BEYOND THE POPULATION PRIOR: RELIABLE AND AMORTIZABLE PERSON REPRESENTATIONS IN MODELS OF HUMAN DECISION-MAKING

Anonymous authors
Paper under double-blind review

ABSTRACT

Most learned models of human decision-making represent a population, not a person. We ask whether a person's departures from such a model persist, improve prediction, and can be inferred quickly. Chess is our testbed: 2.6 million real-game decisions at positions whose difficulty other people rated; person fits use 545 players with at least one hundred each. A model over a frozen population reference infers each person's tempo, dispersion and ability beyond rating. They replicate across date-split halves (length-corrected tempo $r = 0.79$); a year apart agreement is lower; at equal decision counts, tempo and ability show no detectable drop. Per decision the person adds little: the held-out think-time gain, 0.010 nats, misses our 0.02-nat margin; in-game clock history beats static person offsets on timing; ability beyond rating adds accuracy, below margin. An amortized encoder infers held-out players from few decisions, 0.9% short of a fit on all their decisions against a 0.5% margin. Instrumented by the time control, extra time's estimated effect on accuracy is positive for players whose think time it changes, opposite to the within-person association (exclusion restriction checked, not proved). The traits are reliable; their value depends on the baseline; timing is partly amortizable, ability not yet.

1 INTRODUCTION

Two students hand in the same examination score: one works at a steady pace, the other alternates quick answers with long pauses, and at questions of the same difficulty one solves more. A population model that knows each question's difficulty and each student's standing predicts both totals; whether the pace, its consistency (dispersion) and the accuracy beyond standing are worth learning per person is our question. Three questions decide it: will the differences persist months later (reliability); will knowing them improve held-out predictions of the person's next decisions, over the population model and over the person's own recent history (value); and can they be learned from a few observations when someone new arrives (amortization)? A secondary question follows: when someone takes longer and still fails, did the extra time hurt, or did the harder problem make them think longer? The answers matter to anyone who personalises a population model (a simulator of a specific person, an adaptive test, a reward model for individual annotators): if personal differences do not persist, add little, or cannot be learned quickly, a person term has little to offer those uses.

Learned models of human decisions are fitted to a population and scored on held-out people: the person enters as a skill covariate or as hidden context, and where person parameters exist they are fitted where many people answer the same items (Yeung, 2019; Wang et al., 2020), by per-person optimisation, or as a latent inferred from a user's comparisons or past episodes (Section 2). In chess, human-like policies predict the move a player of a given rating would make (McIlroy-Young et al., 2020; Tang et al., 2024; Monroe et al., 2026), the published think-time heads carry no person state (Zhang et al., 2025; Johnson, 2026), and personalised models predict which move, or whether it is an error, not when (McIlroy-Young et al., 2022; Tang et al., 2025; Carlson, 2026; Rokach & Shapira, 2026). This paper measures what those models leave out (Figure 1).

Person terms added to strong population models have gained little per decision (Carlson, 2026; Rokach & Shapira, 2026). A small gain fits two readings: people are interchangeable, or they differ

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

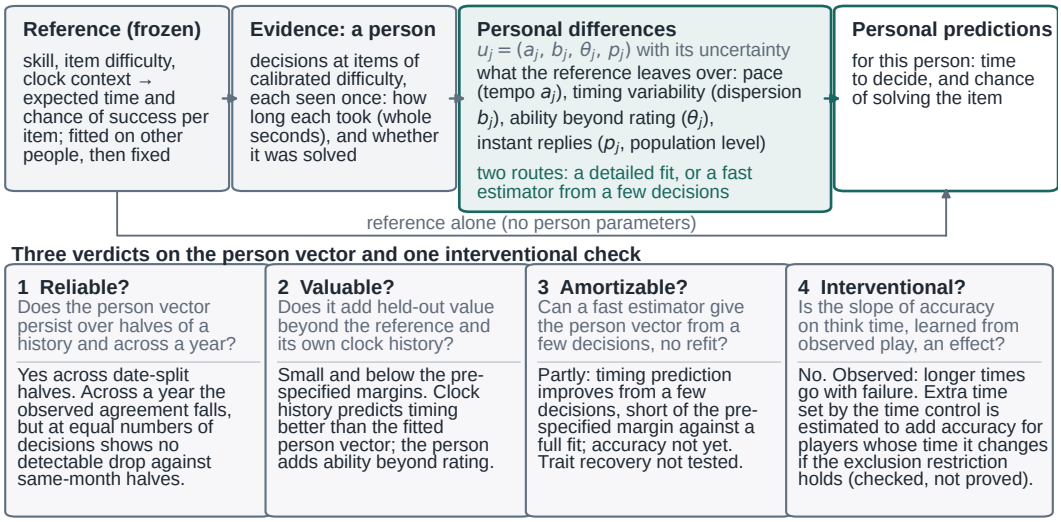

Figure 1: The frozen reference, a person's evidence at calibrated items, the person vector and its predictions; below, Section 5's verdicts: the vector replicates across date-split halves, adds little per decision beyond clock history and is partly amortizable (timing, not yet ability beyond rating).

but the design could not measure the differences reliably, which happens whenever between-person variance is small relative to trial-to-trial noise (Hedge et al., 2018). We call a person-level representation *reliable* when estimates from two date-split halves of a person's history (their decisions in date order) agree, and report its horizon: the agreement that survives an enforced gap.

Separating the two readings requires five properties, each checked, not assumed: (i) the same people make many timed decisions (reliability, amortization); (ii) each decision is at an item whose difficulty was measured on other people (ability beyond rating); (iii) a population model that can be fitted without the person it is scored on (value); (iv) the recording of time is known: whole-second clocks, with the rounding measured (the likelihood); and (v) for the causal check, the time available is set by the game's rules before the position arises. Online chess has all five at scale. A chess puzzle is a position from a real game where one player's mistake left the opponent a single best continuation, usually a win; the site rates its difficulty from later attempts by other people (a median of 354 per puzzle), who know a solution exists, whereas the opponent, the player we study, met the position once, in an ordinary game, not knowing it held one. Locating Lichess puzzles in their source games gives 2.6 million such decisions and 545 players with at least one hundred each; each decision is a *probe*, with a *time channel* (how long it took) and an *accuracy channel* (whether it was solved).

Contributions. The finding is a tension: measurable personal differences persist even where their added value per decision is small, and against the game's own clock history a static person term adds only a small accuracy signal. The paper contributes an evaluation protocol for person representations over a frozen population model and applies it: a per-person posterior over three named traits and a nuisance instant-reply term (Section 4), scored three ways, reliability with its horizon, value per channel against clock history, and amortization, each against a *margin*, a threshold fixed before any fit (a result that clears it is called *material*):

1. **Reliability, with its horizon measured.** The traits replicate across date-split halves (length-corrected tempo $r = 0.79$); a year apart, ability beyond rating agrees only weakly (0.22, below its 0.3 margin at the point estimate), yet a gap removes probes, and at equal probe counts tempo and ability show no detectable one-year drop against same-month halves (Section 5.1).
2. **Against clock history, the fitted vector's static offsets add accuracy, not timing.** Against the population model alone the fitted person vector gains on both channels, below margin; against the game's own *clock history* (a recurrent network over the player's earlier think times in the game, no person identity), each added to the same reference, its static offsets predict timing worse than clock history does, and added on top of that history they raise AUROC by 0.4 points (hundredths

of AUROC), as they do over Allie (Zhang et al., 2025), a published joint model of move and time. A person term scored only against the population model therefore overstates what is missing on timing; a person term learned jointly with clock history is untested (Section 5.2).

3. **Amortized person inference.** A set encoder maps k probes of a held-out player to a Gaussian over the person vector with no per-person optimisation. On held-out players of the training pool (Section 3) it improves on the reference from the first probe (not reproduced by the longer rerun, Section 6), whereas a direct fit to the same few probes is worse than the reference through three probes. At twenty probes it is 0.9% short of a fit that sees all of the person's probes, the scored ones included, missing the 0.5% margin; ability beyond rating is not yet amortized (Section 5.3).
4. **Diagnostic (secondary): the coupling a person model learns is not interventional.** An accuracy model given think time as an input and fitted on observed play learns a negative slope; with the time control as an instrument the estimated effect is positive for the players whose think time it changes, assuming the time control affects accuracy only through think time (the exclusion restriction, checked, not proved) (Section 5.4).

2 Related work

Person representations over population models. Centaur (Binz et al., 2025) is scored on held-out participants with no person parameters. Reward models treat the annotator as hidden context that distributional preference learning recovers only as a mixture (Siththaranjan et al., 2024), and variational preference learning gives each user a Gaussian latent from a few comparisons (Poddar et al., 2024); a set encoder that maps observations to a distribution over a latent is the neural statistician's construction (Edwards & Storkey, 2017), here applied to a person over a frozen population model. Dezfouli et al. (2019) encode a subject's choices into a latent that generates a recurrent model of that subject; ToMnet (Rabinowitz et al., 2018) separates an agent's past episodes (a character embedding) from the current episode (a mental-state embedding), our precedent for splitting the person from the episode's history. Our items are seen once with imported difficulty, an explanatory item-response model with item difficulty supplied, not estimated (De Boeck & Wilson, 2004).

Reliability, think time and time as an instrument. van der Linden (2007) models speed and ability as correlated person traits over items with their own parameters. Laboratory diffusion-model parameters keep their rank order across six months to two years (von Krause et al., 2021); we measure the same quantity at the same gaps on items met once in self-paced play. In chess, response-time distributions have been measured at scale (Sigman et al., 2010), and time tracks the expected improvement from further computation (Russek et al., 2025). In four-in-a-row, per-player planning parameters correlate across sessions days apart (van Opheusden et al., 2023). The clock-aware heads, which predict move delay from the position and both clocks (Johnson, 2026; Siu & Istead, 2026), are population-level, with no per-person parameters. Time controls have instrumented think time: Howard (2024) finds the association negative and the instrumented effect positive on Lichess, and Carow & Witzig (2025) find that tournament-set shorter controls make moves more risk-averse.

Response time and person models. A response-time likelihood improves latent-value estimation for a single subject with few trials (Shvartsman et al., 2024); Jacob et al. (2024) fit a latent search budget over a frozen chess policy per rating and time-control group, with no clocks. In chess, Allie (Zhang et al., 2025) adds a scalar time head with no clock state to its own rating-conditioned transformer policy, and the personalised models of Section 1, from per-player fine-tuning to few-game embeddings (Maia4All), predict which move, or whether it is an error, not when. Dasaka & Bapi (2026) estimate a pooled effect of computation carried forward from the previous move on log response time at puzzle instances (their "amortization" is a cognitive-science usage distinct from amortized inference). None of these works compares a person vector with the game's own clock history per channel; the delta to the concurrent manuscript (Anonymous, 2026) is in Appendix A.

3 Data

Terms used throughout. *Rating*: a skill score updated after every game (players) or every attempt (puzzles); 400 points is the scale on which the rating system turns a gap into expected odds. *Solved*: the move played at the probe is the puzzle's first solution move (alternative wins and the continuation

are not scored; Appendix B). *Ply*: one move by one player. *Time-control class*: the game's speed category (bullet, blitz, rapid, classical; most probes are blitz, a few minutes per player).

Data. The testbed joins public Lichess archives (CC0). The puzzle database gives each puzzle's source game and position, and a Glicko-2 difficulty rating (Glickman, 2001) updated after every later attempt by other people and imported rather than estimated. The monthly game archives give, at that position of the source game, the move the studied player (the one to move) actually made and the whole-second clock readings around it, hence the think time, the outcome and the context. One streaming pass over the 51 monthly archives processed at the freeze joins the two (2,647,029 probes; pipeline, months and study map: Appendix B). To our knowledge no public dataset pairs difficulty calibrated on other people with the same person's time and choice on that item, met unannounced in ordinary play and repeatedly over years. A third archive, with centisecond clocks for overlapping months, is used to measure how whole-second rounding distorts recorded think times (Section 4.2, Appendix C) and, in Appendix V only, for a centisecond refit; no headline number uses it. Calibrated difficulty makes one person's accuracy interpretable, like an exam question whose difficulty was set by other students; because the attempters who set it knew a tactic existed, the rating is an imported covariate, and the reference (Section 4) needs more than the rating gap. Two cohorts carry the tests: a *training pool* of 9,482 players with at least 23 probes, which trains the cross-fitted reference and the encoder, and a *primary cohort* of 545 players with at least one hundred, which carries the person fits. A *player* split holds out whole players for few-shot tests; a *month* split fits through June 2024 and scores July 2024 onward, holding out each player's own later probes (429 primary players).

4 Model, estimator and evaluation protocol

4.1 What we estimate: a person's residual on a fixed population model

A population model integrates the person out as hidden context; we call this population model the reference: it is the expectation for a person we know nothing about (the title's population prior; the Bayesian prior of Section 4.3 is a separate object). In plain terms, for every probe the reference says how long a typical player of these ratings would take, how variable that time is, and how likely the player is to solve it: a location m_i and log-scale s_i for log think time and a logit α_i for solving. These come from gradient-boosted models (Pedregosa et al., 2011) over the outputs of Maia-2 (Tang et al., 2024), a published network trained on online games to predict the move a human of a given rating would play, together with both ratings and the clocks. They are cross-fitted by player, so no player's own probes inform the fitted reference heads (the published Maia-2 backbone's training games may include probe games; Appendix G). On held-out folds it reaches location $R^2 = 0.53$ and accuracy AUROC 0.799; the six Maia-2 outputs (summaries of its move distribution, Appendix G) add 0.045 R^2 and 8.2 AUROC points as a block, so the reference already knows much about the position, and the person vector is scored only on what is left.

The person vector $\boldsymbol{u}_j = (a_j, b_j, \theta_j, p_j)^\top$ is that residual: tempo a_j (does this player take longer or shorter than the reference predicts?), dispersion b_j (are their times more or less erratic?), ability beyond rating θ_j (do they play the puzzle's first solution move more often than their rating predicts, read as tactical detection, an interpretation not separately validated) and the instant-reply term p_j. The instant-reply term is a propensity for premoves (moves queued before the opponent has moved, recorded as zero seconds); its individual values differ between estimators, so they are not read per person (Appendix D); only its spread across players is reported. The traits are parameters of the likelihood, named by role, not a cognitive process model. Player j produces probes $i \in \mathcal{I}_j$ in j's own games, nested in j's dominant time-control class $c(j)$; a probe carries the recorded whole-second think time $y_i \in \{0, 1, \dots\}$, the item's rating R_i, the player's rating E_i, a context vector $\boldsymbol{x}_i$ (the policy's outputs, the clocks, the move number, theme tags and the previous think times) and the outcome f_i, whether the probe was solved. Each context entry describes the position or the game before the decision; the item rating and the theme tags (the puzzle's tactic labels) are later annotations from attempters, so deployment would need items rated in advance (Appendices E, B).

4.2 Observation model: the recorded clock and the solve

The time channel treats a player's true think time as the reference's prediction for that position, made slower or faster by the player's own tempo and more or less variable by their own dispersion.

Formally, v_i is the log of the true think time: the reference predicts it, a_j shifts a player's times by a constant factor, b_j widens or narrows their spread, and $c_{0,c(i)}$ is the intercept of probe i's class:

$$v_i = m_i + c_{0,c(i)} + a_j + \varepsilon_i, \qquad \varepsilon_i \sim \mathcal{N}\big(0,\ \exp(2(s_i + b_j))\big). \tag{1}$$

A positive a_j means slower than predicted, and b_j multiplies the spread of log times by e^{b_j}. In words: we never see the true think time, only the difference of two clock readings that Lichess rounds to the whole second, so the probability of a recorded k adds up the probability of every true duration that could have produced it, weighted by how often each rounding error occurs (an approximation that admits errors of up to one and a half seconds, which two rounded readings cannot produce; a kernel that respects that bound leaves tempo, dispersion and ability nearly unchanged in rank, Appendices C, O). Measured, not assumed, on 2,018,270 plies in both Lichess's whole-second and centisecond archives, the whole-second clock is the centisecond clock rounded half-up on 100.00% of plies, so a think time recorded as k has rounded true time $k + d$, $d \in \{-1, 0, +1\}$, with masses $w_{-1} = 0.116$, $w_0 = 0.766$, $w_{+1} = 0.118$, the same in the one other month tested (June 2020; Appendix C). For $k \geq 1$ the true time lay within half a second of $k + d$; with F_i the normal CDF of v_i and $1 - \pi_i$ the probability that the reply was not a pre-entered instant reply,

$$P(y_i = k) = (1 - \pi_i) \sum_{d \in \{-1,0,+1\}} w_d \left[F_i\big(\log(k + d + \tfrac{1}{2})\big) - F_i\big(\log \max(k + d - \tfrac{1}{2},\ 0^+)\big) \right]. \tag{2}$$

A recorded zero has two sources: a premove, with probability π_i, which the *instant-reply term* p_j (the zero-mass offset of the appendices) raises or lowers, or a considered reply short enough to be recorded as zero, with chance C_i:

$$P(y_i = 0) = \pi_i + (1 - \pi_i)\, C_i, \qquad C_i = w_0\, F_i(\log \tfrac{1}{2}) + w_{+1}\, F_i(\log \tfrac{3}{2}), \tag{3}$$

where $\operatorname{logit} \pi_i$ is linear in a recapture flag, a low-clock flag, a previous zero, the reference's standardised outputs and p_j, chosen by AIC (the zero-mass logit, Appendix D). At $k = 0$ the $d = -1$ case cannot occur and the $d = +1$ window extends down to zero, which gives C_i; with this boundary the probabilities sum to one over $k \geq 0$ (Appendix C). We call Eqs. 2–3 the recording operator.

The accuracy channel is a logistic model whose baseline is the reference's logit. The reference already uses both ratings but may leave part of the rating gap's effect unexplained, so a coefficient β on the gap, in 400-point units, keeps that part out of the person term, ability beyond rating:

$$\operatorname{logit} P(f_i = 1) = \alpha_i + \delta_0 + \beta\, \frac{E_i - R_i}{400} + \theta_j. \tag{4}$$

θ_j shifts the log-odds of solving: a player with positive θ_j solves more often than the reference and the rating gap predict. The estimate of β is near zero; δ_0 is a pooled intercept.

4.3 Estimation: a per-person fit and an amortized encoder

In words: fit what is shared across players, estimate each player's offsets, then pull each estimate toward the class average, more strongly for players with few probes. Each player's vector is drawn from a class-level normal distribution, $\boldsymbol{u}_j \sim \mathcal{N}(\boldsymbol{\mu}_{c(j)}, \boldsymbol{\Sigma})$ (a hierarchical model; priors in Appendix D); the primary estimator is two-stage empirical Bayes, which estimates that distribution from the data and uses it as the prior: a pooled fit of the nuisance terms with person effects at zero, a per-player maximum-likelihood fit with the nuisance fixed, population SDs by a simulated-null moment estimator (an unweighted variant of DerSimonian & Laird, 1986), and shrinkage toward the class mean. A full correlated fit by mean-field ADVI (Kucukelbir et al., 2017; Abril-Pla et al., 2023) agrees on the point estimates of (a_j, b_j, θ_j) but not of p_j (Appendix D). Which traits to read per individual (tempo and ability; dispersion marginally) was fixed before fitting from the expected measurement error (Table 9).

Route 1, the fit above, refits for every new person. Route 2 is a network that reads a few of a new player's decisions and returns a Gaussian over the same person vector, trained toward the route-1 posteriors, with no refit: it leaves the reference frozen and adds an external module, $q_\phi(\boldsymbol{u}_j \mid \{(\boldsymbol{x}_i, R_i, E_i, y_i, f_i)\}_{i \leq k})$, a DeepSets network (Zaheer et al., 2017) over the frozen features with a Gaussian head that feeds Eqs. 1–4 unchanged. It is scored on held-out players at fixed k against (i) the reference; (ii) a zero-probe control, the population class mean; (iii) a point refit, the posterior mode from the same k probes; (iv) a marginalised refit, the point refit averaged over Laplace posterior draws as the encoder is; and (v) the full-information fit on all of the player's probes, the scored ones included, an in-sample ceiling (Appendix J).

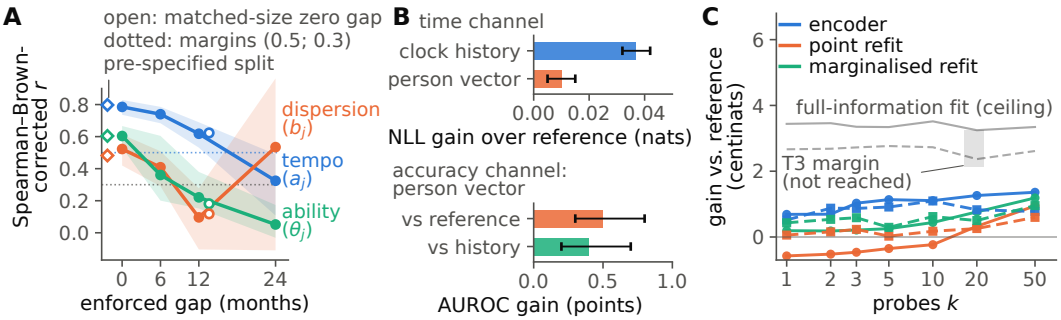

Figure 2: (A) Spearman–Brown-corrected split-half reliability against the enforced gap; open markers: zero-gap values (same-month halves) at matched half size (Appendix K). (B) Held-out-month gain per channel over the comparator in each label. (C) NLL gain over the reference (hundredths of a nat) against k on 370 held-out training-pool players (solid) and 101 primary-cohort players (dashed); the band at $k = 20$ is the T3 margin, within 0.5% of the full-information fit's NLL. Take-home: (A) at matched half size the one-year and zero-gap values of tempo and ability show no detectable difference; (B) clock history gains more on time, the person vector more on accuracy; (C) the encoder improves on the reference from a few probes but misses the band. Sources: Tables 23, 1 and 18.

How the representation is scored: three tests and their margins. Three tests (T1–T3) had their questions and margins fixed before the first fit, each judged alone against its margin (Section 1), no correction across them; intervals are paired player bootstraps. **Reliability (T1):** the split-half correlation of unshrunk per-player estimates from date-split halves, stepped up to full length by the Spearman–Brown formula (Spearman, 1910; Brown, 1910), which assumes that the halves measure the same trait equally well; material if $r \geq 0.5$ for tempo and $r \geq 0.3$ for dispersion and ability beyond rating. **Value (T2):** on months held out from the person fit (Appendix B), material only if all three margins are cleared: time-channel NLL (negative log-likelihood of the recorded time per probe, in nats) falls by 0.02 nats, accuracy AUROC (pooled over players; Section 5.2) rises by 1.0 point, and CRPS (continuous ranked probability score, in seconds) falls by 2%. CRPS is secondary because it scores the latent continuous prediction, not the recorded integer (Appendix D); none is cleared, so the verdict does not depend on that ranking. **Amortization (T3, adapted):** material if, at $k = 20$, the encoder's NLL is within 0.5% of the full-information fit's NLL. The margins were fixed before the first fit as multiples of per-decision gains seen in earlier pilots, not by a power calculation (Appendix H); implementation choices changed after data were seen, so the tests are pre-specified in question and margin, not in every analysis choice (T3's metric and comparator changed); the changes are listed in Appendix F. Of five posterior-predictive checks (Gelman et al., 1996) run before any headline number, one, the share of recorded zeros, fails (Appendix O).

