Reliable detection of longitudinal change in computational models

Taekwan Kim,[1†] Essi Viding,[2*] Quentin J.M. Huys[1*]

Author affiliations:

[1] Applied Computational Psychiatry Lab, Mental Health Neuroscience Department, Division of Psychiatry and Max Planck UCL Centre for Computational Psychiatry and Ageing Research, Queen Square Institute of Neurology, University College London, London, UK

[2] Division of Psychology and Language Sciences, University College London, London, UK

[†] Corresponding author

[*] These authors contributed equally to this work and share senior authorship.

Contact Information:

Dr. Taekwan Kim (tkim.neurosci@gmail.com)

Abstract

Computational models are increasingly used to infer latent cognitive change across development and treatment, yet the reliability of longitudinal change detection remains poorly understood. Here, we present a simulation-based framework for evaluating whether latent change can be detected reliably before data collection. Using reinforcement learning as a testbed, we compared two approaches to change detection: fit-based model comparison and parameter-based thresholding of estimated change parameters. Reliable detection depends not only on effect size and sample size, but also on likelihood geometry, model parsimony, and task design. Fit-based detection provided a more practical basis for validation, whereas parameter-based detection was limited by threshold ambiguity and parameter trade-offs. In regimes with poor detectability, simulation-based task optimisation improved sensitivity without increasing sample size, a finding of particular relevance for studies involving hard-to-recruit populations. Our framework provides a practical route for validating longitudinal change models before their use in developmental or clinical intervention research.

1. Introduction

Computational modelling provides a useful framework for studying cognitive mechanisms because it can decompose high-level cognition into latent processes aligned with neurobiological models. This makes it particularly promising for understanding how cognition changes across development and in response to treatment.[1,2] Yet reliably and sensitively measuring cognitive change in computational models remains challenging. Traditionally, computational modelling of longitudinal change has relied on separate-session fitting, treating different time points as independent individuals.[3-5] This approach is inherently limited because within-subject change is defined as the difference between noisy session-specific estimates, conflating true change with estimation error. Recent studies have therefore moved toward **joint-session fitting**, or change modelling, in which multi-session data are fit within a single computational framework. This allows behaviour to be decomposed into a shared baseline component and a change component, while accounting for within-subject dependence between parameters.[6-8]

However, adopting joint-session models does not guarantee reliable change detection. For change models to support valid inference, added change parameters must produce distinct behavioural signatures that are identifiable from the data. In particular, **fit-based model identifiability** (i.e., whether a target model is correctly preferred over alternatives when fitted to data generated by that model) is a prerequisite for reliable change detection. Although such model recovery analyses are standard in cross-sectional designs, they remain largely overlooked in longitudinal studies. Instead, prior work has mainly relied on **parameter recovery** (i.e., the correspondence between true and recovered change estimates), leaving untested whether detected change is correctly attributed to the intended model.[8-11] Yet parameter recovery alone cannot validate change detection in practice. In real data, true change parameters are unknown, so fitted differences do not by themselves provide a principled criterion for declaring change. These differences may also be difficult to interpret when parameters are weakly identifiable, correlated, or poorly recovered under realistic trial counts and estimation noise.

Fit-based identifiability should therefore be evaluated a priori as part of longitudinal study design. This matters for two reasons: it determines which change regimes allow competing models to be distinguished under a given task design, and whether detectability can be improved when discrimination is poor. The first question concerns **practical identifiability**, the extent to which recoverability varies across parameter regimes.[12] Because change models are by nature more complex than their simpler alternatives, behavioural data should be sufficiently informative to justify this added complexity. When parameter changes induce only subtle behavioural differences, competing models produce similar likelihood surfaces, increasing the risk of misidentification once model complexity is penalised. In longitudinal change models, this problem depends on both the baseline and change regimes (i.e., joint parameter space). Thus, the same magnitude of change may differ in detectability across baseline regimes, and some changes may be more recoverable than others even within the same baseline regime. As a result, pooling estimated changes across heterogeneous participants can be misleading if these regime-dependent constraints are ignored. Second, these constraints also make task design critical. Practical non-identifiability is especially likely when changes of interest are small, as is often the case in children or patients.[13,14] In such cases, improving detection is not only a matter of increasing sample size, but also of increasing the discriminative sensitivity of the behavioural data. Task design optimisation can increase divergence between the likelihood surfaces of competing models.[15,16] A simulation-based framework that maps recovery across the joint parameter space can therefore be used to evaluate how task features influence change detectability and to calibrate tasks so that competing change models generate more distinct behavioural signatures.

Because fit-based identifiability depends on how model fit and parsimony are evaluated, reliable change detection also requires an appropriate model comparison criterion. In other longitudinal research traditions, model comparison often relies on the likelihood ratio test (LRT), particularly when constrained and unconstrained change models are nested within mathematically well-defined model families, such as latent change score models.[9] In computational change modelling, however, candidate models are often derived from distinct cognitive mechanisms and may differ in generative structure, so the competing models are often non-nested. Information criteria may therefore be more suitable for evaluating model parsimony. Even when models are nested, LRT-based inference may still be unreliable because the regularity assumptions of Wilks' theorem can fail near boundaries in joint parameter space, where flat likelihood surfaces and heavier-tailed empirical null

distributions may arise.[17] This can miscalibrate the test and inflate Type I error in favour of spuriously complex models, motivating the need for a dedicated framework for evaluating change detection across model comparison measures.

We use reinforcement learning (RL) as a representative testbed because it is widely used in longitudinal settings and provides a common framework for studying change across development and treatment, although these challenges apply broadly to computational models of longitudinal change in other paradigms. Reward- and loss-based learning are domain-general processes shaped over time by maturation or treatment.[18,19] By quantifying latent RL processes, RL models allow behaviour to be related more selectively to distinct neuromodulatory mechanisms, including dissociable dopaminergic and serotonergic contributions to reward- and loss-related sensitivity.[20] This process-level mapping is useful for longitudinal research because developmental and therapeutic change may affect specific latent processes.[18] Indeed, development of the brain circuits associated with feedback sensitivity and model-based control parallel the age-related changes in these cognitive processes.[18,21-23] Extant evidence also indicates that disrupted learning from reward and loss occurs across several disorders and has been proposed as a target of pharmacological and psychological interventions.[24-30]

Using RL as a testbed, we present a simulation-based framework to evaluate the reliability of longitudinal change detection. We compare two complementary approaches that both estimate change parameters ($\delta\theta$), with second-session (T2) parameters expressed as baseline (T1) parameters plus a change term ($\theta_{T2} = \theta_{T1} + \delta\theta$)[6], but differ in how change is inferred. The first is **fit-based**, selecting among competing change models by assessing model parsimony. The second is **parameter-based**, selecting among models by applying decision thresholds to the estimated change parameters. Using simulations in which the ground truth is known, we evaluate these approaches across effect sizes and sample sizes, focusing on model identifiability, target-change discriminability, and detection power. We then use the same framework to optimise task design to improve sensitivity to the behavioural signatures of change. For interpretability, we develop the framework under maximum likelihood estimation (MLE) before considering how hierarchical or Bayesian approaches may further address between-subject heterogeneity and estimation uncertainty.

2. Results

2.1. Behavioural signatures of longitudinal change in learning

We first asked whether longitudinal changes in two RL parameters produced separable behavioural signatures across sessions in a probabilistic instrumental learning task with block-wise reversals (PILT; Fig. 1A). To isolate the effect of longitudinal change itself, we fixed the baseline parameter regime and task structure across individuals. We manipulated only the session-to-session changes in learning rate ($\delta\alpha$) and reward sensitivity ($\delta\beta$) from small (±0.5) to large (±2.0) increases and decreases, including a no-change (Fig. 1B). This allowed us to examine how recoverability varied across change regimes while minimizing between-subject heterogeneity. Simulated datasets were generated under four candidate models: no-change ($\delta\alpha = 0, \delta\beta = 0$), α-change ($\delta\alpha \neq 0, \delta\beta = 0$), β-change ($\delta\alpha = 0, \delta\beta \neq 0$), and αβ-change ($\delta\alpha \neq 0, \delta\beta \neq 0$).

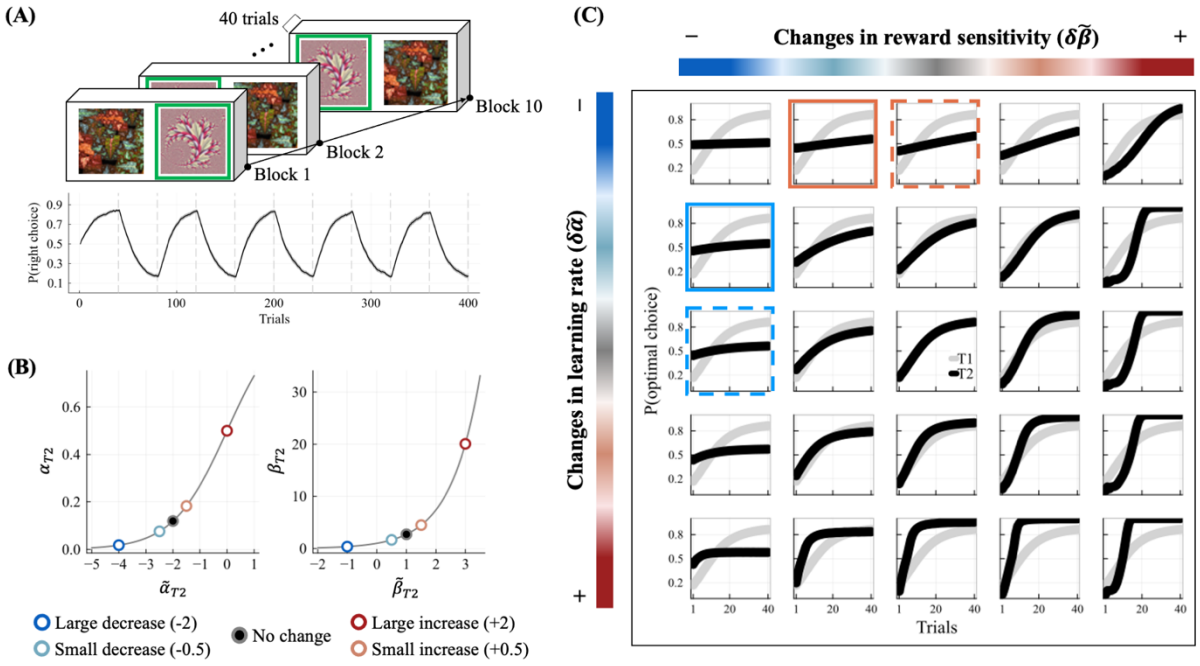

Fig. 1. Reinforcement learning task and longitudinal behavioural changes.

