Neuroscience and Biobehavioral Reviews 148 (2023) 105137

Contents lists available at ScienceDirect

Neuroscience and Biobehavioral Reviews

journal homepage: www.elsevier.com/locate/neubiorev

Individual differences in computational psychiatry: A review of current challenges

Povilas Karvelis [a,*], Martin P. Paulus [b,c], Andreea O. Diaconescu [a,d,e,f]

[a] Krembil Centre for Neuroinformatics, Centre for Addiction and Mental Health (CAMH), Toronto, ON, Canada
[b] Laureate Institute for Brain Research, Tulsa, OK, USA
[c] Oxley College of Health Sciences, The University of Tulsa, Tulsa, OK, USA
[d] Department of Psychiatry, University of Toronto, Toronto, ON, Canada
[e] Institute of Medical Sciences, University of Toronto, Toronto, ON, Canada
[f] Department of Psychology, University of Toronto, Toronto, ON, Canada

ARTICLE INFO

Keywords:
Computational psychiatry
Reliability
Validity
Computational modelling
Individual differences

ABSTRACT

Bringing precision to the understanding and treatment of mental disorders requires instruments for studying clinically relevant individual differences. One promising approach is the development of computational assays: integrating computational models with cognitive tasks to infer latent patient-specific disease processes in brain computations. While recent years have seen many methodological advancements in computational modelling and many cross-sectional patient studies, much less attention has been paid to basic psychometric properties (reliability and construct validity) of the computational measures provided by the assays. In this review, we assess the extent of this issue by examining emerging empirical evidence. We find that many computational measures suffer from poor psychometric properties, which poses a risk of invalidating previous findings and undermining ongoing research efforts using computational assays to study individual (and even group) differences. We provide recommendations for how to address these problems and, crucially, embed them within a broader perspective on key developments that are needed for translating computational assays to clinical practice.

1. Introduction

Computational psychiatry aims to bring precision to the understanding and treatment of mental disorders (Yip et al., 2022; Huys et al., 2021; Hauser et al., 2022). This requires developing tools that can capture clinically relevant individual differences. One of the main approaches to this end is the development of *computational assays*, which combine cognitive tasks with computational models to infer patient-specific disease processes from behavioral or brain data (Stephan and Mathys, 2014). The models are constructed to approximate the underlying neurocomputational mechanisms and are generally fit to data from tasks that are specifically designed to probe symptom-relevant cognitive functions. This provides various *computational measures*, including parameter estimates (e.g., learning rate, prior precision) as well as other dynamic variables (e.g., trial-by-trial prediction errors and beliefs) that aim to characterize the latent computational processes shaping brain activity and behavior. The hope is that these computational measures would capture clinically relevant individual differences, which could open the door for a wide array of applications, ranging from computational phenotyping to treatment response prediction to the development of new treatments and treatment targets (Paulus et al., 2016; Patzelt et al., 2018; Yip et al., 2022; Hauser et al., 2022).

Many methodological advancements have been made towards fulfilling these goals, particularly in the areas of model development, model fitting, and model comparison (e.g., Stephan et al., 2017; Frässle et al., 2021; Smith et al., 2021a). However, basic psychometric properties (*reliability* and *construct validity*) of computational measures have received much less attention (Browning et al., 2020; Paulus et al., 2016). Many studies have investigated which computational measures are associated with clinical ones without knowing the reliability and construct validity of these measures (Fig. 1). If the measures are not reliable enough, spurious associations become more likely than true associations (Hedge et al., 2018). If there is an unrecognized lack of

* Corresponding author.
E-mail address: povilas.karvelis@camh.ca (P. Karvelis).

https://doi.org/10.1016/j.neubiorev.2023.105137
Received 19 January 2023; Received in revised form 4 March 2023; Accepted 14 March 2023
Available online 20 March 2023

construct validity, it leads to misinterpretation of what the measured individual or group differences mean (Yarkoni, 2022).

This is particularly concerning in the context of recent work showing that many commonly used computerized tasks in cognitive sciences, while showing replicable group effects, are not reliable enough for studying individual differences (Rodebaugh et al., 2016; Hedge et al., 2018; Enkavi et al., 2019; Elliott et al., 2020; Kennedy et al., 2022; Nitsch et al., 2022). Furthermore, the validity of many tasks has also been challenged, with studies reporting a lack of correlations among different measures (i.e., different cognitive tasks, self-report) that are meant to capture the same constructs, such as cognitive control (Saunders et al., 2018; Enkavi et al., 2019; Eisenberg et al., 2019; Gärtner and Strobel, 2021; Friedman and Gustavson, 2022), risk preference (Pedroni et al., 2017; Buelow and Barnhart, 2018; Frey et al., 2017), distress tolerance (McHugh et al., 2011), reinforcement learning (RL) (Eckstein et al., 2021, 2022), reliance on visual priors (Grzeczkowski et al., 2017, 2018; Tulver et al., 2019), or positive and negative valence (Peng et al., 2021). Taken together, these results call for a closer examination of reliability and validity of assays in computational psychiatry (Paulus et al., 2016; Hitchcock et al., 2017; Hedge et al., 2020; Enkavi and Poldrack, 2021).

The main goal of this review is to provide a critical assessment of the psychometric properties of computational assays, considering the unique challenges and opportunities that computational methods bring to the table. A more general goal is to provide a broader perspective on milestones towards clinical applicability, clarify key concepts, and suggest a natural prioritization of existing subproblems (Fig. 1). A meta goal here is to encourage viewing current psychometric issues through the prism of long-term clinical translation, so as to facilitate finding more globally optimal ways to address current challenges and to make research more efficient.

Given its primacy, the first and largest part of the review focuses on reliability of computational assays, where we review emerging empirical findings from computational studies. We then turn to construct validity, where (due to a shortage of computational studies) we primarily focus on broader literature investigating behavioral task measures and consider the implications it has for testing computational accounts of mental disorders. Finally, we briefly discuss the intrinsic limitations of cross-sectional clinical validity studies and make a case for the importance of focusing on predictive and longitudinal validity, as well as the need to make the assays more efficient and less burdensome.

2. What is reliability?

Reliability of a measurement is the degree to which it yields similar results when repeated under equivalent conditions (Cook and Beckman, 2006; Mokkink et al., 2010). In other words, it refers to how precisely we can measure what we want to measure. If an instrument is not sufficiently reliable (i.e., it is too imprecise), that undermines all subsequent research goals from construct validity to clinical utility (Fig. 1).

Reliability measures can have different names depending on the measurement protocol. *Test-retest* reliability evaluates the agreement of measurements obtained over repeated administrations of the same instrument to the same individuals. *Internal consistency*, such as *split-half reliability* is concerned with consistency among the different parts of the instrument itself, such as responses on similar trials in a task (e.g., by comparing odd trials vs. even trials) or on similar items in a questionnaire. *Inter-rater reliability* refers to the agreement of ratings between different human raters (e.g., in clinician-rated assessments). *Parallel form reliability* refers to consistency between two interchangeable versions or ways of administering the instrument (e.g., lab-based vs. online-based task completion). In this review, our primary focus will be test-retest reliability of behavioral and computational measures of task performance. We will also briefly comment on test-retest reliability of self-report questionnaires and inter-rater reliability of clinical assessments.

The most widely used reliability index for continuous variables is intraclass correlation coefficient (ICC) (Fleiss, 2011; Koo and Li, 2016; Liljequist et al., 2019). ICC is the ratio of between-individual variance to total variance:

$$ICC = \frac{Variance\ \ between\ \ individuals}{Variance\ \ between\ \ individuals + Variance\ \ between\ \ sessions + Error\ \ variance} \quad (1)$$

Note that the partitioning the variance in this way already reflects that reliability of a measurement might be low if (1) there is too much variance between sessions, (2) there is too much measurement error, and (3) there is too little between-individual variance. We will unpack these factors in detail throughout the paper.

There are two definitions of the ICC for test-retest or inter-rater reliability: *agreement ICC* provides an estimate of absolute agreement or concordance between the measurements and does not tolerate any systematic errors, while *consistency ICC* allows for a systematic offset error (but not scaling factor error) between the measurements.[1] For comparison, Pearson correlation coefficient estimates only a linear relationship between measurements and is not sensitive to any systematic errors, making it less suitable as a measure of reliability (Koo and Li, 2016). As we will see later, however, many studies still use Pearson or even rank correlations to measure reliability (Table 1).

ICC is generally described by using labels *poor*, *fair/moderate*, *good*, and *excellent*. How this maps to the actual values differs depending on whose guidelines are being followed. Fleiss (2011) proposed the different brackets to be: < 0.4 (poor), 0.4–0.59 (fair), 0.6–0.74 (good), and > 0.75 (excellent); while Koo and Li (2016) proposed more conservative brackets: < 0.5 (poor), 0.5–0.75 (moderate), 0.75–0.9 (good), and > 0.9 (excellent). Both of these labelling guidelines are for the most part arbitrary, and studies reviewed here chose one or the other. For consistency we will adopt the more conservative option (Koo and Li, 2016) throughout the review, however, we suggest treating these as ordinal labels and suspending semantic interpretations for the time being. For example, "good" should be read as being better than "moderate" or as indicating a certain range of ICC values, but we should be careful not to conclude that it means "good enough for any further analysis or clinical applications", as it might not be (see section 6 for further discussion).

3. Reliability of cognitive tasks

Despite its importance, so far relatively little effort has gone into investigating reliability of commonly used cognitive tasks in computational psychiatry (Table 1). This might be due to a false assumption that a demonstration of group effects (e.g., evoked by different task conditions) in a task qualifies it to study individual differences in these effects. Quite counterintuitively, tasks that evoke robust group effects tend to have low reliability making them less suitable for studying individual differences (Weir, 2005; Cooper et al., 2017; Hedge et al., 2018). That is because robust group effects require low between-individual variability, while reliability - all else being equal - improves with higher between-individual variability (Eq. (1)). A recent study by Enkavi et al. (2019) has found the median test-retest reliability across a wide range of established self-regulation tasks to be very poor ($ICC = 0.31$), well below the median reliability of self-report questionnaires assessing self-regulation ($ICC = 0.67$). Similar issues have been found in task-based fMRI research. A recent meta-analysis of 90 studies found mean reliability of BOLD response across many common tasks to be poor ($ICC = 0.397$), and across 11 tasks used within the Human Connectome Project and the Dunedin Study to be similarly poor ($ICC = 0.067 -$

[1] note, when considering a linear relationship between the measurements, $x_2 = ax_1 + b$, scaling and offset factors are a and b, respectively

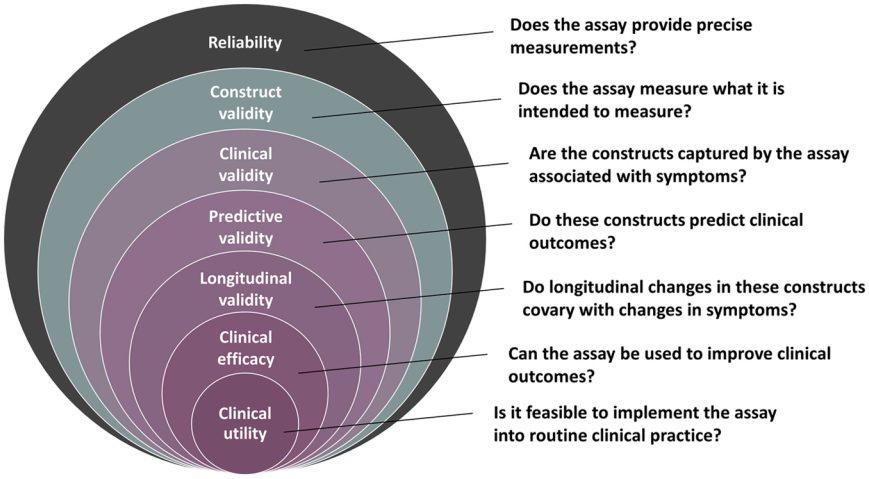

Fig. 1. A hierarchical breakdown of different challenges facing computational assays. Despite their primacy, reliability and construct validity of many computational assays remain largely unexamined. Most research tends to focus on clinical validity by looking for associations between computational and clinical measures. However, for such analyses to be meaningful, it is crucial to first ensure that the computational measures have sufficient reliability and construct validity. Furthermore, clinical validity studies usually focus on cross-sectional associations, which does not by itself provide actionable insights. To get closer to clinical utility it is necessary to study predictive and longitudinal validity. The final steps in clinical translation will require the demonstration of clinical efficacy in randomized controlled trials and clinical utility in a range of real-world settings.

0.485) (Elliott et al., 2020). Recent emerging evidence suggests that many tasks used in computational psychiatry might have poor reliability as well (Table 1).

3.1. Reliability and task design variability

First and foremost, these findings suggest that if reliability is not deliberately optimized for, it will tend to be low. Reliability can be affected by multiple experimental design factors: the number of trials, the duration of inter-stimulus intervals, the number of practice trials, time constrains, overall task difficulty (leading to ceiling or floor effects), the instructions given before the task, or even the population it is tested in, etc. (Henderson et al., 2012; Hitchcock et al., 2017; Cooper et al., 2017; McLean et al., 2018; Plummer et al., 2015; Zorowitz and Niv, 2023). This also means that different versions of tasks that are often referred to by the same name (e.g., Go/No-Go task) can have very different reliabilities, depending on the details of implementation. Because of that, often used statements in the form of 'task X is (un) reliable' might be not very informative, as it generally speaks to a specific implementation of that task. This is particularly relevant in computational psychiatry research, where it is common to adapt task design on study-by-study basis to address a particular research question about a particular clinical group (Nair et al., 2020). Such practice makes it difficult to rely on previous findings for ensuring reliability and calls for making reliability reporting a routine practice (Parsons et al., 2019). Another helpful practice would be to make the task code accessible to other researchers: this would ensure that no important design decisions are left unmentioned and would enable a much more detail comparison among task designs and their effects on reliability.

3.2. Reliability and different ways of deriving task measures

'Task X is (un)reliable' can be a misleading shorthand for another reason. First of all, even within the same task, some measures can have low and some can have high reliability (e.g., Loosen et al., 2022; Waltmann et al., 2022); the reliable ones could still be used for further analysis. Second of all, reliability of measures depends not only on task parameters, but also on how such measures are derived. The typical measures involve simple summary statistics (e.g., averaging reaction times or accuracy across trials) and difference scores (e.g., the differences of averages across conditions). Differences scores tend to have lower reliability as they often mask a portion of between-subject variance (Cronbach and Furby, 1970; Hedge et al., 2018; Draheim et al., 2019). Relying on measures that do not involve difference scores could thus yield better reliability. However, this might make it more difficult to test certain hypotheses that are framed in terms of differential performance across conditions.

Simple averaging across trials has its problems too as it implicitly assumes no uncertainty in such averages by ignoring trial-level variance (Rodebaugh et al., 2016; Haines et al., 2020). This can be addressed by estimating the measures of each individual using linear mixed models that take into account the variance across multiple levels - trial, individual, group, etc. (Rouder and Haaf, 2019; Chen et al., 2021). This approach exploits the hierarchical properties of the data across these levels. For example, in addition to accounting for trial-level variance, it also uses individual-level estimates to compute group priors, which in turn regularize individual-level estimates. This leads to more precise and thus more reliable estimates of the measures, especially, when the number of trials is low (Rouder and Haaf, 2019; Chen et al., 2021). The mixed model approach can even be extended to incorporate both sessions by assuming the measures to be drawn from a multivariate distribution. The correlation in parameter values between sessions (i.e., their reliability) can then be obtained from the covariance matrix; this technique has been shown to further improve reliability estimates (Rouder and Haaf, 2019; Haines et al., 2020; Snijder et al., 2022; Littman et al., 2022).

4. Can modelling task behavior provide more reliable measures?

Regardless of the methods used, behavioral measures remain rather distant proxies of the underlying cognitive processes and thus are not ideal for capturing individual differences in these processes (Haines et al., 2020). An alternative is to model task behavior with generative models that explicitly specify the underlying cognitive process (e.g., RL, Bayesian inference), providing computational measures (parameter estimates and other dynamic variables) of this process. The resulting individual differences in these computational measures could be expected to be much closer to the theoretical constructs and processes of interest (Huys et al., 2016). Furthermore, cognitive modelling can easily address the other issues affecting the conventional measures: it can provide a more complete summary of the data (e.g., by jointly accounting for choice and reaction time data), it can extract valuable information from trail-level variability, especially if it relates to learning (e.g., with trial-by-trial modelling approaches), and in many cases it circumvents the need to use difference scores (Draheim et al., 2019). All of this should allow computationally derived measures to have higher reliability.