5 Results

5.1 Is the person reliable, and over what horizon?

A person representation is worth storing only if it persists; it does across adjacent halves, and across a year the observed agreement falls but, for tempo and ability at matched half size, not detectably. Under the pre-specified date split all three traits clear their margins with intervals clear of them: Spearman–Brown-corrected split-half r is 0.79 for tempo, 0.48 for dispersion and 0.60 for ability beyond rating, against margins of 0.5, 0.3 and 0.3 (Table 13); the two halves meet with almost no gap, so it measures agreement between adjacent chronological halves, which can span years. The zero-gap halves below alternate probes within each calendar month (Appendix K).

An enforced gap between the halves (Figure 2A; Appendix K) lowers the observed values: tempo 0.74 after six months and 0.62 after a year; ability beyond rating 0.36 and then 0.22 $[0.15, 0.38]$, below its 0.3 margin at the point estimate; dispersion 0.10 $[-0.10, 0.28]$ at a year.

A year's gap drops the probes inside it, so each half is smaller (median 68.5, 51.5 and 38 probes at gaps of zero, six and twelve months), and estimates from fewer probes correlate less even if the

person did not change; to separate the two we thin the zero-gap halves to each player's one-year half size. At matched size the one-year value shows no detectable difference from the zero-gap value: the paired difference (one-year minus zero-gap) is $+0.04$ [-0.05, $+0.17$] for ability beyond rating and -0.005 [-0.08, $+0.06$] for tempo (dispersion's matched interval is wide), so, if at matched size the two designs differ only in the gap between the halves, a decline caused by the gap larger than 0.05 (ability) or 0.08 (tempo) is excluded; the two-year point has too few probes per half to interpret. The ability comparison is the weaker one: its matched zero-gap value is unusually low, and rank correlations, which remove that oddity, also detect no change but admit a decline of up to 0.15, so the 0.05 bound holds for the Pearson analysis only (Appendix K). Equivalence is not shown, predictions from a year-old representation were not tested, and ability beyond rating, as observed without matching sizes, agrees only weakly a year apart.

5.2 What does the person add beyond the game's own clock history?

On timing, clock history predicts better than the fitted person vector's static offsets; beyond the comparators tested, the person adds a small, below-margin accuracy signal. Personalising the reference lowers held-out-month time NLL by 0.010 nats per probe (margin 0.02), CRPS by 0.7% (margin 2%) and raises AUROC by 0.5 points (margin 1.0); none is reached, so the test is *not material* (Table 1; margin: Appendix H). Row 3, the concurrent manuscript's fixed-shrinkage tempo offset (Appendix D), gains less.

Per channel, against clock history: a GRU (Cho et al., 2014) over the player's earlier think times and clocks in the game (the reference sees only the two previous think times), with no person and no move-quality history, gains 0.037 nats [0.032, 0.042] over the reference on the time channel, where the person vector gains 0.010 [0.005, 0.015], each added to the same reference (rows 2 and 4). Earlier think times carry the player's pace that day and the game's time pressure, which no stable trait can track. Row 5 is narrower: the person vector's offsets, fitted against the reference, are added unchanged to clock history and worsen its timing prediction (NLL rises by 0.006 nats [0.002, 0.010]), 87% of it from the instant-reply term p_j, which estimators disagree on per person; holding p_j at its population value removes most of it. The comparison is asymmetric, static offsets against a trained sequence model; a person term learned jointly with history was not tested (Appendix X).

On the accuracy channel the order reverses: the person vector raises AUROC by 0.5 points [0.3, 0.8] over the reference, below the 1.0-point margin, and, added on top of clock history, which alone barely moves it, by 0.4 [0.2, 0.7] (row 5 against row 2). Brier, a proper score, falls from 0.1552 to 0.1535 (rows 1 and 4). Because θ_j is constant within a player, it cannot reorder a player's probes: its AUROC gain comes from ranking players, which is what ability beyond rating measures. Over Allie, a published model of move and time, on the blitz probes it covers, θ_j raises AUROC by 0.4 points [0.1, 0.7], below the margin (Table 22). No comparator tested sees the player's past accuracy.

5.3 Can a new person be inferred from a few probes?

For a new person the question is whether inference can be amortized. Partly: timing prediction improves from a few probes, short of the pre-specified margin; the accuracy channel not yet. On 370 held-out players of the training pool, from k randomly drawn probes of each (Figure 2C), the encoder (a three-seed ensemble; intervals exclude seed and fold variability, Appendix J) lowers

Table 1: Held-out prediction (person terms fit through June 2024, scored July 2024 onward; population terms on all months, so retrospective, Appendix B; 429 players, 24,517 probes): NLL per probe (nats, ↓; Eqs. 2–3), CRPS (Gneiting & Raftery, 2007) in seconds (↓), cov.: coverage of the central 50%/90% predictive intervals, AUROC (↑) and Brier (↓) for solving. Bold: numerical best, not a test. Row 5 against row 2: the fitted offsets, not re-estimated, do not help history on timing.

	Model	NLL (nats)	CRPS (s)	cov. 50/90%	AUROC	Brier
(1)	Maia-2 reference (frozen; two previous think times)	1.8660	1.380	0.52/0.91	0.787	0.1552
(2)	+ GRU over within-game clocks (no person)	**1.8290**	**1.359**	0.54/0.92	0.789	0.1546
(3)	+ stable tempo offset, shrinkage 0.1	1.8632	1.376	0.52/0.91	0.787	0.1551
(4)	+ hierarchical person vector, route 1	1.8563	1.370	0.54/0.91	0.792	0.1535
(5)	+ GRU + person vector	1.8345	1.359	0.53/0.91	**0.793**	**0.1531**

NLL, relative to the reference, by 0.3% of the reference's NLL at one or two probes (the one-probe gain is not reproduced by the longer rerun, Section 6) and by 0.45–0.59% from three probes on. A point refit from the same probes is worse than the reference through three probes (Figure 2C). A zero-probe control (the reference plus the population class mean) does no better than the reference, so the gain is not a constant class-level shift; it could still partly reflect a correction of the reference for the positions this player tends to meet rather than the player's traits, which a permuted-player control would separate (not run, Appendix X).

T3, as adapted, asks for the encoder's shortfall from the full-information fit, as a share of that fit's NLL, to be at most 0.5% at twenty probes; it is 0.9% (0.8% on the primary cohort, in parentheses below), so the test is *not material*, and the shortfall is a point value with no interval (Appendix X). The encoder closes 39% (35%) of the reference-to-fit distance, a point refit 10% (11%).

Against the marginalised refit the encoder's NLL is lower on the training-pool players through twenty probes (not for every single seed) and, before correction, distinguishable on the primary cohort only at five and ten probes (Appendix J); the refits hold p_j at its class mean and optimise a different objective, so this compares procedures, not the amortization gap (Cremer et al., 2018).

At twenty probes the encoder shows no material accuracy benefit: AUROC 0.796 against the reference's 0.795 on these players, and 0.805 for the full-information fit (Appendix J). The evidence is predictive (recovery of each trait, or of the posterior given the same k probes, was not tested) and conditional: the encoder never trains on the held-out players, but the population-level quantities it and every comparator use were estimated with them included (Appendices B and X).

5.4 Is the learned speed–accuracy coupling interventional?

A simulator that lets a person think longer needs to know whether the learned coupling is an effect; here it does not identify one. One mechanism makes the association negative even when more time would help: a player who sees the solution stops thinking and plays it, so quick answers are more often right. Some players play the same base time with and without seconds added after each move (the increment), chosen with the game before any position arises. As an instrument (a variable that moves think time for reasons unrelated to the position), it identifies an effect if it moves think time (relevance), is assigned independently of the player's form that day (independence), changes accuracy only through think time (the *exclusion restriction*) and lowers no one's think time (monotonicity, assumed and not tested). Its *compliers* are the players whose think time it raises.

The within-person association: the accuracy channel extended by log think time (Eq. 4 plus $\lambda \log T$; Appendix P) gives about -0.04 in probability per log-unit of think time on the 545-player primary cohort; double machine learning agrees in sign (Table 27). The instrumented effect: the instrument is applied to 5,174 training-pool players (median rating 2,053) with player$\times$month fixed effects; the first stage is strong ($F = 976$), and the local average treatment effect (Angrist et al., 1996) is for the compliers. For them, a log-unit more think time has an estimated effect on solve probability of $+0.084$ [0.055, 0.112], the opposite sign to the within-person -0.04 (a re-specification; the pre-specified one gives a smaller effect of the same sign, Appendix F). The instrument shifts think time by 0.35 log-units, so the per-log-unit scale extrapolates; the reduced form (the change in solve probability from having an increment) is $+0.029$. Cautions (Appendix P): the three estimates (Table 27) come from three samples and two scales, probes exist only where the game produced a puzzle, and the reported specification conditions on a clock floor the increment moves; without the floor it stays positive. The exclusion restriction is checked, not proved: a placebo regression of the player's move quality two own moves earlier on the instrument is not null ($+0.006$ [0.002, 0.011]), and under a small direct effect of the increment (Conley et al., 2012) the compliers' effect stays positive, but a confounder as strong as the strongest measured covariate could remove the reduced-form effect. What this means: the positive sign rests on the exclusion restriction; the main point does not: a slope learned from observed play does not identify the effect of more time.

6 Discussion

What the three tests show together. The introduction asked whether a small per-decision gain means interchangeable people or a design that could not see them. Neither holds: people differ reliably, since the traits replicate across date-split halves, and the design sees them, at calibrated

items against a reference fitted without the person; what is small is the per-decision value of what it sees. Under the illustrative location model of Appendix H, reliability and value reflect one signal-to-noise ratio: stable differences between players are a small share of the variation in a single think time, too little to sharpen one prediction by much, yet over many probes they are estimated precisely, so tempo is reliable yet worth little per decision; the encoder's learning curve departs from that model. Tempo is the most reliable trait, and the time channel is where the encoder gains from a few probes, yet clock history predicts timing better. Ability beyond rating is what neither clock history, the reference nor Allie supplies, yet a year apart it agrees only weakly and is not yet amortized.

What this implies for person representations. Three lessons, each tied to a use named in the introduction (none was tested): (1) for a simulator of a person, an adaptive test or a per-annotator reward model, the comparison baseline sets how much a person term seems to add: against the population model alone, static offsets credit the person with timing information that clock history carries, and only a person term fitted given that history, refitted or jointly trained, can show what timing information is left; (2) for a new person, the timing part is partly amortizable (a few-observation encoder improves timing prediction but falls short of the full fit), the accuracy part not yet; (3) for a model that lets a person think longer, the coupling learned from observed play has the opposite sign to the instrumented effect for the compliers (exclusion restriction checked, not proved), so it does not say what more time would do (Figure 1; Table 2).

Beyond this testbed. Where only repeated timed decisions were available (no calibrated items), timing reliability and a held-out timing gain reappeared and the accuracy channel's gain did not: in a 30-player four-in-a-row pilot (Appendix T; history and amortization not tested). In elite over-the-board play, tempo and dispersion replicate and ability beyond rating does not (Appendix U).

Limitations. *Whole seconds and zeros.* The model over-predicts recorded zeros (11.3% against 8.7%), the instant-reply term differs between estimators per person, the low-clock cut-point was not varied (Appendix D), and the recording operator was measured in only two months, with the same masses (Appendix C.1). *Comparators not run.* Not scored: a person term learned jointly with clock history, a personalised move model (Maia4All), an accuracy-history baseline, and a newcomer's first k probes in date order, where support probes are random; the not-run list (Appendix X) has the rest. *Causal identification.* The sample is selected after the instrument, the placebo is not null, and the exclusion restriction is checked, not proved. *Longer rerun.* A 79-month replication that did not meet pre-specified completeness criteria (Appendix Q) moves several results: the encoder's gain at one probe (gone), the exploratory validity check of ability beyond rating (weaker, largely newly eligible players) and tempo's two-year reliability (higher, largely a shrinking coverage hole); the reliability and value verdicts, T3's verdict and the sign of the instrumented effect hold. *External validity.* The primary cohort is high-volume, strong (median rating 2,376) and mostly blitz, and the conclusions are bounded to it; traits correlate across time-control classes, but the fitted vector does not transfer predictively between them and on accuracy can hurt held-out prediction (Appendix S); the accuracy trait's external validity is weak (Appendix W; untested on ordinary moves); the traits and effect sizes are chess's; no downstream use or refresh after a gap was tested.

7 Conclusion

The population models we build on carry no person parameters; a person is a measurable residual of that model. Reliability: tempo replicates across date-split halves (length-corrected $r = 0.79$); a year apart agreement falls, but at equal probe counts tempo and ability beyond rating are not shown to change. Value: over the population model the fitted vector gains on both channels, below margin; clock history predicts timing better than its static offsets; ability beyond rating adds a small between-player accuracy signal beyond the comparators tested. Amortization: a set encoder improves timing prediction from a few probes, 0.9% short of the full-information fit against a 0.5% margin; ability beyond rating is not yet amortized. Interventional check: the coupling a person model learns from observed play has the opposite sign to the instrumented effect for the compliers, under an exclusion restriction checked, not proved. For learned models of people, a person term's comparator is the strongest within-episode history; next are a jointly learned person term, amortizing ability beyond rating, and a second domain with calibrated items and response times.

REPRODUCIBILITY STATEMENT

The analysis plan (inclusion rules, censoring weights, priors, tests, margins, decision rules, baselines, metrics and posterior-predictive gates) was fixed in a dated internal document before the first fit; every later deviation is listed in Appendix F. Available to reviewers now, in the supplementary material: that plan with its dated change log, the dated note that pre-stated the causal check, the predeclaration of the 79-month replication, the anonymised concurrent manuscript (Anonymous, 2026) for Appendices A and D, and the analysis code, with a README that maps each script to the section or appendix it supports and lists the seeds, the environment, the run order and the inputs not built here: an extraction step (not supplied) that keeps the puzzles' source games from the public Lichess archives, the public Maia-2, Maia-3 and Allie checkpoints, and the four-in-a-row and ordinary-move pilot tables. From the extracted games a game-header script and a builder, whose per-month row counts are checked against the puzzle files (Section 3; Appendix B), rebuild the probe table, and one comparison script the operator table (Appendix C). No end-to-end run from a clean directory was performed; the run order of the earlier model versions that the reported fits read was reconstructed from the scripts' inputs, and of the result files only the recomputations' aggregate outputs are supplied. Released on publication: per-player summaries under random keys. Withheld: the per-probe residual vectors (Ethics statement).

ETHICS STATEMENT

The chess data are public game and puzzle archives released by Lichess under a CC0 licence; the four-in-a-row data of van Opheusden et al. (2023) (Appendix T, released on the Open Science Framework) and the over-the-board data package of Sunde et al. (2026) (Appendix U, deposited in Harvard Dataverse) are public releases used under the terms stated with their publication. No data were collected from participants, no private information was accessed, and the analysis involves no intervention. Because the study uses only public archives and involved no interaction with participants, the authors judged that no ethics-board review was required. Usernames are replaced by pseudonymous identifiers, no named player's estimate appears, and the paper reports cohort-level reliabilities, variance components and predictive scores, never individual profiles. The encoder estimates a person's behavioural parameters from few observations; such a capability would serve adaptive item selection and personalised simulators, but could also support profiling players without their knowledge. Timing is also a fingerprint: from ten probes to one half of a player's history, it identifies them among 545 players at 5 to 19 times chance (Appendix L), so released tables could in principle link pseudonymous accounts. The main estimates are validated at cohort level on one online platform; they do not support decisions about an individual player, and ability beyond rating is not yet recovered from a few probes. We will therefore release per-player summaries only under random keys, withhold per-probe residual vectors, report identification only in aggregate, and recommend that any deployment be disclosed to the players it models.

AI USE STATEMENT

Generative AI tools (Claude, Anthropic, operated through a coding-assistant tool) were used under the authors' direction for the following tasks with required disclosure: implementing methods, in part (the analysis scripts in the supplementary material were written by AI assistants under the authors' direction); cleaning and reformatting datasets (probe-table construction and cohort filtering); and generating synthetic data sets for the simulation study of the recording operator only, so that no reported empirical result rests on synthetic data. We have not used generative AI tools to develop theoretical models or conceptual frameworks, to formulate mathematical claims, to provide ingredients for or write proofs, to propose or refine hypotheses, to design or give feedback on the research methodology or experiments, or to interpret results; translation and qualitative or thematic data analysis are not applicable to this work. Additionally, we used generative AI tools for editing and formatting documents prepared from the authors' specifications, correcting grammar, and debugging the analysis scripts. The authors designed the study, carried out and verified the analyses, and take full ownership of and responsibility for every process involved in the preparation of this document, including its text, claims, code, and artefacts.

REFERENCES

Oriol Abril-Pla, Virgile Andreani, Colin Carroll, Larry Dong, Christopher J. Fonnesbeck, Maxim Kochurov, Ravin Kumar, Junpeng Lao, Christian C. Luhmann, Osvaldo A. Martin, Michael Osthege, Ricardo Vieira, Thomas Wiecki, and Robert Zinkov. PyMC: A modern, and comprehensive probabilistic programming framework in Python. *PeerJ Computer Science*, 9:e1516, 2023.

Theodore W. Anderson and Herman Rubin. Estimation of the parameters of a single equation in a complete system of stochastic equations. *The Annals of Mathematical Statistics*, 20(1):46–63, 1949.

Joshua D. Angrist, Guido W. Imbens, and Donald B. Rubin. Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91(434):444–455, 1996.

Anonymous. Manuscript under review, 2026.

Marcel Binz, Elif Akata, Matthias Bethge, Franziska Brändle, Fred Callaway, Julian Coda-Forno, Peter Dayan, Can Demircan, Maria K. Eckstein, Noémi Éltető, Thomas L. Griffiths, Susanne Haridi, Akshay K. Jagadish, Li Ji-An, Alexander Kipnis, Sreejan Kumar, Tobias Ludwig, Marvin Mathony, Marcelo Mattar, Alireza Modirshanechi, Surabhi S. Nath, Joshua C. Peterson, Milena Rmus, Evan M. Russek, Tankred Saanum, Johannes A. Schubert, Luca M. Schulze Buschoff, Nishad Singhi, Xin Sui, Mirko Thalmann, Fabian J. Theis, Vuong Truong, Vishaal Udandarao, Konstantinos Voudouris, Robert Wilson, Kristin Witte, Shuchen Wu, Dirk U. Wulff, Huadong Xiong, and Eric Schulz. A foundation model to predict and capture human cognition. *Nature*, 644(8078):1002–1009, 2025. doi: 10.1038/s41586-025-09215-4.

William Brown. Some experimental results in the correlation of mental abilities. *British Journal of Psychology*, 3(3):296–322, 1910. doi: 10.1111/j.2044-8295.1910.tb00207.x.

Jason Carlson. Elo-disentangled player-style embeddings for human chess via rating-conditioned residual move model, June 2026. URL `https://arxiv.org/abs/2606.25176v1`. arXiv:2606.25176v1 (version 1, 23 June 2026); later versions are retitled "Matilda: Engine-Agnostic Search with Human Policy Guidance".

Johannes Carow and Niklas M. Witzig. Time pressure and strategic risk-taking in professional chess. *Journal of Economic Behavior & Organization*, 238:107218, 2025. doi: 10.1016/j.jebo.2025.107218.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 2018.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)*, pp. 1724–1734, 2014.

Timothy G. Conley, Christian B. Hansen, and Peter E. Rossi. Plausibly exogenous. *The Review of Economics and Statistics*, 94(1):260–272, 2012. doi: 10.1162/REST_a_00139.

Chris Cremer, Xuechen Li, and David Duvenaud. Inference suboptimality in variational autoencoders. In *Proceedings of the 35th International Conference on Machine Learning (ICML 2018)*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1078–1086. PMLR, 2018. URL `https://proceedings.mlr.press/v80/cremer18a.html`.

Amarnath Dasaka and Raju Surampudi Bapi. The shadow of the past: Amortized inference and belief revision using chess as a model system. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 48, 2026. URL `https://escholarship.org/uc/item/7h38004r`.

Paul De Boeck and Mark Wilson (eds.). *Explanatory Item Response Models: A Generalized Linear and Nonlinear Approach*. Springer, New York, 2004.

Rebecca DerSimonian and Nan Laird. Meta-analysis in clinical trials. *Controlled Clinical Trials*, 7 (3):177–188, 1986.

Amir Dezfouli, Hassan Ashtiani, Omar Ghattas, Richard Nock, Peter Dayan, and Cheng Soon Ong. Disentangled behavioural representations. In *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/e077e1a544eec4f0307cf5c3c721d944-Abstract.html.

Harrison Edwards and Amos Storkey. Towards a neural statistician. In *International Conference on Learning Representations (ICLR 2017)*, 2017.

Andrew Gelman, Xiao-Li Meng, and Hal Stern. Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, 6(4):733–807, 1996.

Mark E. Glickman. Dynamic paired comparison models with stochastic variances. *Journal of Applied Statistics*, 28(6):673–689, 2001.

Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. doi: 10.1198/016214506000001437.

Craig Hedge, Georgina Powell, and Petroc Sumner. The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. *Behavior Research Methods*, 50(3):1166–1186, 2018. doi: 10.3758/s13428-017-0935-1.

Greg Howard. A check for rational inattention. *Journal of Political Economy Microeconomics*, 2 (1):172–199, February 2024. doi: 10.1086/727556. URL https://doi.org/10.1086/727556.

Athul Paul Jacob, Abhishek Gupta, and Jacob Andreas. Modeling boundedly rational agents with latent inference budgets. In *International Conference on Learning Representations (ICLR 2024)*, pp. 10169–10183, 2024. URL https://openreview.net/forum?id=W3VsHuga3j. arXiv:2312.04030.

Thomas Johnson. ChessMimic: Per-rating transformer models for human move, clock, and outcome prediction in online blitz chess, 2026. URL https://arxiv.org/abs/2606.04473. arXiv:2606.04473.

Alp Kucukelbir, Dustin Tran, Rajesh Ranganath, Andrew Gelman, and David M. Blei. Automatic differentiation variational inference. *Journal of Machine Learning Research*, 18(14):1–45, 2017.

David Lopez-Paz and Maxime Oquab. Revisiting classifier two-sample tests. In *International Conference on Learning Representations (ICLR 2017)*, 2017.

Inbal Magar and Roy Schwartz. Data contamination: From memorization to exploitation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 157–165, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-short.18. URL https://aclanthology.org/2022.acl-short.18/. arXiv:2203.08242.

Reid McIlroy-Young, Siddhartha Sen, Jon Kleinberg, and Ashton Anderson. Aligning superhuman AI with human behavior: Chess as a model system. In *Proceedings of the 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD 2020)*, 2020.

Reid McIlroy-Young, Russell Wang, Siddhartha Sen, Jon Kleinberg, and Ashton Anderson. Detecting individual decision-making style: Exploring behavioral stylometry in chess. In *Advances in Neural Information Processing Systems 34 (NeurIPS 2021)*, 2021. arXiv:2208.01366.

Reid McIlroy-Young, Russell Wang, Siddhartha Sen, Jon Kleinberg, and Ashton Anderson. Learning models of individual behavior in chess. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD 2022)*, 2022. arXiv:2008.10086.

