PSYCAT: ROBUST POSTERIOR MIXTURES FOR ADAPTIVE SELF-REPORT ASSESSMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Computerized adaptive testing (CAT) can shorten self-report questionnaires, reducing respondent burden while preserving assessment efficiency. Yet self-report CAT faces two challenges. First, self-report responses have no objectively correct answer, and differences in category interpretation and preferences for extreme or intermediate options can confound the latent trait with response style. Second, maximum-information selection can repeatedly probe one trait, revealing the system's provisional assessment and potentially inducing measurement reactivity. PsyCAT is proposed to systematically address these challenges. Response style is explicit modelled in proposed method in an ordinal measurement model and derives how uncertainty in style increases trait-estimation risk. A Gaussian mixture characterizes the joint posterior over latent traits and response style, separating uncertainty within components from disagreement across components. Shared-attention post-calibration adjusts these posteriors using observed answers, and minimax aggregation produces trait estimates. PsyCAT further controls information leakage by penalizing dependence between next-item selection and alternative response profiles. Across several published questionnaire datasets, PsyCAT achieved the lowest mean endpoint ranked probability score among the evaluated methods. Under controlled latent interventions, PsyCAT tracked response-style changes with little average trait drift in the style-only condition and tracked both variables in the joint condition. In a two-expert study, its selection policy is proven to reduce information leakage.

1 INTRODUCTION

Computerized adaptive testing (CAT) enables efficient psychometric measurement by selecting each question from previous responses and calibrated item-response models. It can reduce questionnaire length and respondent burden while estimating latent traits under a limited item budget, making it particularly relevant to self-report assessment. Applications span cognitive, personality, and symptom measurement (Gibbons et al., 2024; Simms & Clark, 2005; Gibbons et al., 2012).

Unlike objectively scored tests, self-report categories such as "sometimes" and "often" depend on personal interpretation, recall, and reference standards (Schwarz, 1999). Scale design can further shift judgments and encourage extreme or intermediate responding (Schwarz et al., 1985; Bhaktha et al., 2026; Lubbe & Schuster, 2020). This self-report bias is represented as response style. Consequently, the same trait level can yield different responses, while similar responses can reflect different trait levels.

Experimental demand characteristics can affect behavior (Orne, 1962), and measurement itself can alter health-related behavior (Miles et al., 2020). In CAT, repeatedly selecting items from one symptom domain could reveal the system's provisional assessment and affect subsequent reporting (Figure 1). This motivates PsyCAT's penalty on the dependence of next-item selection on alternative response profiles. We evaluate sequence inferability through an expert observation study.

PsyCAT explicitly models response style jointly with latent traits, separating variation in category use from variation in the measured traits. Ambiguous self-report responses increase uncertainty in measurement-model parameters, so a single posterior may omit plausible trait–style relations. A Gaussian mixture retains these alternatives, preserving uncertainty within posteriors and agreement or disagreement across them. Shared-attention post-calibration incorporates information from the

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.

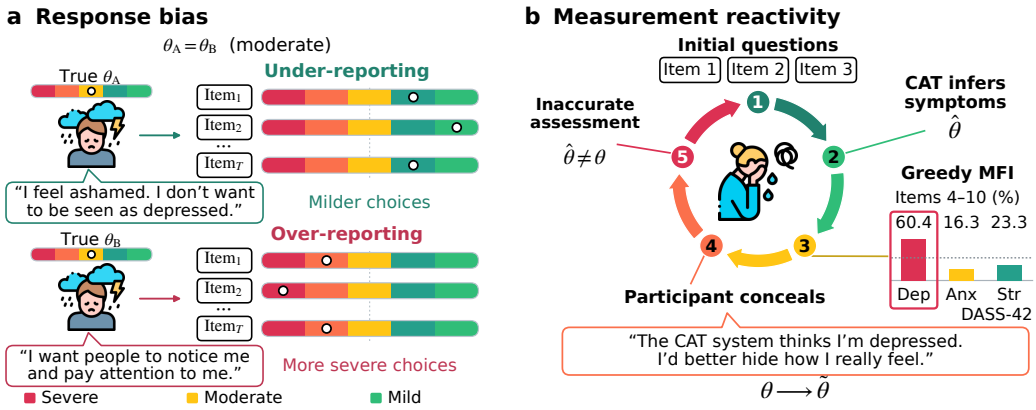

Figure 1: Two challenges in self-report CAT. (a) Response style can produce under- or over-reporting at the same trait level. (b) Repeated domain-specific questions can reveal the system's inference and induce measurement reactivity.

evolving item–answer context, which is particularly relevant when earlier self-report items influence later responses. An information-regularized selection policy limits the dependence of selected questions on response profiles, reducing cues that may induce measurement reactivity.