However, recent computational studies explicitly investigating this idea paint a more complicated picture (Table 1). Many studies report poor to moderate reliabilities of parameter estimates. In many cases it is similar to or even lower than the conventional summary statistics measures (Pike et al., 2022; Shahar et al., 2019; Smith et al., 2021b;

Table 1

Studies investigating test-retest reliability of computational assay measures. Note that 'joint' next to reliability measures refers to modelling the two testing sessions jointly (using bivariate distributions). ICC - intraclass correlation; we also indicate ICC definition with subscripts *a* and *c* for absolute and consistency, respectively, if it is reported. r - Pearson's correlation coefficient, ρ - Spearman's rank correlation coefficient, CCC - concordance correlation coefficient, RL - reinforcement learning, DDM - drift-diffusion model; ML - maximum likelihood, MAP - maximum a posteriori, EB - empirical Bayes, HC - healthy controls, MDD - major depressive disorder, OUD - opioid use disorder, SUD - substance use disorder.

Study	Task	Test-retest reliability		Parameter recovery	Practice effects	Trials	Interval	Sample
		Behavioral measures	Model parameters					
Chung et al., 2017	Gambling task	ρ = .33-.48 (T2) ρ = .36-.37 (T3) ρ = .23-.55 (T4)	Utility theory: ρ = .30-.52 (T2) ρ = .41-.63 (T3) ρ = .32-.65 (T4) (EB)	Not performed	Not investigated	18	3 follow-ups (T2, T3, T4) each 5.5 weeks apart	33 HC 65 MDD
Moutoussis et al., 2018	Associative learning Go/No-Go task	ρ = .15 (T3)	RL: ρ = .08-0.3 (T2) r = .08-0.25 (T3) (EB)	r = .25-.90	Present, but difficult to disentangle from developmental factors	144	2 follow-ups at 6 months (T2), and at 18 months (T3)	61 (T2) 503 (T3) young HC
Price et al., 2019	Dot-probe task	ICC = .001	DDM: ICC = .25 (ML)	Performed, but no correlational analysis presented	Not investigated	300	2 months	70 anxiety
Shahar et al., 2019	Two-stage decision task	r = .28-.33 (non-hier.) r = .33-.40 (hierarchical)	RL: ρ = .16 (ML) ρ = .21 (EB) DDM-RL: ρ = .20 (ML) ρ = .14 (EB)	RL: ρ = 0.53-0.89 DDM-RL: ρ = 0.58-0.99	Not investigated	201	18 months (median)	554 young HC
Brown et al., 2020	Two-stage decision task	Not reported	RL: Dataset 1: r = .13-.40 (ML) r = .39-.46 (EB) r = .72-.89 (EB + joint) Dataset 2: r = .17-.19 (ML) r = .31-.38 (EB) r = .57-.68 (EB + joint) Dataset 3: r = .07-.22 (ML) r = .20-40 (EB) r = .40-.62 (EB + joint)	Not reported (but recovery analysis of reliability results suggests good parameter recoverability)	No significant practice effects in model parameters	Dataset 1: 100 Dataset 2: 200 Dataset 3: 200	Dataset 1: 1 week Dataset 2: 4 months (median) Dataset 3: 18 months (median)	Dataset 1: 38 HC Dataset 2: 242 HC Dataset 3: 541 HC
Konova et al., 2020	Risky decision-making task	Not reported	Power Utility: ICC_a = .70-.72 (OUD) ICC_a = .87-.89 (HC) (ML)	Not performed	Not investigated	120	5 weekly sessions	70 OUD 55 HC

(*continued on next page*)

Table 1 (*continued*)

Ahn et al., 2020	Delay discounting (adaptive design optimization (ADO) and conventional staircase (SC) methods)	Not reported	Hyperbolic discounting model: Dataset 1: CCC = .90 (SC) CCC = .95 (ADO) Dataset 2: CCC = .95 (SC) CCC = .98 (ADO) Dataset 3: CCC = .97 (ADO) (MAP; fixed priors)	Not performed	Potential practice effects	Datasets 1 and 2: 42 Dataset 3: 20	Same day	Dataset 1: 58 HC Dataset 2: 35 SUD Dataset 3: 808 HC
Smith et al., 2021	Approach-avoidance conflict task	Dataset 1: ICC_c = .4-.61 Dataset 2: not reported	Active Inference: Dataset 1 ICC_c = .46-.52 Dataset 2: ICC_c = .54-.84 (MAP; fixed priors)	r = 0.79-0.91	Not investigated	Dataset 1: 60 Dataset 2: 90	Dataset 1: 12 months Dataset 2: 2-3 weeks	Dataset 1: 325 mixed: depression, anxiety, and SUD Dataset 2: 30 HC
Brown et al., 2021	Probabilistic operant learning task	Not reported	RL: r = ~0-.79 (EB + joint)	Performed, but no correlational analysis presented	Not investigated	50	12 weeks	20 HC
Bruder et al., 2021	Delay discounting task (virtual reality and standard lab settings)	Not reported	Hyperbolic discounting model: ICC_a = .34-.91 Temporal discounting DDM: ICC_a = .19-.7 (EB)	Not performed, but cites previous work	Not investigated	288	3 sessions 1-19 days apart	34 HC
Xu & Stocco, 2021	Probabilistic stimulus selection task	Dataset 1: ICC <0.25 Dataset 2: $ICC_a \approx 0$	Adaptive Control of Thought – Rational: Dataset 1: ICC_a = .49-.51 Dataset 2: ICC_a = .42-.43 (MAP; priors based on previous data)	Not performed	Not investigated	60	1 week	Dataset 1: 71 HC Dataset 2: 39 HC

(*continued on next page*)

Table 1 (*continued*)

Weigard et al., 2021	A battery of self-control tasks	Task-general factors of behavioral measures: ICC = 41-.76	Task-general factors of DDM parameters: ICC = .69-.78 (ML)	Not performed	Not investigated	32, local-global; 139, shape matching; 16, directed forgetting; 49, attentional network; 54, Stroop; 54, Simon; 29, task-switching; 24, choice response time	2-8 months	150 HC
Pike et al., 2022	Task 1: Associative learning Go/No-Go task Task 2: Ambiguous midpoint task	Task 1: ICC_a = .18-.50 Task 2: ICC_a = .49	RL (Task 1): ICC_a = ~0-0.2 DDM (Task 2): ICC_a = ~0-0.5 (EB)	Not performed	Not investigated	Not reported	2 weeks	58 HC
Waltman et al., 2022	Probabilistic reversal learning task	ICC_a = .25-.79 (hierarchical) ICC_a = .52-.92 (joint)	RL: ICC_a = 0-.20 (ML) ICC_a = .42-.64 (EB) r = .74-.86 (EB + joint)	r = .76-.93 (EB + joint)	Not investigated	160	1 week	40 HC
Smith et al., 2022	Three-armed bandit task	ICC_c = .15	Active Inference: ICC_c = .25-.54 (MAP; fixed priors)	r = .90-.94	Small to medium session effects	320	12 months	48 HC 83 SUD
Loosen et al., 2022	Predictive-inference task	ICC_a = .60-.80	Bayesian Learner: ICC_a = .41-.68 (ML)	Not performed	Not investigated	200	~3 months	330 HC
Sullivan-Toole et al., 2022	Iowa gambling task	r = .37 (non-hierar.) r = .41 (joint)	RL: r = .36-.65 (EB) r = .64-.82 (EB + joint)	r = .52-.95	Potential practice effects (larger variance between individuals during retest)	100	~1 month	50 HC

(*continued on next page*)

Table 1 (*continued*)

Hitchcock et al., 2022	Self-referential encoding task	ICC_a = ~.43-.63	DDM: ICC_a = ~.08-.62 (EB)	r = .83-.97	Not investigated	118	8 weeks	96 anxiety and depression
Howlett et al., 2022	Rapid assessment of motor processing	Not reported	Proportion-derivative model: ICC_c = .78-.81 (linear regression)	Not performed	Not investigated	20 (note, multiple data points per trial)	1-7 days	66 HC
Mkrtchian et al., 2023	Task 1: Restless four-armed bandit Task 2: Individually calibrated gambling task	Task 1: ICC_c = .46-.67 (non-hierar.) ICC_c = .63-.71 (joint) Task 2: ICC_c = .59-.63 (non-hierar.) ICC_c = .72-.73 (joint)	RL (Task 1): ICC_c = .03-.64 (EB) r = -.01-85 (EB + joint) Prospect Theory (Task 2): ICC_c = .70-.83 (EB) r = .87-.91 (EB + joint)	RL: ICC_a = .10-.97 Prospect Theory: ICC_a = .90-.97 (EB)	Small to medium practice effects for some parameters	Task 1: 200 Task 2: Not reported	2 weeks	50 HC

(Chung et al. (2017), Moutoussis et al. (2018a), Price et al. (2019), Shahar et al. (2019), Brown et al. (2020), Konova et al. (2020), Ahn et al. (2020), Smith et al. (2021a), Brown et al. (2021), Bruder et al.(2021), Xu and Stocco (2021), Weigard et al.(2021), Pike et al.(2022), Waltmann et al. (2022), Smith et al. (2022), Loosen et al.(2022), Sullivan-Toole et al.(2022), Hitchcock et al. (2022a), Howlett et al. (2022), Mkrtchian et al. (2023))

Loosen et al., 2022; Waltmann et al., 2022; Hitchcock et al., 2022b). In other cases it offers only a modest improvement over summary statistics (Price et al., 2019; Mkrtchian et al., 2023; Moutoussis et al., 2018a; Chung et al., 2017; Weigard et al., 2021), and rarely a substantial improvement (Sullivan-Toole et al., 2022; Xu and Stocco, 2021; Smith et al., 2022). Still, some studies achieved better reliability than others and it is important to consider the factors underlying that. Below we cover what we deem to be the most important factors, but see a recent review by Zorowitz and Niv (2023) for a complementary perspective.

4.1. Hierarchical model fitting methods can improve reliability

One of the clearest factors affecting reliability is the approach used for model fitting (Brown et al., 2020; Waltmann et al., 2022). Similar to the idea of using mixed models to account for uncertainty at different levels, hierarchical model fitting with empirical Bayes (EB) - where the parameter estimates at the individual level are informed by the group statistics and vice versa - can lead to better parameter estimates (Huys et al., 2011, 2012; Wiecki et al., 2013; Ahn et al., 2013; Katahira, 2016). This contrasts to maximum likelihood estimation (ML), where parameters of each individual are estimated separately and the uncertainty of such estimates is ignored. Using a RL model of the two-stage task, Brown et al. (2020) found EB to provide more reliable estimates ($r = 0.39 - 0.46$) compared to ML ($r = 0.13 - 0.40$). Similarly, Waltmann et al. (2022) found reliability of RL parameter estimates in the Probabilistic Reversal Learning Task to go from $ICC = 0.20$ using ML to $ICC = 0.42 - 0.64$ using EB. While these results highlight the benefits of using EB for parameter estimation, the resulting reliabilities are still rather poor. Moreover, many of the studies reporting poor reliabilities are already using EB methods (Moutoussis et al., 2018a; Shahar et al., 2019; Brown et al., 2020; Pike et al., 2022; Mkrtchian et al., 2023).

Just like for the behavioral measures, the hierarchical approach can be further extended to incorporate both sessions by assuming the parameter estimates to be drawn from a multivariate distribution (Brown et al., 2020; Sullivan-Toole et al., 2022; Waltmann et al., 2022; Pike et al., 2022; Mkrtchian et al., 2023). Using this method, Brown et al. (2020) and Waltmann et al. (2022) were able to further improve the reliability of parameter estimates, reaching $r = 0.72 - 0.89$ and $r = 0.74 - 0.86$, respectively. Similar improvements were reported by Sullivan-Toole et al. (2022) for a RL model of Iowa Gambling Task, with reliabilities increasing from $r = 0.36 - 0.65$ when modelling the sessions separately to $r = 0.64 - 0.82$ when modelling them jointly. A study by Mkrtchian et al. (2023) reported similar improvements for two other assays: for RL model of restless four-armed bandit task reliability increased from $r = 0.05 - 0.64$ to $r = -0.01 - 0.85$, while for a prospect theory model of a gambling task it increased from $r = 0.72 - 0.84$ to $r = 0.87 - 0.91$. Furthermore, using simulated data with known correlations of parameters between two sessions, Brown et al. (2020) and Waltmann et al. (2022) have demonstrated that this approach provides accurate estimates of correlation between the sessions, while modelling the sessions separately tends to underestimate this correlation. This means that joint modelling of the two sessions provides true improvements in reliability (rather than artificially inflating reliability by biasing values across sessions to be more similar).

This method, of course, is applicable only when there is data for more than one session but it does not help obtain better parameter estimates from a single session. Another caveat is that the reliability estimate derived from the covariance matrix is Pearson correlation, which is not sensitive to systematic errors and thus is not optimal for measuring reliability. Moreover, in their simulation analysis Waltmann et al. (2022) also showed that unlike reliability estimates derived from the covariance matrix, reliability computed directly from jointly estimated parameter values was positively biased. This raises the question whether using these point estimates in any subsequent analysis might introduce biases too.

4.2. Model simulations can provide an upper bound on reliability

A major benefit of computational modelling is that it allows for in silico analyses (Palminteri et al., 2017; Wilson and Collins, 2019). Many aspects of the task design (e.g., trial number, outcome probabilities, etc.), model properties (e.g., collinearity, complexity), and model fitting procedure (eg., ML vs EB), can be systematically studied and optimized through model simulations. Of particular interest in the context of reliability is parameter recovery analysis. Parameter recovery involves first simulating task behavior with a range of model parameter values and then fitting the same model to the simulated data. The correspondence between the true and the recovered model parameters indicates

how reliable parameters would be if the assumed model was a very good approximation of the actual cognitive process and there was no change in performance itself across sessions (Fig. 2). In other words, parameter recoverability provides an upper bound on reliability of measured individual differences. This does not necessarily mean that high recoverability guarantees high reliability but it does mean that low recoverability guarantees low reliability.

As such, parameter recovery results can be particularly informative when there is only one session (retest data is not available) and there is no other way of assessing retest reliability of parameter estimates. Unfortunately, parameter recoverability remains seldom reported, even among studies investigating test-retest reliability of parameter estimates (Table 1). Studies that do report it, however, predominantly use Pearson or rank correlation (e.g., Moutoussis et al., 2018a; Karvelis et al., 2018; Hauke et al., 2022; Smith et al., 2022, 2021b; Shahar et al., 2019; Hitchcock et al., 2022a; Sullivan-Toole et al., 2022). Here we would like to suggest that a more suitable metric of parameter recoverability would be absolute ICC (e.g., see Mkrtchian et al., 2023), which is sensitive to all systematic errors, unlike Pearson or rank correlation.

Note that another way to obtain an upper bound on test-retest reliability is to compute split-half reliability, which, like parameter recovery, is not affected by longitudinal changes in behavior. Less than half of the studies reviewed here estimated split-half reliability (Price et al., 2019; Shahar et al., 2019; Brown et al., 2020, 2021; Bruder et al., 2021; Xu and Stocco, 2021; Hitchcock et al., 2022a; Loosen et al., 2022; Howlett et al., 2022). While most of these studies found split-half reliability to be higher than test-retest reliability, those that used tasks involving trial-by-trial learning found split-half reliability to be similar or even lower for some parameters (Brown et al., 2020, 2021; Loosen et al., 2022). This highlights the fact that important dynamics of trial-by-trial learning might be difficult to preserve when partitioning the trials. Given that many tasks and models in computational psychiatry focus on trial-by-trial learning (RL, hierarchical Gaussian filter, active inference), split-half reliability measures might be difficult to rely on. Furthermore, even in tasks that do not induce strong trial-by-trial learning, split-half analysis comes with potential confounds and depends on how the trials are partitioned (Pronk et al., 2021; Parsons et al., 2019). For example, comparing odd vs even trials - which was the most common method in the reviewed studies - can often lead to overestimation or underestimation of reliability (Pronk et al., 2021; Parsons et al., 2019). A recommended alternative is permutation-based split-half reliability. However, for computational modelling studies this would require refitting the model for each permutation (thousands of times), making it very computationally intensive. For these reasons, here we focused on parameter recovery as a more universal method for obtaining an upper bound on test-retest reliability.