Daniel Monroe, George Eilender, Philip Chalmers, Zhenwei Tang, and Ashton Anderson. Chessformer: A unified architecture for chess modeling. In *International Conference on Learning Representations (ICLR 2026)*, 2026. URL https://arxiv.org/abs/2605.19091.

Jerad H. Moxley, K. Anders Ericsson, Neil Charness, and Ralf T. Krampe. The role of intuition and deliberative thinking in experts' superior tactical decision-making. *Cognition*, 124(1):72–78, 2012. doi: 10.1016/j.cognition.2012.03.005.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12: 2825–2830, 2011.

Sriyash Poddar, Yanming Wan, Hamish Ivison, Abhishek Gupta, and Natasha Jaques. Personalizing reinforcement learning from human feedback with variational preference learning. In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024)*, pp. 52516–52544, 2024. doi: 10.52202/079017-1664. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/5e1c255653eb98cef13f45b2d337c882-Abstract-Conference.html. arXiv:2408.10075.

Neil C. Rabinowitz, Frank Perbet, H. Francis Song, Chiyuan Zhang, S. M. Ali Eslami, and Matthew Botvinick. Machine theory of mind. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning (ICML 2018)*, volume 80 of *Proceedings of Machine Learning Research*, pp. 4218–4227. PMLR, 2018. URL https://proceedings.mlr.press/v80/rabinowitz18a.html. arXiv:1802.07740.

Yarden Rokach and Bracha Shapira. Blunder prediction in chess. *Applied Intelligence*, 56(4):92, 2026. doi: 10.1007/s10489-026-07131-2.

Evan M. Russek, Daniel Acosta-Kane, Bas van Opheusden, Marcelo G. Mattar, and Thomas L. Griffiths. Time spent thinking in online chess reflects the value of computation. *Cognitive Science*, 49(10):e70119, 2025. doi: 10.1111/cogs.70119.

Michael Shvartsman, Benjamin Letham, Eytan Bakshy, and Stephen Keeley. Response time improves Gaussian process models for perception and preferences. In *Proceedings of the Fortieth Conference on Uncertainty in Artificial Intelligence (UAI 2024)*, volume 244 of *Proceedings of Machine Learning Research*, pp. 3211–3226. PMLR, 2024. URL https://proceedings.mlr.press/v244/shvartsman24a.html.

Mariano Sigman, Pablo Etchemendy, Diego Fernández Slezak, and Guillermo A. Cecchi. Response time distributions in rapid chess: A large-scale decision making experiment. *Frontiers in Neuroscience*, 4:60, 2010. doi: 10.3389/fnins.2010.00060.

Anand Siththaranjan, Cassidy Laidlaw, and Dylan Hadfield-Menell. Distributional preference learning: Understanding and accounting for hidden context in RLHF. In *International Conference on Learning Representations (ICLR 2024)*, pp. 27448–27471, 2024. URL https://openreview.net/forum?id=0tWTxYYPnW.

Sherman Siu and Lesley Istead. UniMaia: Steering chess policies with language for human-like play, 2026. URL https://arxiv.org/abs/2605.27767. arXiv:2605.27767.

Charles Spearman. The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101, 1904.

Charles Spearman. Correlation calculated from faulty data. *British Journal of Psychology*, 3(3): 271–295, 1910. doi: 10.1111/j.2044-8295.1910.tb00206.x.

Uwe Sunde, Dainis Zegners, and Anthony Strittmatter. Speed and quality of complex strategic decisions. *Proceedings of the National Academy of Sciences*, 123(20):e2531472123, 2026. doi: 10.1073/pnas.2531472123.

Zhenwei Tang, Difan Jiao, Reid McIlroy-Young, Jon Kleinberg, Siddhartha Sen, and Ashton Anderson. Maia-2: A unified model for human-AI alignment in chess. In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024)*, pp. 20919–20944, 2024. URL https://arxiv.org/abs/2409.20553.

Zhenwei Tang, Difan Jiao, Eric Xue, Reid McIlroy-Young, Jon Kleinberg, Siddhartha Sen, and Ashton Anderson. Learning to imitate with less: Efficient individual behavior modeling in chess, 2025. arXiv:2507.21488.

Wim J. van der Linden. A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika*, 72(3):287–308, 2007. doi: 10.1007/s11336-006-1478-z.

Bas van Opheusden, Ionatan Kuperwajs, Gianni Galbiati, Zahy Bnaya, Yunqi Li, and Wei Ji Ma. Expertise increases planning depth in human gameplay. *Nature*, 618(7967):1000–1005, 2023. doi: 10.1038/s41586-023-06124-2.

Mischa von Krause, Stefan T. Radev, Andreas Voss, Martin Quintus, Boris Egloff, and Cornelia Wrzus. Stability and change in diffusion model parameters over two years. *Journal of Intelligence*, 9(2):26, 2021. doi: 10.3390/jintelligence9020026.

Fei Wang, Qi Liu, Enhong Chen, Zhenya Huang, Yuying Chen, Yu Yin, Zai Huang, and Shijin Wang. Neural cognitive diagnosis for intelligent education systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 6153–6161, 2020.

Chun-Kit Yeung. Deep-IRT: Make deep learning based knowledge tracing explainable using item response theory. In *Proceedings of the 12th International Conference on Educational Data Mining (EDM)*, 2019. arXiv:1904.11738.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Russ R. Salakhutdinov, and Alexander Smola. Deep sets. In *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, 2017.

Yiming Zhang, Athul Paul Jacob, Vivian Lai, Daniel Fried, and Daphne Ippolito. Human-aligned chess with a bit of search. In *International Conference on Learning Representations (ICLR 2025)*, 2025. URL `https://arxiv.org/abs/2410.03893`.

APPENDIX OVERVIEW

The appendices follow the main text. Each opens by naming the question it supports. Appendix A positions the two closest works and the concurrent manuscripts; Appendix B gives the data pipeline from puzzles to source games to clocks, the cohorts, and a study map that names, for every result, its data, the fitted object and the score; Appendices C–F give the recording operator, the likelihood, the history of the observation model and every deviation from the plan; Appendices G–J give the reference, a model-based benchmark for the gain from a known trait, the recovery simulation and the encoder, GRU and Allie details; Appendices K–P give the fixed-gap reliability curve, identification, the ordinary-move transfer, the posterior-predictive checks and the robustness of the instrument. Appendices Q–V are replications and external-validity checks that the main text does not rely on: the 79-month rerun, a second reference model, transfer across time controls, a second game, elite over-the-board play and centisecond clocks. Appendix W collects further descriptive analyses that carry no claim of this paper, and Appendix X lists the analyses not run.

A CLOSEST PRIOR WORK AND CONCURRENT MANUSCRIPTS

This appendix supports the related-work question (Section 2): what the two closest works do, and what this paper adds. Dasaka & Bapi (2026) estimate, by pooled least-squares regression, the effect of computation carried forward from the previous move on log response time at puzzle instances; we treat the person as the estimand and model speed and accuracy jointly. A concurrent manuscript (Anonymous, 2026) audits per-account offset corrections on general decisions; we adopt that logic as a named random-intercept baseline (row 3 of Table 1), whose recipe is restated in full in Appendix D so that it can be checked without the manuscript.

Neither work reports a reliability with a horizon, a per-channel value of the person against clock history, an amortized person posterior or the instrumented effect of think time, which are this paper's results (Table 3); the concurrent manuscript, under review and anonymised, is supplied as supplementary material so that these deltas can be checked.

Table 2: Trait × verdict, per channel (Section 6): split-half r under the pre-specified date split; one-year r under an enforced twelve-month gap (in parentheses, the zero-gap value at matched half size); held-out-month gains over the reference, clock history and Allie, each below its pre-specified margin; units are never compared across channels. †The whole vector's time-channel gain: 0.010 nats over the reference, -0.006 over clock history (Table 1, row 5); clock history's own gain and the full-information ceiling (Appendix H) are not compared. Take-home: tempo is the most reliable, and timing prediction is partly amortizable, but clock history beats it on timing; ability beyond rating adds accuracy but agrees only weakly a year apart and is not yet amortized.

Trait	Channel	Split-half r	One-year r (matched, no gap)	over reference	added to clock history	over Allie
Tempo	time	0.79	0.62 (0.62)	vector†	vector†	–
Dispersion	time	0.48	0.10 (0.12)	vector†	vector†	–
Ability beyond rating	accuracy	0.60	0.22 (0.18)	+0.5 AUROC pts	+0.4 AUROC pts	+0.4 AUROC pts

Shared data. Both works use public data that this paper also uses. The published work above uses puzzle-linked positions located in the public 2019 Lichess game archives, with a 2024 replication cohort; both periods lie within this paper's months, and no estimate of theirs is reused here. The concurrent manuscript (Anonymous, 2026) uses the January–March 2019 Lichess game archives (ordinary decisions, not puzzle probes), also within this paper's months, and the same public four-in-a-row learning cohort as Appendix T, on which it reports that per-participant offsets lower later-session timing loss; the reliabilities and held-out gains of Appendix T are computed here independently with this paper's person model, and no estimate is taken from the manuscript. One data tier is shared with a concurrent submission to this conference that asks a different question: it asks whether a policy network's internal representations carry information about human difficulty and thinking time beyond an engine's search features and the policy's own outputs, and whether that information is used. Both papers draw on the public Lichess puzzle archive, its source games and the Maia-2 policies, and both use the January 2019 archive: it is a development tier here (Appendix B) and that submission's headline month. No data split, trained model, estimate, figure, table or passage of text is shared, and no quantity here is taken from that submission. Two further manuscripts by the authors are under review at other venues on related behavioural questions; one is supplied in anonymised form in the supplementary material, and nothing here depends on either. The one not supplied draws on the same public puzzle archive and Maia-2 policies; no split, fitted model, estimate, figure or text is shared.

B Data: from puzzles to source games to clocks, cohorts and a study map

This appendix supports the data question (Section 3): what a probe is, how it is assembled from public archives, which players carry which test, and which fitted object each result belongs to.

What a probe is. The puzzle database and the monthly game archives, joined, give timed decisions at items of calibrated difficulty by the same pseudonymous players over several years. Educational test logs also pair calibrated items with response times, but there the person knows the item is a test item; to our knowledge no public dataset pairs difficulty calibrated on other people with the same person's time and choice on that item, met unannounced in ordinary play and repeatedly over years. A Lichess puzzle is a position taken from a real rated game at the moment one player has just made a move that allows a forcing continuation for the other. The public puzzle database lists, for each puzzle, its source game and the ply of that move, the position, the solution line (the triggering move followed by the solution moves) and a Glicko-2 rating (Glickman, 2001) that the site updates from later attempts by other people (a median of 354 attempts per puzzle). The player who was to move in the source game met that position once, under their own clock and in an ordinary game, before it became a puzzle. That decision is a *probe*: an item whose difficulty is calibrated on other people's attempts, faced once by one person under real time pressure. For example, a blitz player to move in a position where the opponent's last move allowed a winning combination spends some seconds

Table 3: Elements of the evaluation by work, filled from Section 2 and this appendix: a check marks an element that the work reports; a blank means Section 2 does not attribute it to that work. Take-home: each element exists alone; the new evaluation question is whether a person term is valuable against clock history, with its reliability horizon and amortization, on one protocol.

Work	calibrated items	person-level timing	reliability with horizon	value against clock history, per channel	amortized person posterior	instrumented think-time effect
McIlroy-Young et al. (2022), per-player fine-tuning of a move model						
Maia4All (Tang et al., 2025), few-game embedding for move prediction					embedding	
Allie (Zhang et al., 2025), joint move and time model		population				
Poddar et al. (2024), user latent from a few comparisons					✓	
van der Linden (2007), speed and ability as person traits	estimated	✓				
Howard (2024), time control as an instrument, on moves						✓
Concurrent manuscript (Anonymous, 2026), per-account offsets		✓				
This paper	✓	✓	✓	✓	timing	✓

and either plays the puzzle's first solution move (solved) or not; the puzzle's rating, set later by other people, says how hard that position was. Items are selected: a probe exists only where the site's generator kept a forcing line after an error, a selection on the item side that is not modelled. The player's own rating and the puzzle's rating are on comparable scales, so their gap is a retrospective covariate: the item rating was set by later attempters, independently of the focal player's outcome, and was not available at the decision; this is what lets ability beyond rating be defined for one person, since the difficulty is estimated without that person. Nothing about timing is calibrated: the time channel is scored against the frozen reference of Section 4.1.

From puzzles to source games. Each month's archive of standard rated games is published as PGN with a clock comment after every move, in whole seconds. The archives are streamed, filtered to the games that are the source of a puzzle, and each puzzle's ply is used to locate its triggering move in its game. Source games were located in 51 monthly archives, January 2019 to January 2021 (without January 2020) and January 2023 to March 2025; the 23 months from February 2021 to December 2022 were not processed for the reported table, so it has no probes there (the 79-month rerun of Appendix Q fills most of that interval). Every puzzle has one source game and every located game gives one probe, 2,647,029 in all.

From games to clocks and outcomes. At the puzzle ply the pipeline reads the move the player actually played and the player's clock before and after it. The recorded think time is the player's previous own clock reading minus the clock after the move plus the increment, in whole seconds as the archive records it; the remaining clock, the increment, the move number, the player's two previous own think times and whether a recapture was available are read from the same game, and both players' ratings, the time control and the date from its header. The probe is *solved* when the move played equals the first solution move; this is the accuracy channel. Each row carries a quality flag (missing clock, a puzzle ply beyond the recorded game, a negative think time); flagged rows, each side's first own move and rows whose think time exceeds the remaining clock are excluded from every time likelihood (Appendix D). Maia-2 policy features are computed at the probe position for the training pool of Table 5, which is where the reference is fitted.

Centisecond clocks. A separate public archive carries centisecond clocks for a subset of the same games. It is used twice and never for a headline number: to measure the whole-second recording operator on January 2019 and June 2020 (Appendices C and C.1), and to refit the person model at centisecond resolution on the 13-month overlap of the two archives (Appendix V).

From probes to cohorts and splits. Per-player probe counts define the cohorts of Table 5. The training pool trains the cross-fitted reference, the zero-mass model and the encoder; the primary cohort carries the person fits. The primary cohort is strong (median rating 2,376), mostly blitz (85.8% of its probes) and self-selected by volume, with probes in a median of 28 distinct months per player. A *player* split holds out whole players for the few-shot tests; a *month* split fits through June 2024 and scores each player's own probes from July 2024 on, for the 429 primary players with enough probes on both sides. The puzzle database and the monthly archives are public Lichess releases under the CC0 licence; the supplementary code rebuilds the probe table from the games extracted from them (the extraction step itself is not supplied).

What is fitted on what. Table 4 lists, for each learned component or statistic, the players and months it is fitted on and whether held-out outcomes were excluded, from the records of this appendix and Appendices D, G, J and F and, for the months, from the producing scripts; "not recorded" marks a cell neither states, and no exclusion is claimed there. The population-level quantities are fitted once on all months and enter the held-out-month scoring fixed, so the months are held out for the person terms only; the evaluation is retrospective over a fixed reference, not prospective prediction, and the effect of this reuse was not quantified (Appendix F).

The January 2019 archive (26,352 probes; 26,320 games also in the centisecond archive) is a development tier here and a concurrent submission's headline month (Appendix A).

Study map. Table 6 names, for each result, the data it is computed on, the fitted object it belongs to and the score it is judged by. Every headline test uses one corpus and one clock resolution: Lichess probes with whole-second clocks. The transfer settings differ in data, events, rating range and clock resolution, and each refits its own reference. The cross-setting check uses Lichess players active in two time-control classes, with whole-second clocks and the full rating range (Appendix S). The four-in-a-row pilot uses the learning experiment of a published corpus of a second game, 30 players with recorded response times, no skill scale and no calibrated items, so only its timing channel is comparable (Appendix T). The over-the-board corpus has 276 players rated 2500 or more across tournaments of 2014 to 2017, whole-second clocks and no calibrated items (Appendix U). The centisecond refit uses the same Lichess games on the overlap months (Appendix V).

C THE WHOLE-SECOND RECORDING OPERATOR

Lichess publishes move clocks to the whole second in its standard monthly archives and to the centisecond in a separate archive of the same games. For January 2019, 26,320 games are present in both; 2,018,270 plies have matching moves and clocks in both series (27 plies in a single game differ only in annotation punctuation). The whole-second clock equals the centisecond clock rounded half-up on 100.00% of plies, uniformly across first and later moves and across all four time-control classes; the remainder $\mathrm{clkc}/100 - \mathrm{clk}$ is roughly uniform on $[-0.5, 0.5)$. So the recorded think time, the difference of two independently rounded clocks, differs from the *rounded* true think time by -1, 0 or $+1$ second, and from the true think time by a continuous error in $(-1, 1)$ (Table 7). For each rounding error d the windows of Eq. 2 over $k \geq 0$ partition the positive half-line (at $k = 0$ the $d = -1$ window is empty and the $d = +1$ window runs from zero, which gives C_i in Eq. 3), and the three masses sum to one, so the probabilities of the recorded times sum to one. For $k \geq 1$ the conditional distribution of the centisecond clock difference (lag compensation can make it negative) given a recorded k is close to symmetric within ± 0.47 s of k; for $k = 0$ it is not: mean 0.143 s, 5th/50th/95th percentiles $-0.29/0.00/0.63$ s, with 72.7% of the mass below 0.1 s and 37.7% negative (server-side lag compensation). The centisecond readings are not restricted to a 0.1-s grid (0.29% of values are multiples of 0.1 s) but 98.5% of values end in an odd digit, with a systematic $+0.03$ s offset on freshly issued clocks.

Each side's first move records no clock consumption in 57.7% of cases (a clock-initialisation artefact; berserked players start with half the base time and the header recording it survives only in the centisecond archive), and is excluded from every time likelihood; no puzzle probe in the table is a first move. On June 2020, the only later month used for this operator measurement, the round-half-up match is again 100.00% ($n = 2{,}429{,}506$) and the three discretisation masses differ from January 2019's by at most 0.05 points (every interval containing zero), but the first-move no-consumption share rises to 60.7% ($+3.0$ points [2.4, 3.6]), a drift we flag; June 2021, the archive's last month,

Table 4: Learned components and statistics by the players and months they are fitted on. Held-out players: the 370 training-pool and 101 primary-cohort players of the few-shot tests; held-out months: July 2024 onward for the 429 primary players of Table 1. Take-home: the person terms scored on held-out months are fitted on the fit period only; the encoder's weights are never trained on its held-out players, but the population-level quantities it and the comparators condition on (rows 2–4) were estimated with those players included; the population-level quantities are fitted once on all months, held-out months included, so the evaluation is retrospective over a fixed reference, not prospective prediction.

Component	Players	Months	Held-out outcomes excluded?
Cross-fitted reference heads (Appendix G)	training pool, 9,482 players, cross-fitted by player fold, so no player's own probes inform the heads scored on that player	all months	yes for the scored player (cross-fitted); no for other players' probes, held-out months included
Zero-mass logit, Eq. 5 (Appendix D)	training pool, 9,482 players (419,738 rows), the few-shot held-out fold included, refitted once at the population baseline	all months	no: its rows include the 429 players' held-out-month probes and the few-shot held-out players' probes
Class intercepts, pooled accuracy terms δ_0 and β (Eq. 4), class means and population SDs (the prior)	primary cohort, 545 players, the 101 few-shot held-out players included: a pooled fit with person effects at zero; class means and SDs from the per-player estimates	all months	no: fitted on all months, then held fixed when the person vector is scored on held-out months; the few-shot held-out players are in the fit
Standardisation of the reference outputs	training pool, the zero-mass logit's rows (its covariates); the encoder's inputs: the training pool, all 9,482 players, the held-out fold included	all months	no
Person vector, route 1, held-out-month rows	429 primary players with probes on both sides of the split	fit through June 2024	yes: scored on July 2024 onward only
Encoder and its teacher targets (Appendix J)	teacher: route-1 posteriors of the primary cohort; held-out likelihood term: the training pool; one held-out player fold (101 primary-cohort and 1,896 training-pool players)	not recorded	held-out players: excluded from gradient updates and from the teacher targets; the training players' targets are shrunk toward the class means and SDs of row 3, which include the held-out players; at evaluation the full-information ceiling sees the scored probes
GRU (Appendix J)	429 primary players: 386 for training, 43 for early stopping	fit period through June 2024	yes: no held-out-month probe is seen in training
Allie rows (Appendix J)	published checkpoint, frozen; its time-head offset, scale and excess-zero terms and the accuracy calibration fitted on the fit-period blitz probes (47,671 rows) of the same 429 players	fit period through June 2024 for the fitted terms; the checkpoint's training months: not recorded	yes for the fitted terms; the checkpoint's training games: not recorded

was not tested. Puzzle moves are *less* often instantaneous than ordinary moves, as expected for the critical move of a position. These measurements fix the censoring weights of Eq. 2, the rounding mass C_i of Eq. 3, and the recapture and low-clock terms of Eq. 5.

Triangular kernel (independent clock phases). If the two clock phases are independent and uniform, the true-minus-recorded error has the triangular density $1 - |x|$ on $(-1, 1)$, so $P(y_i = k) = \int f_T(t)\, \max(0, 1 - |t - k|)\, dt$ replaces the three-window mixture of Eq. 2, and for $k = 0$ only the right half is reachable. Against the operator table the triangular kernel's 5th and 95th percentiles (∓ 0.68 s) are wider than the empirical ones (∓ 0.46 to 0.47 s; maximum discrepancy 0.22 s): the two clock reads are more correlated than independence implies. Unlike the three-window mixture of Eq. 2, which ignores where the true time lies within its second, this kernel respects the $(-1, 1)$ bound; neither is exact for correlated reads. Refitting the person model under

Table 5: Cohorts. Probe counts are before quality flags; rating is the per-player median.

Cohort	Players	Probes	Med. rating	Use
Population (all)	901,854	2,647,029	–	descriptives; causal check
$\geq$ 23 probes (training pool)	9,482	423,472	2,027	cross-fitted reference; zero-mass model; encoder training and held-out players; causal check
$\geq$ 100 probes (primary)	545	93,776	2,376	person parameters; reliability; held-out months; identification; validity
$\geq$ 300 probes	43	17,934	2,367	within-person illustration only

Table 6: Study map: which result belongs to which fitted object. The person vector is (a_j, b_j, θ_j) plus the zero-mass offset p_j, reported at the population level. "Accuracy channel" is whether the probe is solved.

