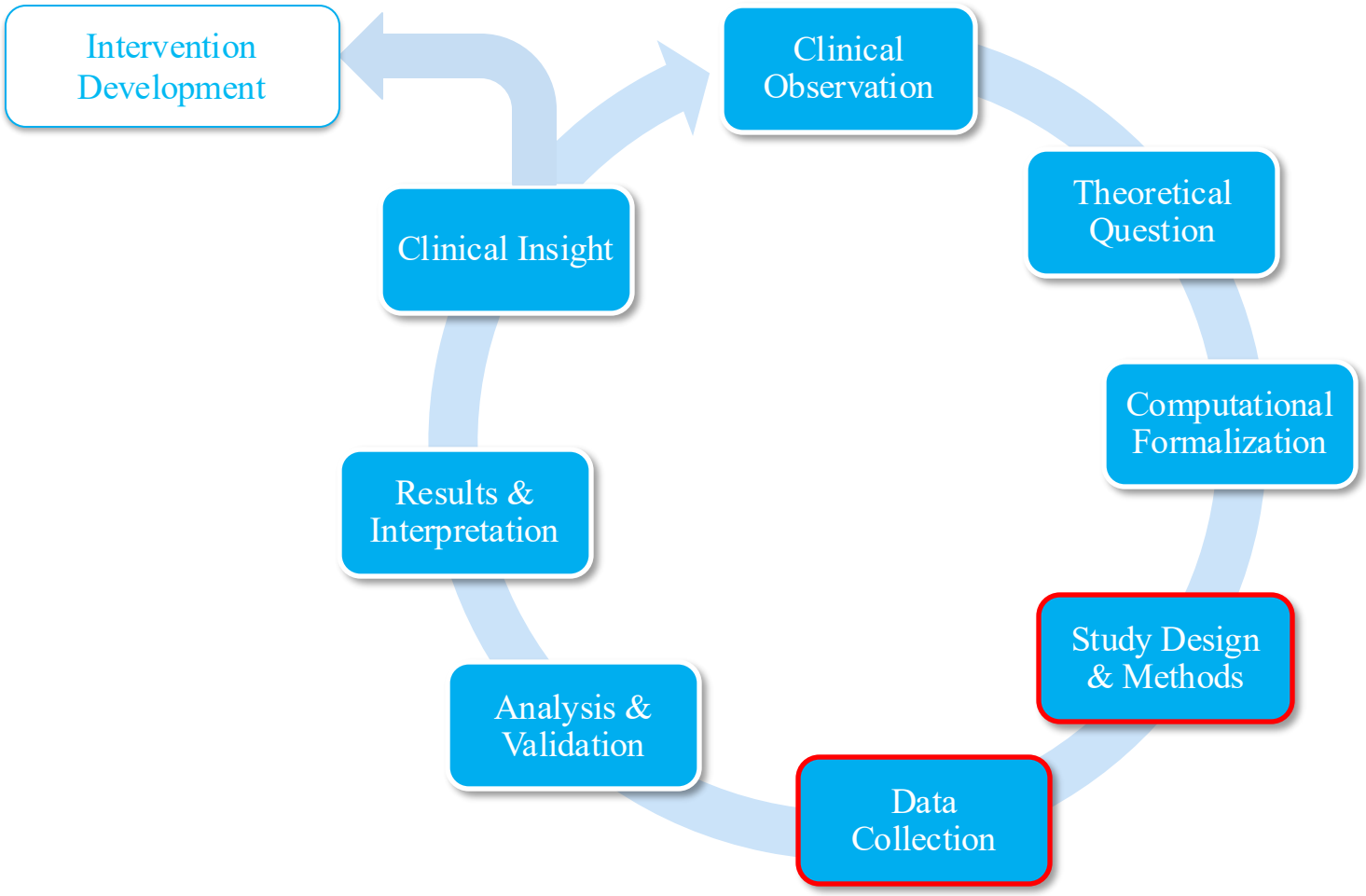
- More helpful: *"Ok, this is panic...again. It is scary but scary doesn't mean dangerous."*

→ Strong prior beliefs (high initial Q-values for threat) ?
→ Smaller δ → weaker update signal ?
→ Low α → slow updating (*resist disconfirming evidence*) ?

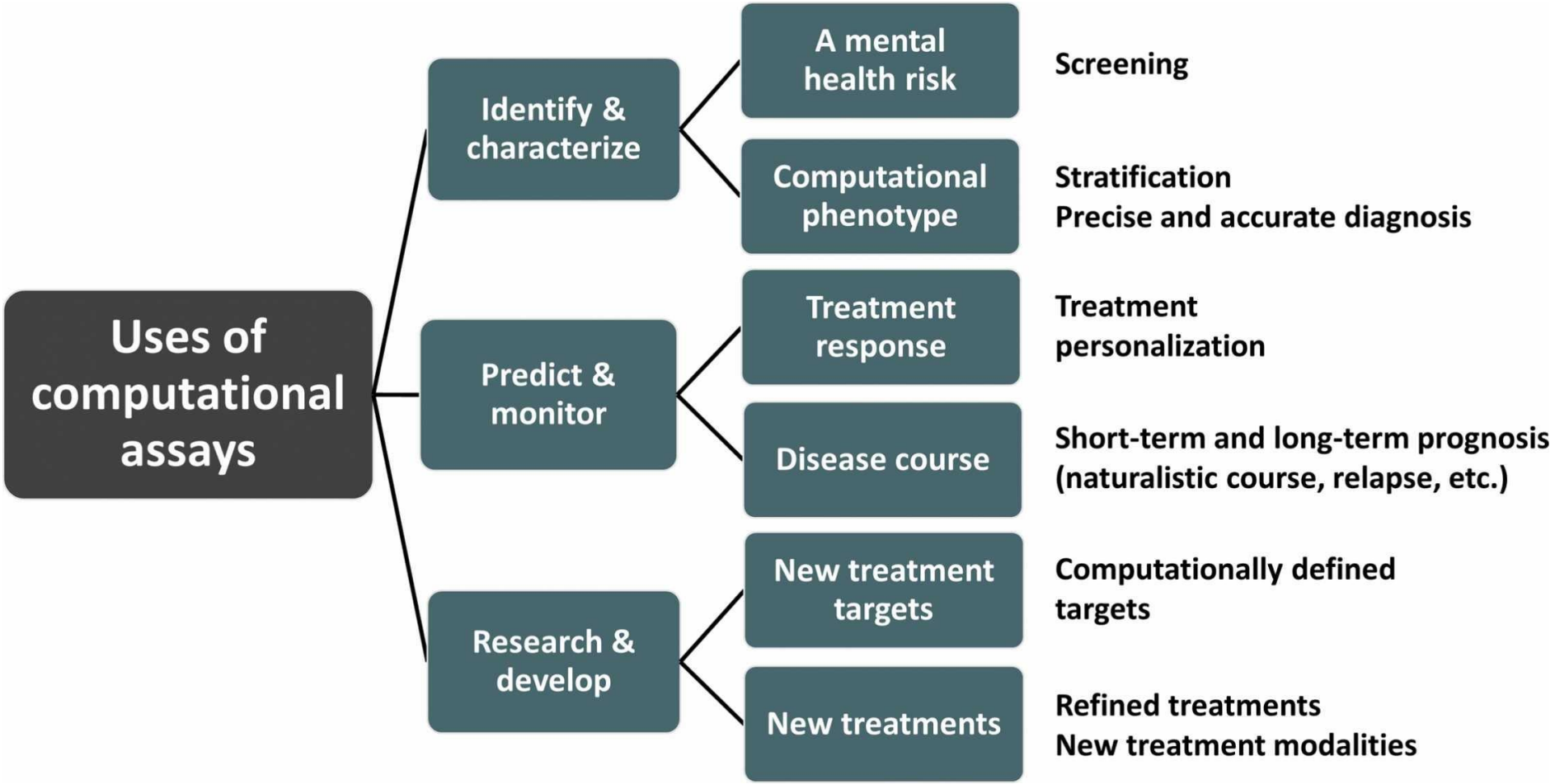
	Patient A	Patient B	Patient C
Higher initial Q-values for threat	✅	❌	❌
Smaller δ	❌	❌	✅
Lower α	❌	✅	✅

Clinical implications?