4.3. Model complexity and parameter collinearity may reduce reliability

Another factor determining reliability of parameter estimates is model complexity. Higher complexity here means more parameters, which means more degrees of freedom. Given the same amount of data, parameter estimates of more complex models will tend to have lower reliability (e.g., Waltmann et al., 2022). This can be further exacerbated by collinearity among parameters. For example, when using EB for model fitting, collinearity can lead to excessive shrinkage and thus poor reliability (Scheibehenne and Pachur, 2015). While, in general, model comparison procedure guards against excessive complexity and collinearity, model simulations can help diagnose and investigate these issues in more detail. Some of the reviewed studies chose to fix model parameters that exhibited collinearity (Brown et al., 2020) or low recoverability (Smith et al., 2021b) in order to improve overall reliability. Note that recoverability analysis can also inform model selection in scenarios where direct model comparison is not possible due to some models relying on additional behavioral data (e.g., Karvelis et al., 2018).

5. Clinically irrelevant changes in task performance pose further challenges for longitudinal testing

So far we have considered how various factors can affect measurement error, and in turn reliability. However, reliability can also be affected by actual changes in task performance across repeated administration of the task (Palminteri and Chevallier, 2018). The challenge here is to separate clinically relevant changes (signal) from clinically irrelevant ones (noise). Generally, task performance is assumed to be primarily driven by trait-like characteristics, which by definition are relatively stable over time; that is what the term 'individual differences' is most often used to refer to (Sackett et al., 2017). However, state-like (e.g., mood, sleepiness, attentiveness) fluctuations over time as well as practice effects due to repeated exposure to the same task can significantly affect task performance, reducing test-retest reliability (Fig. 2).

5.1. The effects of task practice on computational measures are understudied

Practice effects - specifically, improvement in task performance over time - is one of the main confounds in longitudinal studies (Calamia et al., 2012; Scharfen et al., 2018), including developmental research (Anokhin et al., 2022; Lannoy et al., 2021; Sullivan et al., 2017) and clinical trials (Beglinger et al., 2005; Goldberg et al., 2010). The magnitude of practice effects usually plateaus after two sessions, although this can depend on cognitive domain, test-retest interval, the age of participants or their clinical status (Calamia et al., 2012; Scharfen et al., 2018). These improvements are thought to stem from increased

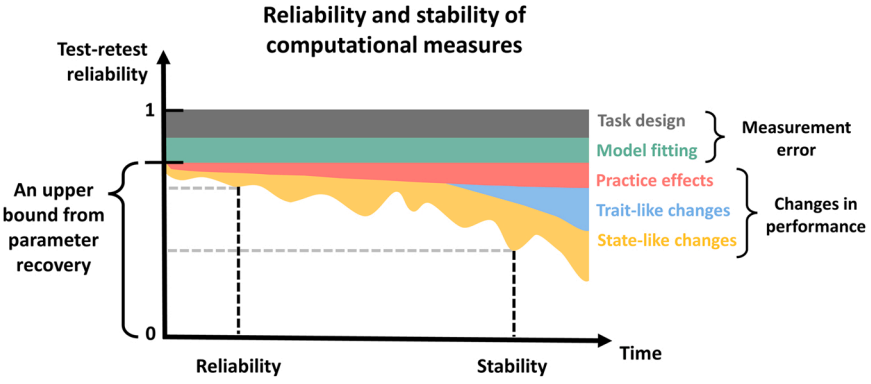

Fig. 2. Test-retest reliability and different sources of variance. To address the reliability issues, different sources of variance must be distinguished and investigated. Task design (e.g., the number of trials, outcome probabilities, stimuli timing, duration and type, task instructions, practice trials, etc.) and model fitting choices (e.g., EB or ML, joint or separate modelling of different testing sessions) contribute to measurement error, while practice effects as well as state-like and trait-like changes can lead to clinically irrelevant changes in task performance. Note that the practice effects tend to plateau, state-like effects can be expected to fluctuate on short timescales (conveyed by the scalloping), while trait-like changes are expected to occur on longer timescales. Parameter recovery analysis can provide a lower bound on measurement error, and thus an upper bound on reliability. Studying longer-term stability of the measured constructs (trait-like changes) must be accompanied by the assessment of reliability in order to account for all other sources of variance.

familiarity with the format of the task, its specific content, and the development of better task-taking strategies (Goldberg et al., 2010). Importantly, practice effects have been shown to change the involvement of different brain regions, suggesting that the neural systems engaged by repeated task performance might differ from those engaged during the initial exposure to the task (Kelly and Garavan, 2005; Chein and Schneider, 2005).

In order to reduce practice effects in longitudinal testing, it is recommended to include more task practice at the baseline to ensure sufficient familiarity with the task format and to use alternate task forms at every session to prevent recall of specific task content (Beglinger et al., 2005). In many task paradigms in computational psychiatry alternate task forms would require to use different cues and stimuli while keeping the task structure the same. More task practice at baseline might entail revealing certain features of the task design that rely on surprise. For example, having reversals of response-outcome contingencies at fixed points in the task will be increasingly less surprising with repeated exposure to the task, leading to strong practice effects, unless such reversals are made less surprising to begin with. However, the caveat here is that with increasing familiarity the individual might adopt a strategy that relies on heuristics rather than on estimation of random parameters, which might substantially deviate from the processes that were originally of interest to the researcher.

In the computational assay studies reviewed here, alternate task forms were not used and practice effects received little attention in general (Table 1). Most studies did not investigate practice effects, one study reported no practice effects (Brown et al., 2020), two studies discussed potential practice effects (Ahn et al., 2020; Sullivan-Toole et al., 2022), one study found significant practice effects but did not report its effect size (Moutoussis et al., 2018a, and two studies found small to medium practice effects (Mkrtchian et al., 2023; Smith et al., 2022). Note that this addresses only group-level practice effects related to initial improvement. In longitudinal settings with repeated task performance we could expect additional practice effects to emerge, for example, due to boredom if the task is experienced as too repetitive and not engaging enough (Agrawal et al., 2022). There could be additional practice effects that affect task-taking strategies in idiosyncratic ways among individuals - e.g., leading to increased between subject variance during retest (Sullivan-Toole et al., 2022). All in all, future research in computational psychiatry would benefit from a more rigorous assessment of practice effects and the challenges it poses for repeated assessment over different timescales.

5.2. The effects of state-like changes on computational measures are understudied

Cognitive performance can also be affected by many state-like factors, including day-to-day fluctuations in mood (Forgas, 2017; Buelow and Suhr, 2013), homeostatic sleep drive and circadian cycle effects (Balter et al., 2022; Schmidt et al., 2007; Blatter and Cajochen, 2007), fluctuations in blood glucose levels (Peters et al., 2020), caffeine intake (Rogers et al., 2013), exercise (Lambourne and Tomporowski, 2010), etc. Furthermore, in addition to mood affecting task performance, engaging with cognitive tasks can in turn affect mood too (Jangraw et al., 2023).

All of this introduces additional noise and further complicates attempts to measure the underlying traits. Only one of the reviewed studies examined such state-like effects (Sullivan-Toole et al., 2022); the authors found that mood intensity (positive or negative) on the day of testing was associated with increased reward learning rate in Iowa gambling task. Future studies would benefit from a more detail investigation of how sensitive different computational assays are to state-like fluctuations and the possibility that some of these fluctuations might be clinically meaningful (Konova et al., 2020).

5.3. Trait-like changes: reliability vs stability

On longer timescales we might expect to also see trait-like changes in task performance (Fig. 2). This creates a distinction between reliability of parameter estimates themselves and temporal *stability* of trait-like mechanisms that these estimates are generally assumed to reflect. Note that in the literature this distinction is not always explicitly made, but we consider it to be worth highlighting. The distinction largely depends on the timescale on which substantial trait-like changes could be expected to occur (note that this may differ in different populations, e.g., due to developmental, aging, or treatment effects). In the studies reviewed here, retesting done on timescales of 6–18 months or longer was usually interpreted and discussed in terms of stability, while anything shorter than that (typically on the order of weeks) was interpreted in terms of reliability. We, therefore, chose 6 months as an operational threshold for the purposes of this review.

In the studies reviewed here, most parameter estimates were found to have poor and some moderate stability (Moutoussis et al., 2018a; Shahar et al., 2019; Smith et al., 2021b, 2022; Chung et al., 2017; Brown et al., 2020). However, these results could also be confounded by low reliability, as none of the studies assessed reliability of parameter estimates (at shorter intervals) within the same sample. While two of the studies (Brown et al., 2020; Smith et al., 2021b) did examine reliability at shorter intervals, it was done using independent samples and different task parameters (e.g., different number of trials) making the comparisons difficult. Future studies seeking to study stability would greatly benefit from the assessment of reliability within the same sample as part of their study design (Heise, 1969).

The goal of stability studies is not so much to find the most stable traits, but instead to find which trait-like changes track one's mental health status or to find traits that capture the underlying vulnerability to develop a disorder and can therefore be used to predict or improve clinical outcomes. None of the above studies found such effects (although see Smith et al., 2022) for an exploratory analysis).

6. What reliability is sufficient?

In this review, we have adopted the classification of ICC values proposed by Koo and Li (2016) (< 0.5 is poor, 0.5–0.75 is fair, 0.75–0.9 is good, and > 0.9 is excellent). While having a standardized labelling system can facilitate communication of findings, we must keep in mind that this and other proposed classifications (Fleiss, 2011; Landis and Koch, 1977) are all arbitrary (Shrout, 1998; Weir, 2005; Hedge et al., 2018). While they capture meaningful ordinal information (i.e., 'good' is better than 'poor'), 'poor' reliability does not necessarily mean that it is not sufficient for clinical applications, while 'good' reliability does not necessarily mean that it is. What is sufficient will depend on many other factors that lie in-between reliability and clinical utility (Fig. 1).

One way to deal with this arbitrariness is to go beyond the qualitative labels and consider quantitative consequences that reliability can have on any subsequent analysis. The clearest and perhaps the most relevant case is correlational analysis: true correlations between the task-based measures (x) and other measures (y; e.g., symptoms or measures from a different task) will be attenuated by their reliability following the equation (Spearman, 1904):

$$r_{observable} = r_{true} * \sqrt{ICC_x * ICC_y} \quad (2)$$

Here we must remember that although task-based measures tend to have much poorer reliability than scale-based measures (e.g., Enkavi et al., 2019), the latter still rarely exhibit excellent reliability. This includes popular scales for schizophrenia (Norman et al., 1996; Peralta et al., 1995), autism (Zander et al., 2016; Hoekstra et al., 2008), anxiety (Maier et al., 1988; Barnes et al., 2002), and depression (Davidson et al., 1986; Carrozzino et al., 2020), as well as the Diagnostic and Statistical Manual of Mental Disorders (DSM-5) (Regier et al., 2013 and the

International Classification of Diseases (ICD-11) (Reed et al., 2018).

The combined effect of less than excellent reliabilities will make it much more difficult to detect true effects and will require larger sample sizes (Baugh, 2002). For example, assuming $ICC = 0.6$ for both measures and the true correlation in the range of $r = 0.3 - 0.5$, the observable correlation would be almost halved to $r = 0.2 - 0.3$ and it would require $1.5 - 3$ times larger sample sizes ($N = 239$ instead of $N = 84$) to detect it (Hedge et al., 2018). Similarly, less than excellent reliabilities will impede efforts to use machine learning models for individual level predictions. While the effects of reliability in this context are less straightforward to intuit and require more empirical investigation, a recent study by Gell et al. (2023) has shown that in the context of predicting behavioral phenotypes from functional connectivity data, phenotypic reliability of $ICC = 0.8$ ("good") may cut prediction accuracy in half (as compared to $ICC = 1$), while $ICC \leq 0.6$ ("fair") can make predictions completely meaningless.

This underscores the importance of being aware of the reliability of one's instruments and the consequences it has on statistical power (Williams and Zimmerman, 1989; Baugh, 2002; Hedge et al., 2018), and subsequently on construct validation and clinical translation (Fig. 1). Improving reliability (instead of collecting larger samples) can make research more efficient and less costly (Nikolaidis et al., 2022). Importantly, for making accurate predictions at the individual level (i.e., to make personalized psychiatry possible) achieving excellent reliability is just as, if not more, important than acquiring large samples (Gell et al., 2023).

7. What is construct validity?

Construct validity refers to the extent to which an instrument measures what it intends to measure (Kane, 2013; Borsboom et al., 2004; Messick, 1984; Cronbach and Meehl, 1955). It can be established in many different ways. First, it can be done based on subjective judgement of the general features of the measure (*face validity*) and how well it captures all facets of the construct (*content validity*). Second, it can be done based on statistical associations with other measures (*criterion validity*), including other measures of the same construct (*convergent validity*) or measures of some other construct taken around similar time (*concurrent validity*) or at a future time (*predictive validity*). One could consider all of these to be different types of validation, all of which support construct validity (Borsboom et al., 2004).

While these are the textbook examples of validity, countless other ways to refer to specific instances of validity exist such as *ecological validity*, which refers to how well a measure captures real-world processes, *neurobiological validity*, which refers to a measured construct being associated with brain function or structure, *clinical validity*, which refers to a measure being associated with clinical measures, *diagnostic validity*, which refers to a measure's ability to differentiate individuals with and without a certain diagnosis, or *longitudinal validity*, which refers to a covariation between longitudinal changes in a measured construct and changes in an outcome of interest, and so on.

More generally speaking, construct validity is established by demonstrating that a measure is associated with some other measure in a way that makes theoretical sense. Validity is thus inherent not to the instrument but to the interpretation and the use of its measures (Kane, 2013). For example, a measure of some cognitive dimension might not be associated with any mental disorders (and thus have no clinical validity), but it might be a valid measure for understanding individual differences in some other context. Unlike reliability, however, validity is not quantifiable (beyond the strength of statistical relationships) and there is no established qualitative labelling system to denote the degree of validity that a measure has. Establishing or disproving validity of a measure is therefore not a straightforward matter. Conveniently, computational modelling approaches (as opposed to strictly conceptual/verbal theorizing), have many desirable properties that make construct validation more manageable: it forces researchers to explicitly specify not only the constructs of interest but also the relationships between them, allowing for quantitative and thus more testable predictions (Grahek et al., 2021).

8. Lack of convergent validity points to the problems of overgeneralization and overinterpretation

In computational psychiatry, most of the effort has gone into studying clinical validity by investigating associations between computational measures and clinician-rated or self-reported symptom measures (concurrent validity) or with diagnostic categories (diagnostic validity) (e.g., Chrysaitis and Seriès, 2022; Kaliuzhna et al., 2019; Katthagen et al., 2022; Pike and Robinson, 2022). Considerable effort has also gone into studying neurobiological validity of computational measures (e.g., Iglesias et al., 2017. However, almost no work has been done on making sure that measures of the same construct (e.g., learning rate, Bayesian priors) have convergent validity across similar tasks (Browning et al., 2020), which one could consider to be a prerequisite for trying to establish clinical validity (Fig. 1).

There is mounting evidence challenging convergent validity of various task measures across multiple cognitive domains. This includes the measures of cognitive control, with a lack of convergent validity among tasks (Noreen and MacLeod, 2015; Hedge et al., 2018; Eisenberg et al., 2019; Gärtner and Strobel, 2021; Whitehead et al., 2020; Raud et al., 2020) and between tasks and self-report (Saunders et al., 2018; Enkavi et al., 2019; Eisenberg et al., 2019), the measures of risk preference, with a lack of convergent validity among tasks (Pedroni et al., 2017; Buelow and Barnhart, 2018) and between tasks and self-report (Frey et al., 2017), the measures of distress tolerance, with a lack of convergent validity between tasks and self-report (McHugh et al., 2011),

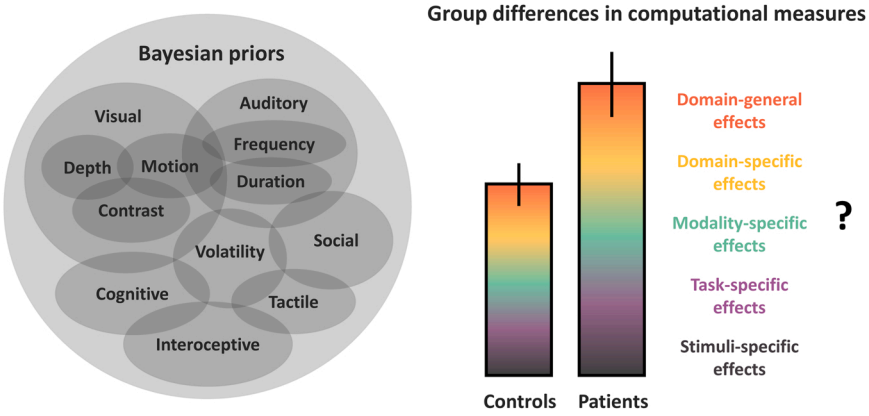

Fig. 3. A lack of convergent validity undermines interpretability. The illustration on the left represents a multitude of largely non-overlapping Bayesian priors as suggested by recent empirical findings showing no common factors for reliance on priors. This poses problems for computational accounts proposing general impairments in Bayesian priors. Computational accounts proposing more specific impairments, however, still face the challenge of appropriately operationalizing such priors. Currently, due to the overreliance on single-task designs, there is a tendency to overgeneralize when interpreting findings from a given task, while the relative contribution of more specific experimental factors often remains unclear - this is depicted in the illustration on the right for group differences (due to it being simpler to illustrate), but the same holds for individual differences.

the measures of reliance on perceptual priors, with a lack of convergent validity among tasks (Grzeczkowski et al., 2017, 2018; Tulver et al., 2019), the measures of sensitivity to positive and negative valence, with a lack of convergent validity among tasks, self-report, and neuroimaging (Peng et al., 2021), and RL model parameter estimates, with a lack of convergent validity among tasks (Eckstein et al., 2022, 2021).