Result	Data	Fitted object	Score (observable)	Reported in
Reference and zero-mass model	training pool, 9,482 players; heads cross-fitted by player fold, zero-mass logit fitted once on all months (Table 4)	frozen contextual reference over Maia-2 features; zero-mass logit (Eq. 5)	held-out NLL, coverage; AUROC, Brier	Apps. D, G
Reliability (T1) and its horizon	primary cohort, 545 players; date-split halves; enforced gaps	per-person maximum-likelihood estimates in each half, unshrunk, with the population quantities of the two-stage empirical-Bayes fit held fixed, over the frozen reference	Spearman–Brown-corrected split-half r against a permutation null	Sec. 5.1; Apps. E, K
Held-out value (T2)	429 primary players; fit through June 2024, scored on later months	person vector as fixed offsets; clock-history GRU; stable offset; Allie (blitz)	NLL of the recorded whole-second time; CRPS (secondary); AUROC, Brier	Sec. 5.2; Table 1; App. J
Amortized posterior (T3)	held-out players: 370 of the training pool, 101 of the primary cohort	set encoder from k probes; posterior-mode and marginalised refits; full-information ceiling	NLL of the recorded time at fixed k	Sec. 5.3; Apps. H, J
Interventional check	slope: primary cohort, 545 players; double ML: training pool, 9,482; instrument: 5,174 of them, clock $\geq$40 s	within-person slope λ; double-ML slope; increment-instrument LATE for its compliers	solved, logit and probability scales	Sec. 5.4; Apps. F, P
Replications and transfer	79-month rerun; Maia-3 reference; two time-control classes; four-in-a-row; over the board; centisecond clocks	the same recipe, refitted per setting	as in the headline tests, per channel	Apps. Q–V

the triangular kernel leaves the tempo, dispersion and ability estimates nearly identical in rank ($r = 0.999$, 0.995, 1.000 with the reported fit) and moves only the excess-zero offset ($r = 0.91$); its gate-5 result is in Appendix O.

C.1 Stationarity of the operator

This check supports the question of what lies between a person's true think time and the record, and how it is measured rather than assumed (Section 4.2). Table 8 gives the June 2020 comparison summarised above. The masses are those of *rounded* true minus recorded think time, which takes exactly the values -1, 0 and $+1$ second; under the independence assumptions of the previous paragraph, the error against the continuous true time is approximately the triangular kernel. Round-

Table 7: Instantaneous-reply mass at centisecond resolution by slice, January 2019 (first moves of each side excluded). Rounded true minus recorded think time on all plies of the month, first moves included, is -1, 0 or $+1$ second with masses $11.64/76.55/11.81\%$ ($n = 2{,}018{,}251$).

slice	n	< 0.1 s	< 1 s
all non-first moves	1,965,611	1.80%	6.15%
player's puzzle move	26,136	1.01%	3.21%
recapture	175,474	10.07%	14.72%
non-recapture	1,790,137	0.99%	5.31%
bullet	18,186	6.06%	24.88%
classical	536,421	0.96%	2.27%

half-up matches on every ply in both months, and each mass moves by less than 0.05 points with an interval containing zero. The two tested months, January 2019 and June 2020, show the same rounding masses. A different share drifts. Each side's first move records no clock consumption in 57.67% of cases in January 2019 and 60.66% in June 2020, a rise of $+2.98$ points [2.41, 3.56]. First moves are excluded from every time likelihood, but the drift shows that this clock-initialisation artefact is not fixed across the archive; we leave it as an open caveat. June 2021 was not tested: its archive had not been scanned when the check was run.

Table 8: The whole-second operator in January 2019 and June 2020. Masses are shares of plies whose rounded true minus recorded think time is -1, 0 or $+1$ second ($n = 2{,}018{,}251$ plies in January 2019); June 2020 masses are reported only as differences from January 2019. Differences are in percentage points with bootstrap intervals.

quantity	January 2019	June 2020	difference
round-half-up match	100.00%	100.0000% [99.99984, 100.0000]	–
plies	2,018,270	2,429,506	
mass $w(-1)$	11.64%	–	+0.0173 [−0.0426, +0.0772]
mass $w(0)$	76.55%	–	+0.0321 [−0.0470, +0.1111]
mass $w(+1)$	11.81%	–	−0.0493 [−0.1095, +0.0109]
first move, no consumption	57.67% [57.25, 58.09]	60.66% [60.26, 61.05]	+2.98 [2.41, 3.56]
plies	52,640	59,999	

D LIKELIHOOD DETAILS, PRIORS, STANDARD ERRORS AND THE ZERO-MASS MODEL

Table 9: Approximate per-person standard errors at whole-second resolution, fixed before any fit. A trait's standard error shrinks with the square root of the player's probe count, and for ability it is largest when solving is near certain; the formulas assume independent probes and the model's residual distribution ($\sigma \approx 1.14$, the raw within-person SD of log think time on the January 2019 table, before any reference; $\mathrm{SE}(a_j) \approx \sigma/\sqrt{n}$, $\mathrm{SE}(b_j) \approx 1/\sqrt{2n}$, $\mathrm{SE}(\theta_j) \approx 1/\sqrt{n\,p(1-p)}$ at $p = 0.73$, $\mathrm{SE}(\lambda_j) \approx \mathrm{SE}(\theta_j)/\sigma$). Take-home: at the primary cohort's probe counts, tempo and ability are precise enough to interpret per player and dispersion marginally (Section 4.3); the instant-reply term, not tabulated, is read only as a spread across players.

probes n	$\mathrm{SE}(a_j)$, log units	$\mathrm{SE}(b_j)$	$\mathrm{SE}(\theta_j)$, logits	$\mathrm{SE}(\lambda_j)$, per log unit (not estimated)
45	0.17	0.105	0.34	≈ 0.29
100	0.11	0.071	0.23	≈ 0.20
300	0.066	0.041	0.13	≈ 0.11

The class intercepts $c_{0,c(i)}$ of Eq. 1 (indexed by the probe's time-control class $c(i)$) and the pooled accuracy terms δ_0 and β of Eq. 4 are fitted once on all months of the primary cohort's probes with the person effects at zero (Table 4); the prior's class $c(j)$ is the player's dominant class. Table 10 lists the frozen priors, used by the ADVI cross-check. The empirical-Bayes estimator uses no fixed hyperpriors: its Gaussian prior is estimated from the data. Its population variances are simulated-null moment estimates, $\hat{\tau}_k^2 = \max\{0,\ s^2(\hat{u}_k) - s^2(\hat{u}_k^{(0)})\}$, where $s^2(\hat{u}_k)$ is the unweighted across-player variance, about the grand mean, of the per-player maximum-likelihood estimates of component k, and $s^2(\hat{u}_k^{(0)})$ is the same variance for estimates refitted to one data set simulated with no person effects ($\tau = 0$) on each player's own design. This is an unweighted analogue of the moment estimator of DerSimonian & Laird (1986), not its inverse-variance-weighted form; because the variance is taken about the grand mean rather than within time-control classes it also contains between-class differences, whose effect is not reported here. Class means are precision-weighted averages of the per-player estimates, and shrinkage weights are $\tau^2/(\tau^2 + \mathrm{SE}_j^2)$ with a non-finite SE mapped to full shrinkage. The null variance is held fixed in the player bootstrap, so the intervals omit its Monte Carlo error. The frozen plan's right-truncation denominator $F_i(\log r_i)$ is not implemented; rows with think time above the remaining clock are excluded instead (532 rows of the full table, unflagged by the build script and masked directly); the exclusion does not supply the truncation denominator, and the fitted mass beyond each remaining clock was not computed.

ADVI cross-check. A full correlated fit by mean-field ADVI under the reported likelihood and reference agrees with the empirical-Bayes point estimates at $r = 0.99$, 0.99 and 0.98 per player for (a_j, b_j, θ_j) and only 0.71 for p_j, whose population SD the two estimators also put far apart. Agreement of point estimates does not validate the posterior variances, which were not compared with exact inference (Appendix X).

Scores. The NLL of Eqs. 2–3 is the proper score for the recorded whole-second integer and is the primary time score. The CRPS reported beside it is that of the continuous predictive (with its point mass at zero) evaluated at the recorded whole-second time, as pre-specified; it scores the latent continuous prediction, not the recorded integer, and it is secondary.

The stable-offset baseline (row 3 of Table 1). This random-intercept baseline follows the per-account offset logic of Anonymous (2026) and is fully specified here. For each player the tempo offset is estimated by per-player maximum likelihood on the fit-period probes, jointly with the dispersion and zero-mass offsets under Eqs. 2–3, with no population prior. One tenth of that raw tempo estimate, a fixed shrinkage factor that is not estimated, is added to the reference's location on every scored probe of the player; the scale and the zero-mass probability stay at the population baseline, and the accuracy channel carries no person term: it is the reference's logit with δ_0 and β, the recalibration row 4 also carries, so on accuracy rows 3 and 4 differ only by θ_j; row 1 is the bare logit α_i, and rows 1 and 3 share AUROC 0.787, so at the printed precision the person vector's accuracy gain over row 1 is its gain over row 3. It is thus a per-person constant on the time channel only, fitted on the training months.

Table 10: Frozen priors for the hierarchical model (ADVI cross-check). The p_j scale was not in the frozen table and was set to the order of magnitude of the first-pass moment estimate.

Parameter	Prior	Basis
pop. SD of a_j	Half-Normal(0.5)	covers the unadjusted between-player log-time SD ($\sqrt{0.38} \approx 0.62$)
pop. SD of b_j	Half-Normal(0.3)	simulation recovers dispersion SDs in $\{0, 0.1, 0.2, 0.3\}$ without artefact
pop. SD of θ_j	Half-Normal(1)	$\mathrm{SE}(\theta_j) \approx 0.34$ logits at $n = 45$ needs a comparable population SD
pop. SD of p_j	Half-Normal(3)	added at the cross-check (change log)
$\boldsymbol{\Omega}$	LKJ(2)	mild shrinkage from ± 1
fixed effects	$\mathcal{N}(0, 1)$	standardised covariates
contextual reference	frozen point estimate	uncertainty not propagated

Zero-mass model. Some replies are predictably instant (a recapture, time pressure, a player who has just replied instantly); the excess-zero logit adds these flags to the reference's three outputs, each standardised (centred and scaled), $(\tilde{m}_i, \tilde{s}_i, \tilde{\alpha}_i)$, and what remains is a person's propensity p_j. The excess-zero probability of Eq. 3 is

$$\operatorname{logit} \pi_i = \pi_{0,c(i)} + \gamma_{\mathrm{rec}} \, \mathrm{rec}_i + \gamma_{\mathrm{clk}} \mathbb{1}_{\mathrm{clock}<20\,\mathrm{s}} + \gamma_0 \mathbb{1}_{y_{i-1}=0} + \gamma_m \tilde{m}_i + \gamma_s \tilde{s}_i + \gamma_\alpha \tilde{\alpha}_i + p_j, \quad (5)$$

with rec_i the recapture-available flag; each coefficient shifts the log-odds of the extra zero component, other inputs held fixed. Eq. 5 was chosen by AIC among six specifications fitted by logistic regression of the zero indicator on the 9,482-player training pool (419,738 rows), with the reported reference's standardised outputs as candidate covariates: class intercepts and the recapture-available flag alone (AIC 179,389); plus a binary low-clock term below 20 s ($-11{,}089$); plus a continuous log-clock term instead ($-16{,}126$); plus "previous own time recorded zero" ($-4{,}747$); low-clock and previous-zero together ($-13{,}675$); and the chosen specification, low-clock, previous-zero and the standardised reference values $(\tilde{m}_i, \tilde{s}_i, \tilde{\alpha}_i)$ ($-51{,}020$), which already summarise clock, increment and Maia-2 predictability. Under Eq. 3 the same specification is refitted by the exact mixture likelihood (Table 11); with C_i carrying the physically necessary near-zero mass, the excess-zero intercepts fall by two to three logits. Through the first two leak-free passes the three reference covariates of this channel had remained the outputs of the second-pass reference, which contained the played-move probability; when this was noticed, the selection and the mixture fit were rerun on the reported reference, and the accuracy-logit coefficient moved from -0.24 to $+0.06$ and the location coefficient from -2.73 to -2.96: the leak had flowed through the accuracy term. The refit also calibrates better: on the primary cohort the population-only plug-in's zero-fraction gap falls from -2.1 to -0.8 points (worst clock bin, below 10 s, from -5.3 to -1.6), and the person fits move by $r = 0.995$ (tempo), 0.996 (dispersion), 1.000 (ability) and 0.82 (excess zeros) against the previous pass (Appendix E).

Table 11: Zero-mass coefficients (Eq. 5) on the 9,482-player training pool: the point-mass version (logistic regression, reported covariates), the mixture version (exact MLE with C_i at the population baseline) of the previous pass, whose reference covariates were the second-pass reference's outputs, and the reported mixture version with the reported reference's outputs.

term	point mass	mixture, previous pass	mixture, reported
π_0 bullet	-4.27	-5.74	-7.01
π_0 blitz	-4.40	-5.95	-6.98
π_0 rapid	-4.22	-5.49	-6.32
π_0 classical	-3.98	-5.09	-5.90
γ_{rec} recapture available	$+0.31$	$+0.55$	$+0.38$
γ_{clk} clock < 20 s	$+0.03$	-0.22	$+0.10$
γ_0 previous own zero	$+0.14$	$+0.12$	$+0.26$
γ_m (standardised location)	-2.16	-2.73	-2.96
γ_s (standardised log-scale)	$+0.29$	$+0.23$	-0.01
γ_α (standardised accuracy logit)	-0.01	-0.24	$+0.06$

Population structure (T6, the plan's descriptive population-structure analysis). The population SDs of the person vector on the primary cohort are 0.13 for tempo and 0.12 for dispersion (log units) and 0.26 for ability beyond rating (logits). Players faster than their context predicts are more accurate than their rating predicts: the tempo–ability correlation $\boldsymbol{\Omega}_{a\theta}$ is -0.13 corrected for attenuation (Spearman, 1904) and -0.09 $[-0.17, -0.01]$ raw. About a third of the raw value is mechanical, because solved decisions are faster: with the solved indicator in the location reference it is -0.06. These are descriptive, not tests.

D.1 Residual-family rows

The family of the residual is not a claim of this paper. Table 12 gives the held-out negative log-likelihood of five families with identical covariates on an 8,000-probe subsample of the January 2019 archive (five-fold cross-validation grouped by game). Within the naive treatment the largest three-parameter gain over the lognormal is 0.0044 nats per probe (sinh-arcsinh, fitted with a separately

written, correctly normalised density), below the provisional floor of Appendix I; within the censored treatment, differences between families are read between censored rows, and paired intervals for them were not computed (Appendix X). Within each recording treatment the generalised-inverse-Gaussian family is close to the lognormal; the Wald family is clearly worse. Naive rows score the continuous density at $y + \frac{1}{2}$ and censored rows the probability of the recorded whole-second integer, so naive and censored rows are on different scales and are not compared numerically. It was not repeated on the primary cohort with person terms (Appendix X).

Table 12: Held-out NLL per probe (nats) by residual family and recording treatment, January 2019, $N = 8{,}000$, difference from the naive lognormal with 95% bootstrap CI. Naive rows score a continuous density at $y + \frac{1}{2}$; censored rows score the probability of the recorded integer; read differences within one treatment only.

Family	Recording	ΔNLL [95% CI]
lognormal	naive	0 (ref.; 3.1638)
lognormal	censored	$+0.0023$ [$+0.0016$, $+0.0029$]
generalised gamma	naive	-0.0037 [-0.0062, -0.0014]
generalised inverse Gaussian	naive	$+0.0055$ [$+0.0014$, $+0.0101$]
generalised inverse Gaussian	censored	-0.0015 [-0.0051, $+0.0021$]
Wald	naive	$+0.0428$ [$+0.0324$, $+0.0542$]
sinh-arcsinh	naive	-0.0044 [-0.0068, -0.0020]
sinh-arcsinh	censored	-0.0069 [-0.0101, -0.0039]

D.2 Item-response calibration of the solve rate

Across ten quantile bins of $(E_i - R_i)/400$ on the full probe table the observed solve rate rises monotonically from 0.515 to 0.838 and the model-implied rate from 0.503 to 0.831 (gate 4). On the January 2019 archive the observed rate by gap bin is 0.855 (below -600 on $R_i - E_i$), falling to 0.473 (above $+400$), and by Maia-2 entropy tercile from 0.880 in the lowest to 0.591 in the highest.

E From the first pass to the frozen observation model and the reported reference

This appendix supports Sections 4.2 and 4.3: how the observation model reached its frozen form, what each revision changed, and the T1 reliability table of the reported fit (Table 13). A *pass* is one complete fit of the person model under one observation model and one reference; the *freeze* is the dated point after which the observation model was not changed, and every later revision is a leak fix or a refit of the zero-mass model, declared in Appendix F.

Table 13: T1: date-split reliability of the unshrunk per-half estimates (545 players, about 86 probes per half; Appendix F): raw split-half r, its Spearman–Brown correction with 95% bootstrap interval, the range of ten within-band permutation nulls, and the verdict. *Inconclusive under the earlier point-mass zero bin (Appendix E). Take-home: all three traits clear their margins across date-split halves; the halves meet with almost no gap, so this split measures agreement between adjacent chronological halves, and its horizon is in Figure 2A.

Component	raw r	corrected r [95% CI]	null range	margin	verdict
a_j tempo	0.660	**0.795** [0.746, 0.834]	$[-0.04, 0.08]$	0.5	material
b_j dispersion	0.320	**0.485** [0.382, 0.571]	$[-0.05, 0.05]$	0.3	material*
θ_j ability	0.430	**0.601** [0.525, 0.666]	$[-0.14, 0.08]$	0.3	material
p_j excess zeros	0.359	0.528	$[-0.04, 0.10]$	–	population SD only

Seven passes were run on the primary cohort, each logged in the change log before the next (Table 14). **v1**: reference without Maia-2 covariates trained on players outside all tiers; zero bin a

point mass with class intercepts and recapture; a latent-scale residual check showed a $+0.22$ offset for nearly every player. **v2**: Maia-2 covariates and player-wise cross-fitting of the reference; class location intercepts; the ADVI cross-check; the offset persisted ($+0.27$) and was shown to be an exact no-op under the two-stage estimator's centring (checked numerically) for any class-level intercept. **Offset resolution**: simulating recorded times from the fitted v2 model reproduced $+0.12$ of the $+0.27$ by the diagnostic's construction; the remainder was a zero fraction twice too high on normal clocks (6.3% observed against 13.4%), so gate 5 was redefined on the recorded scale. **v3**: the zero-mass logit of Eq. 5; clock-bin calibration within 2.7 points; tempo, dispersion and ability numerically invariant ($r > 0.9999999$ against v2); the remaining zero gap traced to the likelihood's pure point mass at zero against a simulator in which sub-half-second times also round to zero (about 4.6 points). **v4**: Eq. 3; likelihood and simulator agree term by term; the per-player discrepancy criterion clears; the zero gap halves again to -2.4 points. The observation model was frozen here. **Leak fix 1**: a check of the covariates' timing showed that the reference's strongest accuracy feature, Maia-2's probability of the move actually played, is post-decision (it equals the solution's probability when the solution move is played); it was replaced by the solution move's probability computed regardless of what was played. **Leak fix 2**: the same check showed the recapture flag defined from the played move to be post-decision as well; it was replaced by whether a recapture was available, and the zero-mass coefficients were refitted. **Eq. 5 refit (reported)**: a later check showed that the zero-mass logit's three reference covariates still read the second-pass reference, whose accuracy logit contained the played-move probability; the zero-mass selection and fit were rerun on the reported reference (Appendix D), and the person model was refitted. Under the three fixes tempo, dispersion and ability move by $r = 0.99$, 0.96 and 0.94 per player against the frozen v4 fit; the accuracy channel of the reference loses 5.5 AUROC points that were the leak; every freeze verdict is unchanged. Total wall-clock of the seven passes was under five hours of CPU.

F Deviations from the plan fixed before fitting

Items up to 17 September are in the plan's change log; later ones are listed here only.

1. T1 correlates the per-player maximum-likelihood estimates of each date-split half, before shrinkage and with the population quantities held fixed, where the plan named posterior means; the plan's change log records this unshrunk statistic from 16 September, the margins were applied to it, and the enforced-gap and matched-size analyses (Appendix K) use it too. T1 thus measures the repeatability of the per-person estimates, not of the shrunk values used for prediction or of the encoder's output; the shrunk values' correlation is not reported.
2. T2's fit period (through June 2024) applies to the person terms; the reference, the zero-mass logit, the class intercepts, the pooled accuracy terms and the prior were fitted once on all months and held fixed (Table 4); a refit of them on the fit period was not run.
3. The two-stage empirical-Bayes estimator, named in the plan as the fallback to NUTS, is the primary estimator for every reported fit; mean-field ADVI under the reported likelihood and reference agrees with it in point estimates at $r = 0.99/0.99/0.98$ per player (posterior variances and coverage were not compared); NUTS chains launched under an earlier observation model were superseded by the later revisions of the model and are not reported.
4. The contextual reference is cross-fitted by player within the 9,482-player training pool (Maia-2 covariates exist only there) rather than trained on an external pool; its log-scale head uses a single held-out split per fold; the accuracy head is trained and scored on the rows with an accuracy outcome; the leak-free refits cap each fold's training pool at 150,000 rows.
5. Two covariates used in the first four passes were post-decision, Maia-2's probability of the move actually played and a recapture flag defined from the played move; both were removed in a later revision and replaced by the solution move's probability and a recapture-available flag. The zero-mass logit's three reference covariates then remained the outputs of the second-pass reference until a later check; they were reselected and refitted on the reported reference. Every main-text fit-dependent number is re-reported on that basis; retained pre-refit columns (Table 22's time columns) and analyses marked as run on the earlier reference are historical diagnostics (Appendix E).

Table 14: The seven passes side by side on the primary cohort: Spearman–Brown-corrected T1 reliabilities, population SDs (moment estimator), the tempo–ability correlation, the reference's held-out-month NLL, the person-vector gain on held-out months, and the recorded-scale gate 5 (mean per-player discrepancy; zero-fraction gap).

	v1	v2	v3	v4 (frozen)	leak fix 1	leak fix 2	Eq. 5 refit (reported)
reference	no Maia-2, external pool	Maia-2, cross-fitted; played-move prob.	as v2	as v2	solution-move prob.	+ recapture-available flag	as leak fix 2
zero bin	point mass; class, recapture	as v1	point mass; Eq. 5	Eq. 3 with Eq. 5	as v4	as v4, refit	reported reference's outputs; reselected
T1 a_j / b_j / θ_j	0.724 / 0.474 / 0.581	0.753 / 0.411 / 0.541	0.753 / 0.411 / 0.541	0.797 / 0.472 / 0.541	0.788 / 0.482 / 0.590	0.786 / 0.464 / 0.601	0.795 / 0.485 / 0.601
b_j verdict	material	incon-clusive	incon-clusive	material	material	material	material
SD *a*/*b*/*θ*	0.103 / 0.102 / 0.242	0.103 / 0.110 / 0.251	0.107 / 0.109 / 0.239	0.132 / 0.121 / 0.239	0.132 / 0.122 / 0.256	0.128 / 0.118 / 0.262	0.134 / 0.121 / 0.262
$\mathbf{\Omega}_{a\theta}$ raw / corrected	-0.204 / -0.313	-0.117 / -0.184	-0.117 / -0.187	-0.103 / -0.161	-0.113 / -0.166	-0.096 / -0.140	-0.092 / -0.133
reference NLL, held-out months	2.037	1.984	1.900	1.845	1.866	1.864	1.866
person gain NLL / CRPS / AUROC	-0.036 / -1.9% / $+0.72$ pt	-0.019 / -0.7% / $+0.22$ pt	-0.013 / -0.7% / $+0.22$ pt	-0.008 / -0.8% / $+0.22$ pt	-0.009 / -0.9% / $+0.46$ pt	-0.012 / -0.6% / $+0.51$ pt	-0.010 / -0.7% / $+0.51$ pt
T2 verdict	not material (NLL only)	not material	not material	not material	not material	not material	not material
gate 5 (recorded scale)	–	$+0.086$; -6.2 pt	$+0.042$; -5.0 pt	$+0.024$; -2.4 pt	$+0.020$; -2.7 pt	$+0.026$; -2.8 pt	$+0.019$; -2.6 pt

6. The zero bin was changed three times before it was frozen (Appendix E); the first pass dropped the plan's low-clock term.