(A) Probabilistic instrumental learning task with block-wise reversals was employed to assess changes in learning. On each trial, individuals chose between options and learned an option with higher expected value (green box). Reward uncertainty was imposed by delivering the better outcome with 70% probability. The choice-outcome contingency was switched across blocks. The lower panel shows the trial-by-trial probability of choosing right, highlighting the block-wise reversals in the higher-valued option. Dashed vertical lines delineate block boundaries. (B) Behaviour was generated using a RL model in which learning rate (α) and reward sensitivity (β) explain choice. We changed the two parameters across sessions by adding values ($\delta\tilde{\alpha}$ and $\delta\tilde{\beta}$) from small (light blue/red) to large (dark blue/red) decreases or increases, with black markers indicating the baseline no-change regime. The two panels illustrate how simulated parameters at second session ($\tilde{\alpha}_{T2}$ and $\tilde{\beta}_{T2}$) in the generative space map onto the internal parameterization in the model after logistic (α_{T2}) and exponential (β_{T2}) transformations. (C) Learning curves of optimal choice probability for each change regime, averaged across blocks. Curves were shown separately for before (gray) and after change (black). Columns and rows correspond to changes in reward sensitivity (δβ) and learning rate (δα), respectively, with color indicating the direction and magnitude of change as in panel (B). Solid boxes indicate the representative regimes where the

behavioural signature of either single-change model (dashed boxes) was difficult to distinguish from those of the generative model.

The two parameters produced distinct effects on behaviour (Fig. 1C). Increasing $\delta\alpha$ (learning rate) accelerated updating and steepened the early learning slope, whereas increasing $\delta\beta$ (reward sensitivity) sharpened choice determinism and increased asymptotic performance. Conversely, decreasing each parameter reduced learning speed and asymptote, respectively. Across most change regimes, joint changes in both parameters preserved these functional signatures in the learning curves. However, this decomposition failed under some regimes where changes in one parameter alone already constrained behaviour, masking the influence of additional changes in the other parameter. For example, when the learning rate decreased substantially, learning curves failed to reach asymptote within the fixed number of trials, making additional changes in reward sensitivity difficult to detect (Fig. 1C; orange boxes). Conversely, when reward sensitivity decreased substantially, choice behaviour became near-flat due to a low asymptote, obscuring differences in early learning slope induced by changes in learning rate (Fig. 1C; blue boxes). Thus, dual-parameter changes did not necessarily produce behaviour that was distinguishable from single-change alternatives when one parameter dominates the observable dynamics.

2.2. Qualitative comparison of fit- and parameter-based change detection

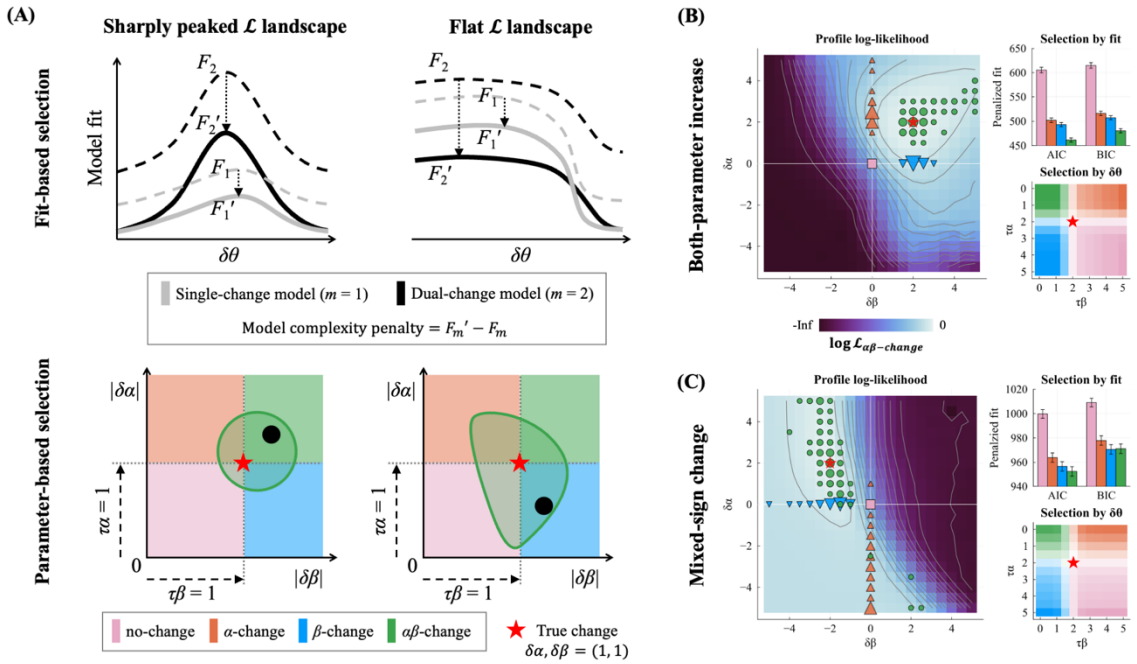

Fig. 2. Decision principles of fit-based and parameter-based model identifications.

(A) *Schematic of model selection principles.* ***Top:*** *Fit-based selection evaluates models by their penalized model fit. Likelihood landscapes for dual-change ($m = 2$, black) and single-change ($m = 1$, grey) models are shown as a function of one change parameter ($\delta\theta$), with the other fixed for the dual-change model. For each model m, penalized fit (F_m') is derived by applying a complexity penalty to the maximum likelihood (F_m), and the model with the higher F_m' is selected. In sharply peaked landscapes (Left), the fit improvement outweighs the penalty, favoring the complex model. In flat landscapes (Right), the improvement is insufficient, favoring the simpler model.* ***Bottom:*** *Parameter-based selection relies on whether estimated magnitudes of $\delta\alpha$ and $\delta\beta$ exceed thresholds (τ_α, τ_β). For a true change at $\delta\alpha, \delta\beta = (1, 1)$ (red star), repeated fits by the generative model produce a distribution of estimates (green-outlined region; example as a black*

dot). Model selection is determined by the quadrant in which estimates fall, defined by whether |δα| and |δβ| exceed their respective thresholds (dashed arrows). Decision regions corresponding to different change models are color-coded: no-change (pink), α-change (orange), β-change (blue), and αβ-change (green). Sharp landscapes yield estimates mainly within the quadrant of the generative model (Left), whereas flat landscapes lead to broad distributions across multiple decision regions, increasing misclassification (Right). ***(B, C)*** *Profile likelihood and model selection results under different change regimes:* ***(B)*** *Both-parameter increase and* ***(C)*** *mixed-sign change.* ***Left:*** *Two-dimensional profile log-likelihood surface over the change-parameter space (δα, δβ), computed under the αβ-change model using data generated from the true parameters (red star), with baseline parameters optimised at each (δα, δβ) point. Overlaid markers indicate parameter estimates obtained by fitting each candidate model to the same dataset. Colours correspond to model identity as in (A), and marker size is proportional to the number of subjects with each estimate.* ***Top right:*** *Model comparison results based on fit-based criteria (AIC, BIC) across candidate change models.* ***Bottom right:*** *Parameter-based model selection across the two-dimensional threshold space (τ_α, τ_β). Colors indicate the most frequently selected model, and color intensity reflects its selection proportion.*

Before quantifying change detection performance, we first illustrate the decision principles of the fit-based and parameter-based approaches schematically (Fig. 2A). The two identification procedures differed fundamentally. Fit-based identification selected the model with the highest penalized fit, depending on whether a more complex model improved fit enough to justify its penalty. By contrast, parameter-based identification selected models according to the location of estimated change parameters relative to predefined thresholds. As illustrated in Fig. 2A, flatter landscapes can lead to different forms of misclassification.

These different principles became clear in representative change regimes (Fig. 2B, C). In a regime with increases in both parameters, the profile likelihood surface was sharply peaked around the true change parameters. The maxima of the simpler models were clearly separated from that of the generative αβ-change model (Fig. 2B, left). Accordingly, fit-based model comparison correctly selected the generative model even under the conservative BIC penalty (Fig. 2B, top right). Parameter-based selection was also robust; estimated changes fell predominantly within the αβ-change decision region across a wide range of thresholds τ_α and τ_β up to 1 (Fig. 2B, bottom right).

By contrast, in a mixed-sign change regime, the profile likelihood formed an elongated ridge reflecting a trade-off between parameters. This ridge spanned the estimates of the generative and simpler models (Fig. 2C, left). Hence, the improvement in fit from the αβ-change model was small relative to the complexity penalty, so fit-based selection under the more conservative criterion favored a simpler β-change model (Fig. 2C, top right). At the same time, fitted parameter estimates were broadly dispersed across the change space, making parameter-based selection highly sensitive to threshold choice and leading to rapid deterioration in recovery over the same threshold range (Fig. 2C, bottom right).

These examples demonstrate that likelihood geometry strongly shapes change detection. While sharply peaked surfaces supported both identification procedures, flatter or ridge-shaped surfaces increased misidentification for different reasons in the two approaches. The fit-based approach depended strongly on the stringency of the chosen penalty, whereas the parameter-based approach was undermined by both variability in fitted estimates and ambiguity in threshold selection.

2.3. Fit-based change detection across change regimes

We then quantified how reliably fit-based criteria detected the true change model across the change space. Using simulations with 30 individuals and 120 observations per individual, we evaluated fit-based detection in two complementary ways: recovery of the true generative model at the intrinsic operating point defined by each criterion, and threshold-free discriminability of a target change against alternatives using ROC analysis (Fig. 3).