How can we make sense of these findings? As we discussed throughout the paper, one obvious culprit is low reliability of task measures. However, validity problems plague even those tasks that demonstrate moderate to good test-retest reliability (Grzeczkowski et al., 2017, 2018; Pedroni et al., 2017; Frey et al., 2017; Saunders et al., 2018; Snijder et al., 2022). Another emerging theme in explaining the lack of validity points to the problem of overgeneralization (Grzeczkowski et al., 2017; Tulver et al., 2019; Whitehead et al., 2020; Dang et al., 2020; Gärtner and Strobel, 2021; Friedman and Gustavson, 2022; Eckstein et al., 2022, 2021; Peng et al., 2021). That is, what is being measured in most cases is likely much more stimuli-specific, task-specific, modality-specific, or domain-specific than is assumed to be. The measures might be capturing only one facet of the assumed construct, or it simply might be confounded by the above-mentioned factors, which in turn can lead to the overinterpretation of findings. To put it yet another way, many of the studied constructs might lack coherence and might be suffering from *jingle fallacy* - an erroneous assumption that measures with the same name are capturing the same processes (e.g., Eisenberg et al., 2019).

8.1. Poor validity implications for testing computational accounts of mental disorders

These issues have significant implications for computational psychiatry research, where a typical study design includes a single cognitive task and tests a rather general hypothesis. For example, the findings of no convergent validity among tasks assessing perceptual priors (Tulver et al., 2019), even when the tasks are from the same narrow subdomain of visual illusions (Grzeczkowski et al., 2017; Cretenoud et al., 2019), means that using such measures to capture the construct of 'reliance on perceptual priors' has little validity. This poses a considerable problem for the Bayesian accounts of autism and schizophrenia, which consider impaired perceptual inference, resulting from over- or under-reliance on priors, to be a core mechanism underlying these disorders (Fletcher and Frith, 2009; Corlett et al., 2009; Pellicano and Burr, 2012; Palmer et al., 2017). It might also explain why attempts to empirically assess these priors have produced a lot of mixed results (Chrysaitis and Seriès, 2022; Katthagen et al., 2022; Sterzer et al., 2018). The theory available for interpretation might simply be too general and the results produced by single-task designs might be too task-specific (Yarkoni, 2022). A few studies that did use a whole battery of tasks, consisting of various visual illusions, found these disorders to be equally susceptible to most and less susceptible only to a couple illusions, suggesting the effects to be illusion-specific (Grzeczkowski et al., 2018; Kaliuzhna et al., 2019; Chouinard et al., 2016).

As Bayesian framework continues to be applied to theorize about increasingly more disorders (Schwartenbeck et al., 2015; Gu and Filbey, 2017; Paulus et al., 2019; Linson et al., 2020; Richards et al., 2020; Karvelis and Diaconescu, 2022; McGovern et al., 2022; Herzog et al., 2022), with a view that each disorder can be understood in terms of a certain kind of maladaptive priors (Fig. 3), sufficient validation of assays for measuring such priors could save a lot of confusion down the line.

8.2. Might computational measures exhibit better convergent validity?

While some were motivated by computational accounts of cognition, the studies lacking convergent validity discussed above have relied solely on behavioral measures. Perhaps computational measures, which are meant to capture latent cognitive variables more directly, would show more convergence among similar tasks? Surprisingly little empirical data exists to answer this question.

A recent study by Eckstein et al. (2022) suggests that computational measures might not necessarily show better convergent validity. The authors investigated convergence of RL model parameters across 3 similar tasks in a large sample (N = 291) and found that most parameters showed none-to-weak correlations across the tasks. Even when assessing shared variance between one parameter in one task and all parameters in a different task, the shared variance was only $< 30\%$ for all parameters and $< 10\%$ for most parameters. This challenges one of the main proposed advantages of theory-driven computational modelling: interpretability of computational measures. The parameters might have a clear theoretical meaning and computational role, but the fact that they do not capture individual differences across similar contexts - i.e., their lack of generalizability - makes it difficult to interpret the meaning of parameter estimates beyond a specific task context (Eckstein et al., 2021).

A recent study by Weigard et al. (2021), however, reported slightly more encouraging results. The authors reanalyzed 8 tasks from Eisenberg et al. (2019) dataset (local-global task, shape matching task, directed forgetting task, attentional network task, Stroop task, Simon task, task-switching task, and choice response time task) using a drift-diffusion model (DDM). Applying a bifactor model (separating task-general and task-specific effects) to analyze parameter estimates revealed that task-general factors of DDM parameters explained 76–86% of variance in parameter estimates across the tasks and were highly reliable/stable when retesting within an interval of 2–8 months ($ICC = 0.69 - 0.78$). However, simple behavioral measures of mean reaction time and accuracy showed similarly good convergence across tasks (with the task-general factor explaining 76–86% of variance) and had similarly good reliability ($ICC = 76 - 78$). Only the difference scores,

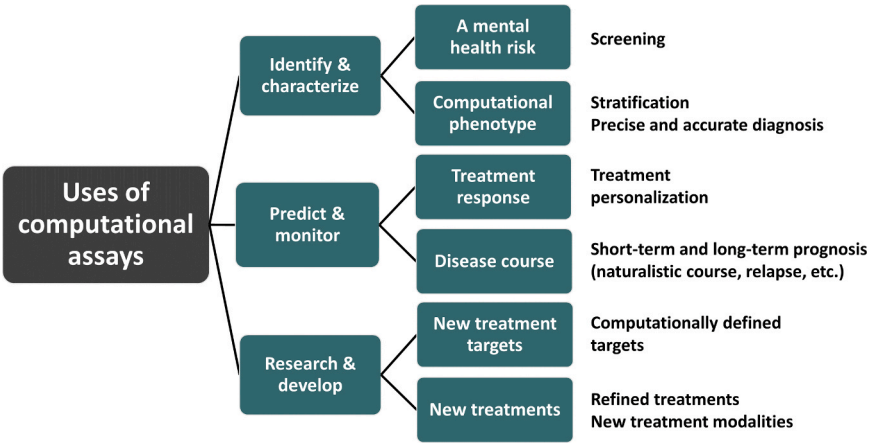

Fig. 4. A graphical summary of the intended uses of computational assays. This summary is based on the most common proposals in the literature (Paulus et al., 2016; Paulus, 2017; Patzelt et al., 2018; Nair et al., 2020; Reiter et al., 2021; Yip et al., 2022; Hauser et al., 2022) and might not be exhaustive. Note that most of these applications require longitudinal study of the disorders. For example, if a computational measure does not show within-individual covariation with changes in symptoms, it cannot be a valid treatment target, be used to monitor disease course, or provide interpretable treatment response predictions. Even computational phenotyping might require longitudinal assessments in order to account for the dynamic nature of mental disorders (Germine et al., 2021; Gueguen et al., 2021).

which are typically used to investigate self-regulation in these tasks, showed poor convergence (29–30% variance explained) and modest reliability (*ICC* = 0.51 − 0.56). This suggests that most of the improvement came from avoiding differences scores while computational modelling did not add much extra value. Nonetheless, task-general drift-rate factor did show a slightly better correlation with self-reported self-control ($r = 0.18$) than did task-general accuracy factor ($r = 0.11$). Another study from the same team also found converging results across three self-control tasks (Stroop, Go/NoGo, and Stop Signal), with task-general drift rate factor accounting for 50% of the variance in this parameter across the tasks (Sripada and Weigard, 2021), although task-general factor results for other parameters and behavioral measures were not reported.

Overall, while still very scarce, preliminary evidence suggests that computational measures might not necessarily offer much improvement over simple behavioral measures in terms of convergent validity. This further cautions against the use of single-task designs due to the dangers of overgeneralization and overinterpretation of findings, even when the task data is modelled (Fig. 3). More empirical studies are needed to investigate convergent validity of computational measures across different tasks and models. This will require using batteries of tasks, ideally in combination with test-retest in order to control for the reliability of the measures (for some examples see: Grzeczkowski et al., 2017, 2018; Frey et al., 2017; Snijder et al., 2022; Weigard et al., 2021).

9. From clinical validity to clinical utility: the importance of longitudinal approaches

Given the problems of reliability and convergent validity of task measures, it is not surprising that studies investigating clinical validity of computational measures (or behavioral measures motivated by computational theories) have produced many mixed results (e.g., Sterzer et al., 2018; Chrysaitis and Seriès, 2022; Pike and Robinson, 2022; Katthagen et al., 2022; Gibbs-Dean et al., 2023). However, even if the reliability and convergent validity issues were addressed, there are other challenges that lay ahead.

Most studies investigating clinical validity rely on cross-sectional analyses: studying how computational and symptom measures vary across a group of people assessed at a single point in time, i.e., studying between-individual or between-group effects. While this approach could eventually be used for identifying mental health risks or for computational phenotyping (Patzelt et al., 2018), it does not by itself inform clinical decision making or lead to improved outcomes. Predicting or monitoring treatment response or disease course as well as developing new and improved treatments or treatment targets (Fig. 4) requires the demonstration that computational measures have predictive and longitudinal validity (Yip et al., 2022). In other words, it requires studying within-subject effects: how computational measures and symptoms vary over time within an individual (e.g., as a result of an intervention or naturally). While between-group and between-individual effects are often assumed to generalize to within-individual effects, in practice it is rarely the case - this is often referred to as *nonergodicity*, *ecological fallacy*, or *Simpson's paradox* (Molenaar and Campbell, 2009; Kievit et al., 2013; Fisher et al., 2018).

9.1. Predictive models based on cross-sectional computational measures

The most common predictive modelling approach involves assessing longitudinal change in clinical measures but using cross-sectional (baseline) computational measures to predict the change; to provide individual-level predictions, machine learning techniques are used. Such approaches have shown some success in predicting various clinical outcomes, such as treatment response (Karvelis et al., 2022; Hauke et al., 2022), relapse after discontinuation of treatment (Berwian et al., 2020), or naturalistic disease course (Frässle et al., 2020; Sebold et al., 2017). The accuracy of individual-level predictions, however, in most cases remains modest. In addition, machine learning validation methods and sample sizes typically used in such studies are likely to give overly optimistic results and not generalize to independent samples (Karvelis et al., 2022). While pursuing this approach remains appealing due to the fact that it requires minimal resources and could be easily integrated into clinical practice, it rests on the assumption that cross-sectional computational measures are sufficient for capturing clinically relevant individual differences. Furthermore, even if high prediction accuracy was achieved, such models provide limited interpretability: they do not tell us *why* certain computational measures are predictive of clinical outcomes or *how* different treatments are affecting the underlying computational processes (Lan and Browning, 2022; Reiter et al., 2021; Nair et al., 2020; Iglesias et al., 2017; Robbins and Cardinal, 2019; Hauser et al., 2022).

9.2. Longitudinal validity and the dynamic nature of mental disorders

A richer understanding of individual differences in how symptoms change over time could be achieved by studying longitudinal validity - also known as *responsiveness* (Liang, 2000; Mokkink et al., 2021) - of computational measures. While reliability and convergent validity issues should be addressed first, several studies have already attempted investigating longitudinal validity. For example, a recent study by Brown et al. (2021) found that symptom improvement in depression following 12 weeks of cognitive behavioral therapy was associated with increase in reward learning rate and loss outcome shift (valuation bias) in a probabilistic operant learning task. Reliability of these parameters, however, was assessed only on a separate small (N = 20) sample of healthy controls, which provided very uncertain estimates of reliability (with 95% credible interval being [−0.35:0.99] and [−0.83:0.96] for each of the parameters). Another recent study by Hitchcock et al. (2022a) found that drift rate regression parameter in the self-referential encoding task covaried with an improvement in a subset of anhedonia-related depression symptoms following 8 weeks of mindfulness training. While this parameter showed rather poor reliability (*ICC* = ∼ 0.5), it was assessed on the same sample using pre- and post-intervention data and thus is likely partially reflective of the clinically meaningful change in this parameter (i.e., the actual reliability assessed on a control group could be expected to be higher). Another recent study investigating people experiencing hallucinations (Kafadar et al., 2022) reported that increase in prior weighing in a conditioned hallucinations task (over a period of ∼ 1 year) was associated with increased frequency of auditory hallucinations in participants daily lives. However, the study did not report test-retest reliability or recoverability of this parameter estimate. Many other attempts to study longitudinal validity of computational measures have lead to null findings (Chung et al., 2017; Moutoussis et al., 2018a; Smith et al., 2021b, 2022).

Most of these studies have used only 2 or 3 widely spaced assessments (ranging from 5.5 weeks to 18 months apart). Considering the dynamic nature of mental disorders, more frequent (and more strategically timed) longitudinal assessments might be needed to capture many important features of the disorders (Borsboom et al., 2013; Sharp et al., 2020; Germine et al., 2021; Gueguen et al., 2021; Hitchcock et al., 2022b; Gauld and Depannemaecker, 2023). A notable recent study by Konova et al. (2020) serves to illustrate the usefulness of this approach for predicting short-term disease course. Studying patients with opioid use disorder, the authors showed that within-individual increase in a computational measure of ambiguity tolerance (assessed on weekly, biweekly, and monthly basis) preceded return to opioid use within the subsequent 1–4 weeks. Importantly, this parameter was not found to be different between patients and controls (no between-group effect) - an example of nonergodicity that we mentioned earlier. Their results are also supported by high test-retest reliability of the parameter estimates (*ICC* = 0.70 − 0.72 in patients and *ICC* = 0.87 − 0.89 in healthy controls. Note that while reliability among patients was lower, the fact that variation in risk tolerance was predictive of opioid use suggest that this

Box 1: Key insights and recommendations for future research

Addressing reliability issues should take the highest priority:

- Reliability analysis and reporting should become a routine practice. This includes both test-retest reliability and parameter recoverability.
- Parameter recoverability provides an upper bound on test-retest reliability and can help disentangle different sources of variance (measurement error vs. change in performance).
- Different sources of variance contributing to low reliability should be investigated more systematically. Currently, practice and state-like effects are particularly understudied.
- Reliability of behavioral and computational measures can be improved by using hierarchical estimation methods (mixed models and empirical Bayes, respectively). Further improvements can be achieved by modelling the two testing sessions jointly, using bivariate distributions.
- Model complexity and collinearity among parameters might negatively affect reliability of parameter estimates. Model simulations can help diagnose these issues.
- Studying long-term stability of computational measures needs to be accompanied by test-retest analysis (on a short timescale) to avoid poor reliability confounding stability results.
- Absolute ICC should be preferred over Pearson or rank correlation for assessing reliability because it is sensitive to all systematic errors. This applies to both test-retest reliability and parameter recovery.

Addressing construct validity should take the second highest priority:

- Convergent validity studies would greatly benefit from controlling for reliability of the measures by incorporating test-retest in their study design.
- Demonstrating convergent validity among computational measures from similar tasks could be seen as a starting goal, with convergence between computational and scale-based measures being a subsequent more challenging goal due to the differences in response processes. However, the former has been studied much less than the latter.

The development of assays should be guided by a long-term vision of their uses:

- To move closer to clinical applications it is necessary to focus on predictive and longitudinal validity. Cross-sectional findings might not always generalize to longitudinal validity (non-ergodicity).
- Predictive models based on cross-sectional computational measures provide limited interpretability, which diminishes their utility. Longitudinal assessments would be much more informative in revealing how computational mechanisms change over time (naturally or in response to treatment).
- Frequent longitudinal assessments might be needed to capture the dynamic nature of many mental disorders.
- To make longitudinal testing more feasible, it is important to focus on making the assays more accessible (smartphone-based), more engaging (gamification), shorter/more efficient (adaptive design optimization, dense sampling of behavior), and integrated with other sources of information (wearable devices and ecological momentary assessments).
- The pursuit of all these subgoals (reliability, longitudinal validity, efficiency, etc.) should be seen as an iterative rather than sequential process.

might be due to a clinically meaningful change in cognition over short periods of time.