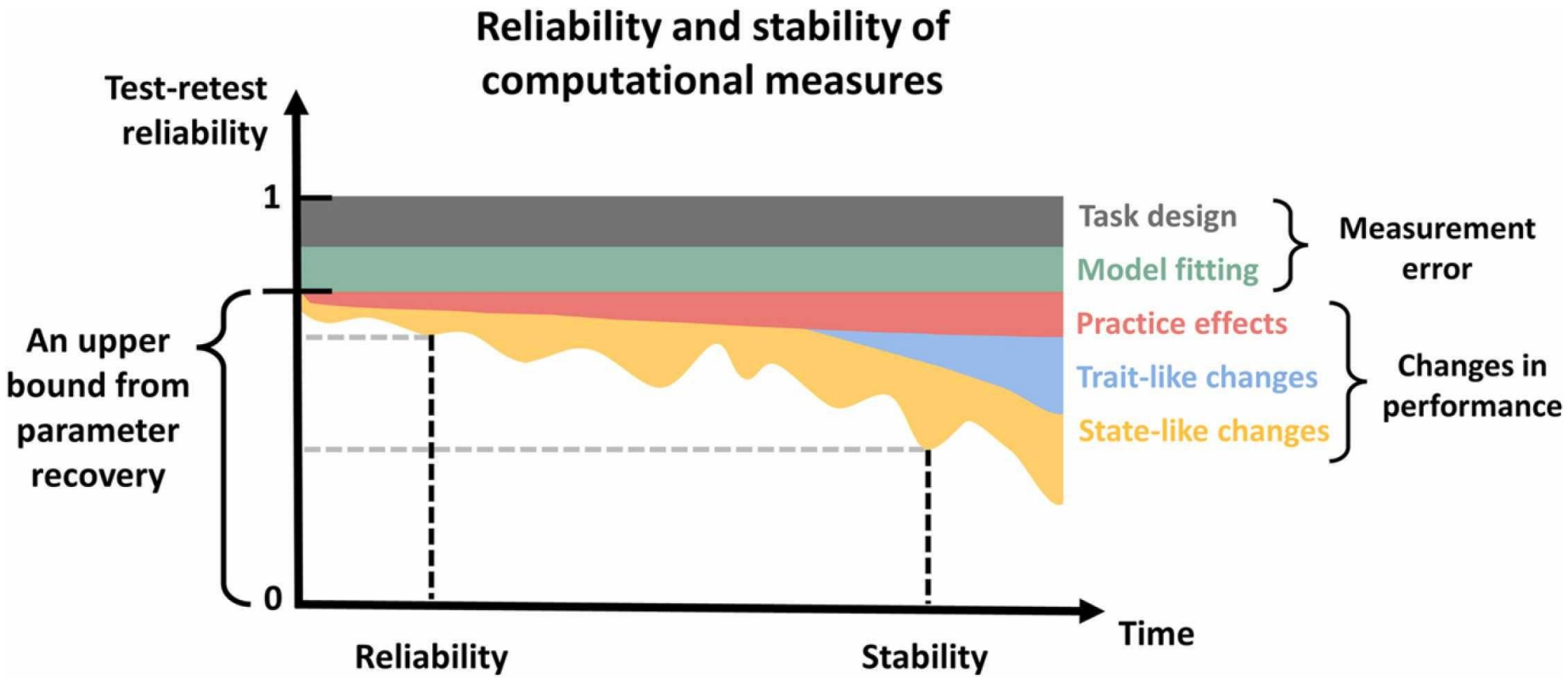
Scientific Lifecycle in Computational Psychiatry



Task/Assay Selection



Sandve et al. (2013) PLoS Comput Biol. doi:10.1371/journal.pcbi.1003285



Sandve et al. (2013) PLoS Comput Biol. doi:10.1371/journal.pcbi.1003285

Task/Assay Selection

Psychometric Properties

- **Test-retest reliability**: Does the task produce consistent measurements over time? Critical for longitudinal studies and tracking treatment effects
- **Internal consistency**: Do different components/trials of the task measure the same construct?
- **Convergent validity**: Does the task correlate with established measures of the same construct?
- **Discriminant validity**: Does the task NOT correlate with measures of different constructs?

Construct Validity:

Is the task capturing the intended psychological or computational phenomenon?

- Does it perform differently in clinical populations versus unselected/healthy controls?
- If no group differences: is the construct truly unaffected, or is the task insensitive?
- If group differences exist: are they specific to your hypothesis or could they reflect general impairment?

Example: Disambiguating effort vs. difficulty

Cognitive, Affective, & Behavioral Neuroscience (2023) 23:290–305
https://doi.org/10.3758/s13415-023-01065-9

RESEARCH ARTICLE

Check for updates

Measuring cognitive effort without difficulty

Hugo Fleming[1,2] · **Oliver J. Robinson**[1] · **Jonathan P. Roiser**[1]

Accepted: 9 January 2023 / Published online: 7 February 2023

Abstract
An important finding in the cognitive effort literature has been that sensitivity to the costs of effort varies between individuals, suggesting that some people find effort more aversive than others. It has been suggested this may explain individual differences in other aspects of cognition; in particular that greater effort sensitivity may underlie some of the symptoms of conditions such as depression and schizophrenia. In this paper, we highlight a major problem with existing measures of cognitive effort that hampers this line of research, specifically the confounding of effort and difficulty. This means that behaviour thought to reveal effort costs could equally be explained by cognitive capacity, which influences the frequency of success and thereby the chance of obtaining reward. To address this shortcoming, we introduce a new test, the Number

Many tasks have multiple possible interpretations

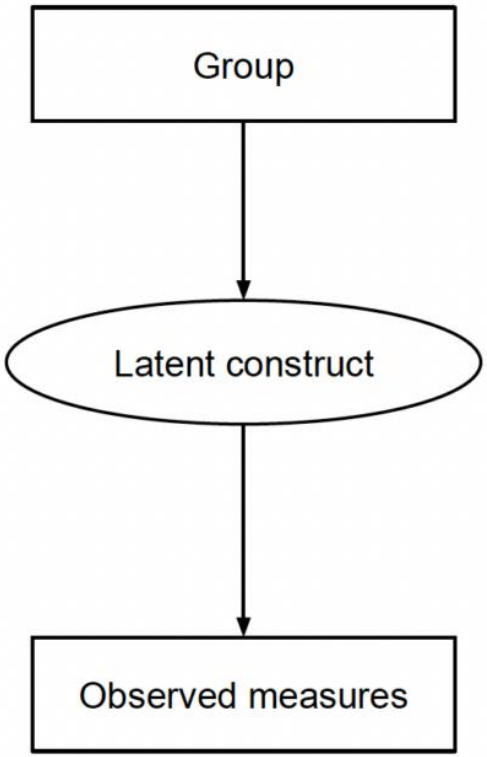
- Influence of task instructions on behavior, e.g. model-based vs. model-free?
- Compensatory strategies in clinical samples?
- Always consider*: what else could produce this pattern of behavior?*

Stimulus Selection

- Disorder-relevant stimuli or generic stimuli?
  - Disorder-relevant: may increase sensitivity but reduce generalizability
  - Generic: better for cross-disorder comparisons but may miss disorder-specific mechanisms
  - Consider using both for convergent evidence
- Consider measurement (non-) invariance! *Does my task/stimuli operate the same way in my sample as in controls/other studies?*

Measurement invariance

All effects of the group variable on the observed measures are mediated by the latent construct.



Measurement non-invariance

The group variable affects the observed measures directly or through some other mechanism.

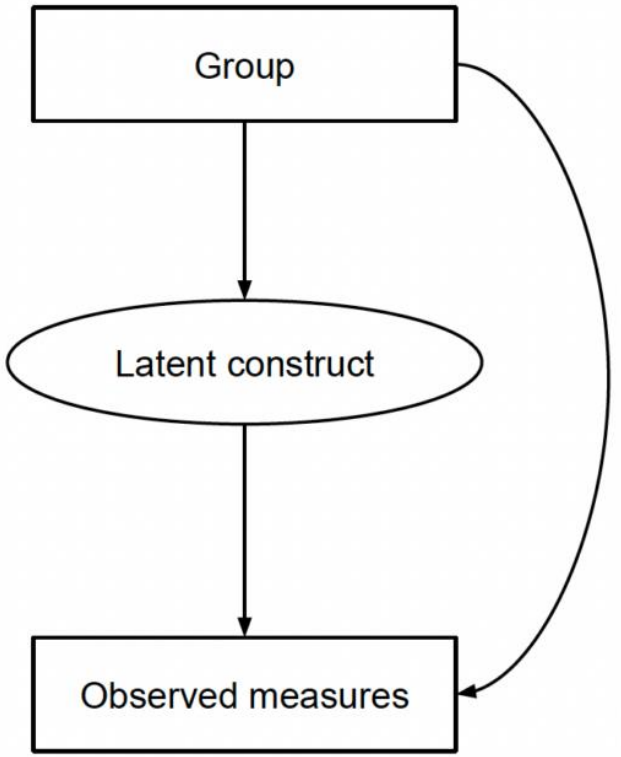
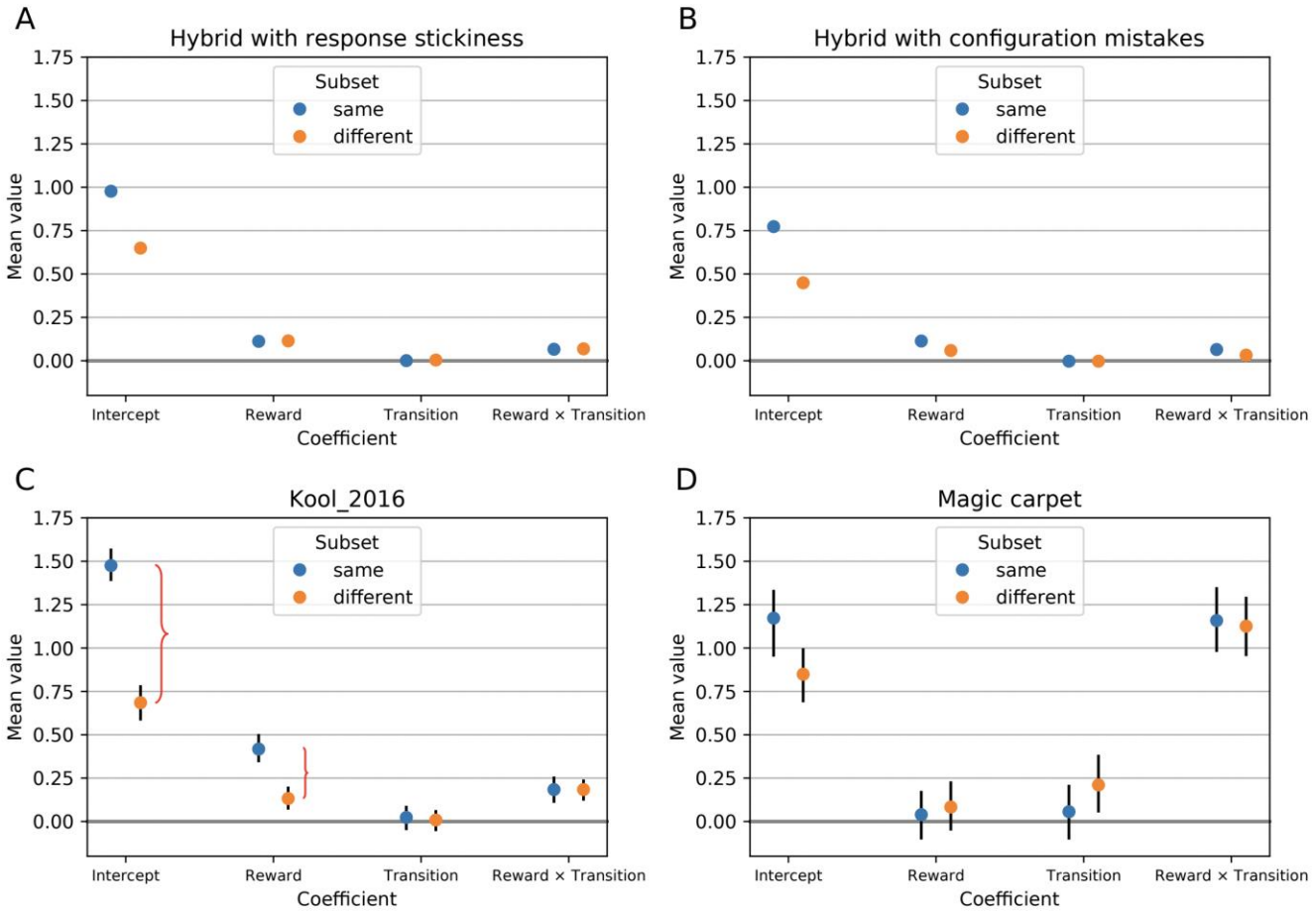