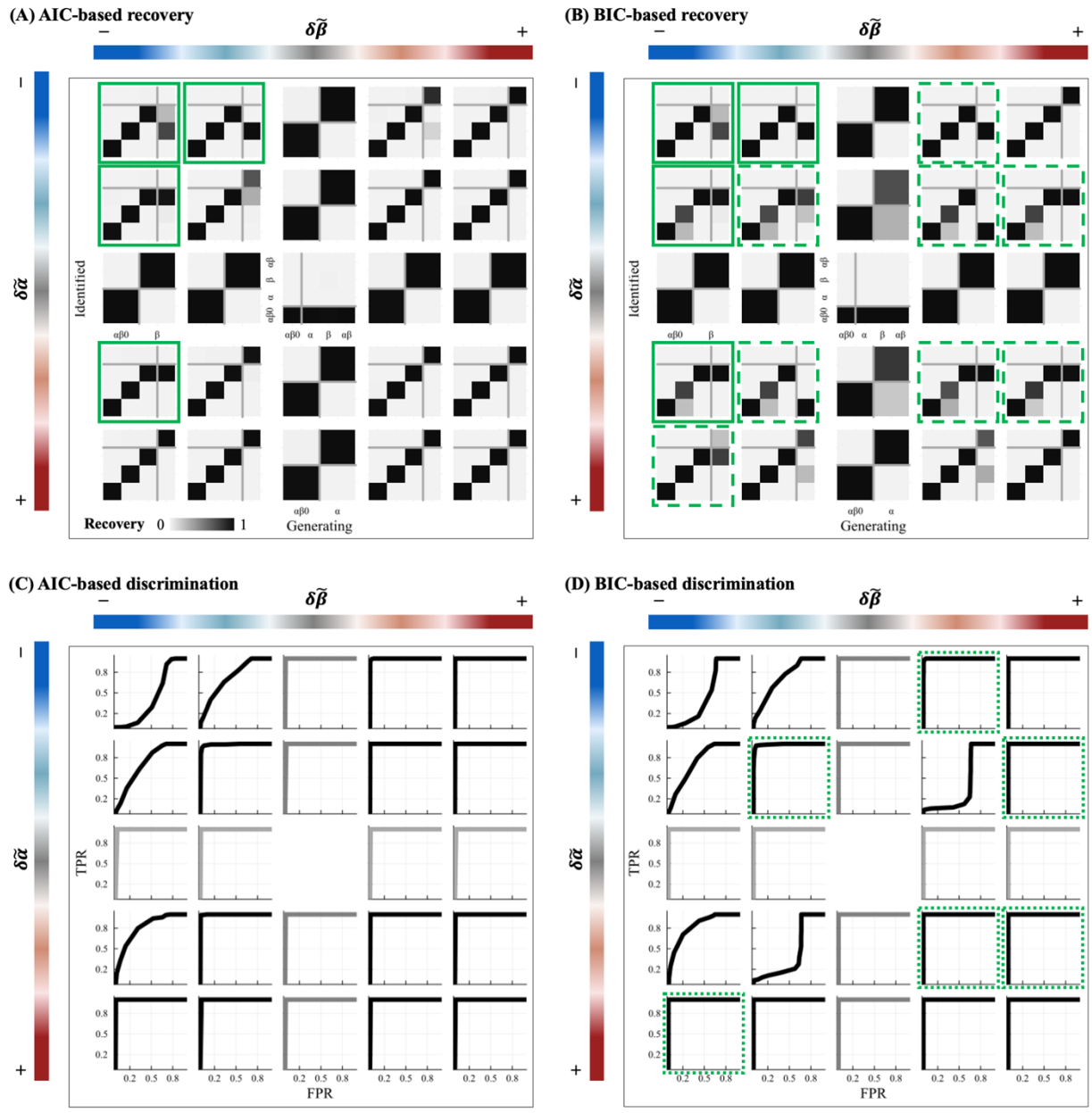

Fig. 3. Recovery and discriminability of change models using fit-based decision statistics.

*Target change detection was evaluated in two complementary analyses: identification of the true generative model at the intrinsically defined single decision threshold **(A, B)**, and discrimination of the generative model from alternative candidates in a threshold-free manner using ROC analysis **(C, D)**. Rows in each panel index δα changes, progressing from large decrease (top) to large increase (bottom). Columns index δβ changes, progressing from large decrease (left) to large increase (right). Top and left colour scales indicate change direction and magnitude as defined in Fig. 1B. Across change regimes, model recovery was evaluated using AIC **(A)** and BIC **(B)**. In each confusion matrix, columns denote the generative models while rows indicate the identified models. For single-change and no-change generative models shown within dual-change regimes, each dataset was generated by constraining the dual-change regime onto its respective model-specific subspace. Change model(s) shown to the right of the grey vertical line in each confusion matrix are treated as a positive class for target change detection. Solid green boxes indicate the regimes*

where pronounced misidentification occurs in AIC-based identification while dashed green boxes indicate significant performance declines in the BIC-based approach. Over the same change space, ROC curves were constructed using AIC ***(C)*** *and BIC* ***(D)****. ROC curves summarize the trade-offs between sensitivity (y-axis) and specificity (x-axis) for detecting the target change. ROC curves were presented in black for αβ-change detection and grey for single-change detection. Dotted green boxes indicate regimes where threshold-free discriminability remained perfect (AUC = 1.0) despite misidentification in the corresponding confusion matrix.*

2.3.1. Model recovery based on model parsimony at the intrinsic decision threshold

Model recovery depended strongly on the magnitude and direction of change, particularly when both parameters changed. Figure 3A and B summarizes fit-based model recovery across the full change space. Each subplot shows a confusion matrix for a specific combination of true $\delta\alpha$ and $\delta\beta$, with perfect recovery appearing as a diagonal pattern. When data were generated with $\delta = 0$, both criteria correctly identified the no-change model. No-change and single-change models were recovered reliably across regimes and criteria, with confusion matrices remaining close to diagonal along the central axes of change space. By contrast, recovery of the $\alpha\beta$-change model varied markedly across the off-axis regimes.

Under AIC-based model selection, the $\alpha\beta$-change model was recovered broadly across regimes involving increases in one or both parameters (Fig. 3A). Misidentification emerged when both change parameters decreased. Specifically, the dual-change model was typically selected as whichever single-change model contained the parameter with larger decrease.

Under BIC-based model selection, the more conservative penalty reduced sensitivity across a broader range of change regimes than AIC-based selection (Fig. 3B). Misidentification occurred not only in both-decrease regimes, but also in small both-increase and mixed-sign change regimes. As under AIC-based selection, these failures typically involved the dual-change model being reduced to the single-change model containing the parameter with larger change. Also, small changes in both parameters sometimes led to misidentification as the no-change model.

Together, these results show that fit-based model identification was constrained not only by effect size, but also by regime-dependent recoverability (i.e., practical non-identifiability). Dual-change models were particularly vulnerable in regimes where behavioural signatures overlapped with those of simpler alternatives.

2.3.2. Threshold-free discriminability under fit-based selection

We next asked whether these failures reflected poor separability between models or merely conservative operating points. ROC curves were constructed using ΔIC, the difference in fit criterion between the target change model and the best-fitting alternative, as the decision statistic. Single-change detection was perfectly discriminable across change regimes and fit criteria (AUC = 1.0), consistent with the robust recovery of the single-change models. In contrast, discriminability of dual-change depended strongly on change regimes (Fig. 3C and D).

Under AIC (Fig. 3C), regimes in which the $\alpha\beta$-change model was correctly recovered also showed perfect discriminability (AUC = 1.0), whereas both-decrease regimes exhibited declined ROC performance, particularly when at least one parameter decreased largely (AUC = 0.4~0.8). Under BIC (Fig. 3D), discriminability was similarly high across much of the change regimes, but dropped sharply under mixed-sign small change regimes where ROC curves showed low sensitivity even at moderate specificity (AUC ≈ 0.4).

A clear dissociation emerged under the BIC-based approach. In many regimes where the αβ-change model was not recovered at the intrinsic threshold (Fig. 3B), ROC performance nevertheless remained perfect (AUC = 1.0; Fig. 3D, dotted green boxes). BIC misidentification reflected conservative thresholding rather than poor separability between the dual-change and alternative models. Thus, fit-based recovery failures could arise from an overly conservative criterion that sacrificed sensitivity to preserve parsimony.

2.4. Parameter-based change detection across change regimes

We next evaluated the parameter-based approach. Unlike fit-based model comparison, parameter-based detection does not provide a unique intrinsic operating point because model selection depends on externally imposed thresholds on estimated change parameters, with no naturally defined criterion for choosing them. For αβ-change detection, this decision rule depends jointly on τ_α and τ_β, such that discriminability is characterized over a two-dimensional threshold space using ROC surfaces. We therefore assessed parameter-based detection in terms of threshold-free detectability and then examined model recovery across representative ranges of threshold space (Fig. 4).

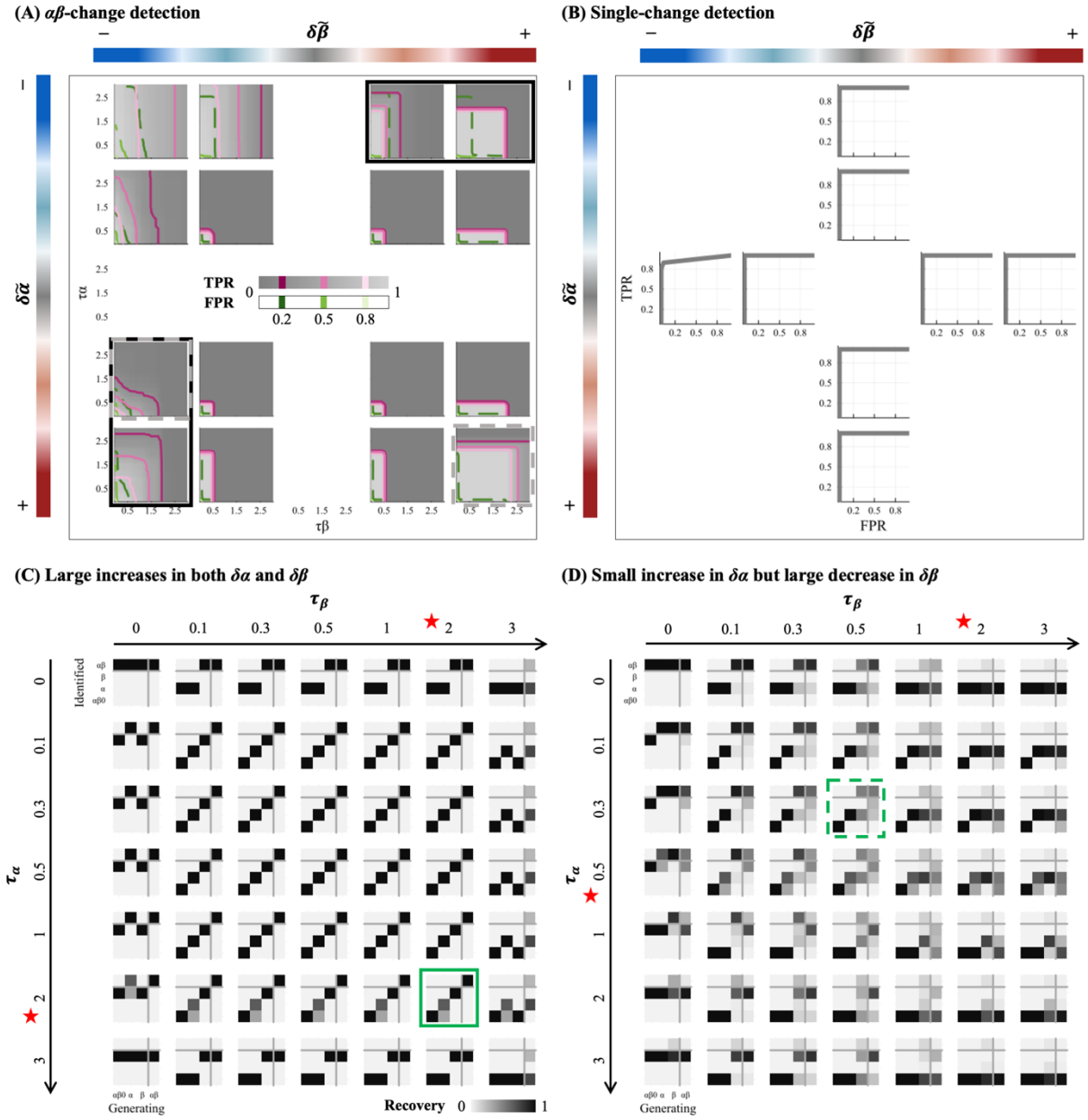

Fig. 4. Model recovery and detectability of change models using the parameter-based model selection.