Despite these encouraging examples, and given the issues of convergent validity among tasks, using a single task might not be sufficient for providing highly accurate insights and predictions for clinical decision making. Instead, longitudinal testing might need to be done with batteries of tasks in order to infer task-general computational deficits (Weigard and Sripada, 2021; Vinckier et al., 2022) or to build a sufficiently rich multidimensional characterization of the disorders (Gueguen et al., 2021).

9.3. *Clinical efficacy, clinical utility, and additional challenges for assay development*

Using retrospective data for studying predictive validity comes with a high risk of overfitting, especially when relying on currently popular validation techniques (Karvelis et al., 2022; Chekroud et al., 2021; Rutledge et al., 2019). A much more convincing test of prediction accuracy will be the demonstration of *clinical efficacy*: showing that clinical decision making guided by computational assays leads to improved outcomes compared to treatment as usual (Paulus and Thompson, 2021; van der Vinne et al., 2021; Kingslake et al., 2017). Importantly, this will come with additional practical challenges of how to best integrate the assays into the clinical workflow (Paulus and Thompson, 2021; Kelly et al., 2019). Finally, even if clinical efficacy can be demonstrated in highly controlled research trials, *clinical utility* of the assays will eventually depend on their cost-effectiveness, ease of use, and its applicability in a range of real-world clinical settings (Hollon et al., 2002; Paulus, 2017).

Although most pertinent to the later stages of assay development, thinking about the deployment and scalability challenges for different assay uses (Fig. 4) can also help inform and optimize the earlier stages of the development (Yip et al., 2022). For example, monitoring disease course, treatment response, or characterizing dynamic computational phenotypes requires longitudinal testing, likely involving batteries of tasks. To make frequent longitudinal testing feasible, it might be necessary to move exclusively to remote testing strategies such as smartphone-based tasks (Gillan and Rutledge, 2021; Zech et al., 2022; Pronk et al., 2022; Howlett et al., 2022). Another underappreciated challenge is engagement. Many of the tasks used in research today are rather lengthy and tedious, which makes it unlikely that patients will adhere to completing them on a regular basis. Task engagement (and user experience more generally) is therefore an important variable to optimize, and will require applying gamification strategies (Vermeir et al., 2020 and adopting patient-centered research frameworks (Germine et al., 2021; Pratap et al., 2020). Gamification can make tasks not only more appealing, but also more efficient (Kucina et al., 2022).

Additional strategies can be used to further increase task efficiency. One novel approach is adaptive design optimization (ADO) (Cavagnaro et al., 2010; Myung et al., 2013; Pooseh et al., 2018; Ahn et al., 2020;

Kwon et al., 2022), which adapts task parameters in real time based on task behavior in order to maximize the informativeness of collected data. A recent study by Ahn et al. (2020) showed that using ADO excellent test-retest reliability ($ICC = 0.97$) in a delay discounting task can be achieved with less than 20 trials (under 2 min of testing). It is interesting to note that ADO could also be synergistic with engagement optimization (cf. dynamic game difficulty balancing in traditional video games), and could reduce ceiling and floor effects in performance, increasing the range of measurable individual differences. Another novel approach for improving efficiency involves dense sampling of behavior (Howlett et al., 2022, 2020). Instead of one or two data points per trial (e.g., response choice or reaction time), tasks could be designed to collect continuous real-time data, resulting in dozens of data points per trial. Using this approach, Howlett et al. (2022) showed that good test-retest reliability of parameter estimates ($ICC = 0.78 - 0.81$) can be achieved under 6 min of testing.

Finally, task efficiency and informativeness of task data could be further increased by combining it with passive data collection from wearable devices (Pratap et al., 2020) and ecological momentary assessments (Pronk et al., 2022). These additional data sources could provide important context for interpreting model parameters and could also help account for potential confounding factors; see Gillan and Rutledge (2021) and Hauser et al. (2022) for related discussions.

10. Discussion

Despite being the bread and butter of individual difference research, psychometric properties of computational assays have so far been understudied. The emerging empirical evidence reviewed here suggests that computational measures obtained from the assays often do not provide much improvement over simple behavioral measures and show similarly poor reliabilities (Table 1). Furthermore, behavioral and computational measures used to test computational accounts of mental disorders show a lack of convergent validity (among themselves and with self-report measures of the same constructs), which mirrors the generalizability crisis in the field of psychology (Yarkoni, 2022). These issues are a major bottleneck for any further development of computational assays (Fig. 1).

Although it may paint a rather dire picture, we hope that this review also provides insights and guidance for how to move forward (see Box 1). The most essential methodological move is to embrace longitudinal testing with batteries of tasks. While this does not automatically solve the problems in question, it allows us to begin studying and addressing them. First, reliability needs to be studied in more detail, distinguishing and accounting for different sources of variance (Fig. 2), including measurement error (stemming from task design and model fitting) and changes in behavior (practice effects, state-like changes, and trait-like changes). Empirical work is needed to get a better sense of the importance of each of these sources. For example, state-like changes and practice effects are currently understudied, while trait-like changes have proven difficult to study due to being confounded by the other sources of variance (Table 1). Note that once the different sources of variance are better understood, it may be worth considering how such effects could be explicitly incorporated into the models rather than treated as uninformative error (e.g., Van Bork et al., 2022).

One important thing to keep in mind is that reliability can be sensitive to any modifications of task design and data analysis methods. Therefore, the concept of declaring sufficient or insufficient reliability for an assay once and for all does not seem possible. In other words, reliability results might not generalize to similar tasks and similar methods. Instead, reporting of reliability must become a routine practice (Parsons et al., 2019). The same goes for parameter recoverability (Palminteri et al., 2017; Wilson and Collins, 2019), which can provide an upper bound on reliability for cross-sectional studies and can also be informative for disentangling different sources of variance in longitudinal studies investigating test-retest reliability. Here we suggest that, just as for reliability, the most informative measure of parameter recoverability would be absolute ICC (instead of Pearson's and rank correlation that are used currently). For a guide on how to report reliability measures see Parsons et al. (2019).

In order to begin addressing the construct validity issues, we suggest to start with studying convergent validity among tasks of varying similarity. This can be seen as the lowest bar for convergent validity because tasks are meant to probe similar response processes. In contrast, self-report might rely on different processes involving self-reflection (Palminteri and Chevallier, 2018; Dang et al., 2020), therefore, seeking convergent validity between tasks and self-report could be seen as more ambitious. In our literature search, we were able to identify only a few studies investigating convergent validity of computational measures across similar tasks (Eckstein et al., 2022; Weigard et al., 2021; Sripada and Weigard, 2021). Many more such studies are needed - ideally including repeated assessments to control for the effects of reliability (Weigard et al., 2021). Due to the lack of such studies most of the literature we presented to support our arguments came from showing a lack of convergent validity among behavioral measures. Note that this is warranted by the fact that many tasks and resulting behavioral measures are often interpreted in computational terms (even without doing any model fitting) and are used to test computational accounts of mental disorders (Chrysaitis and Seriès, 2022; Kaliuzhna et al., 2019; Tulver et al., 2019; Grzeczkowski et al., 2018).

10.1. The big picture: prioritization of subgoals and their joint optimization

Reliability and construct validity constitute the most important part of the review and reflect the need to prioritize these issues. Still, we aimed to provide a wider perspective and to consider other milestones (Fig. 1) and end goals (Fig. 4) of assay development. Being aware of what challenges await at subsequent stages and what properties clinically useful tools need to have, can help make better research decisions at earlier stages.

For example, one simple but significant insight is that most of the milestones require longitudinal data collection, while most of the current research relies on cross-sectional data. On the other hand, when it comes to real-world uses of the assays, longitudinal data collection (especially frequent testing with multiple tasks) becomes very impractical. It is therefore necessary to find ways to make longitudinal data collection more efficient. We have provided some pointers in that regard (gamification, adaptive design optimization, dense sampling techniques, integration with wearable device data, etc.), but more innovative ideas are needed.

The key insight here is that it is necessary to work on many of the subgoals (e.g., reliability, predictive validity, clinical efficacy) in parallel rather than sequentially. For example, there are many ways to improve reliability (Zorowitz and Niv, 2023), but that alone does not guarantee significant improvements in predictive validity and utility of the measures (Finn and Rosenberg, 2021). Speaking in computational terms, optimizing exclusively for reliability can leave us stuck in a local maximum. Furthermore, any assay design changes that would later be done to improve validity or engagement would likely affect reliability too, requiring it to be reevaluated. Thus, it is important to look for solutions that address multiple problems simultaneously and move us closer to the global maximum. That is why here we emphasized the importance of longitudinal designs with batteries of tasks: this allows to simultaneously address any of the subgoals, including reliability, convergent validity, longitudinal validity, and ultimately, clinical utility. In this context, Germine et al. (2021) provide a very helpful sketch of an iterative task development procedure that is aimed at jointly optimizing for psychometric properties, engagement, and accessibility. This approach could be extended to incorporate any other subgoals such as efficiency: while research might benefit from more testing with larger batteries of tasks, real-world applications will require minimal sets of

short tasks that are maximally informative.

10.2. Dealing with temporal and contextual changes in behavior

While the challenges of accounting for temporal and contextual changes in behavior is only beginning to receive serious attention in computational psychiatry (Germine et al., 2021; Gueguen et al., 2021; Hitchcock et al., 2022b), similar challenges have been encountered previously in personality psychology research - this is known as *the personality-situation debate* (Fleeson and Noftle, 2008; Kenrick and Funder, 1988; Epstein, 1979, 1980). In short, this large body of research was concerned with the fact that behavior of an individual tends to vary across time and contexts, making it difficult to infer trait-like characteristics from a single assessment. Perhaps unsurprisingly, it was found that to infer traits that are predictive of future behavior it is necessary to aggregate behavior/responses across measurement occasions and/or different contexts (e.g., Epstein, 1979, 1980).

It may be worthwhile to consider how this body of research can inform assay development (Lilienfeld, 2014; Hitchcock et al., 2017). We could think of different tasks within a battery as providing different contexts and enabling us to build a multidimensional profile of an individual (Gueguen et al., 2021). From here, we could focus on extracting task- or domain-general factors (Weigard and Sripada, 2021; Vinckier et al., 2022), or focus on the variation/differences across contexts as holding important information about the disorder (Hitchcock et al., 2022b). Similarly, there are different ways to consider temporal variation. For example, we might focus on average behavior or its variation across time. However, if the assays are intended to be used for assessing longitudinal changes (for treatment monitoring or short-term risk prediction), the need to aggregate data across many measurement occasions becomes quite impractical. Ideally, the assays should be developed to be sensitive to changes from one instance of measurement to the next. Ultimately, however, the feasibility and clinical utility of each of these approaches is an empirical question. What works best might also differ among different disorders, as some are characterized by more variability than others (Hitchcock et al., 2022b).

Note that clinically meaningful state-like variation introduces another complication in assay development. On one hand, high test-retest reliability is crucially important. On the other hand, test-retest reliability measures assume that the assays are aimed at capturing relatively stable trait-like factors (i.e., state-like variation reduces test-retest reliability). In this review, we discussed two studies where low reliability was partly a result of clinically meaningful short-term changes (Konova et al., 2020; Hitchcock et al., 2022a). In such cases, demonstrating responsiveness (i.e., showing that computational measures are sensitive to within-subject variations in symptoms), or predictive/convergent validity might be more informative than achieving high test-retest reliability. This once again stresses the importance of experimental designs that are based on big-picture thinking and aimed at addressing multiple subgoals simultaneously (Fig. 1).

10.3. Similarities and synergies with idiographic approaches in clinical psychology

It is interesting to note that many ideas central to this review happen to mirror recent trends in clinical psychology (Hayes et al., 2022, 2019; Wright and Woods, 2020; Hofmann et al., 2020; Hofmann and Hayes, 2019). Led by the motivation to personalize psychotherapy, researchers in clinical psychology are focusing on idiographic approaches (studying individual differences) versus nomotheic approaches (studying group averages), with the recognition that the latter do not generalize to the former (Molenaar, 2004). Moreover, similar to what we discuss in the review, there is a strong emphasis on the dynamic nature of mental disorders and the need to build methodological tools for capturing these dynamics at the individual level. Another similarity to what we propose here is an explicit focus on clinical utility: "By focusing on treatment utility as the beginning rather than the end of successful diagnosis, a far more pragmatic and immediately applicable research agenda emerges rather than the 'forever agenda' of endlessly seeking latent disease entities despite year after year of disappointment" (Hayes et al., 2022).

Complementary to what we discussed in the review, this body of research has put a lot more thought into understanding longitudinal individual differences in a clinical context and how to study it. That includes aspects of intervention science (i.e., distilling and personalizing the active ingredients of psychotherapy), mediators of therapeutic effects, the process of change and its non-linear nature. While there exists some work considering how computational constructs could inform psychotherapy (Moutoussis et al., 2018b; Holmes and Nolte, 2019; Nair et al., 2020; Smith et al., 2021c; Pott and Schilbach, 2022; Lohr and Hauke, 2022; Connolly, 2022), there seem to be many more synergistic avenues to explore, particularly with regards to personalization and the development of methodological tools that enable it.

10.4. The concept of validity and the focus on utility

While in this review we have adopted a fairly mainstream view of construct validity (largely in line with Kane, 2013), more nuanced discussions about its theoretical foundations could be had and can be found in the literature (Kane, 2013; Borsboom et al., 2004; Messick, 1984; Cronbach and Meehl, 1955). In particular, some authors have argued that validity can only be established by demonstrating a causal relationship between variation in the construct and variation in its measures (Borsboom et al., 2004). As such, this perspective is less concerned with the intended use and interpretation of the measure, and more with the (ontological) existence of the construct (i.e., it either exists or not). While we agree with Borsboom et al., that causality is a stronger indicator of validity (it is in line with our emphasis on longitudinal validity), we see value in correlational approaches too (e.g., convergent validity) and see no need to adopt a narrower definition of validity. We also see no reason to make strong ontological claims about the constructs of interest. In our world, all models are wrong, but some are useful. The more phenomena a model can explain/predict scientifically and the more it can be leveraged to improve patient outcomes, the better the model and the constructs that it relies on. The hierarchical breakdown of different types of validation that we present (Fig. 1) could therefore be seen as providing different degrees of evidence that the measure captures something meaningful and useful.

The focus on utility is also apparent in our summary of the different assay uses (Fig. 4). Some might object that this leaves out more explanatory goals (Paulus, 2017), i.e., using the assays to build explanatory models of mental disorders. In fact, we consider these goals to be implicit in our summary. For example, computational phenotyping entails computational characterization of a given disorder and identification of the most relevant treatment targets. This could provide useful insights to both clinicians and patients beyond any model predictions. In other words, for it to be useful, computational phenotyping must provide an interpretable and sufficiently intuitive explanatory model of one's condition. In research, building explanatory models for explanation's sake (i.e., basic research) could be considered as an important intermediate stage that provides an ever deeper understanding of various causal dynamics underlying disorders, creating the very possibility of conceptualizing new treatments and treatment targets. However, ultimately, the value of such explanatory models will depend on how actionable they are for improving clinical outcomes (see Weiss and Shanteau, 2021), for a related discussion).

10.5. Limitations

While in this review we focused primarily on behavioral studies, many of the same issues and solutions apply in the context of biomarker research using neuroimaging data, especially in task-based neuroimaging (Blair et al., 2022; Zuo et al., 2019; Milham et al., 2021; Feng

et al., 2022; Haines et al., 2023). However, neuroimaging introduces many additional challenges that would need to be addressed, including motion-related artifacts, physiological noise, generalizability across sites and scanners, cost and duration of scans, etc. (Barch and Mathalon, 2011; Kennedy et al., 2022; Noble et al., 2019; Finn and Rosenberg, 2021; Hu et al., 2022). Finally, we have also not considered the role of animal research, which has and will likely continue to play a significant role in the development and validation of computational models of cognition (Redish et al., 2022).

11. Conclusion

Studying individual differences is challenging, and that is just as true in the field of computational psychiatry. The field has been slow to appreciate the "mundane" aspects of the computational assay development, namely studying their reliability and construct validity (Paulus et al., 2016; Browning et al., 2020). The emerging empirical data suggests that poor reliability and construct validity are common. This poses a risk of invalidating previous findings and undermining ongoing research efforts focused on individual differences and assay development. Cross-sectional single-task designs, which currently dominate the research landscape, are not suitable for addressing these challenges. Instead, the field needs to adopt study designs that assess performance longitudinally on a battery of tasks, and focus on investigating reliability and convergent validity issues. Longitudinal designs will also be needed to move from cross-sectional clinical validity towards clinical utility. Finally, to make longitudinal designs applicable in clinical practice, and to make research more efficient, it is important to develop the assays in a way that minimizes the burden placed on patients.