7. After the zero-mass refit the GRU rows, the ADVI cross-check, the solved-conditioned reference behind $\mathbf{\Omega}_{a\theta}$, the identification test and the triangular-kernel refit were rerun under the reported coefficients; only the time columns of Allie's block in Table 22 (rows 8–9), scored before the refit, remain from the previous pass. The refit moves the reference's held-out time NLL by $+0.0015$ nats, tempo by $r = 0.995$ per player and ability not at all.

8. Right-truncation at the remaining clock is not implemented; rows above the remaining clock are excluded (532 rows unflagged by the build script, masked directly).

9. The per-person speed–accuracy slope λ_j was not estimated; a population λ is (Section 5.4). A first estimate of it, regressed on a residual of a location model that itself conditioned on the solved indicator, was a null by construction and was retracted before the reported estimate.

10. Gate 5 was redefined on the recorded scale with a criterion stated before running; the original latent-scale check is retired.

11. A class-level location intercept $c_{0,c}$ was added to the pooled nuisance fit; it is an exact no-op under the two-stage estimator's centring (checked numerically) for the shrunk person estimates and is kept for interpretability.
12. The p_j population-SD prior for the ADVI cross-check (Half-Normal(3)) was not in the frozen table.
13. The population zero-mass coefficients were refitted once at the population baseline, not iterated against the per-player estimates; the AIC ablation was not repeated under the mixture likelihood.
14. The reported encoder is trained toward the reported person fits at full budget; its teacher is the empirical-Bayes posterior, not NUTS. Earlier T3 evaluations carried four defects, a shrinkage fallback that dropped players from the refit rows only, two under-shrunk refits (a spurious Hessian error, then a between-player floor) and a within-person floor that still over-shrank the few-probe refit and the ceiling; each was fixed and the evaluation rerun, the reported refit and ceiling are posterior modes under the population prior, the encoder was retrained from $k = 1$ in a later revision, and its paired gain over the reference, adopted then, is the primary descriptive contrast reported beside T3's decision rule; the adapted T3 criterion (distance to the full-information fit, Section 4) keeps its original verdict, not material, and later contrasts are exploratory (Appendix J).
15. T3 is an adapted criterion: its metric and comparator changed, its 0.5% margin and $k = 20$ did not. The plan scored the encoder by CRPS and coverage against NUTS posteriors, with a second sub-margin over per-player maximum-likelihood refitting at five probes; the paper scores it by the NLL of the recorded integer against the empirical-Bayes full-information fit, because NUTS was not run for the reported model, so the plan's own comparison was not made. On CRPS the encoder is also further from the full-information fit than the 0.5% margin (Table 20), so the verdict does not depend on the metric; the second sub-margin clears trivially because the unshrunk refit is degenerate (Appendix J). The T3 intervals use fewer player-bootstrap resamples than the plan's convention (the count is in each table's caption).
16. T5, the plan's validity test of ability beyond rating against later rating (Appendix W), was rerun on the reported fits with a player-cluster bootstrap; the first run's interval for the θ_j →accuracy gain was the spread over cross-validation repeats, not a sampling interval, and is withdrawn; that gain is labelled a reliability result.
17. The 9,482-player training pool's person vector was not fitted; population statements on that training pool rest on the cross-fitted reference and the zero-mass model.
18. The causal check restricts to probes with at least 40 s on the clock (75.4% of the training pool), uses year dummies for month and omits theme tags from its direct-adjustment covariates, benchmarks the robustness value on a reduced linear covariate set rather than the gradient-boosted nuisance, and uses the PGN evaluation swing rather than a recomputed engine evaluation for the negative-control outcome; its logit-scale estimates are pooled, not fixed-effects. In a later revision the instrument's specification dropped the two previous think times (descendants of the instrument) and moved from player to player×month fixed effects; the pre-stated specification gave $+0.069$ $[0.050, 0.087]$ with first-stage $F = 2{,}292$, and the negative-control outcome moved from the previous move's quality (which post-error slowing makes non-null, $+0.002$ $[0.001, 0.004]$) to the player's move two own moves (four plies) back, as the dated note's own revision prescribed. The dated note also stated a materiality rule for the instrumented effect on the logit scale; the paper reports the effect on the probability scale and does not score it against that rule. The robustness battery run later, not pre-stated, and its two caveats are reported once, in Appendix P; the exclusion restriction is checked, not proved.
19. The plan was an internal dated document, not a registry; the observation model was revised after seeing the data (Appendix E), and the dispersion verdict changed with it.

G CONTEXTUAL REFERENCE QUALITY AND THE MAIA-2 ABLATION

Table 15 scores the references on the same 419,962 held-out rows of the 9,482-player training pool (pooled over five player folds). Switching from the external training pool to in-cohort cross-fitting

with the same covariates is a wash. The first Maia-2 reference included Maia-2's probability of the move actually played. That feature is post-decision: when the played move is the solution it encodes the outcome, and no forecaster could use it before the move is made. With it, the Maia-2 covariates appeared to add 13.7 AUROC points to the accuracy channel. It was retired, and every Maia-2 covariate of the reported reference is computed from the position before the decision. The six are Maia-2's policy entropy, top-move probability, top-two margin and solution-move probability at the decision position, and its probability of the opponent's actual move and entropy one position earlier. With the leak fixed, the six Maia-2 covariates add 8.2 AUROC points as a block to the accuracy channel and 0.045 to the location R^2 (reported reference against the cross-fitted reference without Maia-2); no attribution to a single covariate is made. The location channel is nearly untouched by either fix (R^2 $0.541 \rightarrow 0.528$). The log-scale head is cross-fitted on held-out residuals with a single additive calibration constant; its 90% coverage on the reference pool is 0.908. The leak-free refits capped each fold's training pool at 150,000 rows (the earlier passes used about 280,000) to bound the fitting time per fold; held-out rows are never subsampled, and the cap is applied identically to every leak-free variant. Cross-fitting guards the reference layer only: the Maia-2 and Maia-3 backbones are frozen as published, and their training games may overlap the probe months, a form of contamination that no player or month split can remove (Magar & Schwartz, 2022).

Table 15: Contextual reference quality on the 9,482-player training pool (pooled held-out folds): location R^2, Gaussian NLL in nats per probe, 90% coverage, and AUROC, Brier and log loss for the accuracy channel (whether the probe is solved). Rows marked retired contain a post-decision feature and are shown for the record.

Reference	R^2	NLL	cov. 90%	AUROC	Brier	log loss
external pool, no Maia-2	0.487	1.086	0.908	0.723	0.172	0.518
cross-fitted in the training pool, no Maia-2	0.482	1.092	0.905	0.717	0.174	0.522
Maia-2 with played-move prob. (retired)	0.541	1.025	0.908	0.854	0.133	0.402
Maia-2 with solution-move prob.; played-move recapture flag (retired)	0.526	1.053	0.909	0.806	0.149	0.454
Maia-2 with solution-move prob.; recapture-available flag (reported)	**0.528**	**1.054**	0.908	**0.799**	**0.151**	**0.461**
Maia-2 contribution (reported − cross-fitted, no Maia-2)	+0.045	−0.038	+0.003	+0.082	−0.022	−0.061

H A MODEL-BASED BENCHMARK FOR THE GAIN FROM A KNOWN TRAIT

This appendix supports the value question (Section 5.2): how large a per-decision gain a stable trait could give if it were known exactly, under the fitted variance components. For a location trait with between-person variance τ^2 and within-person residual variance σ_w^2 around a reference, the predictive distribution of a new probe has variance $\sigma_w^2 + \tau^2$ when nothing is known about the person and σ_w^2 when the trait is known exactly, so the expected gain in log score is $G_\infty = \frac{1}{2}\log(1+\tau^2/\sigma_w^2)$. After k own probes averaged as a sample mean, the posterior variance of the trait is $\tau^2\sigma_w^2/(k\tau^2 + \sigma_w^2)$. More probes improve prediction with diminishing returns toward a finite maximum, and k^* is the probe count at which half of that maximum is reached; to first order the expected gain is

$$\text{gain}(k) \approx G_\infty \frac{k}{k + k^*}, \tag{6}$$

the Spearman–Brown form again, with $k^* = \sigma_w^2/\tau^2$ for that estimator. The value is a property of a Gaussian location model and of that estimator; it is a reference point for the fitted model, not a limit that every estimator is shown to respect.

Tempo, on the reported reference. On the primary cohort the between-person tempo variance is $\tau_a^2 = 0.018$, and the within-person residual variance after the reported reference and each player's own tempo is $\sigma_w^2 = 0.405$ [0.395, 0.415], a residual SD of 0.64. This gives 0.022 nats. The value is not on a latent scale: σ_w^2 is the variance of recorded residuals (the log of the recorded time plus one half, not deconvolved), combined with the model's τ_a^2, so it is an approximate Gaussian

benchmark. The one-way intraclass correlation of the raw residual, which keeps rounding and the zero bin in its denominator, is 0.025 [0.020, 0.031], which gives 0.013 nats, also a Gaussian location-model approximation on recorded residuals, not derived under Eqs. 2–3. The two values differ in their between-person variance (the model's τ_a^2 against the smaller one-way ANOVA component implied by the intraclass correlation), their within-person variances being nearly equal, not by a conversion between latent and recorded scales. Both variance components are computed on the reported reference; the same code run on the earlier reference, which contained a post-decision feature (Appendix E), reproduces the values printed before that fix to the last digit, so only the input lineage differs. Two within-person SDs appear in this paper: Table 9's $\sigma \approx 1.14$ is the raw SD of log think time before any reference, the plan's design number, which an earlier evaluation used as a floor for the refit's standard error (Appendix J; the reported refit is a posterior mode and uses no floor); $\sigma_w = 0.64$, used here, is the residual SD after the reference and the person's tempo.

What the estimators reach. The encoder's gain curve on 370 held-out players fits Eq. 6 with $G_\infty = 0.013$ nats and $k^* = 1.2$ (Figure 2C), at the ANOVA-based benchmark: a consistency check, not a confirmation, since the encoder's asymptote also carries the dispersion and excess-zero terms while the benchmark is for tempo alone. The point refit's curve is negative through $k = 3$, within noise of zero from 5 to 20 and 0.009 at 50, and admits no saturating fit. The full-information ceiling's gain is 0.032–0.035 nats, above the tempo value because it also carries the dispersion and excess-zero terms and because it is fitted on all of the player's probes, the scored ones included. The 0.02-nat margin of the value test lies between the tempo benchmark (a Gaussian benchmark for tempo alone from the ANOVA variance share of recorded residuals, not computed under the recording operator and not a proven limit) and that in-sample gain, so, on that benchmark, clearing it would take timing terms beyond tempo (dispersion and the instant-reply term) to add to it; ability beyond rating enters the accuracy channel, not this time-channel score; the realised 0.010 is below both (Section 5.2). The variance-component value for tempo is therefore a model-based benchmark, not a bound that estimators are shown to respect: a tempo-only empirical-Bayes curve on the primary cohort, scored on the recorded scale against the reference before the leak fix and on random rather than held-out splits, already reaches it within the fitted probe range and its fitted asymptote extrapolates beyond it, past the largest fitted probe count (a descriptive analysis of the same probes, Appendix W; its scale and lineage differ from the held-out rows and its values are not quoted). Neither that asymptote nor the ceiling is compared with the clock-history GRU's gain of 0.037 [0.032, 0.042] over the reference (Section 5.2): the ceiling comes from other players under a player split, a different reference level and in-sample scoring, so an overlap of intervals would support no conclusion either way. The paper claims only that on the time channel clock history beats the fitted person vector.

A failed prediction. The pre-stated prediction that a fitted k^* falls within a factor of two of $\sigma_w^2/\tau_a^2 \approx 23$ fails: k^* is a property of the estimator (its fitted rate is about twenty times that of a sample mean; the encoder also carries other traits and context, so the mechanism is not isolated), whereas G_∞ is a property of the trait. Dispersion and ability add their own terms of the same form.

I Recoverability simulation at whole-second resolution

A simulation fixed in the dated internal plan before any fit generated log times from a four-parameter family that the plan calls sinh-arcsinh (in fact Johnson's S_U family; see the correction below) with a person location SD of 0.3 and a person log-dispersion SD $\tau \in \{0, 0.1, 0.2, 0.3\}$, recorded them to whole seconds with a floor operator ($T \in [k, k+1)$, not the measured ± 1 mixture of Eq. 2) and a 1.6% zero-inflation, and fitted them by the two-stage empirical-Bayes estimator of Section 4.3, cross-checked by mean-field ADVI on a subset. Results (Table 16): $\tau = 0$ is recovered as ≈ 0 with the unclipped estimator straddling zero (no positivity-prior artefact); $\tau = 0.1$ is separated from zero but adds only 0.0007 nats per instance of held-out likelihood; $\tau \geq 0.2$ is clearly recoverable; whether its gain is material depends on the predictive margin. Person location correlates 0.76–0.79 with truth at $n = 45$ regardless of τ and 0.97 at $n = 300$; person log-dispersion correlates 0.71 at $\tau = 0.2$, $n = 45$, rising to 0.95 at $n = 300$; person skew is not recoverable ($r = 0.18$–0.30, population estimate unstable across seeds). The plan's naive (uncensored) comparisons, the scale, τ and skew biases of naive fits and the naive whole-second outlier of the coarsening check, are withdrawn: the naive fits used a density that is not the generator's (see the correction below). A textbook non-degenerate two-parameter shape is worth about 0.06 nats per instance under censoring

against the simulation's two-parameter baseline, which was the S_U member described below rather than a lognormal; the observed real-data family gaps (Table 12) are an order of magnitude smaller, which sets the 0.01–0.02 nats floor below which no family claim is made, a provisional floor until that calibration is repeated with a lognormal baseline. Whole-second censored fits recover scale, skew and τ within about 15% of centisecond fits of the same latent times. Power for a covariate effect of -0.05 on log dispersion at $N = 26{,}000$ is 95% at a 5% false-positive rate. The fitted cohort's population SDs (0.13 for tempo, 0.12 for dispersion) sit at the simulation's marginal-to-material boundary, which is why dispersion's reliability verdict was the one sensitive to the zero model. Repeating the sweep under the measured operator instead of the floor, at the simulation's own sizes ($J = 2{,}000$, $n = 45$, three seeds), lowers $\hat{\tau}$ by 3–4% at $\tau > 0$ and the held-out gain by 5–7% at $\tau \geq 0.2$ and the correlations with truth by at most 1.4 points, and changes none of the verdicts above; the repeat at the exact cohort sizes, which the freeze requires, has not been run (Appendix X).

Correction (later check of the simulation library). The library's generator and its interval-censored likelihood agree with each other, but they define Johnson's S_U family for log time, in which $\delta \operatorname{asinh}((v-\mu)/\sigma) - \epsilon$ is standard normal, not the standard sinh-arcsinh family; with $\epsilon = 0$ and $\delta = 1$ this member is not the lognormal and has much heavier log-time tails, so "lognormal truth" in the plan's study means this member. The library's continuous density carried an extra hyperbolic-cosine factor and does not integrate to one. Only the naive fits (withdrawn above) and the naive sinh-arcsinh fits of Appendix W used that density; the censored fits, the held-out gains, the coarsening and power results and Table 17 use the distribution function and do not depend on it. The generator also caps log time at a fixed ceiling; at the baseline parameters the capped share of draws is small but larger than the code comment states, and it forms an atom that the likelihood does not model. The censored results above therefore show recovery within the family that generated the data, which is heavier-tailed than the plan describes; a repeat with a corrected and tested family has not been run (Appendix X).

Table 16: Recovery of the person log-dispersion SD τ from floor-recorded whole-second data and correlation of per-person estimates with truth. Rows 1–4: $J = 2{,}000$, $n = 45$; censored $\hat{\tau}$, held-out gain and correlations: means of three seeds; naive $\hat{\tau}$: one seed. Rows 5–6: $J = 300$, two seeds. †Withdrawn (text): the naive fits used a density that is not the generator's; kept only as a record of an earlier version of this table.

τ true	$\hat{\tau}$ censored	$\hat{\tau}$ naive†	held-out gain (nats)	$r(\hat{a}, a)$	$r(\hat{b}, b)$
0.0	0.011	0.000	−0.0001	0.79	–
0.1	0.094	0.069	+0.0007	0.79	0.46
0.2	0.196	0.164	+0.0076	0.78	0.71
0.3	0.297	0.253	+0.0233	0.76	0.84
0.2, $n = 100$	–	–	–	0.90	0.86
0.2, $n = 300$	–	–	–	0.97	0.95

I.1 Recovery under the measured operator

This check supports the question of how the person is estimated and which components are individual claims (Section 4.3). Table 17 repeats the sweep of Table 16, recording the simulated times with the measured whole-second operator of Appendix C instead of the floor. Everything else is unchanged: $J = 2{,}000$ players, $n = 45$ probes each, and three seeds per condition. The floor arm reproduces Table 16's $\hat{\tau}$ and held-out gains; two correlations differ in the second decimal (0.7668 against 0.76 and 0.7155 against 0.71). The measured operator lowers every quantity slightly and in the same direction. $\hat{\tau}$ falls by 3–4% relative at $\tau > 0$, and the held-out gain by 5–7% relative at $\tau \geq 0.2$. The correlation of the location estimate with truth falls by 1.0–1.4 points, and that of the dispersion estimate by 0.5–0.6 points; at $\tau = 0.2$ and 0.3 the location intervals narrowly fail to overlap. The bias of either estimate is the same under both operators. No verdict changes: $\tau = 0$ is recovered as null, $\tau = 0.1$ is marginal, and $\tau \geq 0.2$ is clearly recoverable (recoverability, not the 0.02-nat value margin); location is well recovered at every τ, and dispersion is usable once $\tau \geq 0.2$. The rerun used the simulation's own sizes, not the cohort sizes of the reported T3 fits (101 and 370

players). It therefore closes the corresponding item of Appendix X only partially. It uses the same generator and censored likelihood as Table 16, so the correction at the end of Appendix I applies: the recovery shown is within the S_U family that generated the data (no naive fit is involved here).

Table 17: Recovery under the measured operator (m) and the idealised floor operator (f), $J = 2{,}000$ players, $n = 45$, means of three seeds. Held-out gain in nats per instance. $r(\hat{a}, a)$ and $r(\hat{b}, b)$: correlation of per-person location and log-dispersion estimates with truth; –: undefined at $\tau = 0$.

τ true	$\hat{\tau}$ (m / f)	held-out gain (m / f)	$r(\hat{a}, a)$ (m / f)	$r(\hat{b}, b)$ (m / f)
0.0	0.0044 / 0.0107	-0.0004 / -0.0001	0.7793 / 0.7889	–
0.1	0.0900 / 0.0939	$+0.0004$ / $+0.0007$	0.7773 / 0.7878	0.4502 / 0.4561
0.2	0.1904 / 0.1960	$+0.0071$ / $+0.0076$	0.7683 / 0.7804	0.7102 / 0.7155
0.3	0.2886 / 0.2971	$+0.0221$ / $+0.0233$	0.7530 / 0.7668	0.8328 / 0.8388

J Encoder, GRU and Allie details

Set encoder (route 2). Per-probe input: fourteen standardised context values (the reference's m_i, s_i, α_i, puzzle rating, the rating gap, log remaining clock, increment, move number, the previous own log time, the recapture-available flag and four class indicators), $\log(y_i + \frac{1}{2})$ standardised, and f_i. Encoder: per-probe MLP $16 \to 64 \to 64$, masked mean and max pooling (a DeepSets network (Zaheer et al., 2017); the neural statistician of Edwards & Storkey (2017) likewise maps a set of observations to a posterior over a set-level latent), head $128 \to 64 \to 12$ giving the mean, diagonal variance $\boldsymbol{v}$ and rank-one factor $\boldsymbol{w}$ of a Gaussian over $\boldsymbol{u}_j = (a_j, b_j, \theta_j, p_j)$ with $\boldsymbol{\Sigma}_j = \mathrm{diag}(\boldsymbol{v}) + \boldsymbol{w}\boldsymbol{w}^\top$ (the diagonal carries each trait's own uncertainty and the rank-one term the uncertainty the traits share), so it can carry the tempo–ability covariance, the only correlation T6 reports; a rank-one term also constrains the signs of the other correlations; a full-rank head was not run (Appendix X). Mean and max pooling carry no count of probes, so the output does not by construction tighten as evidence is duplicated. Objective: the teacher is the route-1 empirical-Bayes posterior from all of the player's probes, a full-information fit, so the target is not the posterior given the same k probes; the loss is the KL from q_ϕ to that posterior (mean and marginal variance $\mathrm{SE}_j^2\tau^2/(\mathrm{SE}_j^2 + \tau^2)$) on the primary cohort with $k \sim \mathrm{Uniform}\{1, \ldots, 100\}$, plus the decoder-free held-out censored likelihood of the remaining probes on the 9,482-player training pool under Eqs. 2–4 with a sampled $\boldsymbol{u}_j$ (the objective thus matches the teacher's posterior and also scores held-out prediction, and the two can pull apart); weights 1:1, Adam, eight minutes per seed on CPU, three seeds, one held-out player fold (101 primary-cohort and 1,896 training pool players; a five-fold sweep was not run, so the intervals carry no fold variance). Held-out players are excluded from both loss terms, but the population-level quantities that the teacher, the likelihood term and the input standardisation use were estimated with them included (Table 4); the comparators share them, and their removal was not run (Appendix X). A step whose loss or gradient is non-finite is skipped (0.3–0.4% of steps, traced to $\log(e^a - e^b)$ as $a \to b$ in the censored interval and fixed at the source as well). Weight sensitivity at seed 0 (run on the earlier reference only): the likelihood-only objective beats 1:1 by 0.01–0.02 nats at every k and the KL-only objective trails by the same; the KL term is kept because without it the encoder's posterior departs from the hierarchical one (validation KL diverges).