Figure: https://www.the100.ci/2024/01/10/a-casual-but-causal-take-on-measurement-invariance/

Fig. 5: Simulated behaviour of agents and real participants can be influenced by irrelevant changes in stimulus position.



Feher da Silva, C., & Hare, T. A. (2020). *Nature Human Behav*.

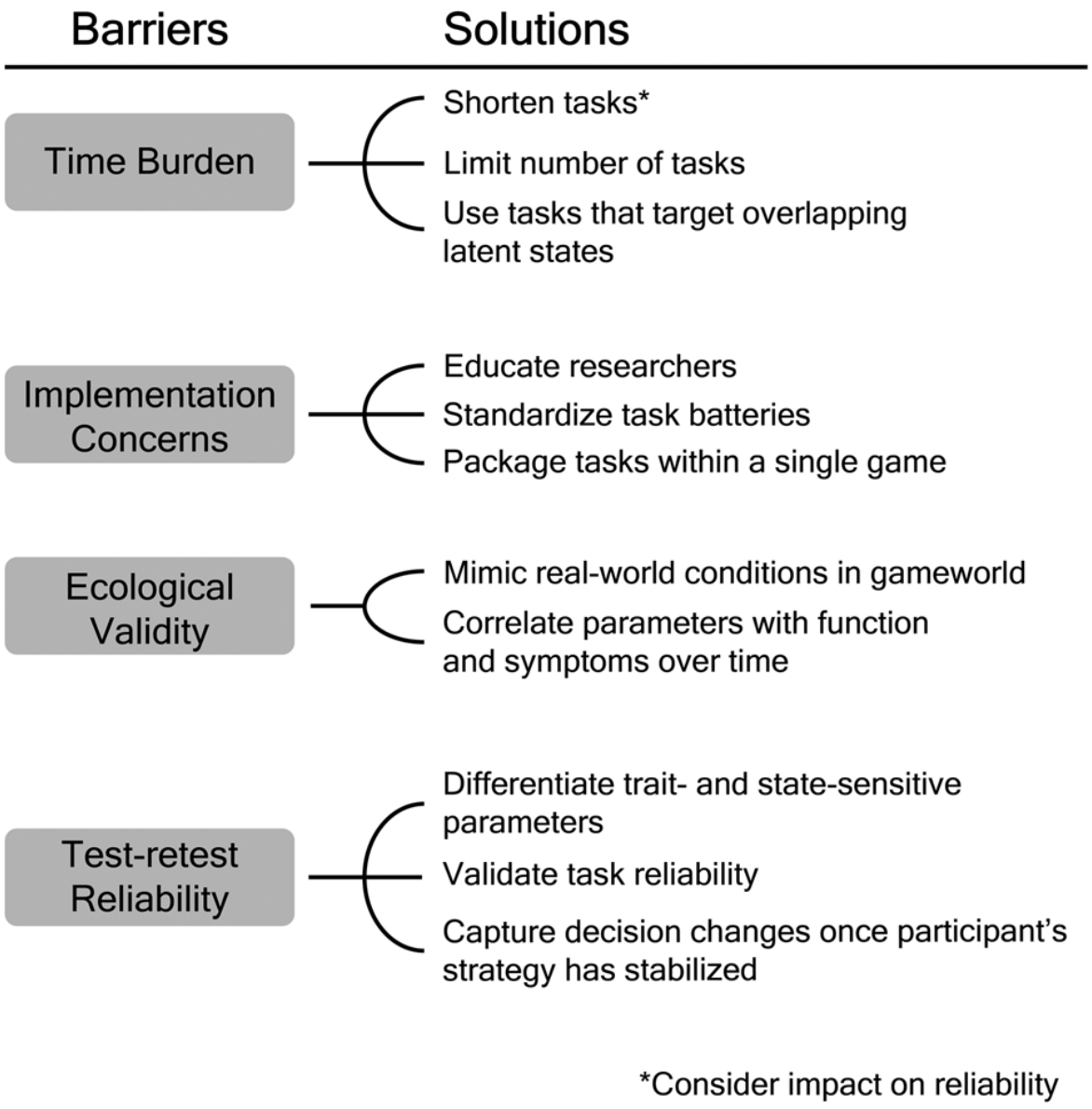
Task Selection: Modeling & Design Considerations

Computational Modeling Considerations

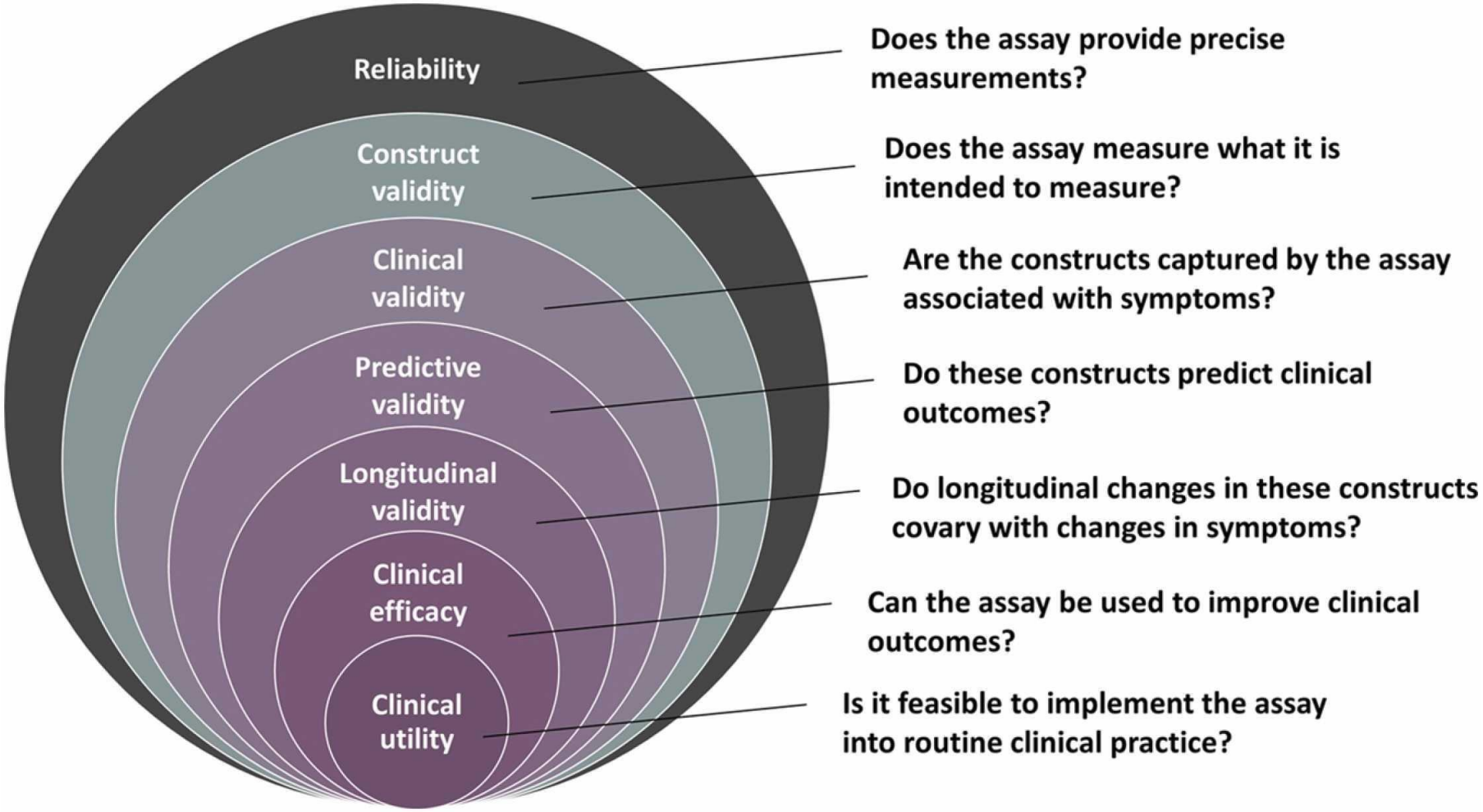
- Model identifiability - can you distinguish between competing models?
- Model comparison strategy
- Parameter recoverability (& simulations validate parameter recovery?)
- Sufficient trial numbers for stable parameter estimation?