CRediT authorship contribution statement

Povilas Karvelis conceptualized the review topic, conducted the literature search, and wrote the first draft. Andreea O. Diaconescu provided supervision and feedback throughout the process. Martin P. Paulus provided feedback for revising the manuscript.

Competing interests

MPP is an advisor to Spring Care, Inc., a behavioral health startup, he has received royalties for an article about methamphetamine in UpToDate. MPP has a consulting agreement with and receives compensation from F. Hoffmann-La Roche Ltd. All other authors have no competing interests to declare.

Data availability

No data was used for the research described in the article.

Acknowledgments

PK is supported by CIHR Fellowship. MPP is supported in part by The William K. Warren Foundation and the National Institute of General Medical Sciences Center Grant Award Number (1P20GM121312) and the National Institute on Drug Abuse (U01DA050989). AOD is supported by the Krembil Foundation.

References

Agrawal, M., Mattar, M.G., Cohen, J.D., Daw, N.D., 2022. The temporal dynamics of opportunity costs: A normative account of cognitive fatigue and boredom. Psychol. Rev. 129 (3), 564.

Ahn, W.-Y., Krawitz, A., Kim, W., Busemeyer, J.R., and Brown, J.W. (2013). A model-based fmri analysis with hierarchical bayesian parameter estimation.

Ahn, W.-Y., Gu, H., Shen, Y., Haines, N., Hahn, H.A., Teater, J.E., Myung, J.I., Pitt, M.A., 2020. Rapid, precise, and reliable measurement of delay discounting using a bayesian learning algorithm. Sci. Rep. 10 (1), 1–10.

Anokhin, A.P., Luciana, M., Banich, M., Barch, D., Bjork, J.M., Gonzalez, M.R., Gonzalez, R., Haist, F., Jacobus, J., Lisdahl, K., et al., 2022. Age-related changes and longitudinal stability of individual differences in abcd neurocognition measures. Dev. Cogn. Neurosci., 101078

Balter, L.J., Matheson, G.J., Sundelin, T., Sterzer, P., Petrovic, P., Axelsson, J., 2022. Experimental sleep deprivation results in diminished perceptual stability independently of psychosis proneness. Brain Sci. 12 (10), 1338.

Barch, D.M., Mathalon, D.H., 2011. Using brain imaging measures in studies of procognitive pharmacologic agents in schizophrenia: psychometric and quality assurance considerations. Biol. Psychiatry 70 (1), 13–18.

Barnes, L.L., Harp, D., Jung, W.S., 2002. Reliability generalization of scores on the spielberger state-trait anxiety inventory. Educ. Psychol. Meas. 62 (4), 603–618.

Baugh, F., 2002. Correcting effect sizes for score reliability: A reminder that measurement and substantive issues are linked inextricably. Educ. Psychol. Meas. 62 (2), 254–263.

Beglinger, L.J., Gaydos, B., Tangphao-Daniels, O., Duff, K., Kareken, D.A., Crawford, J., Fastenau, P.S., Siemers, E.R., 2005. Practice effects and the use of alternate forms in serial neuropsychological testing. Arch. Clin. Neuropsychol. 20 (4), 517–529.

Berwian, I.M., Wenzel, J.G., Collins, A.G., Seifritz, E., Stephan, K.E., Walter, H., Huys, Q. J., 2020. Computational mechanisms of effort and reward decisions in patients with depression and their association with relapse after antidepressant discontinuation. JAMA Psychiatry 77 (5), 513–522.

Blair, R.J.R., Mathur, A., Haines, N., Bajaj, S., 2022. Future directions for cognitive neuroscience in psychiatry: recommendations for biomarker design based on recent test re-test reliability work. Curr. Opin. Behav. Sci. 44, 101102.

Blatter, K., Cajochen, C., 2007. Circadian rhythms in cognitive performance: methodological constraints, protocols, theoretical underpinnings. Physiol. Behav. 90 (2–3), 196–208.

Borsboom, D., Cramer, A.O., et al., 2013. Network analysis: an integrative approach to the structure of psychopathology. Annu. Rev. Clin. Psychol. 9 (1), 91–121.

Borsboom, D., Mellenbergh, G.J., Van Heerden, J., 2004. The concept of validity. Psychol. Rev. 111 (4), 1061.

Brown, V.M., Chen, J., Gillan, C.M., Price, R.B., 2020. Improving the reliability of computational analyses: Model-based planning and its relationship with compulsivity. Biol. Psychiatry.: Cogn. Neurosci. Neuroimaging 5 (6), 601–609.

Brown, V.M., Zhu, L., Solway, A., Wang, J.M., McCurry, K.L., King-Casas, B., Chiu, P.H., 2021. Reinforcement learning disruptions in individuals with depression and sensitivity to symptom change following cognitive behavioral therapy. JAMA Psychiatry 78 (10), 1113–1122.

Browning, M., Carter, C.S., Chatham, C., DenOuden, H., Gillan, C.M., Baker, J.T., Chekroud, A.M., Cools, R., Dayan, P., Gold, J., et al., 2020. Realizing the clinical potential of computational psychiatry: report from the banbury center meeting, february 2019. Biol. Psychiatry 88 (2), e5–e10.

Bruder, L.R., Scharer, L., Peters, J., 2021. Reliability assessment of temporal discounting measures in virtual reality environments. Sci. Rep. 11 (1), 1–16.

Buelow, M.T., Barnhart, W.R., 2018. Test–retest reliability of common behavioral decision making tasks. Arch. Clin. Neuropsychol. 33 (1), 125–129.

Buelow, M.T., Suhr, J.A., 2013. Personality characteristics and state mood influence individual deck selections on the iowa gambling task. Personal. Individ. Differ. 54 (5), 593–597.

Calamia, M., Markon, K., Tranel, D., 2012. Scoring higher the second time around: meta-analyses of practice effects in neuropsychological assessment. Clin. Neuropsychol. 26 (4), 543–570.

Carrozzino, D., Patierno, C., Fava, G.A., Guidi, J., 2020. The hamilton rating scales for depression: a critical review of clinimetric properties of different versions. Psychother. Psychosom. 89 (3), 133–150.

Cavagnaro, D.R., Myung, J.I., Pitt, M.A., Kujala, J.V., 2010. Adaptive design optimization: A mutual information-based approach to model discrimination in cognitive science. Neural Comput. 22 (4), 887–905.

Chein, J.M., Schneider, W., 2005. Neuroimaging studies of practice-related change: fmri and meta-analytic evidence of a domain-general control network for learning. Cogn. Brain Res. 25 (3), 607–623.

Chekroud, A.M., Bondar, J., Delgadillo, J., Doherty, G., Wasil, A., Fokkema, M., Cohen, Z., Belgrave, D., DeRubeis, R., Iniesta, R., et al., 2021. The promise of machine learning in predicting treatment outcomes in psychiatry. World Psychiatry 20 (2), 154–170.

Chen, G., Pine, D.S., Brotman, M.A., Smith, A.R., Cox, R.W., Haller, S.P., 2021. Trial and error: A hierarchical modeling approach to test-retest reliability. NeuroImage 245, 118647.

Chouinard, P.A., Unwin, K.L., Landry, O., Sperandio, I., 2016. Susceptibility to optical illusions varies as a function of the autism-spectrum quotient but not in ways predicted by local–global biases. J. Autism Dev. Disord. 46 (6), 2224–2239.

Chrysaitis, N.A., Seriès, P., 2022. 10 years of bayesian theories of autism: a comprehensive review. Neurosci. Biobehav. Rev., 105022

Chung, D., Kadlec, K., Aimone, J.A., McCurry, K., King-Casas, B., Chiu, P.H., 2017. Valuation in major depression is intact and stable in a non-learning environment. Sci. Rep. 7 (1), 1–9.

Connolly, P., 2022. Instability and uncertainty are critical for psychotherapy: how the therapeutic alliance opens us up. Front. Psychol. 12, 6171.

Cook, D.A., Beckman, T.J., 2006. Current concepts in validity and reliability for psychometric instruments: theory and application. Am. J. Med. 119 (2), 166–e7.

Cooper, S.R., Gonthier, C., Barch, D.M., Braver, T.S., 2017. The role of psychometrics in individual differences research in cognition: A case study of the ax-cpt. Front. Psychol. 8, 1482.

Corlett, P.R., Frith, C.D., Fletcher, P.C., 2009. From drugs to deprivation: a bayesian framework for understanding models of psychosis. Psychopharmacology 206 (4), 515–530.
Cretenoud, A.F., Karimpur, H., Grzeczkowski, L., Francis, G., Hamburger, K., Herzog, M. H., 2019. Factors underlying visual illusions are illusion-specific but not feature-specific. J. Vis. 19 (14), 12.
Cronbach, L.J., Furby, L., 1970. How we should measure "change": Or should we? Psychol. Bull. 74 (1), 68.
Cronbach, L.J., Meehl, P.E., 1955. Construct validity in psychological tests. Psychol. Bull. 52 (4), 281.
Dang, J., King, K.M., Inzlicht, M., 2020. Why are self-report and behavioral measures weakly correlated? Trends Cogn. Sci. 24 (4), 267–269.
Davidson, J., Turnbull, C.D., Strickland, R., Miller, R., Graves, K., 1986. The montgomery-åsberg depression scale: reliability and validity. Acta Psychiatr. Scand. 73 (5), 544–548.
van der Vinne, N., Vollebregt, M.A., Rush, A.J., Eebes, M., van Putten, M.J., Arns, M., 2021. Eeg biomarker informed prescription of antidepressants in mdd: a feasibility trial. Eur. Neuropsychopharmacol. 44, 14–22.
Draheim, C., Mashburn, C.A., Martin, J.D., Engle, R.W., 2019. Reaction time in differential and developmental research: A review and commentary on the problems and alternatives. Psychol. Bull. 145 (5), 508.
Eckstein, M.K., Wilbrecht, L., Collins, A.G., 2021. What do reinforcement learning models measure? interpreting model parameters in cognition and neuroscience. Curr. Opin. Behav. Sci. 41, 128–137.
Eckstein, M.K., Master, S.L., Xia, L., Dahl, R.E., Wilbrecht, L., Collins, A.G., 2022. The interpretation of computational model parameters depends on the context. Elife 11, e75474.
Eisenberg, I.W., Bissett, P.G., ZeynepEnkavi, A., Li, J., MacKinnon, D.P., Marsch, L.A., Poldrack, R.A., 2019. Uncovering the structure of self-regulation through data-driven ontology discovery. Nat. Commun. 10 (1), 1–13.
Elliott, M.L., Knodt, A.R., Ireland, D., Morris, M.L., Poulton, R., Ramrakha, S., Sison, M. L., Moffitt, T.E., Caspi, A., Hariri, A.R., 2020. What is the test-retest reliability of common task-functional mri measures? new empirical evidence and a meta-analysis. Psychol. Sci. 31 (7), 792–806.
Enkavi, A.Z., Poldrack, R.A., 2021. Implications of the lacking relationship between cognitive task and self-report measures for psychiatry. Biol. Psychiatry.: Cogn. Neurosci. Neuroimaging 6 (7), 670–672.
Enkavi, A.Z., Eisenberg, I.W., Bissett, P.G., Mazza, G.L., MacKinnon, D.P., Marsch, L.A., Poldrack, R.A., 2019. Large-scale analysis of test–retest reliabilities of self-regulation measures. Proc. Natl. Acad. Sci. 116 (12), 5472–5477.
Epstein, S., 1979. The stability of behavior: I. on predicting most of the people much of the time. J. Personal. Soc. Psychol. 37 (7), 1097.
Epstein, S., 1980. The stability of behavior: Ii. implications for psychological research. Am. Psychol. 35 (9), 790.
Feng, C., Thompson, W.K., Paulus, M.P., 2022. Effect sizes of associations between neuroimaging measures and affective symptoms: A meta-analysis. Depress Anxiety 39 (1), 19–25.
Finn, E.S., Rosenberg, M.D., 2021. Beyond fingerprinting: Choosing predictive connectomes over reliable connectomes. NeuroImage 239, 118254.
Fisher, A.J., Medaglia, J.D., Jeronimus, B.F., 2018. Lack of group-to-individual generalizability is a threat to human subjects research. Proc. Natl. Acad. Sci. 115 (27), E6106–E6115.
Fleeson, W., Noftle, E., 2008. The end of the person–situation debate: An emerging synthesis in the answer to the consistency question. Soc. Personal. Psychol. Compass 2 (4), 1667–1684.
Fleiss, J.L., 2011. Design and Analysis of Clinical Experiments. John Wiley & Sons,.
Fletcher, P.C., Frith, C.D., 2009. Perceiving is believing: a bayesian approach to explaining the positive symptoms of schizophrenia. Nat. Rev. Neurosci. 10 (1), 48–58.
Forgas, J.P., 2017. Mood effects on cognition: Affective influences on the content and process of information processing and behavior. Emot. Affect Hum. Factors Hum. -Comput. Interact. 89–122.
Frässle, S., Marquand, A.F., Schmaal, L., Dinga, R., Veltman, D.J., Van der Wee, N.J., van Tol, M.-J., Schöbi, D., Penninx, B.W., Stephan, K.E., 2020. Predicting individual clinical trajectories of depression with generative embedding. NeuroImage: Clin. 26, 102213.
Frässle, S., Aponte, E.A., Bollmann, S., Brodersen, K.H., Do, C.T., Harrison, O.K., Harrison, S.J., Heinzle, J., Iglesias, S., Kasper, L., et al., 2021. Tapas: an open-source software package for translational neuromodeling and computational psychiatry. Front. Psychiatry 12, 680811.
Frey, R., Pedroni, A., Mata, R., Rieskamp, J., Hertwig, R., 2017. Risk preference shares the psychometric structure of major psychological traits. Sci. Adv. 3 (10), e1701381.
Friedman, N.P., Gustavson, D.E., 2022. Do rating and task measures of control abilities assess the same thing? Curr. Dir. Psychol. Sci., 09637214221091824
Gärtner, A., Strobel, A., 2021. Individual differences in inhibitory control: A latent variable analysis. J. Cogn. 4 (1).
Gauld, C., Depannemaecker, D., 2023. Dynamical systems in computational psychiatry: A toy-model to apprehend the dynamics of psychiatric symptoms. Front. Psychol.
Gell, M., Eickhoff, S.B., Omidvarnia, A., Kueppers, V., Patil, K.R., Satterthwaite, T.D., Mueller, V.I., and Langner, R. (2023). The burden of reliability: How measurement noise limits brain-behaviour predictions.bioRxiv, 2023-2102.
Germine, L., Strong, R.W., Singh, S., Sliwinski, M.J., 2021. Toward dynamic phenotypes and the scalable measurement of human behavior. Neuropsychopharmacology 46 (1), 209–216.
Gibbs-Dean, T., Katthagen, T., Tsenkova, I., Ali, R., Liang, X., Spencer, T., Diederen, K., 2023. Belief updating in psychosis, depression and anxiety disorders: A systematic review across computational modelling approaches. Neurosci. Biobehav. Rev., 105087
Gillan, C.M., Rutledge, R.B., 2021. Smartphones and the neuroscience of mental health. Annu. Rev. Neurosci. 44, 129.
Goldberg, T.E., Keefe, R.S., Goldman, R.S., Robinson, D.G., Harvey, P.D., 2010. Circumstances under which practice does not make perfect: a review of the practice effect literature in schizophrenia and its relevance to clinical treatment studies. Neuropsychopharmacology 35 (5), 1053–1062.
Grahek, I., Schaller, M., Tackett, J.L., 2021. Anatomy of a psychological theory: Integrating construct-validation and computational-modeling methods to advance theorizing. Perspectives on. Psychol. Sci. 16 (4), 803–815.
Grzeczkowski, L., Clarke, A.M., Francis, G., Mast, F.W., Herzog, M.H., 2017. About individual differences in vision. Vis. Res. 141, 282–292.
Grzeczkowski, L., Roinishvili, M., Chkonia, E., Brand, A., Mast, F.W., Herzog, M.H., Shaqiri, A., 2018. Is the perception of illusions abnormal in schizophrenia? Psychiatry Res. 270, 929–939.
Gu, X., Filbey, F., 2017. A bayesian observer model of drug craving. JAMA Psychiatry 74 (4), 419–420.
Gueguen, M.C., Schweitzer, E.M., Konova, A.B., 2021. Computational theory-driven studies of reinforcement learning and decision-making in addiction: What have we learned? Curr. Opin. Behav. Sci. 38, 40–48.
Haines, N., Kvam, P.D., Irving, L., Smith, C., Beauchaine, T.P., Pitt, M.A., Ahn, W.-Y., and Turner, B., 2020, Learning from the reliability paradox: How theoretically informed generative models can advance the social, behavioral, and brain sciences. PsyArXiv, psyarxiv. com/xr7y3.
Haines, N., Sullivan-Toole, H., Olino, T., 2023. From classical methods to generative models: Tackling the unreliability of neuroscientific measures in mental health research. Biol. Psychiatry.: Cogn. Neurosci. Neuroimaging.
Hauke, D., Roth, V., Karvelis, P., Adams, R., Moritz, S., Borgwardt, S., Diaconescu, A., Andreou, C., 2022. Increased belief instability in psychotic disorders predicts treatment response to metacognitive training. Schizophr. Bull.
Hauser, T.U., Skvortsova, V., De Choudhury, M., Koutsouleris, N., 2022. The promise of a model-based psychiatry: building computational models of mental ill health. Lancet Digit. Health.
Hayes, S.C., Hofmann, S.G., Stanton, C.E., Carpenter, J.K., Sanford, B.T., Curtiss, J.E., Ciarrochi, J., 2019. The role of the individual in the coming era of process-based therapy. Behav. Res. Ther. 117, 40–53.
Hayes, S.C., Ciarrochi, J., Hofmann, S.G., Chin, F., Sahdra, B., 2022. Evolving an idionomic approach to processes of change: Towards a unified personalized science of human improvement. Behav. Res. Ther., 104155
Hedge, C., Powell, G., Sumner, P., 2018. The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. Behav. Res. Methods 50 (3), 1166–1186.
Hedge, C., Bompas, A., Sumner, P., 2020. Task reliability considerations in computational psychiatry. Biol. Psychiatry.: Cogn. Neurosci. Neuroimaging 5 (9), 837–839.
Heise, D.R., 1969. Separating reliability and stability in test-retest correlation. Am. Sociol. Rev. 93–101.
Henderson, D., Poppe, A.B., Barch, D.M., Carter, C.S., Gold, J.M., Ragland, J.D., Silverstein, S.M., Strauss, M.E., MacDonald III, A.W., 2012. Optimization of a goal maintenance task for use in clinical applications. Schizophr. Bull. 38 (1), 104–113.
Herzog, P., Kube, T., Fassbinder, E., 2022. How childhood maltreatment alters perception and cognition–the predictive processing account of borderline personality disorder. Psychol. Med. 1–18.
Hitchcock, P., Niv, Y., Radulescu, A., Sims, C.R., 2017. Translating a reinforcement learning task into a computational psychiatry assay: Challenges and strategies. CogSci.
Hitchcock, P.F., Britton, W.B., Mehta, K.P., Frank, M.J., 2022a. Self-judgment dissected: A computational modeling analysis of self-referential processing and its relationship to trait mindfulness facets and depression symptoms. Cogn., Affect., Behav. Neurosci. 1–19.
Hitchcock, P.F., Fried, E.I., Frank, M.J., 2022b. Computational psychiatry needs time and context. Annu. Rev. Psychol. 73, 243–270.
Hoekstra, R.A., Bartels, M., Cath, D.C., Boomsma, D.I., 2008. Factor structure, reliability and criterion validity of the autism-spectrum quotient (aq): a study in dutch population and patient groups. J. Autism Dev. Disord. 38 (8), 1555–1566.
Hofmann, S.G., Hayes, S.C., 2019. The future of intervention science: Process-based therapy. Clin. Psychol. Sci. 7 (1), 37–50.
Hofmann, S.G., Curtiss, J.E., Hayes, S.C., 2020. Beyond linear mediation: Toward a dynamic network approach to study treatment processes. Clin. Psychol. Rev. 76, 101824.
Hollon, D., Miller, I.J., Robinson, E., et al., 2002. Criteria for evaluating treatment guidelines. Am. Psychol. 57 (12), 1052–1059.
Holmes, J., Nolte, T., 2019. "Surprise" and the bayesian brain: implications for psychotherapy theory and practice. Front. Psychol. 10, 592.
Howlett, J.R., Thompson, W.K., Paulus, M.P., 2020. Computational evidence for underweighting of current error and overestimation of future error in anxious individuals. Biol. Psychiatry.: Cogn. Neurosci. Neuroimaging 5 (4), 412–419.
Howlett, J.R., Larkin, F., Touthang, J., Kuplicki, R.T., Lim, K.O., Paulus, M.P., 2022. Rapid, reliable mobile assessment of affect-related motor processing. Behav. Res. Methods.
Hu, Z., Zhang, Z., Liang, Z., Zhang, L., Li, L., Huang, G., 2022. A new perspective on individual reliability beyond group effect for event-related potentials: A multisensory investigation and computational modeling. NeuroImage, 118937.