The contributions are threefold.

- **Explicit and robust trait–style modelling.** PsyCAT represents response style directly, derives its induced estimation risk, and provides theoretical justification for Gaussian-mixture and minimax aggregation.
- **Information-controlled adaptive selection.** PsyCAT limits information leakage from selected questions, reducing cues that may induce measurement reactivity.
- **Multi-level empirical evaluation.** Experiments assess predictive performance, trait–style separation, and reduced inferability of selected question sequences.

2 Related Work

Self-report CAT connects adaptive item selection with uncertainty arising from response styles. The following review relates these foundations to PsyCAT's estimation and selection design.

Computerized adaptive testing. CAT updates a latent-trait posterior after each response and selects items until a budget or precision criterion is met. Classical rules maximize Fisher information or reduce posterior risk, with EAP reporting the posterior mean. The graded response model (GRM) links ordered responses to latent traits, typically estimated by marginal likelihood (Samejima, 1969; Bock & Aitkin, 1981). Learned policies include neural reinforcement learning in NCAT (Zhuang et al., 2022), adaptive test structures in UATS (Yu et al., 2024), examinee–item relations in RECAT (Wang et al., 2025), and Transformer-based response prediction (Rudolph et al., 2026).

Response styles and calibrated ensembles. Scaled-threshold models represent extreme and intermediate response styles through threshold spacing (Lubbe & Schuster, 2020). Expectation–maximization estimates latent-variable models (Dempster et al., 1977), while bootstrap refitting captures calibration variation (Efron, 1979). Mixture density networks parameterize conditional mixtures with neural outputs (Bishop, 1994). PsyCAT combines Gaussian calibration components, shared-attention post-calibration, and minimax trait reporting.

3 Method

Building on adaptive selection and response-style calibration, PsyCAT couples trait–style inference with information-controlled item selection. Sections 3.1–3.3 define the objective, explicit style model, and Gaussian-mixture uncertainty; Sections 3.4–3.6 develop shared-attention post-

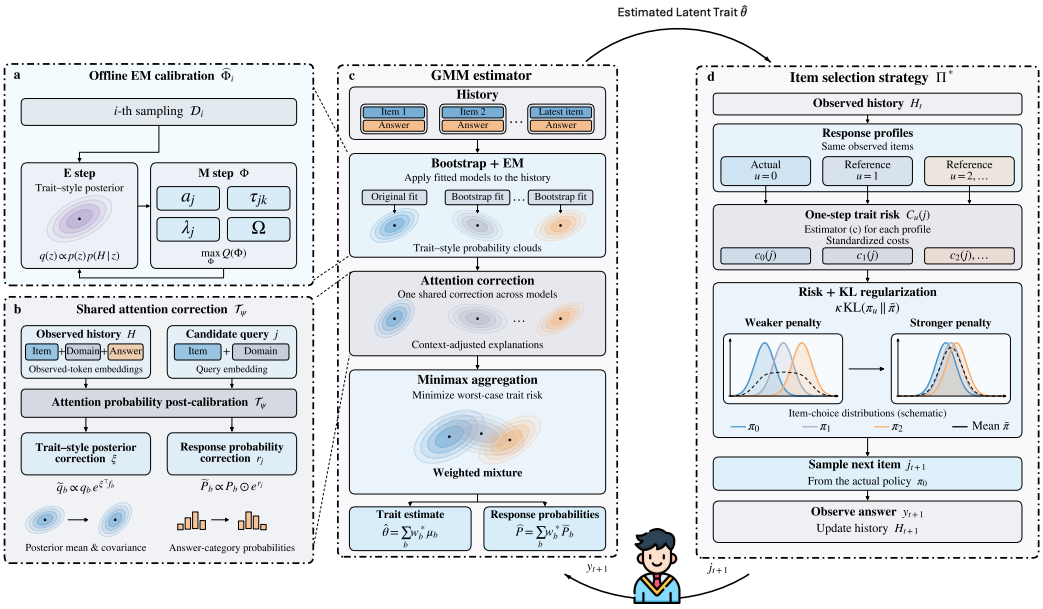

Figure 2: PsyCAT overview. (a) Offline EM calibrates the response-style model on resampled training records. (b) Shared attention corrects trait–style posteriors and response probabilities. (c) Bootstrap calibrations and minimax aggregation produce trait estimates and response predictions. (d) Information-regularized selection chooses the next item; the observed response updates the history.

calibration, minimax reporting, and information-regularized selection. Figure 2 summarizes the estimation and selection cycle.