Practical Design Issues

- Task length and participant burden
- Learning/practice/instruction effects
- Counterbalancing
- Engagement and motivational – ***gamification?***
- Multi-site consistency (software, implementation, hardware)
- Sample size requirements for planned analyses
- Missing data



Benrimoh et al. (2023). *Molecular Psychiatry.*



Sandve et al. (2013) PLoS Comput Biol. doi:10.1371/journal.pcbi.1003285

Example: Grief-LEARN Study

- Tested two probabilistic reversal learning models* via R *hBayesDM:*

  1. Rescorla-Wagner/temporal difference model with the parameters
  – **β (inverse temperature), α_{reward}, $\alpha_{punishment}$**

  2. Experience-weighted attraction model with the parameters
  – **β (inverse temperature)**, **rho (ρ), phi (Φ)**

- We hypothesized that PGD involves decision-making biased toward past experience at the expense of new information, as captured by **ρ.**

- Alternatively, differences in reversal learning might be due to people with integrated grief being (1) faster to "forget" the past (**Φ**), (2/3) more sensitive to losses and/or rewards (**$\alpha_{punishment}$** / **α_{reward}**) or (4) more goal-directed decision-making (**β**).

*Den Ouden (2013)

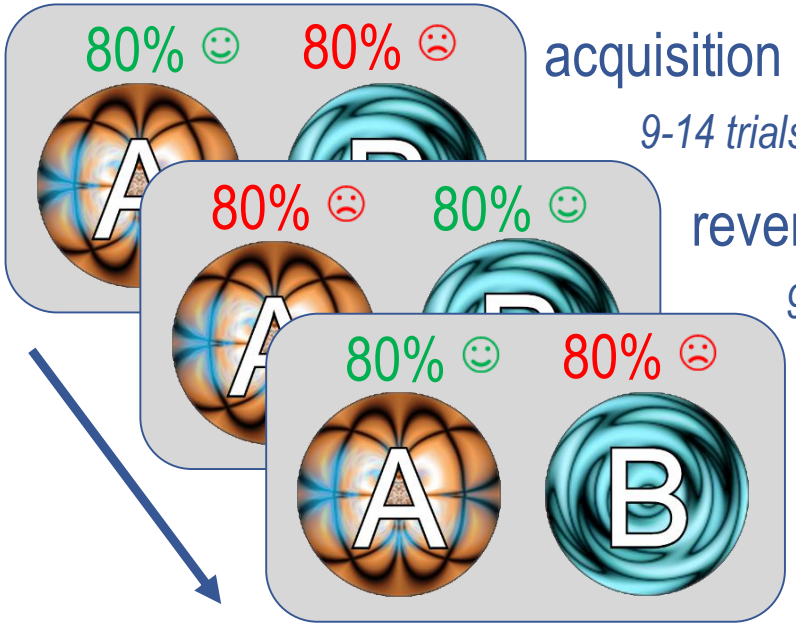
PRL Advantages

- Technical:
  - Test-retest reliability, psychometric properties, other papers using these models with this task (shared code)

- Theoretical:
  - Well-established effects in other populations
  - Well-established neural correlates/computations
  - Translational value
  - Computational models map to clinical phenomena & theory
    - Probabilistic
    - Perseverative errors
    - ***Q: What do people do when the thing they wanted & thought would happen, didn't happen?***

Case Study: Design Decisions Matter

Probabilistic Reversal Learning Task (Cools et al., 2002)



acquisition phase

9-14 trials

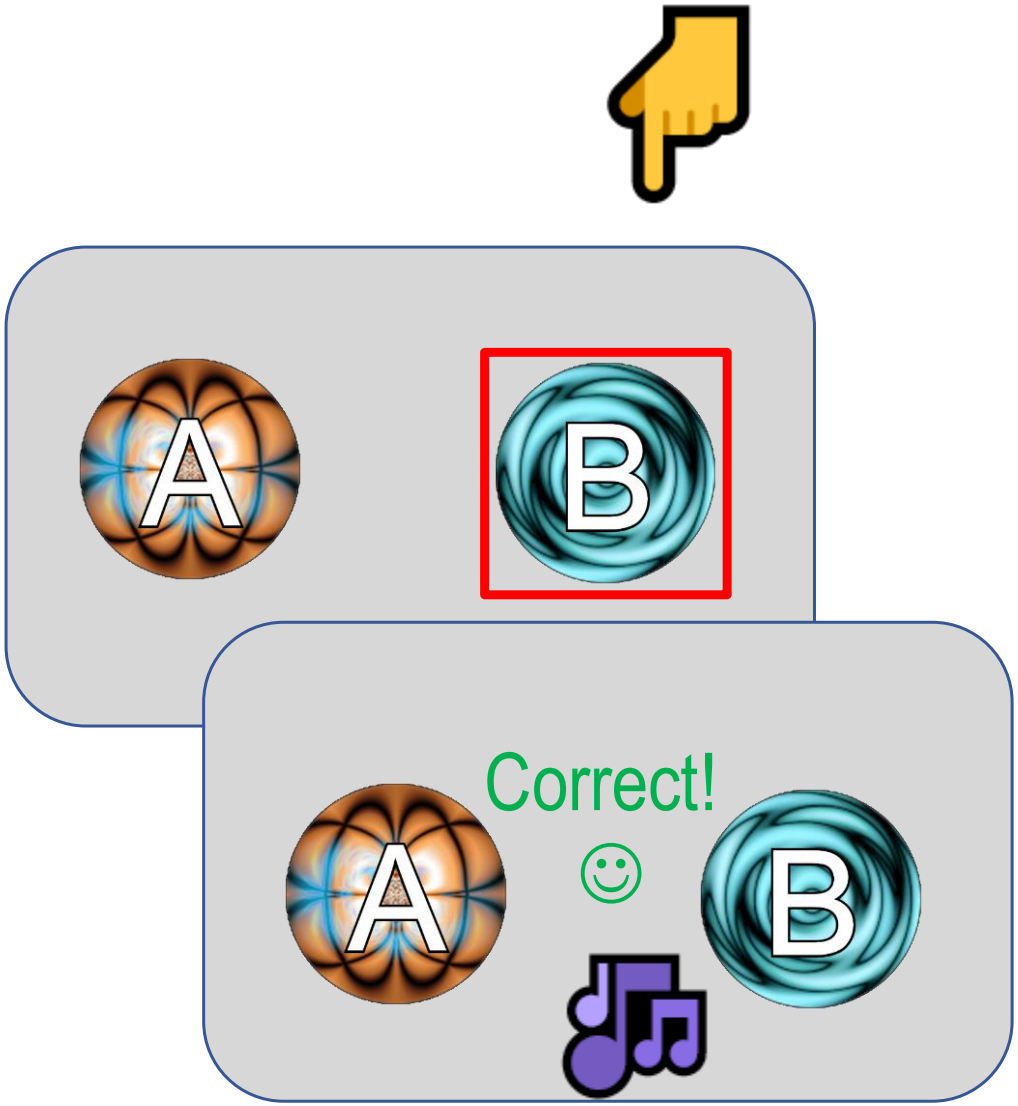
reversal phase #1

9-14 trials

reversal phase #2...9

9-14 trials

**adapted for MRI response hardware*


A
B
Correct!
A
B

Issue #1: Response Deadline Effects

Design Question:

- **fMRI version**
  - 2-second response deadline, auto-advance if no response (missed trials possible)

- **Online version**
  - Response required before continuing (no missed trials possible)

(Third option: no deadline for responses at all)

Does this difference affect comparability of computational modeling results?

**with credit to Blair Shevlin*

Issue #1: Response Deadline Effects

It depends on your model.

Likely minimal impact on:

- Simple choice models (e.g., Rescorla-Wagner) that only model which option is chosen:
  - Focus on how beliefs update based on outcomes, not the decision-making process itself.
  - Some missing trials okay.

CAVEAT: If number of missed trials substantially impacts both the cognitive process under investigation and model performance, then this design choice will be more problematic.

- Especially if clinical group vs. controls have big difference in % missed trials or % missed trials is correlated with something you care about.

*with credit to Blair Shevlin

Issue #1: Response Deadline Effects

It depends on your model.

Likely minimal impact on:

- Simple choice models (e.g., Rescorla-Wagner) that only model which option is chosen:
  - Focus on how beliefs update based on outcomes, not the decision-making process itself.
  - Some missing trials okay.

Likely major impact on:

- Response time models (Drift Diffusion) for the decision process.
  - Collapsing boundary effects (urgency → less evidence needed)
  - Changes the decision-making process.

Different deadline structures → different cognitive processes → incomparable parameters between versions of the "same task".

**with credit to Blair Shevlin*

Issue #1: Response Deadline Effects – Solution/Compromise

Goal:
Maintain comparability while preventing online participant disengagement.

Add "soft deadline" to online task:

- Feedback provided after 2s (*"TOO SLOW!"*)
  - Participants learn response window but still have to respond on every trial.
  - As with missed responses, responding too slowly = no feedback on your choice.

Issue #2: Temporal Separation between Choice and Feedback?

Design Question:

(**Choice → Immediate Feedback**) or (**Choice → Delay → Feedback**)?