Huys, Q.J., Cools, R., Gölzer, M., Friedel, E., Heinz, A., Dolan, R.J., Dayan, P., 2011. Disentangling the roles of approach, activation and valence in instrumental and pavlovian responding. PLoS Comput. Biol. 7 (4), e1002028.
Huys, Q.J., Eshel, N., O'Nions, E., Sheridan, L., Dayan, P., Roiser, J.P., 2012. Bonsai trees in your head: how the pavlovian system sculpts goal-directed choices by pruning decision trees. PLoS Comput. Biol. 8 (3), e1002410.
Huys, Q.J., Maia, T.V., Frank, M.J., 2016. Computational psychiatry as a bridge from neuroscience to clinical applications. Nat. Neurosci. 19 (3), 404–413.
Huys, Q.J., Browning, M., Paulus, M.P., Frank, M.J., 2021. Advances in the computational understanding of mental illness. Neuropsychopharmacology 46 (1), 3–19.
Iglesias, S., Tomiello, S., Schneebeli, M., Stephan, K.E., 2017. Models of neuromodulation for computational psychiatry. Wiley Interdiscip. Rev.: Cogn. Sci. 8 (3), e1420.
Jangraw, D.C., Keren, H., Sun, H., Bedder, R.L., Rutledge, R.B., Pereira, F., Thomas, A.G., Pine, D.S., Zheng, C., Nielson, D.M., Stringaris, A., 2023. A highly replicable decline in mood during rest and simple tasks. Nat. Hum. Behav. 2397–3374.
Kafadar, E., Fisher, V.L., Quagan, B., Hammer, A., Jaeger, H., Mourgues, C., Thomas, R., Chen, L., Imtiaz, A., Sibarium, E., et al., 2022. Conditioned hallucinations and prior overweighting are state-sensitive markers of hallucination susceptibility. Biol. Psychiatry 92 (10), 772–780.
Kaliuzhna, M., Stein, T., Rusch, T., Sekutowicz, M., Sterzer, P., Seymour, K.J., 2019. No evidence for abnormal priors in early vision in schizophrenia. Schizophr. Res. 210, 245–254.
Kane, M.T., 2013. Validating the interpretations and uses of test scores. J. Educ. Meas. 50 (1), 1–73.
Karvelis, P., Diaconescu, A.O., 2022. A computational model of hopelessness and active-escape bias in suicidality. Comput. Psychiatry 6, 1.
Karvelis, P., Seitz, A.R., Lawrie, S.M., Seriès, P., 2018. Autistic traits, but not schizotypy, predict increased weighting of sensory information in bayesian visual integration. ELife 7.
Karvelis, P., Charlton, C.E., Allohverdi, S.G., Bedford, P., Hauke, D.J., Diaconescu, A.O., 2022. Computational approaches to treatment response prediction in major depression using brain activity and behavioral data: A systematic review. Netw. Neurosci. 1–52.
Katahira, K., 2016. How hierarchical models improve point estimates of model parameters at the individual level. J. Math. Psychol. 73, 37–58.
Katthagen, T., Fromm, S., Wieland, L., Schlagenhauf, F., 2022. Models of dynamic belief updating in psychosis—a review across different computational approaches. Front. Psychiatry 13.
Kelly, A.C., Garavan, H., 2005. Human functional neuroimaging of brain changes associated with practice. Cereb. Cortex 15 (8), 1089–1102.
Kelly, C.J., Karthikesalingam, A., Suleyman, M., Corrado, G., King, D., 2019. Key challenges for delivering clinical impact with artificial intelligence. BMC Med. 17 (1), 1–9.
Kennedy, J.T., Harms, M.P., Korucuoglu, O., Astafiev, S.V., Barch, D.M., Thompson, W. K., Bjork, J.M., Anokhin, A.P., 2022. Reliability and stability challenges in abcd task fmri data. NeuroImage 252, 119046.
Kenrick, D.T., Funder, D.C., 1988. Profiting from controversy: Lessons from the person-situation debate. Am. Psychol. 43 (1), 23.
Kievit, R.A., Frankenhuis, W.E., Waldorp, L.J., Borsboom, D., 2013. Simpsonas paradox in psychological science: a practical guide. Front. Psychol. 4, 513.
Kingslake, J., Dias, R., Dawson, G.R., Simon, J., Goodwin, G.M., Harmer, C.J., Morriss, R., Brown, S., Guo, B., Dourish, C.T., et al., 2017. The effects of using the predict test to guide the antidepressant treatment of depressed patients: study protocol for a randomised controlled trial. Trials 18 (1), 1–10.
Konova, A.B., Lopez-Guzman, S., Urmanche, A., Ross, S., Louie, K., Rotrosen, J., Glimcher, P.W., 2020. Computational markers of risky decision-making for identification of temporal windows of vulnerability to opioid use in a real-world clinical setting. JAMA Psychiatry 77 (4), 368–377.
Koo, T.K., Li, M.Y., 2016. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J. Chiropr. Med. 15 (2), 155–163.
Kucina, T., Wells, L., Lewis, I., de Salas, K., Kohl, A., Palmer, M., Sauer, J.D., Matzke, D., Aidman, E., and Heathcote, A. (2022). A solution to the reliability paradox for decision-conflict tasks.
Kwon, M., Lee, S.H., Ahn, W.-Y., 2022. Adaptive design optimization as a promising tool for reliable and efficient computational fingerprinting. Biol. Psychiatry.: Cogn. Neurosci. Neuroimaging.
Lambourne, K., Tomporowski, P., 2010. The effect of exercise-induced arousal on cognitive task performance: a meta-regression analysis. Brain Res. 1341, 12–24.
Lan, D., Browning, M., 2022. What can reinforcement learning models of dopamine and serotonin tell us about the action of antidepressants? Comput. Psychiatry 6, 1.
Landis, J.R., Koch, G.G., 1977. The measurement of observer agreement for categorical data. biometrics 159–174.
Lannoy, S., Pfefferbaum, A., LeBerre, A.-P., Thompson, W.K., Brumback, T., Schulte, T., Pohl, K.M., De Bellis, M.D., Nooner, K.B., Baker, F.C., et al., 2021. Growth trajectories of cognitive and motor control in adolescence: How much is development and how much is practice? Neuropsychology.
Liang, M.H., 2000. Longitudinal construct validity: establishment of clinical meaning in patient evaluative instruments. Med. care 38 (9), II–84.
Lilienfeld, S.O., 2014. The research domain criteria (rdoc): An analysis of methodological and conceptual challenges. Behav. Res. Ther. 62, 129–139.
Liljequist, D., Elfving, B., Skavberg Roaldsen, K., 2019. Intraclass correlation–a discussion and demonstration of basic features. PloS One 14 (7), e0219854.
Linson, A., Parr, T., Friston, K.J., 2020. Active inference, stressors, and psychological trauma: A neuroethological model of (mal) adaptive explore-exploit dynamics in ecological context. Behav. Brain Res. 380, 112421.
Littman, R., Hochman, S., and Kalanthroff, E., 2022, Reliable affordances: A generative modeling approach for test-retest reliability of the affordances task.
Lohr, C., Hauke, G., 2022. Piloting the update: The use of therapeutic relationship for change. a free energy account. Front. Psychol. 1306.
Loosen, A.M., Seow, T., and Hauser, T.U. (2022). Consistency within change: Evaluating the psychometric properties of a widely-used predictive-inference task.
Maier, W., Buller, R., Philipp, M., Heuser, I., 1988. The hamilton anxiety scale: reliability, validity and sensitivity to change in anxiety and depressive disorders. J. Affect. Disord. 14 (1), 61–68.
McGovern, H., De Foe, A., Biddell, H., Leptourgos, P., Corlett, P., Bandara, K., Hutchinson, B.T., 2022. Learned uncertainty: The free energy principle in anxiety. Front. Psychol. 13.
McHugh, R.K., Daughters, S.B., Lejuez, C.W., Murray, H.W., Hearon, B.A., Gorka, S.M., Otto, M.W., 2011. Shared variance among self-report and behavioral measures of distress intolerance. Cogn. Ther. Res. 35 (3), 266–275.
McLean, B.F., Mattiske, J.K., Balzan, R.P., 2018. Towards a reliable repeated-measures beads task for assessing the jumping to conclusions bias. Psychiatry Res. 265, 200–207.
Messick, S., 1984. The psychology of educational measurement. ETS Res. Rep. Ser. 1984 (1), i–55.
Milham, M.P., Vogelstein, J., Xu, T., 2021. Removing the reliability bottleneck in functional magnetic resonance imaging research to achieve clinical utility. JAMA Psychiatry 78 (6), 587–588.
Mkrtchian, A., Valton, V., Roiser, J.P., 2023. Reliability of decision-making and reinforcement learning computational parameters. Comput. Psychiatry 7 (1), 30–46.
Mokkink, L., Terwee, C., de Vet, H., 2021. Key concepts in clinical epidemiology: Responsiveness, the longitudinal aspect of validity. J. Clin. Epidemiol. 140, 159–162.
Mokkink, L.B., Terwee, C.B., Patrick, D.L., Alonso, J., Stratford, P.W., Knol, D.L., Bouter, L.M., de Vet, H.C., 2010. The cosmin study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. J. Clin. Epidemiol. 63 (7), 737–745.
Molenaar, P.C., 2004. A manifesto on psychology as idiographic science: Bringing the person back into scientific psychology, this time forever. Measurement 2 (4), 201–218.
Molenaar, P.C., Campbell, C.G., 2009. The new person-specific paradigm in psychology. Curr. Dir. Psychol. Sci. 18 (2), 112–117.
Moutoussis, M., Bullmore, E.T., Goodyer, I.M., Fonagy, P., Jones, P.B., Dolan, R.J., Dayan, P., in Psychiatry Network Research Consortium, N., 2018a. Change, stability, and instability in the pavlovian guidance of behaviour from adolescence to young adulthood. PLoS Comput. Biol. 14 (12), e1006679.
Moutoussis, M., Shahar, N., Hauser, T.U., Dolan, R.J., 2018b. Computation in psychotherapy, or how computational psychiatry can aid learning-based psychological therapies. Comput. Psychiatry (Camb., Mass.) 2, 50.
Myung, J.I., Cavagnaro, D.R., Pitt, M.A., 2013. A tutorial on adaptive design optimization. J. Math. Psychol. 57 (3–4), 53–67.
Nair, A., Rutledge, R.B., Mason, L., 2020. Under the hood: using computational psychiatry to make psychological therapies more mechanism-focused. Front. Psychiatry 11, 140.
Nikolaidis, A., Chen, A.A., He, X., Shinohara, R., Vogelstein, J., Milham, M., and Shou, H. (2022). Suboptimal phenotypic reliability impedes reproducible human neuroscience.bioRxiv.
Nitsch, F.J., Lüpken, L.M., Lüschow, N., Kalenscher, T., 2022. On the reliability of individual economic rationality measurements. Proc. Natl. Acad. Sci. 119 (31), e2202070119.
Noble, S., Scheinost, D., Constable, R.T., 2019. A decade of test-retest reliability of functional connectivity: A systematic review and meta-analysis. Neuroimage 203, 116157.
Noreen, S., MacLeod, M.D., 2015. What do we really know about cognitive inhibition? task demands and inhibitory effects across a range of memory and behavioural tasks. PLoS One 10 (8), e0134951.
Norman, R.M., Malla, A.K., Cortese, L., Diaz, F., 1996. A study of the interrelationship between and comparative interrater reliability of the saps, sans and panss. Schizophr. Res. 19 (1), 73–85.
Palmer, C.J., Lawson, R.P., Hohwy, J., 2017. Bayesian approaches to autism: Towards volatility, action, and behavior. Psychol. Bull. 143 (5), 521.
Palminteri, S., Chevallier, C., 2018. Can we infer inter-individual differences in risk-taking from behavioral tasks? Front. Psychol. 9, 2307.
Palminteri, S., Wyart, V., Koechlin, E., 2017. The importance of falsification in computational cognitive modeling. Trends Cogn. Sci. 21 (6), 425–433.
Parsons, S., Kruijt, A.-W., Fox, E., 2019. Psychological science needs a standard practice of reporting the reliability of cognitive-behavioral measurements. Advances in Methods and Practices in. Psychol. Sci. 2 (4), 378–395.
Patzelt, E.H., Hartley, C.A., Gershman, S.J., 2018. Computational phenotyping: using models to understand individual differences in personality, development, and mental illness. Personal. Neurosci. 1.
Paulus, M.P., 2017. Evidence-based pragmatic psychiatry—a call to action. JAMA Psychiatry 74 (12), 1185–1186.
Paulus, M.P., Thompson, W.K., 2021. Computational approaches and machine learning for individual-level treatment predictions. Psychopharmacology 238 (5), 1231–1239.
Paulus, M.P., Huys, Q.J., Maia, T.V., 2016. A roadmap for the development of applied computational psychiatry. Biol. Psychiatry.: Cogn. Neurosci. neuroimaging 1 (5), 386–392.
Paulus, M.P., Feinstein, J.S., Khalsa, S.S., 2019. An active inference approach to interoceptive psychopathology. Annu. Rev. Clin. Psychol. 15, 97.