(A, B) *Threshold-free detectability of target change in the parameter-based approach was evaluated using ROC surfaces for dual-change detection over a two-dimensional threshold space and ROC curves under single-change regimes. Top and left colour scales indicate change direction*

and magnitude of the parameters as defined in Fig. 1B. ***(A)*** *In the ROC surfaces, TPR for αβ-change detection was visualized as heatmap intensity, with brighter grey indicating higher values. Contour lines were overlaid for TPR and FPR values of 0.8 (light pink/green), 0.5 (medium pink/green), and 0.2 (dark pink/green). Two representative regimes, highlighted with dashed grey boxes, were selected to examine parameter-based model recovery: one showing good discriminability (TPR ≥ 0.5 along the low FPR contour) and the other showing poor discriminability (TPR < 0.5 at some points along that contour). Solid black boxes indicate the regimes showing poor discriminability.* ***(B)*** *Over the single-change regimes, ROC curves were constructed by thresholding one of the change parameters. Parameter-based model recovery under the representative regimes:* ***(C)*** *one with large increases in both parameters, and* ***(D)*** *the other one with small δα increase alongside large δβ decrease. Rows in each panel index thresholds to |δα| (τ_α), progressing from 0 (top) to 3 (bottom). Columns index thresholds to |δβ| (τ_β), progressing from 0 (left) to 3 (right). In each confusion matrix, columns index the generative models while rows index the identified models. Identification performance was evaluated exploratorily over two-dimensional threshold pairs (τ_α, τ_β). For each regime, green boxes indicate the largest threshold pair at which the generative αβ model remained predominantly recovered; a solid-line marks this pair where the alternative models were also correctly recovered, whereas a dashed-line marks this pair where recovery of alternative models failed. Red stars indicate the true change parameters.*

2.4.1. Threshold-free discriminability under parameter-based detection

Parameter-based discriminability of the αβ-change model varied strongly across the threshold space (Fig. 4A). ROC surfaces summarized how TPRs and FPRs varied across pairs of decision thresholds. Good discriminability was observed across much of the change regimes, but mixed-sign regimes involving a large decrease in one parameter showed poor performance: even threshold pairs achieving low FPR (0.2) often failed to yield TPR at the intermediate level (0.5). In these regimes, parameter-based dual-change detection could appear favourable when thresholds happened to align with the true change magnitudes; however, these are unknown in empirical settings.

For single-change detection, ROC curves were constructed using a one-dimensional threshold on either change parameter. Unlike dual-change detection, this setting did not require resolving decision ambiguity across a two-dimensional threshold space. Both α- and β-change models exhibited high discriminability against no-change across change regimes, with AUCs exceeding 0.9 (Fig. 4B).

2.4.2. Model recovery based on parameter estimates across the two-dimensional threshold space

Parameter-based recovery was evaluated at each single threshold pair. Single-change recovery against no-change was reliable for all single-change regimes (Fig. S1). For dual-change detection, we asked whether the threshold pairs that showed good discriminability in the ROC surfaces also supported correct recovery of the change models at the two representative regimes. In the regime with large increases in both parameters, all change models were correctly recovered across a broad range of threshold space ($0 < \tau_\alpha \leq 2$ and $0 < \tau_\beta \leq 2$; Fig. 4C), consistent with its good discriminability (Fig. 4A, right dashed box).

By contrast, in the regime with a small increase in δα but a large decrease in δβ (Fig. 4A, left dashed box), recovery of alternative models failed even at the largest threshold pair at which the generative model remained predominantly recovered ($(\tau_\alpha, \tau_\beta) = (0.3, 0.5)$; Fig. 4D). This failure arose mainly because data generated from the β-change were frequently misidentified as α-change. Rather than remaining close to zero, δα estimates often exceeded non-zero thresholds, indicating that part of the δβ signal was absorbed into the δα estimate.

Thus, parameter-based recovery was limited not only by threshold dependence but also by cross-parameter spillover. Even when local separability appeared acceptable in ROC surfaces, structural trade-offs between parameters could still undermine correct model attribution.

2.5. Detection sensitivity for single-parameter changes

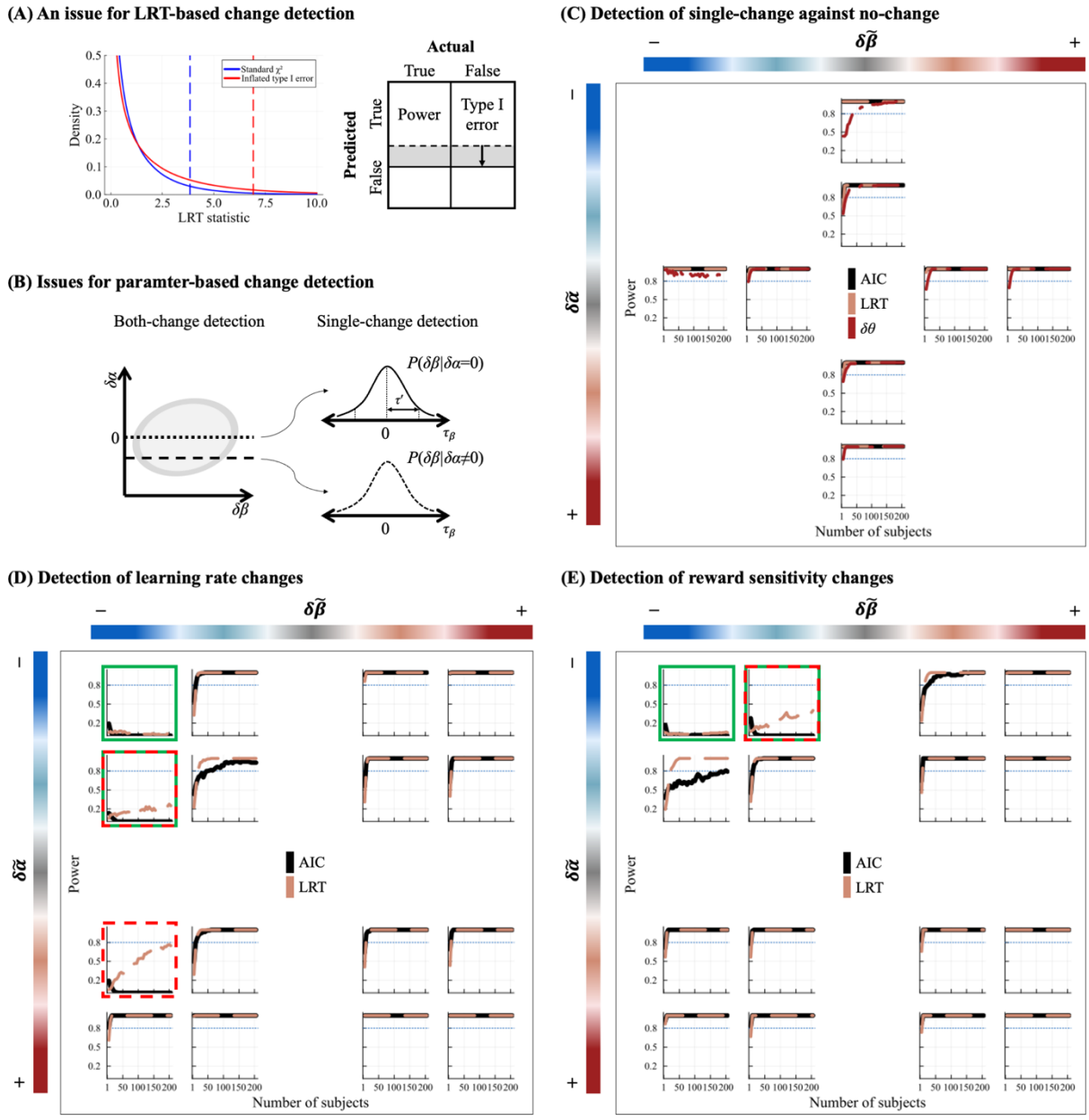

Fig. 5. Power and limitations of single-parameter change detection.

Two limitations of LRT- and parameter-based approaches are illustrated. ***(A) Left:*** *Distributions of the LRT statistic for a nested comparison differing by one parameter (df = 1), under the standard asymptotic χ^2 null (blue) or a mis calibrated distribution (red). Dashed vertical lines indicate the p = 0.05 critical values.* ***Right:*** *Distortion of statistical power under type I error inflation.* ***(B)*** *For the parameter-based approach, FPR-based thresholding is only practically definable for single-change detection against a no-change null because the decision threshold (τ′) can be calibrated from the empirical null distribution under the no-change model (here, FPR = 0.05; upper right). When the competing model also involves change in another parameter, the null distribution is no longer anchored to a no-change baseline, making such threshold calibration impractical for empirical studies (lower right).* ***(C-E)*** *Power curves are shown as a function of the number of subjects (x-axis) for each change regime. Top and left colour scales indicate change direction and*

magnitude of the parameters as defined in Fig. 1B. Single-parameter change detection was evaluated via pairwise classifications: single-change vs. no-change, dual-change vs. β-change for detecting α-change, and dual-change vs. α-change for detecting β-change. ***(C)*** *In single-change regimes, power is shown for three different methods (AIC-, LRT-, and the parameter-based).* ***(D, E)*** *In dual-change regimes, the AIC- and LRT-based power are shown. The target power level of 0.8 is indicated by the blue dotted horizontal line. Green boxes indicate the regimes where change detection power did not reach the target in the both approaches, while red boxes indicate the regimes where LRT-based detection was more sensitive than underpowered AIC-based detection.*

We next examined how sample size affected the power to detect single-parameter change. Change detection power was defined as the probability of correctly detecting a target parameter change when data were generated from a corresponding change model. We considered two scenarios: 1) detecting an α- or β-change against no-change, and 2) detecting a target single-change within data generated by the αβ-change model. Because these comparisons remained nested, LRT could be used as a conventional benchmark alongside AIC-based selection. Because LRT relies on regularity assumptions that may fail in regimes with low Fisher information, its power should be interpreted cautiously in flatter likelihood landscapes (Fig. 5A). For the fit-based approach, we report AIC-based power because this criterion was more sensitive than BIC in the preceding analyses. For parameter-based detection, a practically meaningful false-positive threshold could be defined only against a no-change null (Fig. 5B), so parameter-based power was evaluated only for single-change versus no-change.