Choice → Immediate Feedback:
+ More naturalistic, maintains typical learning dynamics
+ Ensures everyone completes full task w/in scan sequence duration
– Can't separate decision & feedback (temporal overlap in BOLD responses)

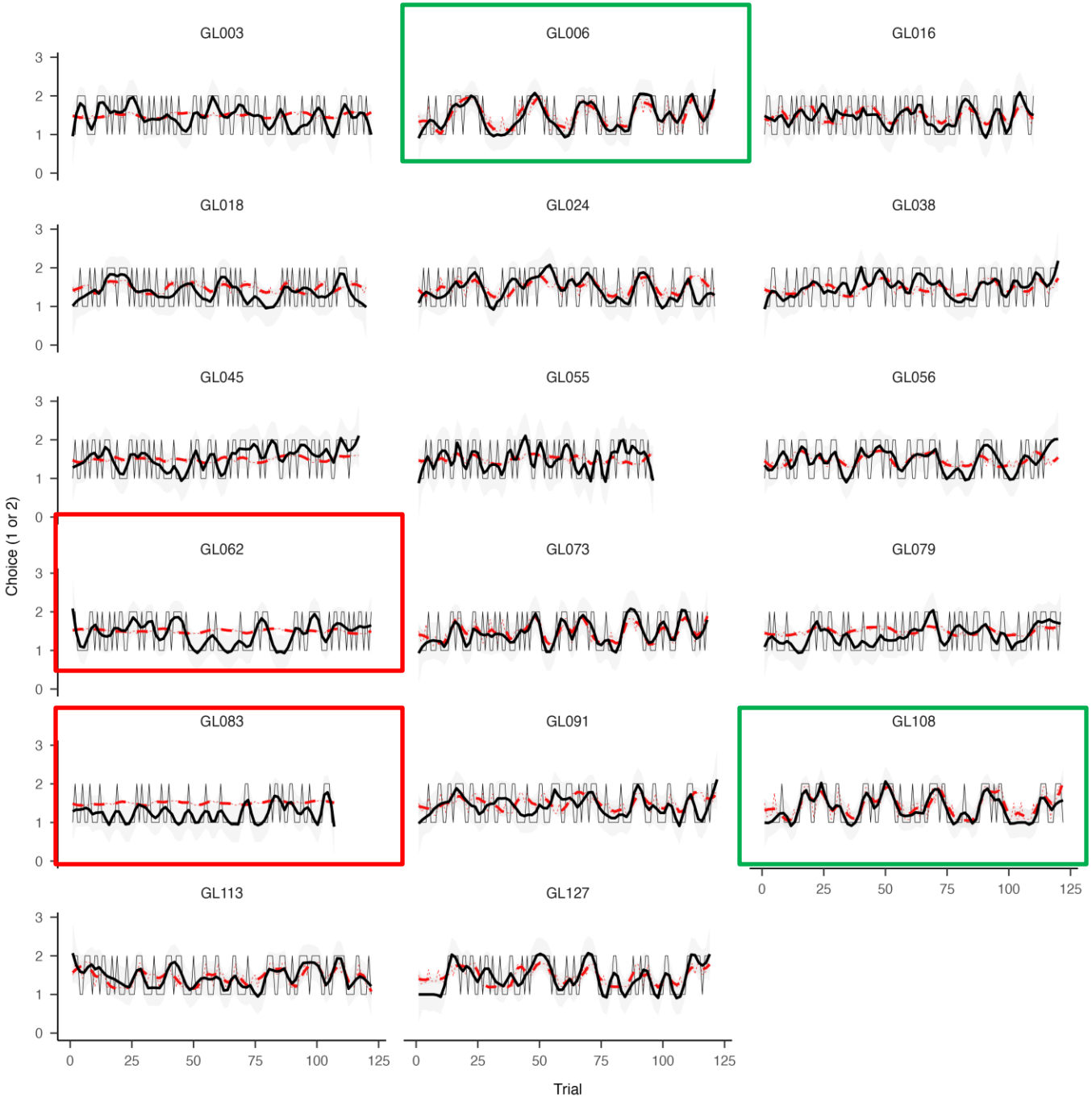
Choice → Delay → Feedback:
+ Can isolate neural correlates of choice, anticipation, feedback
+ Better BOLD deconvolution
– Temporal delay changes learning and belief updating (impairs RL)
– Less ecologically valid?

What is my priority?

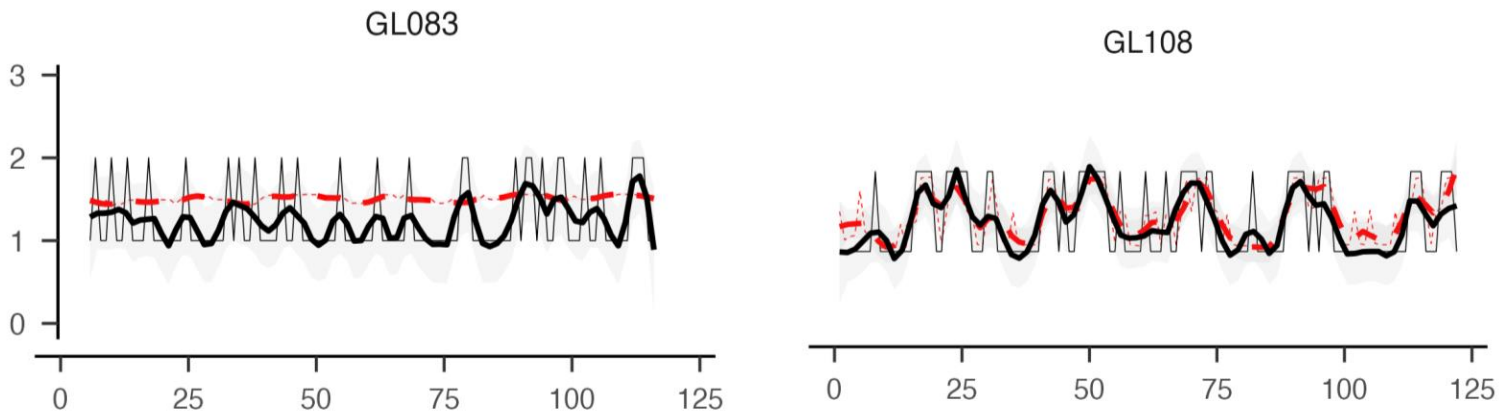
PRL Disadvantages & Other Issues

- **Not ideal:**
  - Is this grief specifically or general distress/psychopathology?
  - Is this kind of learning (short-term, unidimensional, probabilistic) even relevant to grief, which unfolds over time?

- **Unexpected issues:**
  - Difficulty comprehending “probabilistic” aspect
    - Post-task questions
  - People w/higher PGD have noisier performance:
    - Affects parameter recovery & model fit to actual data


GL003
GL006
GL016
GL018
GL024
GL038
GL045
GL055
GL056
GL062
GL073
GL079
GL083
GL091
GL108
GL113
GL127
Choice (1 or 2)
Trial
0
1
2
3
0
25
50
75
100
125

Posterior Predictive Checks: Two Subjects



What are some of the experimental design tensions or tradeoffs you might encounter in your own study?

What advantages/disadvantages would you prioritize?

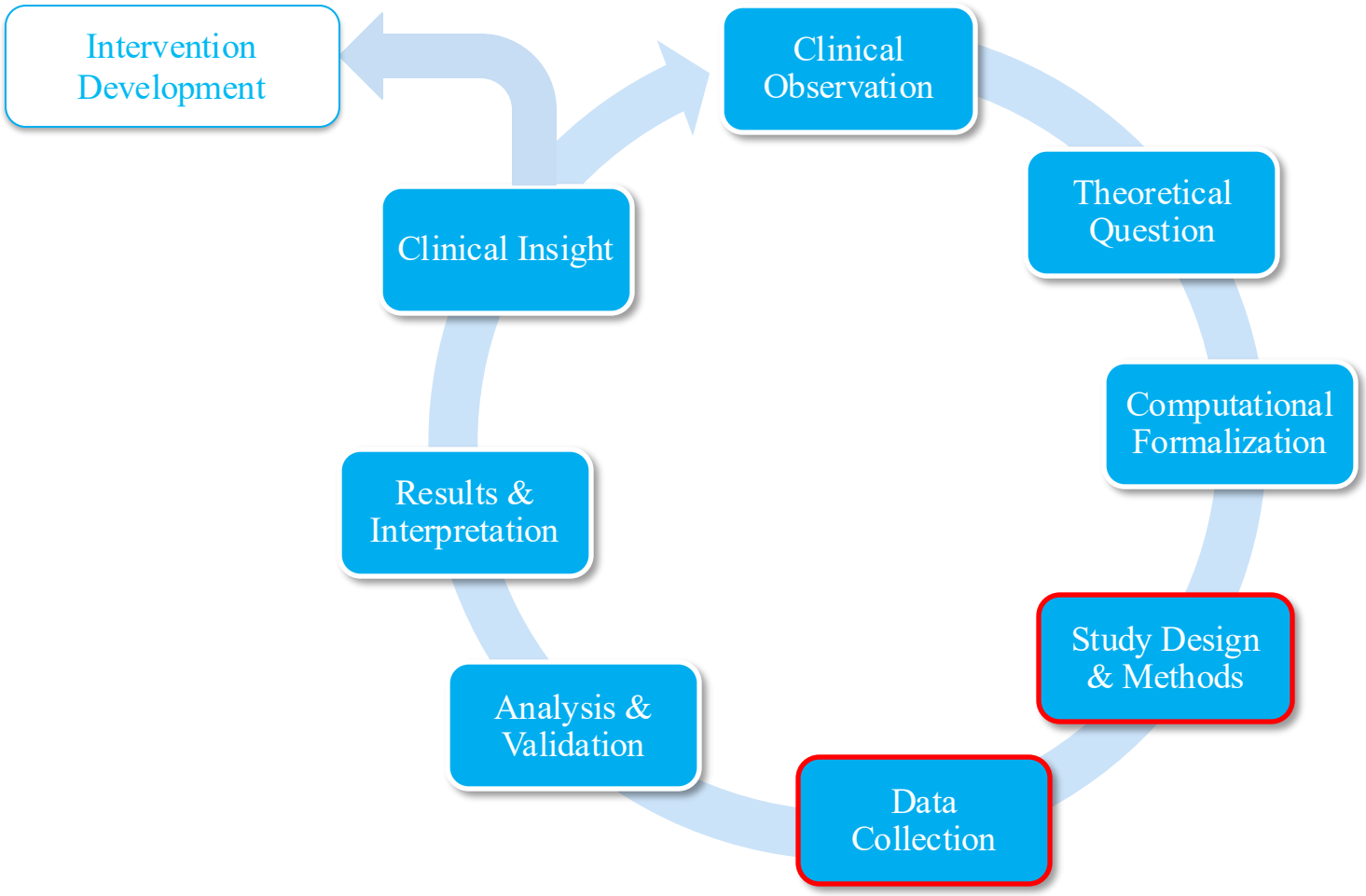
Takeaway:

- Goal: balance psychometric quality, theoretical validity, practical feasibility, & model appropriateness and reliability.

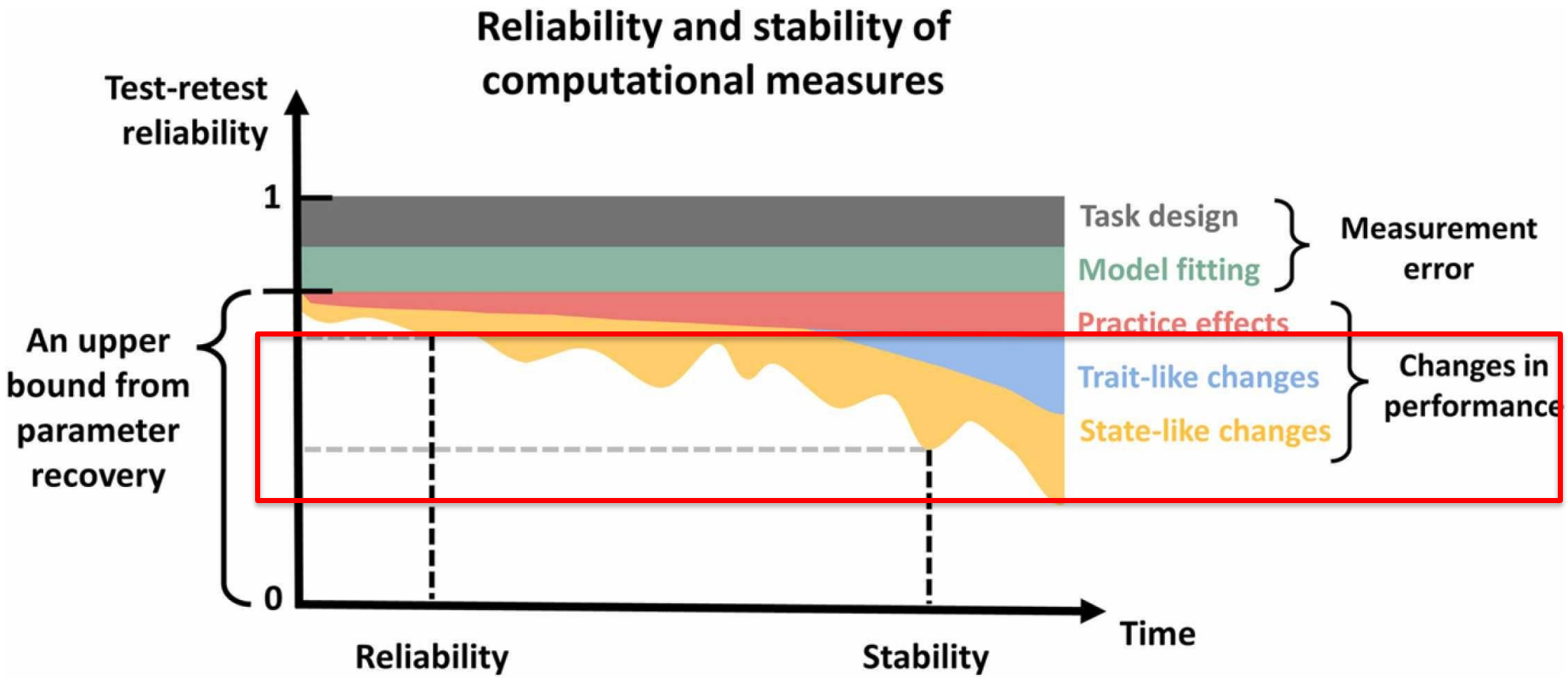
- No task is perfect, but it should be "good enough" *for your purposes*.
  - Be aware of alternative explanations → measure those factors.
  - Post-task checks
  - Be explicit about limitations and triangulate with multiple measures (*convergent validity of your latent construct*)

- **There is NO substitute for piloting & feedback!**

Scientific Lifecycle in Computational Psychiatry



Participants



Sandve et al. (2013) PLoS Comput Biol. doi:10.1371/journal.pcbi.1003285

Who's in the study?

- Clinical assessment vs. self-report?
  - Categorical vs. continuous?
- Inclusion/exclusion criteria?
- "Healthy" controls ?
- Is WNL *within normal limits* or *we never looked*?
- Ecological validity vs. experimental control?
  - Covariate selection
  - Medication
  - Comorbidity (psychiatric and medical)
  - State effects (standardize vs. not?)
  - Sensitivity analysis

When people are engaged, data quality is better.

When participants feel we are invested in them, they are more invested in us.

- Interactions with participants before, during, and after the study
  - Anticipate anticipatable occurrences
    - Clarify research vs. treatment
    - Grief interview → crying
    - fMRI → anxiety
    - Ask about suicidality → disclosed SI (need protocol)
- Understanding motivation to participate
- Framing & vocabulary/jargon
  - Is it really a DEFICIT, or is it that what they're doing doesn't match demands of current environment?

Issue #1:

- **I want to do an actual clinical assessment, but the full SCID and CAPS-5 interview were lengthy & sometimes poorly tolerated (in a past study)**

  – *What is the information I need to get from this assessment?*
  - Differential diagnosis of PGD vs. other conditions
  - Rule out exclusionary diagnoses

  – *What resources do I have?*
  - Just me (and at 15% effort)

Solution:

- 1–2-hour interview:
  – SCI-PG
  – SCID-5 modules for current PTSD, MDD, SUD + lifetime psychosis & bipolar disorder

Hi Saren,

Thank you for sharing the preliminary results of the Grief-LEARN Study.

…

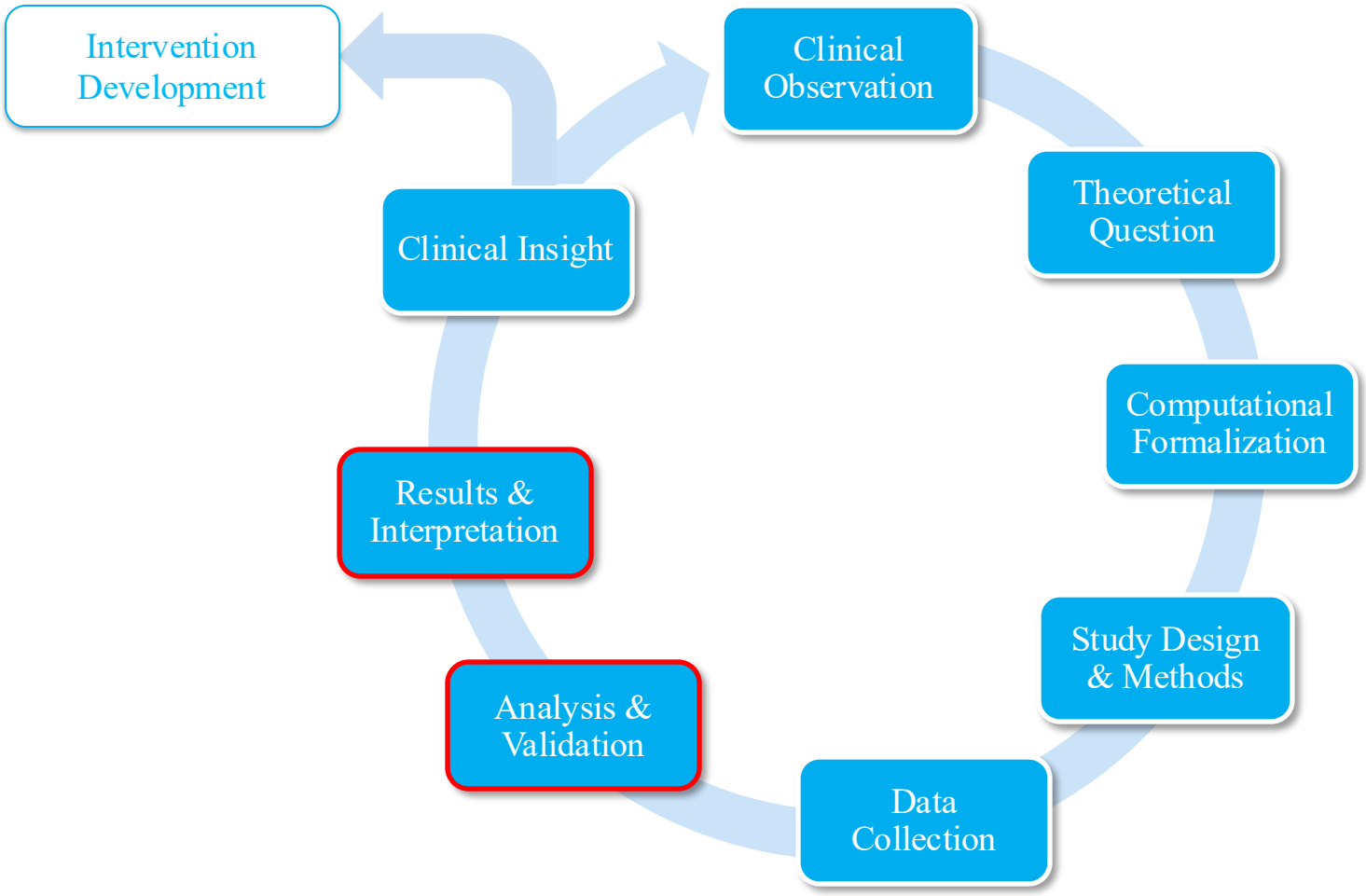
Interacting with you on this study was a process that led to a greater self-awareness of what I was going through. It helped me express my feelings as I had not spoken to anyone about how I was feeling and helped me more than you'll ever know.