T3 evaluation. The plan's T3 has a second criterion besides the distance to full inference: a margin over per-player refitting at five probes. It clears trivially, because the unshrunk per-player refit is degenerate at twenty probes or fewer, so the first criterion, which is not met, decides T3. We score a fixed set of held-out players with at least 51 probes, so that every player is scored at every k (370 of the training pool fold, 101 of the primary cohort); a random k of the player's probes are encoded and up to 150 of the rest scored (the same split for every estimator). Estimators on the same k probes: the reference alone; the stable offset with shrinkage 0.1; the unshrunk per-player maximum-likelihood refit; the point refit (posterior mode), which maximises the frozen likelihood of the k probes plus the log of the Gaussian population prior $\mathcal{N}(\boldsymbol{\mu}_{c(j)}, \tau^2)$ jointly over (a_j, b_j) and separately over θ_j, with p_j held at its class mean (its population variance is about 1,400 times tempo's, so a few-probe estimate of it would be unconstrained by the prior); the full-information fit (in-sample ceiling), the

posterior mode from all of the player's probes, the scored ones included; and the encoder (a three-seed Monte-Carlo ensemble of 60 posterior samples). The unshrunk refit is numerically degenerate at low k: for a few players the excess-zero logit diverges on one to ten probes and drives one held-out probability to underflow, so its mean NLL is astronomical (up to 10^{62} at $k = 3$); the stable offset inherits part of it. Both are reported as computed; the point refit (posterior mode) is the fair classical comparison. A zero-probe control, the reference plus the population class-mean person vector, is worse than the reference alone on the training pool at every k ($+0.0070$ to $+0.0082$ nats, intervals excluding zero) and tied with it on the cohort, so the encoder's gain net of it (-0.014 to -0.022 nats on the training pool, -0.005 to -0.011 on the cohort) is not a constant training pool-level recalibration. The marginalised refit (Laplace), the point refit averaged over 60 draws of a Laplace approximation about the same mode, as the encoder is marginalised, improves on the point refit at every k and shrinks the encoder's edge over it by roughly half on the training pool (-0.005 nats at $k = 1$ and $k = 20$), where the three-seed ensemble's edge excludes zero through $k = 20$ and closes by $k = 50$; on the primary cohort its interval excludes zero only at $k = 5$ and 10, unadjusted over seven probe counts, so there the encoder is not distinguishable from a marginalised refit. The marginalised ceiling moves by at most 0.0011 nats and the $k = 20$ gap closure by under one point (training pool 38%, cohort 34%). Four defects of earlier evaluations were identified and fixed before the reported run. The shrinkage fallback for a non-finite raw estimate multiplied zero by a non-finite value, so the shrunk and full-information refits returned no value for some players (the classical-class players of the training pool, whose class mean of the excess-zero term is undefined in a class of one), and the bootstrap silently averaged those rows over fewer players than the other rows, which flattered the refit by about 0.05 nats; the refit's per-player standard error from a finite-difference Hessian could be spuriously small on a near-degenerate few-probe fit, under-shrinking it; and the first correction floored that error at $\tau/\sqrt{k}$, the between-player SD, about nine times too small for tempo, so the refit still lost to its own prior mean by 0.03 nats at $k = 2$–5. A later evaluation floored the error instead at the within-person $\sigma/\sqrt{k}$ with Table 9's $\sigma = 1.14$, the raw SD before any reference and about 1.8 times the post-reference residual SD, and let the excess-zero logit, whose few-probe estimate is heavy-tailed, enter with a non-zero weight; that refit still lost to its own prior mean by $+0.007$ to $+0.008$ nats at $k = 1$–10 with intervals excluding zero, and the same floor shrank the full-information posterior (a median of 72 probes, where $1.14/\sqrt{72} \approx \tau_a$) by $+0.004$ nats at $k = 20$, which read as a 0.8% gap. The reported evaluation replaces the shrunk refit and the ceiling by posterior modes, which need no standard error and no floor and cannot diverge; it uses the pooled mean for a class with fewer than three players, scores every estimator on one common player set per k (all 370 and all 101) and reports the paired gains (Table 19). The refit's loss is then $+0.005$ to $+0.006$ nats at $k = 1$–3 only and the ceiling is 0.0015 nats better at $k = 20$ (2.241 against 2.243), so the encoder's gap to it is 0.9% on the training pool and 0.8% on the cohort. The encoder never calls the refit code, so its paired gain over the reference is unchanged to four decimals. The evaluation is re-run once more on the reported encoder itself (trained from $k = 1$), so the encoder-minus-refit column of Table 19 is paired on the same checkpoints as encoder-minus-reference throughout the table. The accuracy channel is a near wash across the four non-degenerate estimators (AUROC 0.79–0.81 at every k, the rating gap dominating), and CRPS, coverage and AUROC at $k = 20$ are in Table 20.

Training seeds. The encoder is trained with three seeds under the same eight-minute wall-clock budget, so the seeds stop after different numbers of steps; the reported encoder is their ensemble. Every interval in Tables 18 and 19 resamples players, not training seeds or player folds, so it excludes training-seed and fold variability, and no deterministic reproduction is claimed. Table 21 evaluates each seed's checkpoint alone against the marginalised refit on the same players and splits (evaluation only; in each run every non-encoder row reproduces the reported run to within 10^{-6} nats). On the primary cohort the advantage at $k = 5$ and 10 appears in one of three training seeds and in their ensemble: seed 0 changes sign at $k = 10$, and seeds 0 and 1 are not distinguishable from the marginalised refit at any of the three probe counts. On the training pool seed 2 excludes zero at $k = 5$, 10 and 20, seed 1 at $k = 5$ only and seed 0 at none.

Provenance and cost. The encoder above is trained toward the posteriors of the reported person fits at the full budget of three seeds and eight minutes each on CPU; a GPU was measured two to three times slower for this small model, whose training step is bound by per-player Python work. Encoders trained toward the frozen v4 fit, whose reference contained the two retired features

Table 18: T3: NLL of the recorded time in nats per probe on a fixed set of held-out players after encoding k probes, three seeds, every estimator on the same players (top: the 370 training pool players; bottom: the 101 primary-cohort players); the paired differences are in Table 19. Reported person fits, full budget; refit and ceiling at the posterior mode under the population prior.

Estimator	$k = 5$	$k = 10$	$k = 20$	$k = 50$
9,482-player training pool, 370 players				
contextual reference alone	2.276	2.282	2.274	2.292
stable offset, shrinkage 0.1	2.732	2.305	2.269	2.285
per-player maximum likelihood, unshrunk	1×10^{22}	3×10^{14}	54.3	2.291
point refit (posterior mode) from the k probes	2.280	2.284	2.270	2.282
set encoder, route 2	**2.265**	**2.271**	**2.261**	**2.278**
full-information fit (in-sample ceiling)	2.243	2.247	2.241	2.258
primary cohort, 101 players				
contextual reference alone	1.949	1.953	1.947	1.972
stable offset, shrinkage 0.1	2.125	1.955	1.943	1.966
per-player maximum likelihood, unshrunk	9×10^{15}	1×10^{3}	2.046	1.989
point refit (posterior mode) from the k probes	1.949	1.951	1.945	1.966
set encoder, route 2	**1.940**	**1.942**	**1.939**	**1.964**
full-information fit (in-sample ceiling)	1.921	1.925	1.924	1.946
fitted law, encoder on the training pool: $G_\infty = 0.013$ nats, $k^* = 1.2$ (reference values 0.013, 0.022)				
point refit: no saturating fit				

Table 19: T3: paired bootstrap differences in NLL (nats per probe, 95% intervals, 300 resamples over players) at every k on the fixed player sets of Table 18: encoder minus reference, encoder minus point refit (posterior mode) and refit minus reference on the training pool set; encoder minus reference and refit minus reference on the cohort set. Negative is better for the first-named estimator.

9,482-player training pool, 370 players			
k	encoder − reference	encoder − refit	refit − reference
1	−0.007 [−0.012, −0.002]	−0.013 [−0.018, −0.008]	+0.006 [+0.003, +0.009]
2	−0.007 [−0.011, −0.003]	−0.012 [−0.016, −0.009]	+0.005 [+0.002, +0.008]
3	−0.010 [−0.014, −0.007]	−0.015 [−0.018, −0.011]	+0.005 [+0.001, +0.008]
5	−0.011 [−0.016, −0.007]	−0.015 [−0.018, −0.012]	+0.004 [−0.002, +0.008]
10	−0.011 [−0.014, −0.008]	−0.013 [−0.017, −0.011]	+0.002 [−0.001, +0.006]
20	−0.013 [−0.017, −0.008]	−0.009 [−0.012, −0.007]	−0.003 [−0.010, +0.002]
50	−0.014 [−0.021, −0.007]	−0.004 [−0.008, +0.000]	−0.009 [−0.018, −0.001]
primary cohort, 101 players			
k	encoder − reference	refit − reference	
1	−0.005 [−0.011, +0.001]	−0.001 [−0.006, +0.004]	
2	−0.009 [−0.016, −0.003]	−0.002 [−0.007, +0.003]	
3	−0.009 [−0.016, −0.002]	−0.002 [−0.007, +0.003]	
5	−0.009 [−0.016, −0.003]	−0.000 [−0.006, +0.004]	
10	−0.011 [−0.017, −0.005]	−0.002 [−0.008, +0.005]	
20	−0.008 [−0.013, −0.003]	−0.003 [−0.009, +0.004]	
50	−0.008 [−0.014, −0.003]	−0.006 [−0.014, +0.000]	

(Appendix E), and toward the previous pass's fits give the same picture under the corrected evaluation (paired gains over the reference of -0.013 and -0.012 nats at $k = 5$, -0.017 and -0.016 at $k = 20$); the first evaluation had shown the leaky encoder 1–2% ahead of an under-shrunk refit, and neither is reported as the method's result. Per player on one core (medians over the 370 players), the three-seed ensemble with 60 posterior draws takes 1.3 ms at $k = 20$, the earlier shrunk refit, a joint L-BFGS fit of tempo, dispersion and excess zeros with numerical standard errors plus the ability fit, 27 ms, and a full-information refit on a median of 72 probes 23 ms (the point refit (posterior mode), which needs no standard errors, was not timed). The teacher is the empirical-Bayes

Table 20: Held-out players, $k = 20$ probes encoded (route 2): 370 held-out players of the 9,482-player training pool, the rest of each player's probes scored; NLL per probe, CRPS in seconds, PIT coverage at 50/90%, AUROC for whether the probe is solved. The ceiling is fitted on all of the player's probes, the scored ones included. Bold: the numerical best, not a test.

Model	NLL (nats)	CRPS (s)	cov. 50/90%	AUROC
Maia-2 reference	2.274	2.793	0.50/0.90	0.795
+ point refit from the k (posterior mode)	2.270	2.798	0.51/0.90	0.796
+ marginalised refit (Laplace)	2.266	–	–	–
+ set encoder from the k, route 2	**2.261**	2.790	0.52/0.90	0.796
+ full-information fit (in-sample ceiling)	2.241	2.759	0.52/0.91	0.805

Table 21: T3 by training seed: encoder minus the marginalised refit (Laplace), NLL of the recorded time in 10^{-3} nats per probe with 95% player-bootstrap intervals (300 resamples), on the fixed held-out players of Table 18. Negative favours the encoder. Unadjusted over the probe counts shown.

	primary cohort, 101 players			training pool, 370 players		
checkpoint	$k = 5$	$k = 10$	$k = 20$	$k = 5$	$k = 10$	$k = 20$
seed 0	-1.4 [-7.7, $+4.4$]	$+2.5$ [-3.2, $+8.0$]	$+0.2$ [-5.3, $+5.8$]	-2.6 [-6.5, $+1.7$]	-0.5 [-4.8, $+3.5$]	$+0.4$ [-3.8, $+4.7$]
seed 1	-1.7 [-5.9, $+2.0$]	-0.0 [-4.0, $+3.5$]	$+0.3$ [-4.0, $+3.6$]	-3.8 [-6.6, -0.5]	-2.2 [-5.6, $+1.2$]	-0.0 [-3.3, $+3.3$]
seed 2	-6.6 [-11.4, -2.5]	-4.9 [-9.2, -1.1]	-2.9 [-7.8, $+1.3$]	-9.5 [-13.3, -6.3]	-7.1 [-10.5, -3.4]	-5.0 [-8.1, -1.8]
ensemble (reported)	-6.2 [-10.5, -2.2]	-4.7 [-9.2, -1.2]	-3.3 [-7.5, $+0.6$]	-8.8 [-11.9, -5.9]	-6.7 [-9.9, -3.3]	-5.0 [-8.0, -1.8]

posterior, not NUTS. An earlier free-latent set autoencoder trained by the ELBO on one fold with the first reference is superseded by the encoder above and is not reported.

Table 22: Held-out months (as Table 1) on the 22,737 blitz probes that Allie covers, the joint speed–accuracy comparator: Allie (Zhang et al., 2025), a published chess model that predicts both the move and the time a human takes, with no clock state, its time head with a log-normal residual, and a fitted calibration of its move head's probability of the solution with the two ratings as the accuracy logit, then with θ_j added (row 7 against row 6, the comparison of Section 5.2); rows 8–9 are the reference and the person vector on the same probes. NLL is not comparable with Table 1 (different probes); the time columns of rows 8–9 were scored before the final zero-mass refit (Appendix F). Bold: the numerical best, not a test. Take-home: row 7 against row 6: ability adds accuracy over Allie too, below the 1.0-point margin.

	Model	NLL (nats)	CRPS (s)	cov. 50/90%	AUROC	Brier
(6)	Allie time head + residual; calibrated move head + ratings for accuracy	1.841	1.257	–	0.789	0.1574
(7)	Allie accuracy logit + θ_j	–	–	–	**0.793**	0.1558
(8)	Maia-2 reference	1.803	1.206	–	0.785	0.1564
(9)	+ hierarchical person vector	**1.792**	**1.199**	–	0.790	**0.1547**

GRU. A one-layer GRU (hidden 64) over the player's own within-game sequence of $(\log(y + \frac{1}{2}), \text{zero indicator}, \text{remaining clock})$ up to the probe, concatenated with the reference outputs, predicting the same censored lognormal and accuracy logit through additive offsets, trained on 386 of the 429 players with early stopping on the remaining 43, both on fit-period rows only (through June 2024), so no held-out-month probe is seen in training, and scored on the same 24,517 held-out-month probes as Table 1 (three seeds, averaged; 98.98% of probes matched to their game's move sequence). Three variants: the GRU alone, with the stable offset and with the person vector, which enters as fixed additive offsets of its empirical-Bayes estimates, not as learned inputs. Randomised-PIT coverage at 50/80/90/95% is 0.536/0.826/0.915/0.954 for the GRU alone and 0.525/0.818/0.905/0.948 with the person vector. Paired bootstrap over players: the GRU alone gains -0.037 nats [-0.042, -0.032] over the reference, with a small CRPS gain and a small AUROC gain, and more on the time channel than the person vector (Table 1); adding the person vector to the GRU raises NLL by $+0.0055$ [$+0.0016$, $+0.0095$] and raises AUROC by $+0.62$ points [0.38, 0.91] over the reference, $+0.40$ [0.18, 0.67] over the GRU alone and $+0.12$ points [0.06, 0.19] over the person vector alone;

the GRU with the stable offset instead scores 1.8286 nats, the lowest time NLL of any row, 0.0060 [0.0020, 0.0101] below the GRU with the person vector. Holding p_j at its population value leaves AUROC unchanged by construction and lowers the time NLL by 0.0048 nats [0.0031, 0.0068], about 87% of the $+0.0055$ cost, so most of the cost comes from the excess-zero term. In the first GRU run a column-name collision fed the reference through the zero-mass standardisation of an earlier reference and inflated its reference row by 0.005 nats, which read as a second evaluator; fixed, the GRU script reproduces the main evaluator's reference row to nine digits on the same rows. The GRU takes the zero-mass logit as an input and was retrained on the reported coefficients after the Eq. 5 refit; under the previous coefficients the person vector's cost on top of the GRU had been $+0.0002$ [-0.0039, $+0.0043$], indistinguishable from zero, and the refit moves the GRU-alone row by -0.0016 nats.

Allie. The published Allie checkpoint's think-time prediction at the probe (mean absolute error 2.10 s [2.04, 2.17] on the held-out months) is the location of a lognormal whose offset, scale and excess-zero terms are fitted on the fit-period blitz probes (47,671 rows) and scored on the held-out-month blitz probes (22,737 time rows, 22,749 accuracy rows) of the same 429 players; its tokenizer covers all 70,452 blitz probes. Its move head's probability of the solution move, alone or with the two ratings, is the accuracy logit after a fitted calibration. Allie's top move equals the played move on 66.1% of probes and the solution on 65.9%. Intervals: time NLL Allie 1.841 [1.816, 1.867] against reference 1.803 [1.774, 1.830]; accuracy AUROC Allie with ratings 0.789 [0.782, 0.796] against reference 0.785 [0.778, 0.793] and reference with the person vector 0.790 [0.783, 0.797]. Adding θ_j with unit coefficient to Allie's calibrated accuracy logit gives 0.793 [0.785, 0.800], a paired gain of $+0.0039$ [0.0014, 0.0069] (Brier -0.0016 [-0.0027, -0.0008]); adding the whole person vector as free covariates gives $+0.0025$ [-0.0008, $+0.0059$], so the gain is carried by θ_j. Allie's time rows do not depend on the reference; the time columns of Table 22's reference rows were scored before the Eq. 5 refit, which leaves every accuracy column unchanged (θ_j and α_i do not use the zero-mass model).

K FIXED-GAP RELIABILITY (51 MONTHS)

This appendix supports the question of how reliable the person is, and over what horizon (Section 5.1). The pre-specified T1 split cuts each player's date-sorted history at its median rank, so the gap between the halves is not controlled and is usually small. We therefore refit split-half reliability with the gap fixed by design. Each player's probes are split at that player's own median date, leaving an empty window of 6, 12 or 24 months centred on it; the zero-gap condition alternates probes within each calendar month. The split point is set without reference to the outcome. Everything else is the pre-specified T1 fit: the estimator, the inclusion masks, the leak-free reference, the Spearman–Brown correction and the player bootstrap. As a reproduction check, the pre-specified split rerun under this code gives tempo 0.7976 against the reported 0.7949, dispersion 0.4822 against 0.4845, ability 0.6042 against 0.6015 and excess zeros 0.5243 against 0.5281, each inside the reported interval; the small differences come from ties between same-day probes at the boundary.

At a near-zero gap reliability matches the pre-specified split (Table 23). As the gap grows, the Spearman–Brown-corrected r of tempo and of ability beyond rating falls. Tempo clears its 0.5 margin at 0, 6 and 12 months and falls below it at 24 months, 0.325 [0.107, 0.490]. Ability's interval already spans its 0.3 margin at 6 months, its point estimate is below the margin at 12 months (0.221 [0.152, 0.380]), and at 24 months it is close to null (0.052 [-0.031, 0.173]).

Half size, not only the gap. The halves shrink as the gap grows, because probes inside the empty window are dropped: the median half falls from 68.5 probes at zero gap to 38.0 at 12 months and 10.0 at 24 months, and a Spearman–Brown-corrected r falls with half size even when nothing changes over time. The 12-month comparison was therefore repeated at matched half size: each player's zero-gap halves were subsampled to that player's 12-month half sizes (same estimator, masks, minimum of five probes per half and correction; 511 players in both arms; 20 subsample draws; 2,000 paired player-bootstrap resamples). At matched half size the one-year value does not differ from the zero-gap value for either trait: the paired difference is $+0.04$ [-0.05, $+0.17$] for ability beyond rating (0.22 at one year against 0.18 at zero gap) and -0.005 [-0.08, $+0.06$] for tempo (0.62 against 0.62). Rank correlations also detect no difference (differences -0.04 [-0.15, $+0.06$] and

-0.002 [-0.07, $+0.06$]), and neither do dispersion and the excess-zero offset, whose intervals are wide; the two analyses agree only on no detected change, since the rank interval for ability beyond rating admits a decline of up to 0.15. The decline from zero gap to one year in Table 23 shows no gap effect beyond the smaller halves under this matching (which does not establish that half size explains it); the same code reproduces the table's zero-gap and one-year point estimates exactly. Two caveats. The comparison is a bound, not an equivalence: it excludes a gap-induced decline larger than about 0.05 for ability beyond rating and 0.08 for tempo in the Pearson analysis (about 0.15 and 0.07 in the rank analysis), and the intervals are percentile player-bootstrap intervals, which are asymmetric here. And the size-matched zero-gap value of ability beyond rating, 0.18, is below what the Spearman–Brown formula projects from the full halves at the median matched size, so the matched subsample does not behave as the same players with fewer probes; the half sizes are heterogeneous, a minimum of five probes per half applies, and 1.5% of the matched halves were capped at the available size. A per-player projection and the matched value without the five-probe players were not computed (Appendix X). The fitted slopes and crossing gaps below are a four-point descriptive summary, not corrected for half size, and are not horizon estimates: tempo -0.020 and ability -0.022 per month, with crossings extrapolated at about 16 and 11 months.

The 24-month point is thinner still. The probe table has no probes dated 2021-02 to 2022-12, 23 consecutive months, so many players' windows fall into the hole. Eligible players drop from 511–545 at the shorter gaps to 236–237 at 24 months, so the 24-month values mix temporal separation with estimation noise, and dispersion's 24-month interval (0.535 [-0.110, 0.963]) is uninformative. T1 measures consistency over about a year at the half sizes available, not multi-year stability. The same curve on the extended table is in Appendix Q.

Table 23: Spearman–Brown-corrected split-half r [bootstrap interval] by enforced gap between halves. n: players with both halves fitted; half: median probes per half. Slope: least-squares fit over the four fixed gaps, per month, not corrected for half size; its interval is a percentile interval over bootstrap replicates in which each gap condition's players were resampled independently, not paired across gaps, so it ignores the players shared between gaps and is descriptive only (the per-condition resampling seed also depends on the process's string hash, so the bootstrap intervals of this table are not bit-reproducible). Crossing: gap at which the fitted line reaches the margin, a descriptive extrapolation, not a horizon estimate; the size-matched comparison is in the text. Excess zeros (p_j) carry no pre-specified margin; their crossing is at $r = 0.3$. –: not reported here.

gap	n	half	tempo a_j	dispersion b_j	excess zeros p_j	ability θ_j
pre-specified split	–	–	0.795 [0.746, 0.834]	0.485 [0.382, 0.571]	0.528	0.601 [0.525, 0.666]
0 months	545	68.5	0.785 [0.736, 0.829]	0.523 [0.419, 0.613]	0.551 [0.462, 0.624]	0.604 [0.531, 0.666]
6 months	537	51.5	0.740 [0.683, 0.786]	–	–	0.360 [0.200, 0.605]
12 months	511	38.0	0.618 [0.531, 0.690]	0.096 [−0.104, 0.277]	–	0.221 [0.152, 0.380]
24 months	236–237	10.0	0.325 [0.107, 0.490]	0.535 [−0.110, 0.963]	0.263 [0.040, 0.436]	0.052 [−0.031, 0.173]
slope			−0.0199 [−0.0293, −0.0127]	+0.0001 [−0.0284, 0.0195]	−0.0114 [−0.0214, −0.0031]	−0.0220 [−0.0280, −0.0158]
margin			0.5	0.3	–	0.3
crossing			16.4 months	–	20 months	10.9 months

K.1 Reliability of person statistics on ordinary moves (pilot)

Before the probe fits, split-half reliability of per-player residual statistics was measured on a table of ordinary blitz moves from 2019 with centisecond clocks, with one focal player's moves taken per game. The table is external to this pipeline: it was built earlier, and the supplied code does not rebuild it. Its policy features come from Maia-3, not Maia-2, and its pooled least-squares adjustment of log time includes the probability of the move actually played and the recorded clock, both post-decision

under the standard of Appendix G. The share of residual variance explained by player identity and the split-half correlations computed on this table are therefore descriptive only. No claim of this paper rests on them, and the pilot appears only as the open markers of Figure 3.