Single-change detection against no-change was generally robust across methods (Fig. 5C). For both α- and β-change from T1 to T2, power typically reached the target value of 0.8 by $N = 30$. The only exception was the parameter-based approach in the regime with a large $\delta\alpha$ decrease, which required a slightly larger sample ($N = 35$).

By contrast, detecting a target single-change within dual-change regimes depended strongly on the accompanying change in the non-target parameter. For α-change detection, both AIC and LRT remained underpowered in both-decrease regimes involving a large δβ decrease (Fig. 5D, green boxes). Symmetrically, β-change detection by both methods was underpowered when accompanied by a large δα decrease (Fig. 5E, green boxes). Thus, even when the target parameter changed, strong change in a non-target parameter could mask the behavioural signature of the target parameter.

AIC- and LRT-based power were broadly similar across most regimes but often diverged when the change in the target parameter was less than that in the non-target parameter. For example, when δα-change was small in either direction and δβ decreased a large amount, LRT reached the target power at the largest sample size whereas AIC did not (Fig. 5D, red boxes). A similar pattern was observed for small β-decrease detection when δα decreased a large amount (Fig. 5E, red box). Thus, even for single-change detection, power depended not only on the size of the target change but also on its interaction with concurrent change in the other parameter.

2.6. Detection sensitivity for dual-parameter change

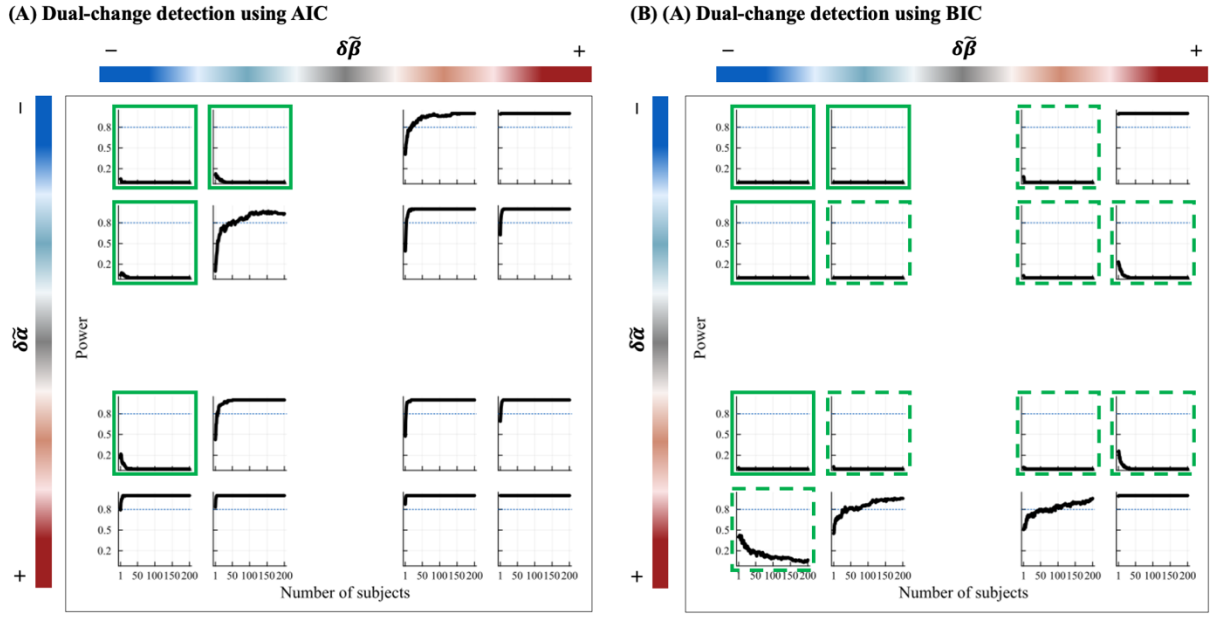

Fig. 6. Power to detect the generative change using the fit-based decision statistics.

*Power curves drawn using AIC-based **(A)** and BIC-based **(B)** decision statistics are presented. Top and left colour scales indicate change direction and magnitude of the parameters as defined in Fig. 1B. Green boxes indicate the regimes where pronounced underpower occurs in AIC-based power curves (solid-line) or significant performance declines in BIC-based power curves (dashed-line).*

We then evaluated power to detect dual-parameter change using a four-way comparison among all change models. Power was defined as the probability of correctly recovering the generative model under *αβ*-change regimes. Because this setting involved simultaneous comparison among multi-class and often non-nested alternatives, dual-change detection was assessed exclusively using fit-based criteria.

Detection power for the *αβ*-change model closely mirrored the recovery results (Fig. 3A, B). Regimes in which the αβ-change model was misidentified also exhibited AIC- or BIC-based power curves that remained near zero across the tested sample range (Fig. 6, green boxes). Thus, increasing sample size alone was insufficient to improve detection. Detection power was constrained not only by effect size or sample size but also by practical non-identifiability in specific change regimes.

Across regimes, both AIC and BIC showed high power in many both-increase regimes, with performance improving as effect size and sample size increased. However, BIC was consistently more conservative, yielding lower sensitivity and a broader set of underpowered regimes than AIC (Fig. 6). These results indicate that the detectability of dual-change depends heavily on where the expected developmental or intervention-related change lies in the change space and on the model comparison criterion used to evaluate it.

2.7. Improving change detection through task design optimisation

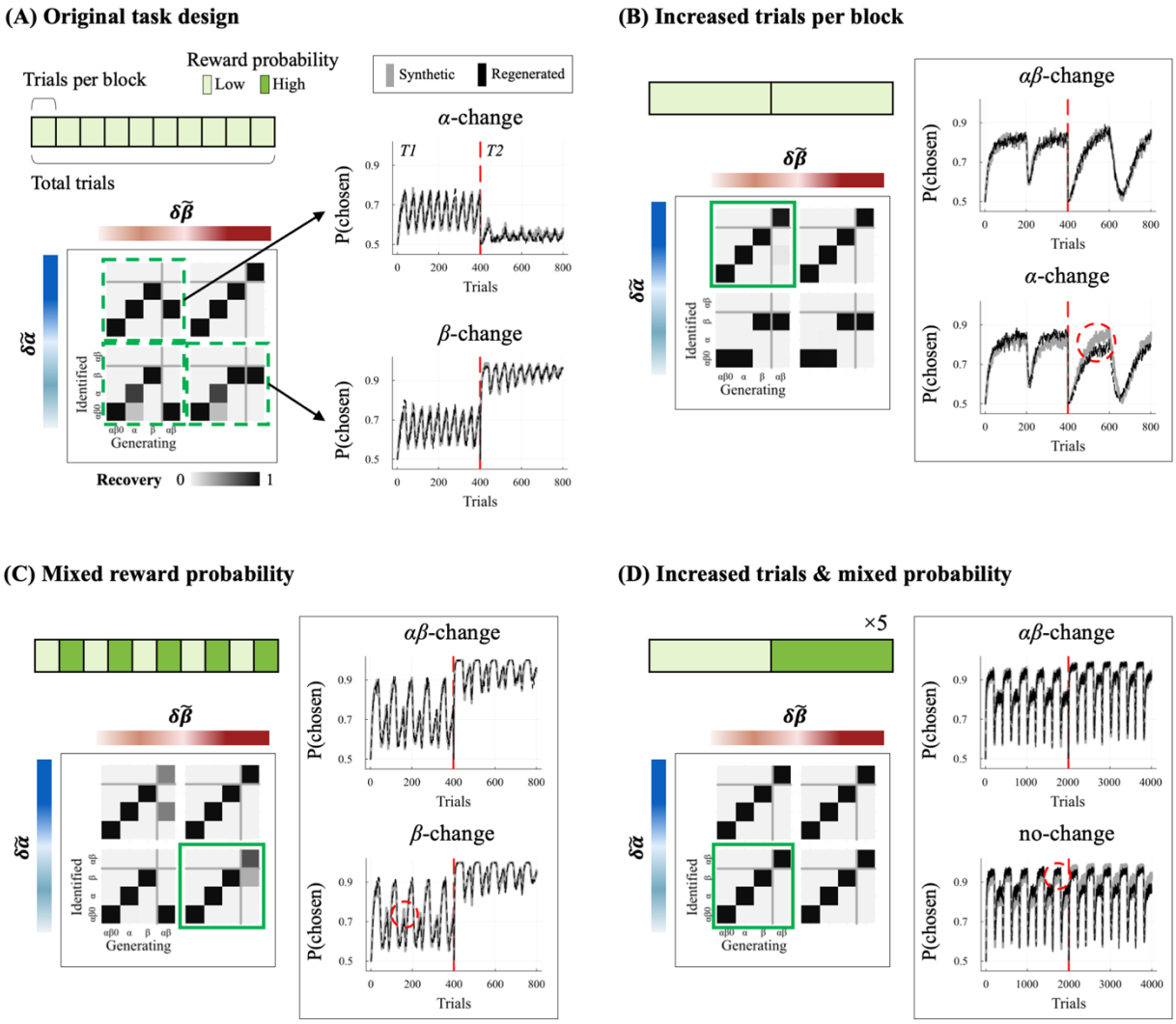

Fig. 7. Improving model identifiability through task design modification.

Four task designs and their effects on BIC-based model recovery are shown for representative regimes with decreased δα and increased δβ. Top and left colour scales indicate change direction and magnitude of the parameters as defined in Fig. 1B. In each panel, the upper left schematic illustrates the task design, including the number of trials per block, the total number of trials, and reward probability across blocks (light/dark green indicates low/high reward probability, respectively). The lower left plot shows model recovery under corresponding task design. The right plots compare synthetic behaviour from the true generative model (grey solid-line) with behaviour regenerated from fitted models (black dashed-line), shown as the probability of the chosen action across trials. The red vertical line marks the boundary between two sessions. ***(A)*** *Original task design. The original design consisted of 40 trials per block with a fixed low reward probability. The highlighted regimes (dashed boxes) indicate cases in which data generated from the dual-change model were misidentified under the original design. The right plots show that behaviour regenerated from the incorrectly selected single-change models closely matched the synthetic behaviour.* ***(B-D)*** *Modified task designs. Task design was modified by increasing the number of trials per block while holding the total number of trials constant* ***(B)****, alternating reward probability across blocks* ***(C)****, or combining both modifications with more total trials* ***(D)****. Solid boxes indicate*

regimes that were misidentified under the original design but correctly recovered after each task modification. Red circles indicate behavioural signatures that became more distinguishable under the modified design.