Thank you!!

Are there any special considerations for your sample population?

What about your sample (other than the thing you're interested in) could impact their participation & data?

Scientific Lifecycle in Computational Psychiatry



Staying on top of your “researcher degrees of freedom”

Variability as Signal AND Noise

Why experimental variation in neuroimaging should be embraced

Gregory Kiar ✉, Jeanette A. Mumford, Ting Xu, Joshua T. Vogelstein, Tristan Glatard & Michael P. Milham

Nature Communications **15**, Article number: 9411 (2024) | Cite this article

Abstract

In a perfect world, scientists would develop analyses that are guaranteed to reveal the ground truth of a research question. In reality, there are countless viable workflows that produce distinct, often conflicting, results. Although reproducibility places a necessary bound on the validity of results, it is not sufficient for claiming underlying validity, eventual utility, or generalizability. In this work we focus on how embracing variability in data analysis can improve the generalizability of results. We contextualize how design decisions in brain imaging can be made to capture variation, highlight examples, and discuss how variability capture may improve the quality of results.

Kiar et al. 2024 *Nat Commun*. doi:10.1038/s41467-024-53743-y

Variability as Signal AND Noise

Data Selection	—	Is the data population representative to answer the posed question? Have multiple permutations of the input data been shown to produce similar results? Are the results robust across measurement equipment (e.g., scanner manufacturer)?
Expert Opinion	—	Do any choices in the design rely on subjective interpretation or preference? Are any manual steps (e.g., quality control) or data filtering process applied? Are the findings robust to the selection of priors (e.g. brain parcellation)?
Analytic Decisions	—	Are there other valid statistical approaches to answer the posed question? Are there other configurations of confounding variables that could be considered? Are there other possible processing steps or configurations that could be used?
Tool Selection	—	Could the processing pipeline be constructed in another software library? Are there other relevant measures or modalities that could be considered?
Computational Infrastructure	—	Have the findings been replicated across computer systems (hardware or software)? Are the results consistent across different versions of the tested libraries or data?
Numerical State	—	Are the findings robust to data-specific (e.g., acquisition) noise? Are the results deterministic, or dependent upon a specific random state? Have the numerical stability and error properties of tools been tested?

Pre-Analytic

(Post-) Analytic

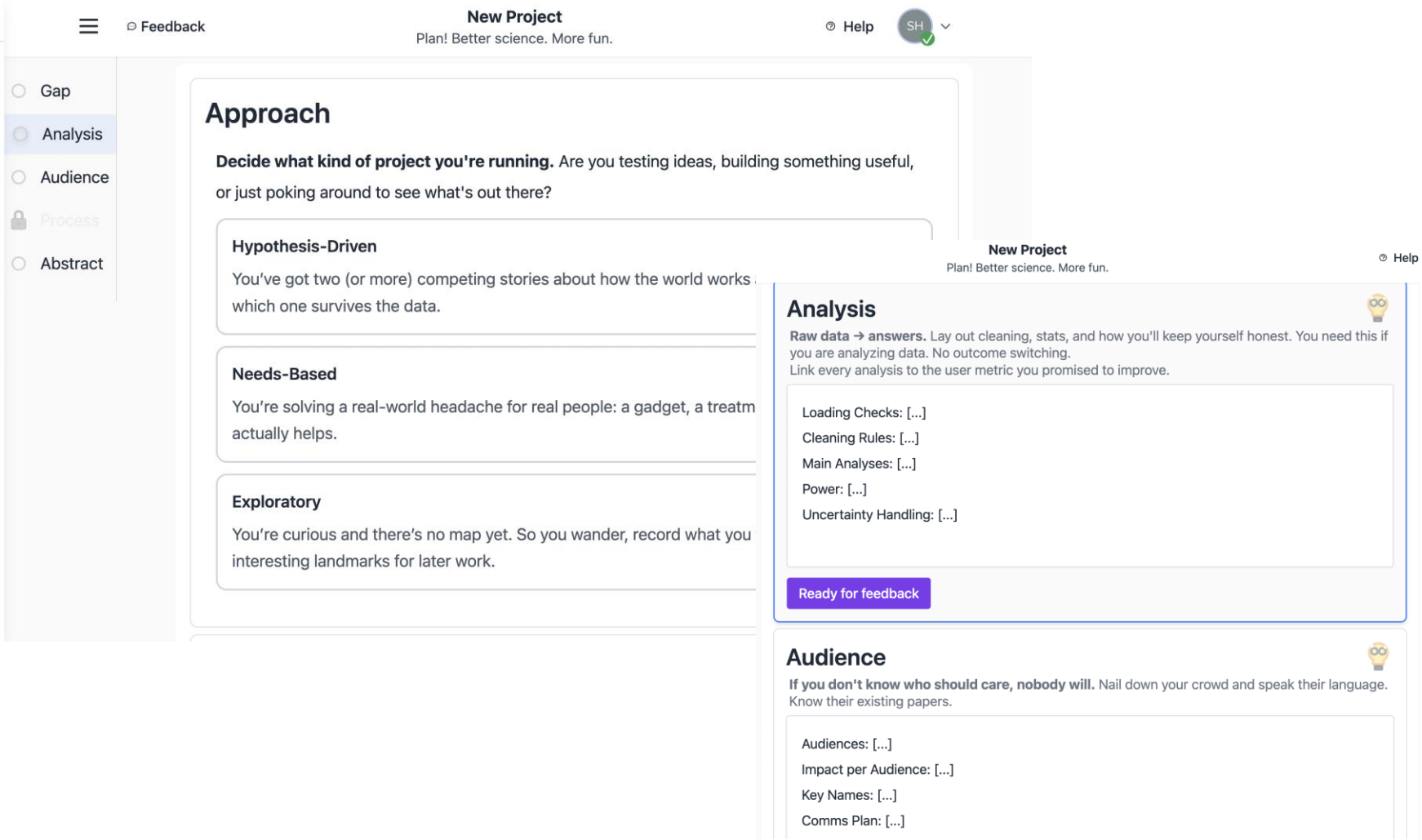
Kiar et al. 2024 *Nat Commun*. doi:10.1038/s41467-024-53743-y