Which clock the table records. A read-only check on the table's non-test rows (993,626 rows, 92,817 games, 5,079 players) settles the clock convention. Over 697,249 consecutive own-move pairs of the same player, the recorded centisecond clock equals the player's previous own clock minus the think time plus the increment for a share of 1.000 of pairs (residual within 0.005 s), with or without an increment, so it is the clock after the move, with the increment added. The alternatives fit almost no pairs: the two pre-move forms hold for shares of 0.012 and 0.010, and the form without the increment for none of the pairs with a positive increment. Because the first identity partly restates how the table defines think time, the informative evidence is the rejection of the alternatives. The pilot's clock covariate is therefore read after the decision.

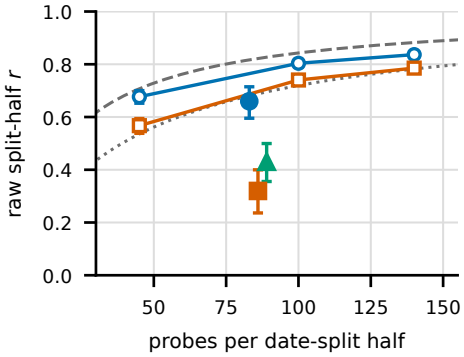

Figure 3: Split-half reliability against probes per half. Filled: the probe cohort at 86 per half (circle tempo, square dispersion, triangle ability; raw r, 95% intervals). Open: location (circles) and dispersion (squares) of the ordinary-move pilot, computed on the external table with post-decision covariates and shown as description only. Grey: Spearman–Brown curves from each table's move-level intraclass correlation (dashed: the pilot table; dotted: the probe table).

L IDENTIFICATION DETAIL

Within-person signature. Matched pool: twenty other primary-cohort players of the same rating band and dominant time control (same-band fallback for 54 of 545 players). Energy-distance self/other ratio median 0.497 [0.267, 0.832], 79.8% below one; symmetric KL 0.972 [0.797, 1.180], 54.1% below one. Observed mean self-distance 0.443 (symmetric KL) and 0.0217 (energy) against permutation means 0.498 and 0.0426; none of 200 permutations reached either value. Nearest-neighbour identification among 545 players from a summary vector (chance 0.183%): 0.84% at $k = 10$, 0.99% at $k = 20$, 1.76% at $k = 45$, 3.49% with the whole half. These residuals are computed under the reported reference and class intercepts; under the second-pass reference the ratio was 0.510 and the whole-half identification 2.20%.

M IDENTIFICATION AND THE IDENTITY-SENSITIVE OUTLOOK

This appendix supports the discussion question of what an identity-sensitive use of a person representation would require, and what the present evidence does not supply (Section 6). The identification result already in the paper is in Appendix L. Its self/other ratio compares the raw log-think-time residual distributions of a player's date-split halves by energy distance; its nearest-neighbour identification uses a summary vector. The self/other ratio has median 0.50; the observed mean self-distance was below all 200 permutation values ($p < 1/200$). Nearest-neighbour identification among 545 players, where chance is 0.183%, reaches 0.84%, 0.99% and 1.76% with 10, 20 and 45 probes per query and 3.49% with the whole half.

The test here is complementary to that result, not a replication of it. It uses a different distance: a diagonal Gaussian distance between the fitted empirical-Bayes posteriors of the halves. It also uses a different k axis: the rank cut-off, not the number of probes per query. Its numbers are therefore not comparable with those of Appendix L. With the full person vector $(a_j, b_j, p_j, \theta_j)$ the true player ranks first for 2.20% of queries, in the top five for 6.06% and in the top ten for 9.91% (Table 24). Timing alone (a_j, b_j) carries most of this signal: 1.10%, 4.59% and 7.34%. Ability beyond rating alone is markedly weaker: 0.37%, 0.92% and 2.94%.

For impostor scoring, a claimed player's first-half posterior is compared with the class-level model without a person offset, by a log-likelihood ratio on k probes from the second half; impostors come from the same rating band. The AUROC rises with k, from 0.576 [0.565, 0.588] at $k = 5$ to 0.685 [0.662, 0.706] with the whole half. The true-positive rate stays at 2.2–4.8% at a 1% false-positive rate and at 15.9–23.5% at a 10% false-positive rate. The signal is real but weak at every k tested.

Turning this outlook into an identity or anomaly claim would require five things this evidence does not supply. It would require a chosen operating point, an impersonation base rate, an evasion or adversarial model, a stylometry baseline (McIlroy-Young et al., 2021), and a labelled anomaly set for evaluation. It would also require probe sets drawn jointly across the time and accuracy channels; these tests resample the pools independently. We report these numbers as an outlook only.

Table 24: Identification from fitted posteriors (top-k rank, percent [bootstrap interval]; chance for top-1 is 0.183%) and impostor scoring by probes scored (k), same-rating-band impostors. Not comparable with the summary-vector nearest-neighbour identification of Appendix L.

features	top-1	top-5	top-10
full $(a_j, b_j, p_j, \theta_j)$	2.20 [1.10, 3.49]	6.06 [4.22, 8.26]	9.91 [7.52, 12.48]
timing (a_j, b_j)	1.10 [0.37, 2.02]	4.59 [2.94, 6.42]	7.34 [5.14, 9.55]
ability (θ_j)	0.37 [0.00, 0.92]	0.92 [0.18, 1.83]	2.94 [1.65, 4.40]
chance	0.183	–	–
probes k	AUROC	TPR at 1% FPR	TPR at 10% FPR
5	0.576 [0.565, 0.588]	2.2	15.9
20	0.622 [0.606, 0.638]	3.8	19.6
100	0.659 [0.638, 0.679]	4.4	22.6
whole half	0.685 [0.662, 0.706]	4.8	23.5

N TRANSFER FROM ORDINARY MOVES (T4, EXPLORATORY)

The plan's T4 asked whether a person's tempo on ordinary moves predicts their probe tempo (material if $r \geq 0.5$); it was run after the freeze and is exploratory. The 545 primary-cohort players' own moves in the same 51 months (3,427,069 moves at ply 15 or later, excluding the puzzle move and each side's first move) received Maia-2 features at the decision position, an ordinary-move reference of the probe recipe (five player folds; $R^2 = 0.42$, 90% coverage 0.90), a fresh zero-mass fit and the two-stage empirical-Bayes person fit ($\tau_a = 0.135$, $\tau_b = 0.089$ against the probes' 0.134 and 0.121). Table 25 lists the transfer coefficients; the few-shot rows shrink a k-probe tempo estimate toward the ordinary-move value instead of the class mean with the same weight, isolating the prior mean (dispersion and excess-zero terms fixed at their full-sample values in both arms). Caveat: the ordinary-move table's capture, recapture and check flags are computed from the move played, the post-decision covariates removed from the probe side in a later revision, and were not corrected.

O POSTERIOR-PREDICTIVE CHECKS: RECORDED-SCALE DETAIL

The five gates of Section 4.3: recorded think-time mean and percentiles by (1) rating band, (2) puzzle-rating band, (3) time-control class; (4) solve rate by rating-gap decile, which must be monotone; (5) a recorded-scale check that simulates recorded times from the fitted model 200 times and passes only if the mean per-player discrepancy in $\log(k + \frac{1}{2})$ is below 0.03 and the zero-fraction gap below one point, both in absolute value (gate 5 replaced a latent-scale check, Appendix F).

Table 25: T4: transfer of person parameters from ordinary moves to probes (545 players; player-bootstrap 95% intervals; correction for attenuation by each side's reliability for the cross-domain rows and by Spearman–Brown for the split-half row; date halves anchored on each player's probes). Puzzle-prominence contrast: per player, mean $\log(y + \frac{1}{2})$ residual at probes minus that at ordinary moves, each against its own reference.

Quantity	estimate [95% CI]	corrected
tempo a_j, ordinary moves against probes (Pearson)	0.63 [0.58, 0.68]	0.71
dispersion b_j, ordinary moves against probes	0.50 [0.43, 0.56]	0.59
held out: ordinary tempo, half 1 → probe tempo, half 2	0.53 [0.46, 0.59]	–
held out: probe tempo, half 1 → probe tempo, half 2	0.66 [0.60, 0.71]	–
few-shot NLL gain, ordinary-move prior mean, $k = 5$ (nats)	-0.0031 [-0.0037, -0.0026]	–
$k = 10$	-0.0024 [-0.0029, -0.0019]	–
$k = 20$	-0.0015 [-0.0019, -0.0012]	–
puzzle-prominence contrast, split-half r	0.43 [0.34, 0.51]	0.60

Under the reported model, 200 simulated replicates of every recorded time on the primary cohort (92,696 rows, 544 players; one player with a non-finite excess-zero estimate excluded) give the comparisons in Table 26. Gates 1–3 pass on shape; gate 4 passes; gate 5's per-player criterion passes ($+0.019$) and its zero-fraction criterion fails (-2.6 points, from -2.8 before the Eq. 5 refit), in every class but classical and most on low clocks. Under the triangular kernel of Appendix C, refitted under the reported reference and coefficients, the per-player discrepancy is $+0.021$ and the zero gap -1.7 points, the smallest of any variant: a real but partial improvement, and the remaining excess is consistent with the two clock reads being more correlated than independence implies, or with zero-inflation structure the covariates of Eq. 5 do not carry. The original latent-scale diagnostic (mean standardised residual $+0.32$ observed) is reproduced at $+0.28$ by simulation from the fitted model, so most of it is the diagnostic's construction.

Table 26: Recorded-scale posterior-predictive check for the reported model (200 replicates): zero fraction observed against simulated by slice, and the two gate-5 criteria.

slice	rows	obs. zeros	sim. zeros
all	92,696	8.67%	11.26%
blitz	79,599	9.55%	12.25%
bullet	789	17.49%	19.45%
rapid	11,782	2.44%	4.45%
classical	526	2.28%	1.93%
clock $<$ 10 s	9,051	30.29%	35.03%
clock $\geq$ 10 s	83,645	6.33%	8.69%

Gate 5: mean per-player discrepancy in $\log(k + \frac{1}{2})$ $+0.019$ (criterion $|\cdot| < 0.03$: pass); zero-fraction gap -2.60 points (criterion $|\cdot| < 1.0$ point: fail).

P Robustness of the increment instrument

This appendix supports the question of whether thinking longer causes better decisions (Section 5.4). For a laboratory study of deliberation and chess move quality, see Moxley et al. (2012). This appendix reports the robustness battery of the causal check; the battery was not pre-stated, and the pre-stated specification and every change to it are listed in Appendix F. In the instrument and double-ML estimates, $\log T$ is $\log(y_i + \frac{1}{2})$ of the recorded whole-second think time, so a recorded zero enters as $\log \frac{1}{2}$. The within-person slope of the accuracy channel extended by $\lambda \log T$ is $\lambda = -0.20$ logits [-0.23, -0.18] (Table 27, row 1). The instrument is applied to 5,174 players of the training pool (median rating 2,053) who play both an increment and a no-increment variant of the same base time. Every instrumented effect below is the probability-scale LATE of the increment instrument, a local effect for its compliers, an unidentified subpopulation of those players, and with a continuous treatment a weighted average of their marginal effects; it is not a population-wide effect of think time

on accuracy. The negative-control outcome throughout is the quality of the decision two own moves (four plies) back, regressed on the adjusted residual of think time (the negative-control-outcome row of Table 28). Two selection issues are not resolved by the battery: the reported 40 s clock floor conditions on the remaining clock, which the increment itself moves, so the no-floor row is the one free of that clock conditioning; and a probe exists only where the game produced a puzzle position, so the analytic sample is selected after the instrument and no balance table of item covariates by instrument arm was tabulated (Appendix X).

No variation of the LATE with the time-control class was detected (Table 28); the classical-class estimate is imprecise. It is $+0.0818$ [0.0460, 0.1177] in blitz with first-stage $F = 698.9$, and $+0.0690$ [-0.0231, 0.1612] in rapid with $F = 93.4$; classical is too weak to inform ($F = 1.1$). No bullet pool enters the sample: every pool's base time sits at or above the blitz boundary. Cochran's Q gives $p = 0.963$ ($I^2 = 0\%$), and the pooled inverse-variance estimate, 0.0801, lies inside every class's interval. The dominant increment arm, a 180 s base with a 2 s increment, gives $+0.0817$ [0.0412, 0.1223] against the main $+0.0836$ [0.0549, 0.1123]; minor arms with F above 10 are also positive and of similar size. Across eight clock floors from none to 180 s the LATE stays within 0.070–0.084 and the first-stage F stays above 10; only the 180 s interval includes zero. This stability is read together with the two caveats below: the raw instrument is associated with the quality of the decision two own moves back, and the reduced form's robustness value is below the strongest measured covariate's partial R^2, though a plausibly exogenous bound holds. Permuting the instrument within its pools 200 times gives a first-stage F with median 0.50 and maximum 12.30, against the real 975.8.

The controls of the causal check are gathered in Table 28: the pre-stated negative-control outcome is null ($+0.001$ [0.000, $+0.002$]), the positive control is under-powered ($+0.012$ [-0.007, $+0.033$]), and a negative-control exposure, think time two moves back, still predicts the outcome ($+0.012$ [0.009, 0.016]), a further sign that occasion-level confounding is present in the observational stream. The four-lag window of the negative-control outcome, the mean quality of moves $t-2$ to $t-5$, is $+0.0003$ [-0.0004, $+0.0009$] on 58,614 probes of 545 players. The battery also leaves caveats, and we report them together. First, the raw increment indicator is associated with the quality of the decision two own moves back: $+0.0063$ [0.0017, 0.0109] ($n = 39{,}511$, 375 players). The reported negative control, which regresses the same quality on the adjusted residual of think time rather than on the instrument, stays null ($+0.001$ [-0.000, $+0.002$]); the association with the instrument is a caveat, not a passed check. This is consistent with occasion-level confounding of the assignment, or with an effect of the chosen increment on the earlier move itself; the diagnostic does not separate the two. Second, the Cinelli–Hazlett robustness value of the instrument's reduced form is 0.050, below the partial R^2 of the strongest measured covariate, 0.118. The reduced form is thus less robust than the double-ML slope, whose robustness value is 0.36. The plausibly-exogenous bound of Conley et al. (2012) allows the increment a direct effect on the probability of solving, in probability per unit of the instrument: a positive direct effect of the increment of 0.029 would bring the point LATE to zero and one of 0.019 its interval's lower end; the observational slope implies a benchmark of the opposite sign, which would raise the LATE, so it does not bound the violations that matter. The exclusion restriction is checked, not proved.

Table 27: Effect of a log-unit of think time on solving the item: the within-person slope of the accuracy channel extended by $\lambda \log T$ (Eq. 4), double machine learning (Chernozhukov et al., 2018), and the increment instrument (player×month effects; the reported 40 s clock floor conditions on a post-instrument variable; other floors, the placebo and the negative and positive controls are in Table 28, Appendix P). Intervals: player-cluster bootstrap or cluster-robust standard errors; AR, weak-instrument-robust (Anderson & Rubin, 1949). Take-home: the within-person slope and double machine learning are negative; the instrumented effect for the compliers is positive, under an exclusion restriction that is checked, not proved.

	Estimand, estimator	probes / players	estimate [95% CI]
1	within-person slope λ, logits (Eq. 4 extended by $\lambda \log T$)	92,796 / 545	-0.203 [-0.228, -0.178]
2	double ML, partially linear, player FE, probability	316,586 / 9,482	-0.048 [-0.050, -0.045]
3	increment instrument, first stage on $\log T$ ($F = 976$)	173,680 / 5,174	$+0.348$ (SE 0.011)
4	LATE, probability, clock ≥ 40 s (AR [0.056, 0.112])	173,680 / 5,174	$+0.084$ [0.055, 0.112]

Table 28: Robustness of the increment-instrument LATE (probability scale). n: probes / players. The main specification is the $\geq$40 s clock floor. The controls and the last block are diagnostics; the placebo row is a coefficient, not a LATE, and it and the robustness value must be read together with the plausibly-exogenous bound. Class rows keep only players with both arms in that class, so they do not partition the main sample.

specification	n	F	LATE [interval]
class: blitz	76,319 / 2,624	698.9	+0.0818 [0.0460, 0.1177]
class: rapid	49,200 / 2,144	93.4	+0.0690 [−0.0231, 0.1612]
class: classical	7,404 / 517	1.1	+0.0432 [−0.6556, 0.7419]
class: bullet	–	–	absent from the sample
heterogeneity (Cochran's Q)			$p = 0.963$, pooled 0.0801
dominant arm (180 s + 2 s)	59,211 / 2,142	539.9	+0.0817 [0.0412, 0.1223]
floor: none	251,779 / 5,503	1,808.0	+0.0773 [0.0581, 0.0965]
floor: $\geq$20 s	207,479 / 5,348	1,340.1	+0.0754 [0.0522, 0.0987]
floor: $\geq$30 s	188,557 / 5,256	1,212.6	+0.0790 [0.0535, 0.1045]
floor: $\geq$40 s (main; AR [0.056, 0.112])	173,680 / 5,174	975.8	+0.0836 [0.0549, 0.1123]
floor: $\geq$60 s	147,717 / 5,006	659.0	+0.0809 [0.0448, 0.1170]
floor: $\geq$90 s	116,305 / 4,704	407.3	+0.0723 [0.0266, 0.1179]
floor: $\geq$120 s	92,928 / 4,318	288.3	+0.0705 [0.0181, 0.1228]
floor: $\geq$180 s (AR [0.004, 0.173])	62,684 / 2,984	99.7	+0.0826 [−0.0013, 0.1666]
negative-control outcome: quality at $t-2$ on adjusted T	58,655 / 545	–	+0.001 [0.000, +0.002]
positive control: error at $t-1$ on $\log T_t$	58,656 / 545	–	+0.012 [−0.007, +0.033]
negative-control exposure: $\log T_{t-2}$ on solving	58,656 / 545	–	+0.012 [0.009, 0.016]
permutation, 200 draws		median 0.50, max 12.30	–
placebo (quality two own moves back)	39,511 / 375	–	+0.0063 [0.0017, 0.0109]
reduced-form robustness value			0.050 (benchmark 0.118)
plausibly-exogenous bound: direct effect zeroing LATE			0.029 (lower end: 0.019)

Q 79-MONTH REPLICATION

This appendix bears on the reliability horizon, on amortization and on the limitations (Section 6); the main text keeps the 51-month numbers throughout.

What was added, and the gate. The rerun adds 28 months, for 79 months and 4,806,330 located probes against 2,647,029; the primary cohort grows from 545 to 1,676 players (Table 29). Two pre-specified completeness criteria governed promotion to the main text. Criterion (a) required the whole evaluation chain to complete, including the own-move and game-prefix extractions that the GRU, Allie and T4 rows need for the new months. Criterion (b) required that no step of the evaluator chain fail silently. Both failed. The own-move table was not rebuilt, so the GRU row first ran at 32.8% join coverage; and two evaluator steps failed without signalling it and were corrected afterwards (the GRU step's output location and the T3 step's pass criterion, which had accepted an all-NaN result; the T3 numbers below come from the $k = 0$ control evaluator, and Table 1 reads an untouched directory). A later rebuild of the own-move table raised the GRU join coverage from 32.8% to 99.06% and gave the causal controls on the primary cohort a genuine 79-month sample; the Laplace refit was also re-run. Criterion (a) is still not fully met, because T4 was never part of the chain, and criterion (b) cannot be passed after the fact. The verdict is a replication appendix.

What did not move. The pre-specified split's reliabilities reproduce: tempo, dispersion and ability reach 0.782, 0.497 and 0.659, against 0.795, 0.485 and 0.601, all above their margins. The held-out gain is −0.00896 nats [−0.01242, −0.00564] against −0.0096, and it is again not material. The increment instrument keeps its sign and size: the local effect is +0.0717 [0.0549, 0.0886] with $F = 2{,}776$, against +0.084 [0.055, 0.112] with $F = 976$, and the within-person slope stays negative. Gate 5 still fails, with a zero-mass gap of −1.967 points against −2.596 (Appendix O). Adding the person to the GRU still costs time-channel likelihood, and θ_j still adds accuracy AUROC to Allie with the rating gap, +0.0037 [0.0022, 0.0054] against +0.0039.

What moved. T3 moves most (Appendix J). The 79-month encoder is one seed, not three: the replication launcher did not pass its seed argument, so its three members were trained with the same default seed along one identical training path and differ only in the step at which the wall-

clock limit stopped them (their recorded run summaries); the 79-month T3 rows therefore carry no seed variability, unlike the 51-month ensemble. At one probe the encoder is no longer ahead of the reference ($+0.0020$ and $+0.0055$ nats on the training pool and the cohort), and its gain over the reference reaches 0.5% of the reference's NLL from five probes; this is not the T3 margin, which concerns the distance to the ceiling. At twenty probes it closes 52% of the gap to the full-information ceiling on both the training pool and the cohort, against 39% and 35%; the remaining distance, 0.68% and 0.72%, is still above the 0.5% margin. The $k = 0$ control now loses to the bare reference on both the training pool and the cohort, where at 51 months it was tied on the primary cohort. Against the marginalised refit (Laplace), the encoder's edge on the primary cohort holds at 5 of 7 probe counts, up from 2. The few-shot law refits with $G_\infty = 0.0239$ nats and $k^* = 5.24$, against 0.013 and 1.2. The clock-floor row at 180 s now excludes zero, $+0.0549$ [0.0116, 0.0982] (Appendix P). The exploratory validity check weakens: Spearman's correlation of first-half θ_j with later rating falls to $+0.0369$ [-0.0157, $+0.0948$]. Composition drives most of this: of the 1,161 eligible players, 744 (64%) are newly eligible, with a null correlation of their own, 0.012 [-0.061, $+0.084$]. Last, tempo's reliability at a 24-month gap rises from 0.325 to 0.535 and clears its 0.5 margin (Appendix K). This is largely a coverage artefact: the 51-month table has no probes from 2021-02 to 2022-12, 23 consecutive months, and the new months shrink that hole to five months, 2021-06 to 2021-10. Ability's 24-month point rises from 0.052 to 0.126 but stays below its 0.3 margin. Reliability is not gap-invariant.

Table 29: 51-month (main text) and 79-month (replication) values. Training pool: players with at least 23 probes; cohort: the primary cohort, at least 100 probes. Brackets give the reported interval.