Finally, we examined whether practical non-identifiability could be mitigated by modifying task design rather than increasing sample size. We focused on three mixed-sign regimes where the generative $\alpha\beta$-change model was consistently misidentified under the original task design (40 trials per block, fixed 70% reward probability; Fig. 7A).

Under the original design, regenerated behaviour from the incorrectly selected single-change models closely matched behaviour generated by the true dual-change model (Fig. 7A). In the regime with a large decrease in $\delta\alpha$ but a small increase in δβ, the model was misidentified as α-change because the short block length did not allow asymptotic differences associated with β to emerge reliably. Conversely, in the regime with a small decrease in δα but a large increase in δβ, the model was misidentified as β-change because the stable reward environment allowed differences in learning speed to be absorbed by changes in reward sensitivity.

We therefore tailored task modifications to the specific behavioural signature that was being masked. Increasing the number of trials per block (40 to 200) selectively improved recovery when the missing signal was asymptotic performance driven by β-change (Fig. 7B). Alternating reward probabilities across blocks between 70% and 85% improved recovery when the missing signal was the early learning slope driven by α-change (Fig. 7C). Combining both modifications enhanced the recovery when both limitations co-occurred (Fig. 7D).

Thus, poor change detection in these regimes did not reflect an unavoidable lack of information but a mismatch between task design and the behavioural signatures needed to distinguish competing change models. Simulation-based task optimisation could therefore improve detection sensitivity without increasing sample size.

3. Discussion

Reliable longitudinal change detection in computational models requires more than adding change parameters to a joint-session model. In the present study, detectability depended jointly on the size and direction of change, the model comparison criterion, and the task design used to express behavioural signatures of change. Using RL as a representative testbed, we found that fit-based and parameter-based approaches behaved differently across the same change space. Fit-based detection asked whether behavioural data justified a more complex change model, whereas parameter-based detection relied on whether estimated changes crossed externally imposed thresholds. Across both approaches, change detection was strongly regime-dependent, and increasing sample size alone was often insufficient when models were practically non-identifiable. Simulation-based task redesign improved detectability in regimes where the original task failed to separate competing change models. Together, these findings suggest that longitudinal change modelling should be treated as a detection problem to be validated a priori.

A central contribution of the present work is to distinguish between two conceptually different ways of inferring longitudinal change. Although both approaches begin with the same fitted change parameters, they differ in what is treated as evidence. Fit-based detection asks whether a target change model provides a better account of behaviour than simpler or alternative models after accounting for model complexity. Parameter-based detection instead asks whether estimated changes fall into threshold-defined regions of change-parameter space. Our analyses of likelihood geometry clarified why these approaches can diverge. When likelihood landscapes were sharply peaked, both procedures supported the generative model. When likelihoods were flat or ridge-shaped, however, they failed for different reasons. Fit-based detection became sensitive to the strictness of the complexity penalty, whereas parameter-based detection became unstable because estimates spread broadly across change space and classification depended on threshold choice. Therefore, recovering a parameter value and identifying the correct change mechanism are not the same inferential goal.[31,32]

This distinction has direct implications for model validation. Fit-based identifiability provides the more practical foundation for change inference because it asks whether behavioural data support a target change model over plausible alternatives. By contrast, parameter-based model selection requires externally imposed thresholds that are not naturally defined by the framework itself. This difficulty was especially clear for dual-change detection. Although ROC surfaces could characterize threshold-free discriminability, practical classification still required selecting threshold pairs without a principled criterion. Furthermore, threshold regions that appeared favourable did not always guarantee correct model recovery due to cross-parameter spillover. Even in the single-change case, where thresholding was more tractable and discriminability was high, practical decisions still depended on how thresholds were set relative to the null. For this reason, parameter-based summaries may be useful as descriptive supplements, but they are less suitable than fit-based model comparison as the primary tool for validating longitudinal change.

The present framework also shows that reliable change detection depends on how model parsimony is evaluated. AIC and BIC embody different assumptions about how complexity should be penalised. In our simulations, AIC and BIC often agreed in easy regimes but diverged in weakly discriminable parts of the change space. AIC was more sensitive, whereas BIC more often collapsed dual-change models onto simpler alternatives or, in the weakest regimes, onto no-change. Importantly, ROC analyses showed that these failures were not always due to poor separability. In several regimes, BIC failed to recover the generative model even when threshold-free discriminability remained perfect, indicating that the criterion was overly conservative. These findings show that criterion choice is an integral part of longitudinal change inference.

These results also clarify the place of the LRT. It remains a useful benchmark for nested comparisons, such as pairwise detection of single-parameter change, but it is not a general solution for computational change modelling. Many comparisons of interest are non-nested or multiclass, and even in nested settings LRT can be unreliable when regularity conditions fail in flat likelihood landscapes. Accordingly, LRT is informative in restricted settings but less suitable as a general framework for evaluating change detectability across computational change models.

One of the clearest findings across our analyses was that change detection is intrinsically regime-dependent. The same magnitude of change was not equally detectable everywhere in the joint parameter space.

In particular, dual-change models became difficult to distinguish from single-change models when one parameter dominated the observable behavioural dynamics and effectively absorbed the contribution of the other. In our RL example, this occurred especially in both-decrease and mixed-sign regimes, where one form of change compressed behaviour enough to obscure the other. This pattern illustrates practical non-identifiability: the limitation is not that the model cannot express change, but that the available behavioural data under a given design are insufficiently informative to distinguish the intended change mechanism. An important implication is that estimated change should not be pooled across participants as if all individuals were equally informative, because identical true changes may differ in detectability across regions of the joint baseline-change space.

A major practical implication of the present work is that poor detectability need not be accepted as fixed. In several regimes, increasing sample size alone did not rescue dual-change detection because the task itself failed to express the behavioural signatures needed to separate competing models. By contrast, targeted task modifications improved identifiability at the same sample size by amplifying the specific signature that had been masked. For instance, increasing the number of trials per block enhanced detectability of reward sensitivity change by allowing asymptotic differences to emerge, whereas alternating reward contingencies across blocks enhanced detectability of learning rate change by increasing sensitivity to learning dynamics. This shifts the problem of low power away from sample size alone and toward a broader notion of design sensitivity. In longitudinal studies, especially in developmental and treatment settings where recruitment and follow-up are costly, task optimisation may therefore be more realistic and efficient than relying solely on larger samples. Computational models are especially well suited to this kind of simulation-based optimisation because they provide a mechanistic account of how task features interact with latent processes to generate behaviour.

These issues are especially relevant in the developmental and treatment settings that motivate computational change modelling. Developmental trajectories are not always monotonic: some cognitive processes improve during childhood and adolescence, then stabilise or decline, yielding mixed-sign or even joint-decrease patterns across interacting latent processes.[18,33] Clinical change may be similarly heterogeneous. During illness progression, multiple processes may deteriorate together; during treatment, some processes may improve while others remain impaired or worsen.[26,27,34] Such regimes are where latent change can become hardest to detect. This makes a priori validation particularly valuable for developmental and treatment studies, which often involve recruitment and testing challenges, attrition over follow-up, and substantial costs per participant. In these settings, a simulation-based framework can be used not only to estimate the sample size needed to reach target power, but also to test whether the intended task and change model are capable of detecting the kinds of latent change that are theoretically or clinically expected.

Several limitations should be noted. First, the present framework was developed using a simple two-parameter RL model under MLE. Although this simplicity helped isolate the principles governing change detectability, more complex models may show stronger parameter trade-offs and richer forms of non-identifiability. Second, we fixed the baseline regime to isolate the role of change itself. In empirical settings, baseline heterogeneity and estimation uncertainty will also affect detectability. Hierarchical or Bayesian approaches may help model these sources of variability, but they do not remove the need to validate whether a task can distinguish the targeted change mechanisms. Third, the present framework focused on two-session designs. Extending it to three or more time points will require careful choices about how change is parameterised and whether the added flexibility is itself identifiable.

In summary, reliable longitudinal change detection in computational models cannot be assumed from model structure alone. Detectability depends on whether change produces distinct behavioural signatures, whether those signatures are recoverable under an appropriate model comparison criterion, and whether the task is designed to express them clearly enough for the expected effect sizes and sample sizes. By framing computational change modelling as a problem of identifiability, discriminability, and design sensitivity, the present work provides a practical route for validating change models before they are deployed in developmental or treatment studies. Rather than asking only whether a model can represent change in principle, researchers should also ask whether that change can be detected reliably under the conditions of the study they intend to run.

4. Materials and Methods

4.1. Behavioural task

The PILT assessed how individuals learned action values from stochastic reward and loss outcomes (Fig. 1A). On each trial, individuals chose between two options to reveal an associated outcome. Outcomes were reward (+1) or loss (-1), delivered probabilistically. The higher valued option was rewarded with 70% probability, and the optimal choice reversed across blocks. The task comprised 10 blocks of 40 trials each. To capture within-subject behavioural change in a minimal longitudinal design, individuals completed the same task in two sessions, allowing session-to-session shifts in latent processes to be estimated.