Example: Bayesian ideal observer model – Stop Signal task

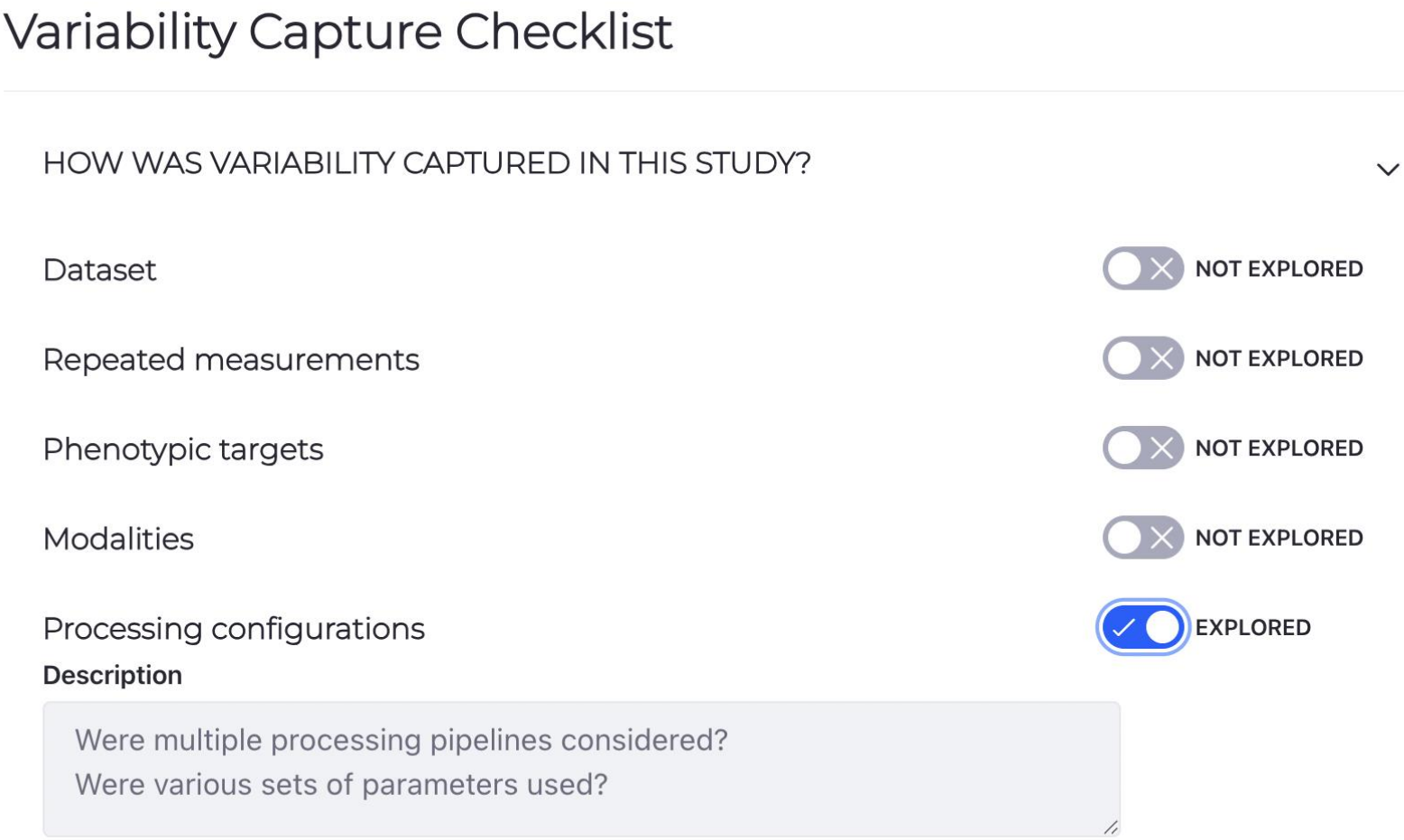
Model variants (ways to implement a single model):
Fixed belief (a=1.0, pm = .25, scale = 10)
MolPsych paper parameters (a=.99, pm = .25, scale = 10)
MolPsych approach (find best a, but pm = .25, scale = 10)
univK23 (fit on full sample, best full range parameters)
univK23 (best but prior mean constrained to .10-.40 range)
Group x State (fit on GxS combos, best full range parameters)
Group x State (best but prior mean constrained to .10-.40 range)
Subject x State (best full range)
Subject x State (pm .10-.40, others free)
Subject x State (a = .6-.95, pm = .25 or .1-.4, or...?)

How will you adjudicate between models?

PlanYourScience.com



NMIND Variability Capture Checklist



https://www.nmind.org/variability-capture-checklist

HOW WAS VARIABILTY USED IN THE ANALYSES?

Stability of features/effect size — USED

Description

Such as measures of test-retest reliability (e.g. discriminability), the number of significant digits in results, variance in model performance, or measures of group differences.

Repeated Measures Analysis — NOT USED

Meta-analysis — NOT USED

Meta-study — NOT USED

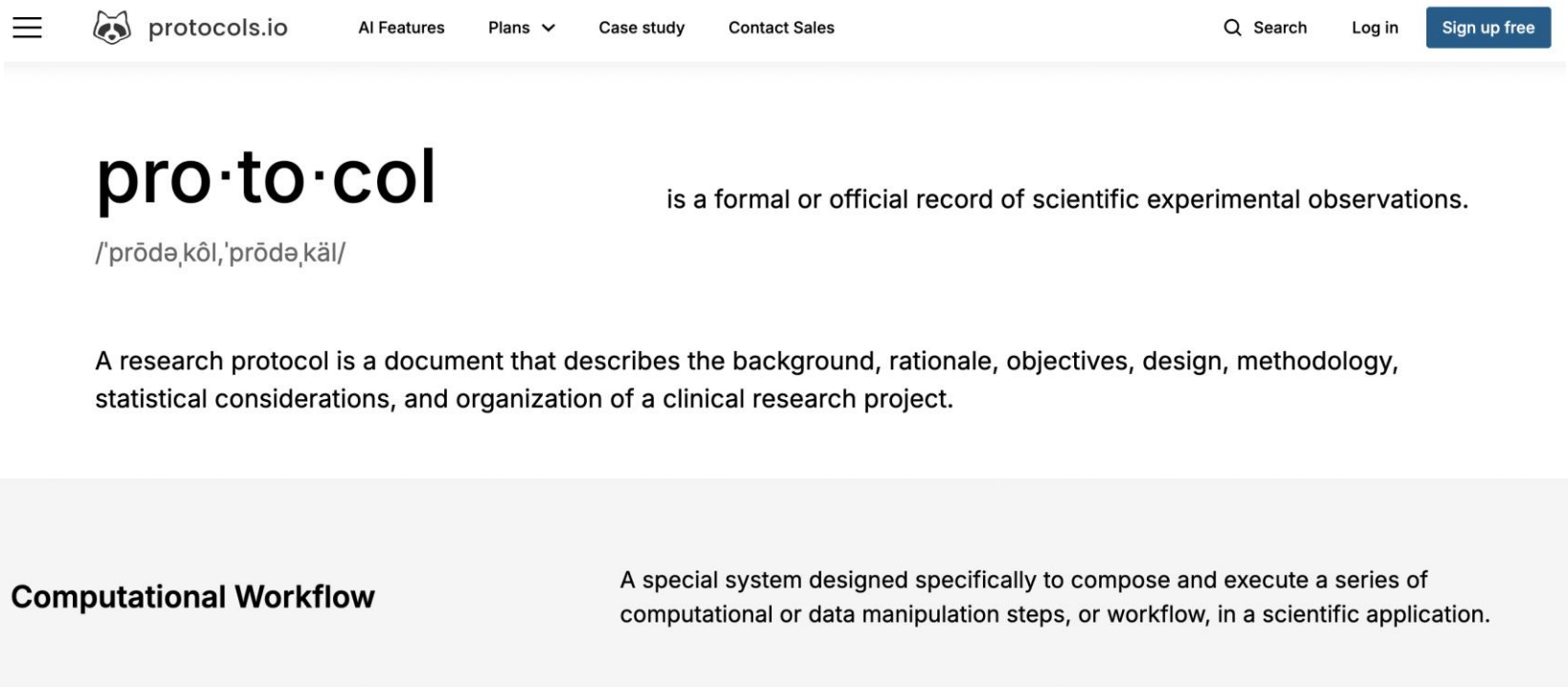
Dataset augmentation — NOT USED

Other — NOT USED

REMAINING BIASES & IMPACT

Potential remaining sources of bias

protocol.io


protocols.io
AI Features
Plans
Case study
Contact Sales
Search
Log in
Sign up free
pro·to·col
/ˈprōdəˌkôl,ˈprōdəˌkäl/
is a formal or official record of scientific experimental observations.
A research protocol is a document that describes the background, rationale, objectives, design, methodology, statistical considerations, and organization of a clinical research project.
Computational Workflow
A special system designed specifically to compose and execute a series of computational or data manipulation steps, or workflow, in a scientific application.

Conclusion/Summary

“Meaningful Science” questions to ask yourself

- **What <u>specific</u> clinical phenomenon am I studying?**
  - *“low anticipatory social reward” vs. “psychosis”*
- **Is my goal to predict, explain, or describe?**
- **Why does this computational construct matter for this phenomenon?**
  - *How does parameter X relate to symptom Y mechanistically?*
  - *What theory am I testing?*
- **What are the competing explanations?**
  - *Can my design/models distinguish them? (Can I show a double dissociation?)*
  - *What would it mean if I do not find the hypothesized effects? How can I make sure a null finding is still informative?*
- **Why is it important to answer this question?**
  - *Theory: Would this support/refute a mechanistic account? If this is true, so what?*
  - *Practice: How would this inform treatment, prediction, or understanding?*

“Robust Science” Questions to ask yourself

- **Can I measure this construct reliably in this population using this method and approach?**
  - *Psychometrics*
  - *Task appropriateness*
  - *Model appropriateness and performance in other populations*

- **Is my design adequate to test my hypothesis?**
  - *Power, sample size, controls*
  - *Context-specific or domain-general effect? Temporal aspects?*
  - *Confounds/covariates*

- **How will I validate my approach?**
  - *Parameter recovery, model recovery, predictive validity*

- **Can others (& future-me) reproduce/replicate and build on my work?**
  - *Data/code sharing, documentation, version control, reporting standards*

Aspirational goals for robust and meaningful computational psychiatry research in a nutshell:

- ❑ Starts with clinical observation or theoretical question.
- ❑ Understands what the brain is computing & how the computation is implemented neurobiologically.
- ❑ Formalizes mechanistic hypotheses in computational terms.
- ❑ Assay tests those specific mechanisms.
- ❑ Assay/measures have adequate psychometric properties.
- ❑ Makes intentional choices (experimental control / ecological validity).
- ❑ Uses valid & appropriate stimuli, interventions, controls.
- ❑ Has purposeful inclusion/exclusion criteria.
- ❑ Appreciates the impact of participation context on participants.
- ❑ Validates the modeling approach.
- ❑ Adequately powered for meaningful effects.
- ❑ Relates parameters to specific symptoms/outcomes.
- ❑ Interprets results in context of existing theory/data.
- ❑ Reproducible and transparent methods (documentation!)
- ❑ Generates testable predictions for future work & refines theory.
- ❑ Produces clinically-actionable insights.

What questions do you have?

Acknowledgements

- Laura Berner, Ph.D.
  - Blair Shevlin, Ph.D.
  - Berner Lab members
- Center for Computational Psychiatry faculty & members

- Grief-LEARN Study:
  - Ruiyuan Guo, MA
  - Riley Macks, B.A.
  - Adriana Feder, M.D.
  - Daniela Schiller, Ph.D.
  - Mary-Frances O'Connor, Ph.D. & NOGIN members