Pedroni, A., Frey, R., Bruhin, A., Dutilh, G., Hertwig, R., Rieskamp, J., 2017. The risk elicitation puzzle. Nat. Hum. Behav. 1 (11), 803–809.
Pellicano, E., Burr, D., 2012. When the world becomes 'too real': a bayesian explanation of autistic perception. Trends Cogn. Sci. 16 (10), 504–510.
Peng, Y., Knotts, J.D., Taylor, C.T., Craske, M.G., Stein, M.B., Bookheimer, S., Young, K. S., Simmons, A.N., Yeh, H.-W., Ruiz, J., et al., 2021. Failure to identify robust latent variables of positive or negative valence processing across units of analysis. Biol. Psychiatry.: Cogn. Neurosci. Neuroimaging 6 (5), 518–526.
Peralta, V., Cuesta, M., De Leon, J., 1995. Positive and negative symptoms/syndromes in schizophrenia: reliability and validity of different diagnostic systems. Psychol. Med. 25 (1), 43–50.
Peters, R., White, D., Cleeland, C., Scholey, A., 2020. Fuel for thought? a systematic review of neuroimaging studies into glucose enhancement of cognitive performance. Neuropsychol. Rev. 30 (2), 234–250.
Pike, A.C., Robinson, O.J., 2022. Reinforcement learning in patients with mood and anxiety disorders vs control individuals: A systematic review and meta-analysis. JAMA Psychiatry.
Pike, A.C., Tan, K., Ansari, H.J., Wing, M., and Robinson, O.J. (2022). Test-retest reliability of affective bias tasks.
Plummer, P., Grewal, G., Najafi, B., Ballard, A., 2015. Instructions and skill level influence reliability of dual-task performance in young adults. Gait Posture 41 (4), 964–967.
Pooseh, S., Bernhardt, N., Guevara, A., Huys, Q.J., Smolka, M.N., 2018. Value-based decision-making battery: A bayesian adaptive approach to assess impulsive and risky behavior. Behav. Res. Methods 50 (1), 236–249.
Pott, J., Schilbach, L., 2022. Tracking and changing beliefs during social interaction: Where computational psychiatry meets cognitive behavioral therapy. Front. Psychol. 5812.
Pratap, A., Neto, E.C., Snyder, P., Stepnowsky, C., Elhadad, N., Grant, D., Mohebbi, M.H., Mooney, S., Suver, C., Wilbanks, J., et al., 2020. Indicators of retention in remote digital health studies: a cross-study evaluation of 100,000 participants. NPJ Digit. Med. 3 (1), 1–10.
Price, R.B., Brown, V., Siegle, G.J., 2019. Computational modeling applied to the dot-probe task yields improved reliability and mechanistic insights. Biol. Psychiatry 85 (7), 606–612.
Pronk, T., Molenaar, D., Wiers, R.W., Murre, J., 2021. Methods to split cognitive task data for estimating split-half reliability: A comprehensive review and systematic assessment. Psychon. Bull. Rev. 1–11.
Pronk, T., Hirst, R.J., Wiers, R.W., Murre, J.M., 2022. Can we measure individual differences in cognitive measures reliably via smartphones? a comparison of the flanker effect across device types and samples. Behav. Res. Methods 1–12.
Raud, L., Westerhausen, R., Dooley, N., Huster, R.J., 2020. Differences in unity: The go/no-go and stop signal tasks rely on different mechanisms. NeuroImage 210, 116582.
Redish, A.D., Kepecs, A., Anderson, L.M., Calvin, O.L., Grissom, N.M., Haynos, A.F., Heilbronner, S.R., Herman, A.B., Jacob, S., Ma, S., et al., 2022. Computational validity: using computation to translate behaviours across species. Philos. Trans. R. Soc. B 377 (1844), 20200525.
Reed, G.M., Sharan, P., Rebello, T.J., Keeley, J.W., ElenaMedina-Mora, M., Gureje, O., LuisAyuso-Mateos, J., Kanba, S., Khoury, B., Kogan, C.S., et al., 2018. The icd-11 developmental field study of reliability of diagnoses of high-burden mental disorders: results among adult patients in mental health settings of 13 countries. World Psychiatry 17 (2), 174–186.
Regier, D.A., Narrow, W.E., Clarke, D.E., Kraemer, H.C., Kuramoto, S.J., Kuhl, E.A., Kupfer, D.J., 2013. Dsm-5 field trials in the united states and canada, part ii: test-retest reliability of selected categorical diagnoses. Am. J. Psychiatry 170 (1), 59–70.
Reiter, A.M., Atiya, N.A., Berwian, I.M., Huys, Q.J., 2021. Neuro-cognitive processes as mediators of psychological treatment effects. Curr. Opin. Behav. Sci. 38, 103–109.
Richards, K.L., Karvelis, P., Lawrie, S.M., Seriès, P., 2020. Visual statistical learning and integration of perceptual priors are intact in attention deficit hyperactivity disorder. PloS One 15 (12), e0243100.
Robbins, T.W., Cardinal, R.N., 2019. Computational psychopharmacology: a translational and pragmatic approach. Psychopharmacology 236 (8), 2295–2305.
Rodebaugh, T.L., Scullin, R.B., Langer, J.K., Dixon, D.J., Huppert, J.D., Bernstein, A., Zvielli, A., Lenze, E.J., 2016. Unreliability as a threat to understanding psychopathology: The cautionary tale of attentional bias. J. Abnorm. Psychol. 125 (6), 840.
Rogers, P.J., Heatherley, S.V., Mullings, E.L., Smith, J.E., 2013. Faster but not smarter: effects of caffeine and caffeine withdrawal on alertness and performance. Psychopharmacology 226 (2), 229–240.
Rouder, J.N., Haaf, J.M., 2019. A psychometrics of individual differences in experimental tasks. Psychon. Bull. Rev. 26 (2), 452–467.
Rutledge, R.B., Chekroud, A.M., Huys, Q.J., 2019. Machine learning and big data in psychiatry: toward clinical applications. Curr. Opin. Neurobiol. 55, 152–159.
Sackett, P.R., Lievens, F., Van Iddekinge, C.H., Kuncel, N.R., 2017. Individual differences and their measurement: A review of 100 years of research. J. Appl. Psychol. 102 (3), 254.
Saunders, B., Milyavskaya, M., Etz, A., Randles, D., Inzlicht, M., Vazire, S., 2018. Reported self-control is not meaningfully associated with inhibition-related executive function: A bayesian analysis. Collabra: Psychol. 4, 1.
Scharfen, J., Peters, J.M., Holling, H., 2018. Retest effects in cognitive ability tests: A meta-analysis. Intelligence 67, 44–66.
Scheibehenne, B., Pachur, T., 2015. Using bayesian hierarchical parameter estimation to assess the generalizability of cognitive models of choice. Psychon. Bull. Rev. 22 (2), 391–407.
Schmidt, C., Collette, F., Cajochen, C., Peigneux, P., 2007. A time to think: circadian rhythms in human cognition. Cogn. Neuropsychol. 24 (7), 755–789.
Schwartenbeck, P., FitzGerald, T.H., Mathys, C., Dolan, R., Wurst, F., Kronbichler, M., Friston, K., 2015. Optimal inference with suboptimal models: addiction and active bayesian inference. Med. Hypotheses 84 (2), 109–117.
Sebold, M., Nebe, S., Garbusow, M., Guggenmos, M., Schad, D.J., Beck, A., Kuitunen-Paul, S., Sommer, C., Frank, R., Neu, P., et al., 2017. When habits are dangerous: alcohol expectancies and habitual decision making predict relapse in alcohol dependence. Biol. Psychiatry 82 (11), 847–856.
Shahar, N., Hauser, T.U., Moutoussis, M., Moran, R., Keramati, M., Consortium, N., Dolan, R.J., 2019. Improving the reliability of model-based decision-making estimates in the two-stage decision task with reaction-times and drift-diffusion modeling. PLoS Comput. Biol. 15 (2), e1006803.
Sharp, P.B., Miller, G.A., Dolan, R.J., Eldar, E., 2020. Towards formal models of psychopathological traits that explain symptom trajectories. BMC Med. 18 (1), 1–8.
Shrout, P.E., 1998. Measurement reliability and agreement in psychiatry. Stat. Methods Med. Res. 7 (3), 301–317.
Smith, R., Badcock, P., Friston, K.J., 2021a. Recent advances in the application of predictive coding and active inference models within clinical neuroscience. Psychiatry Clin. Neurosci. 75 (1), 3–13.
Smith, R., Kirlic, N., Stewart, J.L., Touthang, J., Kuplicki, R., McDermott, T.J., Taylor, S., Khalsa, S.S., Paulus, M.P., Aupperle, R.L., 2021b. Long-term stability of computational parameters during approach-avoidance conflict in a transdiagnostic psychiatric patient sample. Sci. Rep. 11 (1), 1–13.
Smith, R., Moutoussis, M., Bilek, E., 2021c. Simulating the computational mechanisms of cognitive and behavioral psychotherapeutic interventions: Insights from active inference. Sci. Rep. 11 (1), 10128.
Smith, R., Taylor, S., Stewart, J.L., Guinjoan, S.M., Ironside, M., Kirlic, N., Ekhtiari, H., White, E.J., Zheng, H., Kuplicki, R., et al., 2022. Slower learning rates from negative outcomes in substance use disorder over a 1-year period and their potential predictive utility. Comput. Psychiatry 6, 1.
Snijder, J.-P., Tang, R., Bugg, J., Conway, A.R., and Braver, T. (2022). On the psychometric evaluation of cognitive control tasks: An investigation with the dual mechanisms of cognitive control (dmcc) battery.
Spearman, C., 1904. The proof and measurement of association between two things. Am. J. Psychol. 15 (1), 72–101.
Sripada, C., Weigard, A., 2021. Impaired evidence accumulation as a transdiagnostic vulnerability factor in psychopathology. Front. Psychiatry 12, 627179.
Stephan, K.E., Mathys, C., 2014. Computational approaches to psychiatry. Curr. Opin. Neurobiol. 25, 85–92.
Stephan, K.E., Schlagenhauf, F., Huys, Q.J., Raman, S., Aponte, E.A., Brodersen, K.H., Rigoux, L., Moran, R.J., Daunizeau, J., Dolan, R.J., et al., 2017. Computational neuroimaging strategies for single patient predictions. Neuroimage 145, 180–199.
Sterzer, P., Adams, R.A., Fletcher, P., Frith, C., Lawrie, S.M., Muckli, L., Petrovic, P., Uhlhaas, P., Voss, M., Corlett, P.R., 2018. The predictive coding account of psychosis. Biol. Psychiatry 84 (9), 634–643.
Sullivan, E.V., Brumback, T., Tapert, S.F., Prouty, D., Fama, R., Thompson, W.K., Brown, S.A., Cummins, K., Colrain, I.M., Baker, F.C., et al., 2017. Effects of prior testing lasting a full year in ncanda adolescents: contributions from age, sex, socioeconomic status, ethnicity, site, family history of alcohol or drug abuse, and baseline performance. Dev. Cogn. Neurosci. 24, 72–83.
Sullivan-Toole, H., Haines, N., Dale, K., Olino, T., et al., 2022. Enhancing the psychometric properties of the iowa gambling task using full generative modeling. Comput. Psychiatry 6 (1), 189–212.
Tulver, K., Aru, J., Rutiku, R., Bachmann, T., 2019. Individual differences in the effects of priors on perception: a multi-paradigm approach. Cognition 187, 167–177.
Van Bork, R., Rhemtulla, M., Sijtsma, K., Borsboom, D., 2022. A causal theory of error scores. Psychol. Methods.
Vermeir, J.F., White, M.J., Johnson, D., Crombez, G., Van Ryckeghem, D.M., 2020. The effects of gamification on computerized cognitive training: systematic review and meta-analysis. JMIR Serious Games 8 (3), e18644.
Vinckier, F., Jaffre, C., Gauthier, C., Smajda, S., Abdel-Ahad, P., LeBouc, R., Daunizeau, J., Fefeu, M., Borderies, N., Plaze, M., et al., 2022. Elevated effort cost identified by computational modeling as a distinctive feature explaining multiple behaviors in patients with depression. Biol. Psychiatry.: Cogn. Neurosci. Neuroimaging.
Waltmann, M., Schlagenhauf, F., Deserno, L., 2022. Sufficient reliability of the behavioral and computational readouts of a probabilistic reversal learning task. Behav. Res. Methods 1–22.
Weigard, A., Sripada, C., 2021. Task-general efficiency of evidence accumulation as a computationally defined neurocognitive trait: Implications for clinical neuroscience. Biol. Psychiatry Glob. Open Sci. 1 (1), 5–15.
Weigard, A., Clark, D.A., Sripada, C., 2021. Cognitive efficiency beats top-down control as a reliable individual difference dimension relevant to self-control. Cognition 215, 104818.
Weir, J.P., 2005. Quantifying test-retest reliability using the intraclass correlation coefficient and the sem. J. Strength Cond. Res. 19 (1), 231–240.
Weiss, D.J., Shanteau, J., 2021. The futility of decision making research. Stud. Hist. Philos. Sci. Part A 90, 10–14.
Whitehead, P.S., Brewer, G.A., Blais, C., 2020. Reliability and convergence of conflict effects. Exp. Psychol.
Wiecki, T.V., Sofer, I., Frank, M.J., 2013. Hddm: Hierarchical bayesian estimation of the drift-diffusion model in python. Front. Neuroinformatics 14.
Williams, R.H., Zimmerman, D.W., 1989. Statistical power analysis and reliability of measurement. J. Gen. Psychol. 116 (4), 359–369.
Wilson, R.C., Collins, A.G., 2019. Ten simple rules for the computational modeling of behavioral data. Elife 8, e49547.

Wright, A.G., Woods, W.C., 2020. Personalized models of psychopathology. Annu. Rev. Clin. Psychol. 16, 49–74.
Xu, Y., Stocco, A., 2021. Recovering reliable idiographic biological parameters from noisy behavioral data: the case of basal ganglia indices in the probabilistic selection task. Comput. Brain Behav. 4 (3), 318–334.
Yarkoni, T., 2022. The generalizability crisis. Behav. Brain Sci. 45.
Yip, S.W., Barch, D.M., Chase, H.W., Flagel, S., Huys, Q.J., Konova, A.B., Montague, R., Paulus, M., 2022. From computation to clinic. Biol. Psychiatry Glob. Open Sci.
Zander, E., Willfors, C., Berggren, S., Choque-Olsson, N., Coco, C., Elmund, A., Moretti, A.H., Holm, A., Jifält, I., Kosieradzki, R., et al., 2016. The objectivity of the autism diagnostic observation schedule (ados) in naturalistic clinical settings. Eur. Child Adolesc. Psychiatry 25 (7), 769–780.
Zech, H.G., Reichert, M., Ebner-Priemer, U.W., Tost, H., Rapp, M.A., Heinz, A., Dolan, R. J., Smolka, M.N., Deserno, L., 2022. Mobile data collection of cognitive-behavioral tasks in substance use disorders: Where are we now? Neuropsychobiology 1–13.
Zorowitz, S., Niv, Y., 2023. Improving the reliability of cognitive task measures: A narrative review. Biol. Psychiatry.: Cogn. Neurosci. Neuroimaging.
Zuo, X.-N., Xu, T., Milham, M.P., 2019. Harnessing reliability for neuroscience research. Nat. Hum. Behav. 3 (8), 768–771.