4.2. Reinforcement learning model

The Rescorla-Wagner (RW) model was used to parameterize choices with α and β. Within each session $S \in \{T1, T2\}$, the action value Q_t of a chosen action a_t was updated on each trial t based on a prediction error (PE), defined as the difference between the $\tilde{\beta}$-scaled reward and the current expected value:

$$PE_t = \beta_S r_t - Q_t(a_t),$$
$$Q_{t+1}(a_t) = Q_t(a_t) + \alpha_S PE_t.$$

Here, $r_t \in \{+1, -1\}$ denotes the obtained reward. Action values were initialized at zero at the beginning of each session. Choices were generated using a softmax policy over current action values,

$$P(a_t) = \frac{\exp\big(Q(a_t)\big)}{\sum_a \exp\big(Q(a)\big)}.$$

For numerical stability, we optimised unconstrained parameters ($\tilde{\alpha}$ and $\tilde{\beta}$) after transforming them to the native space via a sigmoid and an exponential transform, respectively (Fig. 1B):

$$\alpha = \frac{1}{1 + \exp(-\tilde{\alpha})},$$
$$\beta = \exp\big(\tilde{\beta}\big).$$

We modeled longitudinal parameter change across sessions using additive change terms ($\delta\tilde{\alpha}$ and $\delta\tilde{\beta}$) such that $\tilde{\alpha}_{T2} = \tilde{\alpha}_{T1} + \delta\tilde{\alpha}$ and $\tilde{\beta}_{T2} = \tilde{\beta}_{T1} + \delta\tilde{\beta}$. This yielded four candidate change models:

No-change model: $\delta\tilde{\alpha} = 0, \delta\tilde{\beta} = 0,$
α-change model: $\delta\tilde{\alpha} \neq 0,\ \delta\tilde{\beta} = 0,$
β-change model: $\delta\tilde{\alpha} = 0, \delta\tilde{\beta} \neq 0,$
$\alpha\beta$-change model: $\delta\tilde{\alpha} \neq 0,\ \delta\tilde{\beta} \neq 0.$

This model set was used to test whether session-to-session behavioural change was better explained by shifts in both, either, or neither of the parameters.

4.3. Synthetic dataset generation

For each change condition, we simulated datasets for 200 subjects and repeated the full procedure across 120 independent experiments. Baseline parameters were fixed at $\tilde{\alpha}_{T1} = -2$ and $\tilde{\beta}_{T1} = 1$. Changes were evaluated over a grid of $\delta\tilde{\alpha}, \delta\tilde{\beta} \in \{0, \pm 0.5, \pm 1, \pm 1.5, \pm 2\}$, corresponding to no change and to small, moderate, moderately large, and large increases or decreases. These baseline and change parameters were used to generate synthetic datasets for simulation-based evaluations of model identifiability, target-change discriminability, and detection

power. We generated choices across both sessions under the task structure using the true parameters from each change model.

4.4. Model identifiability

Model parameters were fitted using MLE with a trust-region Newton optimizer and forward-mode automatic differentiation. We fit all change models to each generated behavioural dataset. Model identifiability was then evaluated using two complementary procedures: one by model fit, which selected the model with the best penalized fit, and the other by parameter thresholds, which assigned discrete model labels by applying a prespecified decision rule to the change parameters fitted by the full-change model (i.e., $\alpha\beta$-change).

4.4.1. Model fit metrics

For model comparison, we computed Akaike's Information Criterion (AIC) and Bayesian Information Criterion (BIC) from the maximized log-likelihood ($\mathcal{L}$). Since parameters were fit jointly across sessions, we defined the effective number of observations as the total number of trials across sessions. For model identification tests, group-averaged AIC and BIC were computed as

$$\overline{AIC} = 2k - \frac{2}{N}\sum_i \log \mathcal{L}_i,$$

$$\overline{BIC} = k\ \log(m) - \frac{2}{N}\sum_i \log \mathcal{L}_i,$$

where k denotes the number of model parameters, m is the number of observations across sessions, and i indexes subjects ($i = 1, \ldots, N$). We used the subject-averaged $\mathcal{L}$ to keep the criterion comparable across sample sizes.

In fit-based identification, we selected the best-fitting model for each simulated dataset via $\overline{AIC}$ and $\overline{BIC}$ separately. For each change regime, identifiability was quantified by model recovery rates summarized in a confusion matrix, where columns and rows indexed the generative and identified models, respectively. Entries reported the proportion of experiments in which each generative model was identified as each candidate model. In dual-change regimes, model recovery was summarized in a 4-by-4 confusion matrix where the $\alpha\beta$-change model served as the generative model. The three remaining columns corresponded to datasets synthesized under nested alternatives, specifically projections onto single-change or no-change regimes. In single-change regimes, recovery reduced to a binary comparison between single-change and no-change models and was summarized in a 2-by-2 confusion matrix. When neither parameter changed, only the no-change model was defined.

4.4.2. Parameter change metrics

Parameter-based model identification required an explicit discretization rule to translate continuous change estimates into model labels. We first fit the full-change model to recover subject-level estimates of $\delta\tilde{\alpha}$ and $\delta\tilde{\beta}$. For each experiment, we used the absolute mean estimated change parameters across subjects ($|\overline{\delta\tilde{\alpha}}|$ and $|\overline{\delta\tilde{\beta}}|$) as the decision statistics.

A key difficulty of the parameter-based approach is that, unlike fit-based model selection, its decision criterion is not naturally defined. Declaring change requires introducing thresholds τ_α and τ_β, and model identification can depend strongly on how these thresholds are chosen. Therefore, we considered threshold selection as an operating point problem. For single-change detection against no-change, we defined thresholds using a false positive rate (FPR) criterion analogous to conventional hypothesis testing. Thresholds were chosen so that the probability of incorrectly declaring change under the no-change generative model was 0.05.

This FPR-based threshold is practically definable only for single-change detection against no-change. In this discrimination, the negative class is well defined because data generated under trivial between-session change (i.e., $\delta\theta \approx 0$) can be used to determine how far estimated changes deviate from zero under the null. This threshold can be used as a decision boundary for declaring single-parameter change. However, this procedure does not extend straightforwardly to comparisons in which the negative class is itself another change model.

When discriminating a dual-change model from a single-change model, the distribution of the negative model is no longer anchored to a no-change null. Thus, an FPR-based threshold cannot be defined in the same way.

A further complication arises for dual-change detection. Besides the absence of a uniquely defined negative model distribution, it is not straightforward to select a single threshold for four-way classification because it would require defining operating points for a multi-class decision boundary. In this multi-class case, there is no single FPR on a decision surface because operating points depend on which alternative model is treated as the negative class. Accordingly, we used the FPR-based threshold only for single-change detection against no-change. For four-way classification, we instead reported model recovery across an exploratory grid of threshold pairs $(\tau_\alpha, \tau_\beta) \in \{0, 0.1, 0.3, 0.5, 1, 2, 3\}$ to assess robustness to threshold selection. We summarized the parameter-based model recovery in confusion matrices, as in the fit-based approach. Given the FPR-based or predefined thresholds, we assigned discrete model labels for each experiment according to the decision rule:

$$Change\ model_{\delta\tilde{\alpha}\neq 0,\delta\tilde{\beta}\neq 0} = \begin{cases} \alpha\beta\text{-}change, & if\ \left|\overline{\delta\tilde{\alpha}}\right| > \tau_\alpha\ and\ \left|\overline{\delta\tilde{\beta}}\right| > \tau_\beta \\ \alpha\text{-}change, & if\ \left|\overline{\delta\tilde{\alpha}}\right| > \tau_\alpha\ and\ \left|\overline{\delta\tilde{\beta}}\right| \le \tau_\beta \\ \beta\text{-}change, & if\ \left|\overline{\delta\tilde{\alpha}}\right| \le \tau_\alpha\ and\ \left|\overline{\delta\tilde{\beta}}\right| > \tau_\beta \\ no\text{-}change, & if\ \left|\overline{\delta\tilde{\alpha}}\right| \le \tau_\alpha\ and\ \left|\overline{\delta\tilde{\beta}}\right| \le \tau_\beta \end{cases},$$

$$Change\ model_{\delta\tilde{\alpha}\neq 0,\delta\tilde{\beta}=0} = \begin{cases} \alpha\text{-}change, & if\ \left|\overline{\delta\tilde{\alpha}}\right| > \tau_\alpha \\ no\text{-}change, & if\ \left|\overline{\delta\tilde{\alpha}}\right| \le \tau_\alpha \end{cases},$$

$$Change\ model_{\delta\tilde{\alpha}=0,\delta\tilde{\beta}\neq 0} = \begin{cases} \beta\text{-}change, & if\ \left|\overline{\delta\tilde{\beta}}\right| > \tau_\beta \\ no\text{-}change, & if\ \left|\overline{\delta\tilde{\beta}}\right| \le \tau_\beta \end{cases}.$$

4.4.3. Profile likelihood visualization

To illustrate the decision principles underlying fit-based and parameter-based model identification, we visualized the profile log-likelihood over the two-dimensional change-parameter space. This analysis was used as a qualitative demonstration of decision principles. For each subject, we computed the profile log-likelihood of the αβ-change model on a grid of $\delta\tilde{\alpha}, \delta\tilde{\beta} \in \{-5, -4.5, \ldots, 5\}$, while optimising the baseline parameters at each pair of change parameters. We averaged these log-likelihood surfaces across subjects to obtain a group-level landscape for visualization. Maximum likelihood estimates from each change model were overlaid on this landscape to show how fitted parameters from different models were positioned relative to the likelihood surface. The $\alpha\beta$-change model was represented by an optimum in the grid, whereas the α- and β-change models were restricted optima on the corresponding axes. The no-change model was located at the origin.

4.5. Evaluation of discriminability for target change

We performed ROC analyses to assess change detection performance across decision thresholds. To construct binary classification problems, we relabeled the four change models into positive and negative classes according to one of three detection targets: $\alpha\beta$-, α-, and β-change. For each detection target, the positive class comprised datasets generated by the corresponding change model whereas the negative class comprised datasets generated by the alternative models.

For fit-based change detection, we used an information criterion (IC) as the decision statistic. Within each experiment, we averaged the IC across subjects for each candidate change model. For a given detection target, let M_+ denote the target change model and let M_- denote the set of alternative models. We then defined

$$\Delta IC = \min(IC_{M_-}) - IC_{M_+},$$

where $\min(IC_{M_-})$ is the smallest IC among the alternative models. Thus, larger ΔIC values indicate stronger evidence in favour of the target change model relative to the best-fitting alternative model. ROC curves were

obtained by classifying an experiment as positive when $\Delta IC > \tau_{IC}$, where τ_{IC} was varied from −150 to 120 in steps of 0.2. The same procedure was applied separately to AIC and BIC.

For parameter-based change detection, ROC analyses followed the same threshold-based decision rule used for parameter-based model identification. For $\alpha\beta$-change detection, classification required the two thresholds because change had to be evaluated jointly in both parameters. Accordingly, we constructed ROC surfaces over a two-dimensional grid of $(\tau_\alpha, \tau_\beta) \in \{0, 0.1, \dots, 20\} \times \{0, 0.1, \dots, 20\}$, examining how true positive rate (TPR) and FPR changed across this decision space. By contrast, single-change detection depended on only one threshold for the corresponding change parameter. For these cases, ROC curves were constructed by sweeping a single threshold τ_α or $\tau_\beta \in \{0, 0.1, \dots, 20\}$ under the corresponding decision rule.