Quantity	51 months	79 months
Located probes	2,647,029	4,806,330
Training pool: players / probes	9,482 / 423,472	24,464 / 1,206,505
Cohort: players / probes	545 / 93,776	1,676 / 342,022
At least 300 probes: players / probes	43 / 17,934	230 / 113,825
Held-out-eligible cohort players	429	1,133
T1 corrected r: a_j / b_j / θ_j / p_j	0.795 / 0.485 / 0.601 / 0.528	0.782 / 0.497 / 0.659 / 0.556
24-month gap: a_j / θ_j	0.325 / 0.052	0.535 / 0.126
θ_j decay slope, per month	-0.0220	-0.0221
T2 held-out gain, nats	-0.0096	-0.00896 [-0.01242, -0.00564]
Encoder − reference, $k = 1$, training pool, nats	-0.007 [-0.012, -0.002]	$+0.0020$ [-0.0014, $+0.0053$]
Encoder gain at $k = 5$, training pool	0.50%	0.53%
Ceiling distance at $k = 20$, training pool / cohort	0.9% / 0.8%	0.68% / 0.72%
Gap closure at $k = 20$, training pool / cohort	39% / 35%	52% / 52%
$k = 0$ control − reference, cohort, nats	tied	$+0.0107$–0.0115
Edge over Laplace refit, cohort	2 of 7 k	5 of 7 k
Few-shot law G_∞ / k^*	0.013 / 1.2	0.0239 / 5.24
Local effect, clock $\geq$ 40 s	$+0.084$ [0.055, 0.112]	$+0.0717$ [0.0549, 0.0886]
First-stage F	976	2,776
Local effect, clock $\geq$ 180 s	$+0.083$ [-0.001, 0.167]	$+0.0549$ [0.0116, 0.0982]
Within-person slope λ, logits	-0.203	-0.18928
Negative control, quality at $t - 2$	$+0.001$ [-0.000, $+0.002$]	$+0.00082$ [-0.00010, $+0.00178$]
Gate 5 zero-mass gap, points	-2.596	-1.967
GRU + person − GRU, NLL, nats	$+0.006$ [$+0.002$, $+0.010$]	$+0.00262$ [$+0.00026$, $+0.00496$]
Allie + θ_j − Allie + gap, AUROC	$+0.0039$ [0.0014, 0.0069]	$+0.0037$ [0.0022, 0.0054]
Spearman(θ_j, later rating)	$+0.1114$ [0.0286, 0.1926]	$+0.0369$ [-0.0157, $+0.0948$]

R MAIA-3 REFERENCE REPLICATION

This appendix supports Sections 5.1 and 5.2, whether the person is reliable and what it adds per decision, by changing the human-like policy inside the reference. We replace Maia-2 with Maia-3 (Chessformer; Monroe et al., 2026) in the leak-free contextual reference. The folds, hyperparameters, recapture covariate, person-model recipe, evaluator and margins are held fixed. This is a partial replication. The zero-mass (clock-rounding) channel stays on the Maia-2 reference and was not refitted, so the two columns of Table 30 are not two independent pipelines. The check is appendix-level. It does not replace the Maia-2 numbers in the main text. Neither backbone is retrained here: as for Maia-2 (Appendix G), Maia-3's training games may overlap the probe months, and cross-fitting guards only the reference layer built on it.

Maia-3 is the stronger accuracy-channel model: the reference's accuracy AUROC rises from 0.7987 to 0.8244 on the same pool (Table 30). The person layer does not move with it. On the same 545 players, every T1 margin (tempo at least 0.5, dispersion and ability at least 0.3) is still met. The held-out gain of the full person vector is -0.0100 nats against -0.0096 under Maia-2, and it clears none of the three materiality margins under either reference. Per-player traits from the two fits agree at raw r 0.946–0.990. One quantity shifts. The tempo–ability correlation $\mathbf{\Omega}_{a\theta}$ is -0.0916 $[-0.171, -0.014]$ under Maia-2 and -0.0777 $[-0.158, 0.002]$ under Maia-3, so its interval now touches zero. We read this as noise at $n = 545$, not as a reference-dependent effect. Both references support the same weak, possibly null, link between tempo and ability.

Table 30: Maia-2 versus Maia-3 in the contextual reference (partial replication; zero-mass channel held on Maia-2 in both columns; $n = 545$ players). T1 rows are Spearman–Brown-corrected split-half r; their bootstrap intervals exist only for the raw r and are not shown. $\mathbf{\Omega}_{a\theta}$ is the raw correlation with its interval. T2 rows are full person vector minus reference, with paired-bootstrap intervals.

Quantity	Maia-2	Maia-3
Reference accuracy AUROC	0.7987	0.8244
T1 tempo a_j	0.795	0.804
T1 dispersion b_j	0.485	0.509
T1 zero-mass shift p_j	0.528	0.558
T1 ability θ_j	0.601	0.588
$\mathbf{\Omega}_{a\theta}$ (raw)	-0.092 $[-0.171, -0.014]$	-0.078 $[-0.158, 0.002]$
T2 ΔNLL (nats)	-0.0096 $[-0.0146, -0.0051]$	-0.0100 $[-0.0145, -0.0055]$
T2 ΔAUROC	$+0.0051$ $[0.0026, 0.0080]$	$+0.0043$ $[0.0023, 0.0067]$
T2 clears all margins	no	no

S Cross-setting (time-control) transfer

This appendix supports the discussion question of how far a person representation carries beyond the conditions it was fitted in (Section 6), using players active in two Lichess time-control classes. The finding has two halves, and they belong together. A person's traits correlate across settings. The fitted vector does not transfer predictively. The main numbers are 79-month appendix numbers, which did not meet the pre-specified completeness criteria (Appendix Q); the thinner 51-month table is shown beside them. The roster is drawn from the full probe population, because the primary cohort has too few players with enough probes in two classes. Samples are small by construction: $n = 53$–168 players per pair at 79 months.

Correlation first. At 79 months tempo, the zero-mass shift and ability correlate positively across classes in all three tested pairs, with every interval excluding zero; dispersion clears zero only in the best-powered pair, blitz and rapid (Table 31). Prediction second. On the time channel the held-out gain of a transferred vector over the reference is null: at 51 months every direction's interval spans zero, and at 79 months 11 of 12 direction-by-comparison cells do. The exception runs the wrong way: from classical to rapid the transferred vector is worse than the reference, $+0.0532$ $[+0.0126, +0.0928]$ nats per probe. On the accuracy channel the 51-month table suggested that a cross-class vector beats a within-class fit, but that comparison gave the cross-class fit more training data. With training size equalised at 79 months, none of those gains survives, and the transferred vector is significantly worse than no person vector in half the directions (for example blitz to bullet, -0.0488 $[-0.0760, -0.0242]$ AUROC). The frozen class intercepts already carry the population-level setting shift (bullet -0.224 to classical -0.056 log-time units), which is consistent with, not proof of, the null time-channel result. Traits relate across settings; on the time channel no transfer benefit was detected, and one direction (classical to rapid) is significantly worse than the reference; on the accuracy channel transfer hurts in some directions; the setting's own fit carries the predictive value.

Table 31: Raw cross-class correlation of each person parameter. 79 months: player-bootstrap intervals; n is the number of players per pair. 51 months: point estimates only, on thinner samples. Values corrected for attenuation are not shown because the within-class ceilings are small-sample.

Parameter	Class pair	n (79m)	79 months	51 months
Tempo a_j	blitz–rapid	100	0.386 $[0.176, 0.574]$	0.352
	bullet–blitz	53	0.503 $[0.177, 0.727]$	0.590
	rapid–classical	168	0.551 $[0.419, 0.670]$	0.286
Zero-mass p_j	blitz–rapid	100	0.250 $[0.052, 0.431]$	0.008
	bullet–blitz	53	0.286 $[0.018, 0.531]$	0.234
	rapid–classical	168	0.324 $[0.153, 0.475]$	0.277
Ability θ_j	blitz–rapid	100	0.551 $[0.402, 0.671]$	0.388
	bullet–blitz	53	0.412 $[0.167, 0.608]$	0.222
	rapid–classical	168	0.186 $[0.132, 0.464]$	0.343
Dispersion b_j	blitz–rapid	100	0.334 $[0.179, 0.476]$	0.021
	bullet–blitz	53	0.208 $[-0.031, 0.444]$	-0.385
	rapid–classical	168	0.105 $[-0.081, 0.317]$	0.193

T SECOND DOMAIN: A FOUR-IN-A-ROW PILOT

This appendix supports the discussion question of how far the person representation carries beyond chess (Section 6). It is a pilot ($n = 30$ players) without calibrated items: the timing verdicts replicate; a move-quality channel is reported, but it is not ability beyond skill, and its verdict does not replicate. The corpus is the four-in-a-row data of van Opheusden et al. (2023): 67,331 human moves, all with a response time. We use its learning experiment, 40,154 moves by 30 players over 5 sessions each. The frozen population reference is a leave-one-subject-out engine policy that excludes the player's own lagged response time (held-out R^2 of log response time 0.2786). Tempo a_j, dispersion b_j and move quality θ_j are offsets from it; move quality is whether the player chose the reference policy's modal move. Split-half reliability uses alternating games within a session, so the halves share no move and no game, and each half is fitted independently. With 30 players a single move-level permuted-label draw is far from zero, about 0.3 for every parameter (Table 32); one draw does not characterise the null distribution, and every reliability below must be read against it. Against that floor, tempo and dispersion are reliable (0.966 and 0.831) and improve held-out prediction on later sessions by $+0.0556$ nats per move $[+0.0292, +0.0843]$. Move-quality reliability is observed, but its incremental predictive value is unresolved: its reliability (0.725, same floor) comes with a null held-out gain, and as a person summary it adds almost nothing over the raw training match rate (partial r $+0.087$). This pilot uses no skill scale (no rating accompanies the released moves we use, and the win-rate proxy has no reliable between-player variance) and no calibrated items (only 98 of 29,700 positions are seen by 10 or more players), so move quality here is not ability beyond skill. No quantity in this appendix is taken from other manuscripts; the concurrent manuscript uses the same cohort (Appendix A).

Table 32: Four-in-a-row pilot ($n = 30$ players). Reliability is Spearman–Brown-corrected split-half r at gap 0; its bootstrap intervals exist only for the raw r and are not shown. The floor is the same statistic after permuting player labels (a single permutation draw per trait, at move and at game level, not a null distribution). Held-out gain is the reduction in NLL per move on sessions 4–5, fitted on sessions 1–3, with a paired player-bootstrap interval.

Parameter	Reliability	Floor (move / game)	Held-out gain (nats/move)
Tempo a_j	0.966	0.324 / 0.349	timing $+0.0556$ $[+0.0292, +0.0843]$
Dispersion b_j	0.831	0.315 / 0.218	
Move quality θ_j	0.725	0.295 / 0.145	quality $+0.0032$ $[-0.0007, +0.0078]$

U ELITE OVER-THE-BOARD EXTERNAL VALIDITY

This appendix supports the discussion question of how far the person representation carries beyond online play (Section 6), using the elite over-the-board corpus of Sunde et al. (2026): 217,403 moves in 3,685 games by 276 players rated 2500 or more, 2014 to 2017. Its supplied engine evaluations agree with our own recomputation at Pearson 0.9885. Accuracy is a no-blunder indicator (win-probability loss below a threshold set in the fit script). Tempo and dispersion clear their margins under all five split designs, including tournament-disjoint halves and enforced 6- and 12-month gaps: a_j 0.740–0.836, b_j 0.723–0.808 (Table 33). Ability beyond rating is below its 0.3 margin at the point estimate under all five (0.030–0.266), and no design establishes reliability above it; its per-player measurement reliability is 0.345. The zero-mass shift p_j is not identified over the board ($\hat{\tau}_p = 0.0000$, mean standard error 32.4); there is no premove. The person layer's held-out gain on later tournaments is real but not material, -0.0083 nats per move $[-0.0139, -0.0034]$ against the 0.02-nat margin, close to the Lichess -0.0096. Person contrasts do not depend on the assumed recording operator: tempo estimates correlate at 0.9914 between the uniform and the Lichess-measured rounding masses.

The increment instrument's sign reversal reappears only within the 8 events that contain both time-control regimes, at about half the Lichess size: $+0.0427$ $[+0.0026, +0.0829]$ with player and tournament fixed effects ($F = 18.8$), against the Lichess LATE of $+0.084$ for that instrument's compliers and an observational within-person slope of -0.0104. This estimate carries two caveats: the negative control fails in every specification, at $+0.0193$ $[+0.0030, +0.0356]$ within the same mixed events, and only 9.2% of the instrument's variance survives tournament fixed effects. It corroborates the direction and does not confirm it independently. The corpus also limits the comparison structurally: no calibrated items, clocks exactly whole-second (maximum rounding residual 2.27×10^{-12} s), no fast mass, first-stage F 18.8–43.2 against 976 on Lichess, and ratings of 2500 or more by construction.

Table 33: Over-the-board split-half reliability (Spearman–Brown-corrected r, 2,000-replicate player bootstrap intervals; "–" where no interval was reported). Margins: $a_j \geq 0.5$, b_j and $\theta_j \geq 0.3$.

Split design	n	Tempo a_j	Dispersion b_j	Ability θ_j
Rank-median date	234	0.804 $[0.736, 0.852]$	0.765 $[0.684, 0.826]$	0.180 $[-0.049, 0.359]$
Tournament-disjoint	180	0.750 $[0.660, 0.816]$	0.723 $[0.617, 0.806]$	0.030 $[-0.294, 0.269]$
Within-event ceiling	–	0.836 $[0.779, 0.882]$	0.808 $[0.740, 0.857]$	0.160 (–)
6-month gap	46	0.740 $[0.525, 0.857]$	0.760 $[0.597, 0.869]$	0.259 (–)
12-month gap	34	0.756 $[0.585, 0.855]$	0.740 $[0.540, 0.873]$	0.266 (–)

V CENTISECOND REPLICATION ON THE OVERLAP MONTHS

This appendix supports Section 4.2, what lies between a person's true think time and the record (Appendix C), by refitting the person model at centisecond resolution on the 13-month overlap of the two archives. Alignment covers 97.06% of primary-cohort probe rows, and every aligned think time matches the reported one. The samples are small (Table 34). The held-out gain keeps its sign and order of magnitude at both resolutions, close to the main-text 0.0096 nats, but both intervals include zero at 67 players. The dispersion component τ_b is matched. Split-half reliability moves in both directions: tempo is lower and dispersion higher at centisecond resolution. This supports the whole-second operator only on the tested months; the June 2021 month was not tested.

W FURTHER DESCRIPTIVE ANALYSES (NOT PART OF THE CLAIMS)

The analyses in this appendix were run on the same probes and are descriptive; none of this paper's claims rests on them. The within-person residual variance net of each player's own tempo, and a tempo-only empirical-Bayes gain curve on random splits, were also computed: the variance is recomputed on the reported reference in Appendix H, and the curve is given there without values.

Table 34: Whole-second (WS) versus centisecond (CS) fits on the overlap months. T1 is Spearman–Brown-corrected split-half r (bootstrap intervals exist only for the raw r and are not shown); T2 is the held-out reduction in NLL per probe, trained on January–October 2019 and tested on later months, with bootstrap interval.

Quantity	WS	CS
T1 tempo a_j (n)	0.552 (98)	0.455 (100)
T1 dispersion b_j	0.259	0.469
T2 gain (nats/probe)	$+0.0117$ $[-0.0015, +0.0252]$	$+0.0094$ $[-0.0035, +0.0229]$
τ_b	0.152 $[0.000, 0.248]$	0.155 $[0.053, 0.216]$

Sequence pattern after the puzzle move. For the 2019 months the player's next two own moves after the puzzle were extracted (239,513 probes, matching the probe table's clocks exactly). The puzzle move stands out from the player's own game and the next own move is shorter, with small graded rating slopes: $+0.22$, $+0.19$ and $+0.23$ log units per 1,000 rating points for the puzzle move's within-game z-score and the drops to the next one and two own moves (player-cluster bootstrap intervals exclude zero), against a residual interquartile range of 1.3–1.6; the categorical "long-then-short" pattern ($z > 1$ and next move below the game median) is flat at 11.6–11.9% across the middle rating bands and lower (9.3%) at the top. Per player (285 players with at least 23 usable 2019 probes, date split within 2019, raw r): mean z-score 0.395 [0.311, 0.479], mean first drop 0.137 [0.021, 0.254], mean second drop 0.099 [-0.034, 0.236]; permutation nulls 0.006, 0.007, 0.022.

Shape of the residual distribution by rating. On the puzzle probes, pooled naive fits of the residual log time by rating band show a population gradient: the lognormal scale falls at the top band, so the strongest players' residual distribution is narrower. Per player, only the lognormal location and scale are reliable. The sinh-arcsinh asymmetry rows of this analysis (a gradient toward zero and a per-player replication of $r = 0.01$) are withdrawn: they used the simulation library's continuous density, which is not a normalised sinh-arcsinh density (Appendix I). Censored and naive fits of the generalised inverse Gaussian score different objects (an integer probability and a continuous density), so their NLL difference is not a predictive gain; the top two bands' censored parameters sit at a boundary and are not reported as shape estimates.

External validity of ability beyond rating (exploratory). First-half θ_j against the later rating level: Spearman $\rho = 0.11$ [0.03, 0.19] over 487 players; cross-validated ΔR^2 for the rating change $+0.002$, below its 0.01 margin. No convergent test on ordinary moves was run, so the external validity of the trait is weak (Section 6).

X NOT-RUN LIST

Format: *analysis — status — reason*.

- NUTS on the primary cohort — not run under the reported likelihood; the chains launched under an earlier observation model were superseded by the later revisions of the model — mean-field ADVI cross-check reported instead.
- Time columns of Allie's block (rows 8–9 of Table 22) under the zero-mass refit — not rerun — their reference time columns were scored before the Eq. 5 refit, whose size is in Appendix F; every accuracy column is unchanged (Appendix J).
- T4 transfer (ordinary-move tempo) — run after the freeze, exploratory (Appendix N) — its ordinary-move reference carries post-decision capture and check flags.
- Reliability at a fixed temporal gap — run after the freeze — removed from this list; the curve is in Appendix K and Figure 2A.
- Player-level Glicko-2 rating deviation — not available — the game archives carry point ratings only; the puzzle items' own deviation has median 79.

- Time-control-class splits — run after the freeze — removed from this list; the instrument by class is in Appendix P and person-vector transfer across classes in Appendix S.
- Recovery simulation under the measured operator — partially closed — run at the simulation's own sizes (Appendix I.1); the repeat at the exact cohort sizes is not run.
- Operator stationarity in June 2021 — not run — its archive was not scanned (Appendix C.1).
- A personalised move model (for example Maia4All's probability of the solution move from k games of the player) as a row on the accuracy channel of Table 1 — not run — it could be scored exactly as Allie's move head is, on the same AUROC and Brier; the accuracy claim is therefore scoped to the population and clock-history models compared.
- Residual-family rows on the primary cohort; centisecond coarsening refit; training-pool person-vector fit; Set Transformer ablation; full-rank ADVI; a full-rank encoder head; repeated encoder training across folds and seeds with intervals (a five-fold sweep with seed-level variance); encoder weight sweep under the reported fits — not run — outside the compute budget of this revision; their effect on the verdicts is unmeasured; the encoder's seed and fold variability and its objective sensitivity bear on the uncertainty of T3.
- Paired intervals for family differences within the censored treatment — not computed — Table 12 gives each family's difference from the naive lognormal only, and naive and censored rows are on different scales (Appendix D.1).
- The likelihood-only encoder objective under the reported reference — not run — in a single-seed check on the earlier reference it scored 0.01–0.02 nats better than the reported $1{:}1$ objective at every k (Appendix J); the KL term was kept so that the encoder's posterior stays close to the hierarchical one, and the likelihood-only variant was not re-scored under the reported reference.
- An exact per-instance posterior on the k probes (NUTS or quadrature over the latent with p_j free) as the amortization-gap comparator — not run — the marginalised refit of Appendix J holds p_j at its class mean, so the encoder's edge over it may reflect that comparator's shrinkage rather than a negative gap.
- A non-transductive ceiling (a full-information fit on all of the player's probes except the scored ones, a non-optimistic T3 comparator), and the whole-vector ceiling on the held-out-month probes of the 429 primary players with held-out months, with a paired history-minus-ceiling interval — not run — the reported ceiling sees the scored probes and is measured on other players, so it is not compared with clock history's gain.
- A clock-history baseline that also sees within-game move-quality history; a GRU that takes the person vector as an input; person offsets re-estimated conditional on the GRU; a tempo-only random intercept with estimated shrinkage; a sequence model over the player's earlier games — not run — Table 1 row 5 adds the vector as fixed offsets fitted against the reference, and row 3 uses a fixed shrinkage of 0.1.
- The within-person slope λ on the instrument's own 5,174-player sample with player$\times$month effects and on the probability scale; a balance table of pre-decision item covariates by instrument arm; the LATE with the item's rating and reference logit as covariates — not run — the reported estimates come from three samples (Section 5.4).
- Convergent validity of ability beyond rating on ordinary moves with a single best continuation; its reliability within a single time control; its correlation with the player's time-control mix and rating deviation — not run — the ordinary-move table of Appendix N transferred tempo and dispersion only.
- A per-player Spearman–Brown projection at the one-year half sizes, the size-matched comparison without the five-probe players, and BCa intervals for the curve — not run — Appendix K reports percentile player-bootstrap intervals and a mean-of-draws matched value.
- A second domain with calibrated items and response times (a knowledge-tracing corpus) — not run — the four-in-a-row pilot has no calibrated items; it tests timing and an uncalibrated move-quality channel, not ability beyond skill (Appendix T).
- Re-estimation of a person vector from new probes after a gap — not run — the fixed-gap curve (Appendix K) measures reliability across a gap, not recovery after one.

- A two-sample test of person-conditioned simulated think times against recorded ones, conditional on the recorded moves (Lopez-Paz & Oquab, 2017) — not run — held-out proper scores do not establish the fidelity of generated times; the posterior-predictive gates of Appendix O compare slices, and their known failure is the zero-fraction gap.
- ChessMimic (Johnson, 2026) and UniMaia (Siu & Istead, 2026) as further clock-aware population baselines — not run — both are population-level; the GRU over the game's own clocks is the clock-history population baseline of Table 1.
- An engine-based success label that credits alternative winning continuations (a played move that wins but is not the puzzle's first solution move) — not run — the outcome scores only the puzzle's designated first move (Section 3, Appendix B).
- A comparison of the empirical-Bayes posterior variances with exact inference (posterior coverage) — not run — the ADVI cross-check compares point estimates only (Appendix D).
- A censored-survival treatment of flagged and timed-out decisions — not run — rows whose think time exceeds the remaining clock, and flagged rows, are excluded rather than modelled as censored (Appendices B and D).
- A permuted-player encoder control (a held-out player's posterior formed from k probes of another player) and a marginalised population-only control — not run — they would separate person information from a player-level correction of the reference; the zero-probe control of Section 5.3 excludes only a constant class-level shift.
- A paired interval on the encoder's shortfall from the full-information fit at $k = 20$ — not computed — the paired intervals of Appendix J cover the encoder against the reference and the refits only.
- A test of the instrument's monotonicity (whether an increment lowers any player's think time) — not run — monotonicity is assumed (Section 5.4).
- Coverage of the encoder's Gaussian against the posterior given the same k probes — not run — the encoder is scored on held-out prediction only (Appendix J).
- A within-player (stratified) AUROC — not computed — θ_j is constant within a player and leaves each player's order unchanged.
- The pilot gains behind the margins — not reported — they were measured under other population models than the reported reference (Section 4.3).
- Items added after a later check of the code (each a re-run, none started): the recoverability simulation with a corrected and tested sinh-arcsinh or lognormal family, including the naive comparisons, the family-gap calibration and the naive asymmetry fits of Appendix W (Appendix I); the few-shot tests with the population prior, class means, intercepts, zero-mass logit and input standardisation re-estimated without the held-out players (Section 5.3, Table 4); the 79-month encoder with three genuine seeds (Appendix Q); a benchmark with a deconvolved within-person variance (Appendix H); fixed-gap slope intervals from a player-paired bootstrap with a fixed seed (Appendix K); the population variances centred within time-control classes (Appendix D) — not run — the affected statements are scoped or withdrawn in place; whether any number would change is unmeasured.