For each pair of thresholds, we computed TPR and FPR as the proportions of experiments generated from the positive and negative classes that were classified as positive, respectively. In ROC surfaces, both TPR and FPR were defined over the two-dimensional threshold space. TPR was visualized as heatmap intensity with brighter colours indicating higher values. Contour lines were overlaid for both TPR and FPR at low (0.2), intermediate (0.5), and high (0.8) levels to summarize the sensitivity-specificity trade-off across the threshold space.

4.6. Power estimation

We quantified change detection power as the probability of detecting change when change was present (i.e., TPR). Following the decision rules described in the ROC analysis section, power was computed for two complementary procedures. The decision thresholds used here were identical to those defined in the identifiability analyses. For the fit-based approach, we used the threshold employed for best-fitting model selection (i.e., $\tau_{IC} = 0$). For the parameter-based approach, we used the threshold corresponding to a predefined FPR of 0.05 for single-change detection against no-change. As described above, parameter-based power was not evaluated for dual-change detection or for comparisons of dual-change versus single-change models because decision thresholds could not be practically defined in those settings.

Accordingly, power for $\alpha\beta$-change detection was estimated only for the fit-based approach. For single-change detection, power was estimated not only for the two complementary procedures but also using LRT at Type I error of 0.05, as a conventional benchmark for comparisons between full and nested models.

Power curves were constructed as a function of sample size and effect size, where the latter was indexed by the magnitude and direction of parameter change. We repeated simulations across 120 independent experiments to estimate power. Power was reported separately for each change-detection target.

4.7. Task design modification

To examine whether practical non-identifiability could be mitigated through task design rather than increased sample size alone, we conducted a simulation-based task optimisation analysis in representative change regimes where the generative $\alpha\beta$-change model was consistently misidentified under the original task design. The original design comprised 40 trials per block with a fixed reward probability of 70%, yielding 400 trials per session (Fig. 7A). We focused on mixed-sign regimes with decreased $\delta\alpha$ and increased $\delta\beta$, in which dual-change data were misidentified as simpler single-change models under BIC-based model comparison. These regimes were selected because they illustrated distinct failures to express behavioural signatures under the original task.

Task modifications were designed to amplify the behavioural signatures that were masked in each regime. When misidentification was attributed to insufficient asymptotic separation associated with β-change, we increased the number of trials per block from 40 to 200 while holding the total number of trials constant (Fig. 7B). When misidentification was attributed to weak expression of learning rate differences, we alternated reward probabilities across blocks between 70% and 85% to increase sensitivity to learning dynamics (Fig. 7C). To test whether both limitations could be addressed simultaneously, we also evaluated a design combining the two

modifications, with increased trials per block and alternating reward probabilities across blocks. In the combined design, the number of blocks was not reduced, yielding 2,000 trials per session (Fig. 7D).

For each modified task design, synthetic datasets were regenerated under the same change regimes and evaluated using the same fitting and model recovery procedures described above. Recovery was assessed using BIC-based model comparison to test whether task modification improved identification of the generative $\alpha\beta$-change model at the same sample size. To interpret the source of improvement qualitatively, we also compared synthetic behaviour generated under the true model with behaviour regenerated from the fitted competing models.

Data availability

The behavioural datasets generated are available upon request to the lead contact.

Code availability

All the necessary codes to reproduce the results are available at github.com/huyslab/longitudinalModel.

Acknowledgements

We are grateful for the Applied Computational Psychiatry Lab members for detailed input on early versions of this manuscript. We especially thank Isabel Berwian for constructive suggestions.

Financial disclosure

This work was funded by the National Institute for Health and Care Research University College London Hospitals Biomedical Research Centre (NIHR UCLH BRC).

QJMH has obtained fees and options for consultancies for Aya Technologies and Alto Neuroscience. QJMH has received research grant funding from Carigest S.A., German Research Foundation, Koa Health, National Institute of Health Research, Swiss National Science Foundation, Wellcome Trust. The other authors declare no competing interests. We acknowledge support by the NIHR UCLH BRC and NIHR MH-TRC MHM.

References

1 Pezzoli, P. *et al.* Challenges and solutions to the measurement of neurocognitive mechanisms in developmental settings. *Biol Psychiatry Cogn Neurosci Neuroimaging* **8**, 815-821 (2023).

2 Huys, Q. J., Maia, T. V. & Frank, M. J. Computational psychiatry as a bridge from neuroscience to clinical applications. *Nat Neurosci* **19**, 404-413 (2016).

3 Reiter, A. M. F. *et al.* Preference uncertainty accounts for developmental effects on susceptibility to peer influence in adolescence. *Nat Commun* **12**, 3823 (2021).

4 Reiter, A. M. F. *et al.* Self-reported childhood family adversity is linked to an attenuated gain of trust during adolescence. *Nat Commun* **14**, 6920 (2023).

5 Malamud, J. *et al.* The selective serotonin reuptake inhibitor sertraline alters learning from aversive reinforcements in patients with depression: Evidence from a randomized controlled trial. *Psychol Med* **54**, 2719-2731 (2024).

6 Schurr, R., Reznik, D., Hillman, H., Bhui, R. & Gershman, S. J. Dynamic computational phenotyping of human cognition. *Nat Hum Behav* **8**, 917-931 (2024).

7 Norbury, A., Hauser, T. U., Fleming, S. M., Dolan, R. J. & Huys, Q. J. M. Different components of cognitive-behavioral therapy affect specific cognitive mechanisms. *Sci Adv* **10**, eadk3222 (2024).

8 Norbury, A., Dercon, Q., Hauser, T. U., Dolan, R. J. & Huys, Q. J. M. Learning training as a cognitive restructuring intervention. *Biol Psychiatry Cogn Neurosci Neuroimaging* (2025).

9 Parsons, S. & McCormick, E. M. Limitations of two time point data for understanding individual differences in longitudinal modeling - What can difference reveal about change? *Dev Cogn Neurosci* **66**, 101353 (2024).

10 Brandmaier, A. M., Lindenberger, U. & McCormick, E. M. Optimal two-time point longitudinal models for estimating individual-level change: Asymptotic insights and practical implications. *Dev Cogn Neurosci* **70**, 101450 (2024).

11 Malamud, J. & Huys, Q. J. M. Distancing alters the controllability of emotional states by affecting both intrinsic stability and extrinsic sensitivity. *eLife* (2025).

12 Wieland, F.-G., Hauber, A. L., Rosenblatt, M., Tönsing, C. & Timmer, J. On structural and practical identifiability. *Curr Opin Syst Biol* **25**, 60-69 (2021).

13 Cuijpers, P. *et al.* Psychotherapy for depression across different age groups: A systematic review and meta-analysis. *JAMA Psychiatry* **77**, 694-702 (2020).

14 Landa, R. J., Gross, A. L., Stuart, E. A. & Faherty, A. Developmental trajectories in children with and without autism spectrum disorders: The first 3 years. *Child Dev* **84**, 429-442 (2013).

15 Valentin, S. *et al.* Designing optimal behavioral experiments using machine learning. *eLife* **13** (2024).

16 Myung, J. I. & Pitt, M. A. Optimal experimental design for model discrimination. *Psychol Rev* **116**, 499-518 (2009).

17 Self, S. G. & Liang, K. Y. Asymptotic properties of maximum-likelihood estimators and likelihood ratio tests under nonstandard conditions. *J Am Stat Assoc* **82**, 605-610 (1987).

18 Hauser, T. U., Will, G. J., Dubois, M. & Dolan, R. J. Annual research review: Developmental computational psychiatry. *J Child Psychol Psychiatry* **60**, 412-426 (2019).

19 Nussenbaum, K. & Hartley, C. A. Reinforcement learning across development: What insights can we draw from a decade of research? *Dev Cogn Neurosci* **40**, 100733 (2019).

20 Mkrtchian, A. *et al.* Differential associations of dopamine and serotonin with reward and punishment processes in humans: A systematic review and meta-Analysis. *JAMA Psychiatry* **82**, 818-829 (2025).

21 Palminteri, S., Kilford, E. J., Coricelli, G. & Blakemore, S. J. The computational development of reinforcement learning during adolescence. *PLoS Comput Biol* **12**, e1004953 (2016).

22 Decker, J. H., Otto, A. R., Daw, N. D. & Hartley, C. A. From creatures of habit to goal-directed learners: Tracking the developmental emergence of model-based reinforcement learning. *Psychol Sci* **27**, 848-858 (2016).

23 Moutoussis, M. *et al.* Change, stability, and instability in the Pavlovian guidance of behaviour from adolescence to young adulthood. *PLoS Comput Biol* **14**, e1006679 (2018).

24 Huys, Q. J., Pizzagalli, D. A., Bogdan, R. & Dayan, P. Mapping anhedonia onto reinforcement learning: A behavioural meta-analysis. *Biol Mood Anxiety Disord* **3**, 12 (2013).

25 Kim, T. *et al.* Neurocomputational model of compulsivity: Deviating from an uncertain goal-directed system. *Brain* **147**, 2230-2244 (2024).

26 Pike, A. C. & Robinson, O. J. Reinforcement learning in patients with mood and anxiety disorders vs control individuals: A systematic review and meta-analysis. *JAMA Psychiatry* **79**, 313-322 (2022).

27 Sullivan-Toole, H. *et al.* Reward-specific learning parameters change across normative adolescent development and are blunted in youth with high risk for depression. *J Child Psychol Psychiatry* (2025).

28 Michely, J., Eldar, E., Erdman, A., Martin, I. M. & Dolan, R. J. Serotonin modulates asymmetric learning from reward and punishment in healthy human volunteers. *Commun Biol* **5**, 812 (2022).

29 Dercon, Q. *et al.* A core component of psychological therapy causes adaptive changes in computational learning mechanisms. *Psychol Med* **54**, 327-337 (2024).

30 Brown, V. M. *et al.* Reinforcement learning disruptions in individuals with depression and sensitivity to symptom change following cognitive behavioral therapy. *JAMA Psychiatry* **78**, 1113-1122 (2021).

31 Wilson, R. C. & Collins, A. G. Ten simple rules for the computational modeling of behavioral data. *eLife* **8** (2019).

32 Danwitz, L., Mathar, D., Smith, E., Tuzsus, D. & Peters, J. Parameter and model recovery of reinforcement learning models for restless bandit problems. *Comput Brain Behav* **5**, 547-563 (2022).

33 Brod, G., Werkle-Bergner, M. & Shing, Y. L. The influence of prior knowledge on memory: A developmental cognitive neuroscience perspective. *Front Behav Neurosci* **7**, 139 (2013).

34 Rossberg, J. I. *et al.* Mechanisms of change and heterogeneous treatment effects in psychodynamic and cognitive behavioural therapy for patients with depressive disorder: A randomized controlled trial. *BMC Psychol* **9**, 11 (2021).