3.1 Task and Decision Objectives

Respondent i has latent traits $\theta_i \in \mathbb{R}^D$ and response style s_i, with $\theta_{i,d}$ denoting the trait in domain d. Each item j in pool $\mathcal{Q}$ measures domain d_j and has ordered categories $1, \ldots, K_j$. History $H_t = \{(j_r, y_r)\}_{r=1}^t$ records each administered item j_r and response y_r; $\mathcal{S}_t = \{j_r\}_{r=1}^t$ contains answered items. PsyCAT's policy Π^* selects from $\mathcal{Q} \setminus \mathcal{S}_t$, balancing expected trait-risk reduction against information conveyed about response profiles.

Definition 1 (Self-report adaptive assessment)**.** For respondent i and fixed budget T, choose a trait estimator $\mathcal{E}$ and adaptive policy Π to minimize expected trait-estimation error:

$$\min_{\mathcal{E}, \Pi} \ \mathbb{E}_\Pi \left[\| \mathcal{E}(H_T) - \theta_i \|_2^2 \right]. \tag{1}$$

PsyCAT jointly infers traits and style, then marginalizes style for trait reporting.

The expectation covers latent traits, style, responses, and randomized selection under the assessment model. PsyCAT addresses this objective through a staged estimator and one-step posterior-risk surrogates with information regularization.

3.2 Explicit Response Style Modelling

For a fixed respondent and step, write $H = H_t$, $\mathcal{S} = \mathcal{S}_t$, and $z = (\theta, s) \in \mathbb{R}^{D+1}$. Subsequent posteriors condition on observed answers; measurement-model training records remain fixed during assessment.

Traits describe the measured content; style describes how respondents use ordered categories. We extend the GRM by letting style expand or contract threshold spacing:

$$\begin{aligned} \tau_{jk}(s) &= \bar{\tau}_j + e^{\lambda_j s} (\tau_{jk} - \bar{\tau}_j), \\ \Pr(Y_j > k \mid \theta, s) &= \sigma\big(a_j \theta_{d_j} - \tau_{jk}(s)\big). \end{aligned} \tag{2}$$

Here Y_j is the item response, σ is the logistic function, and $k = 1, \ldots, K_j - 1$ indexes ordered baseline thresholds τ_{jk} with mean $\bar{\tau}_j$. Discrimination $a_j > 0$ controls sensitivity to θ_{d_j}; $\lambda_j > 0$ controls style sensitivity. Positive s expands threshold spacing; negative s contracts it. Traits shift position relative to thresholds, whereas style changes their spacing (Samejima, 1969; Lubbe & Schuster, 2020).

Adjacent cumulative probabilities give the response probability $p(y \mid j, z)$. The prior is $p(z) = \mathcal{N}(z \mid 0, \Omega)$, with fitted correlation matrix Ω having unit diagonal. A calibration $\Phi = \{a_j, \tau_{jk}, \lambda_j, \Omega\}$ learned from training responses defines the joint posterior

$$q(z) \propto p(z) \prod_{(j,y) \in H} p(y \mid j, z). \tag{3}$$

Whereas $p(y \mid j, z)$ predicts responses, q describes trait–style combinations supported by observed answers. Their dependence allows style uncertainty to affect trait inference; reporting marginalizes over s.

3.3 Gaussian mixture model

To quantify the trait-estimation error attributable to unobserved style, we compare estimation with and without style information, holding answers and fitted parameters fixed. Under q, $\mathbb{E}_q$, Var_q, and Cov_q denote expectation, variance, and covariance. A covariance matrix records coordinate variances on its diagonal and pairwise covariation elsewhere; its trace tr sums the variances.

The posterior mean $\hat{\theta}_q = \mathbb{E}_q[\theta]$ minimizes squared error when style is unknown; observing style gives $\hat{\theta}_q(s) = \mathbb{E}_q[\theta \mid s]$. Their risks, summed over trait domains, are

$$R(q) = \mathbb{E}_q\left[\|\theta - \hat{\theta}_q\|_2^2\right] = \sum_{d=1}^{D} \mathrm{Var}_q(\theta_d) = \mathrm{tr}\,\mathrm{Cov}_q(\theta), \tag{4}$$

$$R(q \mid s) = \mathbb{E}_q\left[\|\theta - \hat{\theta}_q(s)\|_2^2 \mid s\right] = \sum_{d=1}^{D} \mathrm{Var}_q(\theta_d \mid s) = \mathrm{tr}\,\mathrm{Cov}_q(\theta \mid s). \tag{5}$$

Definition 2 (Excess risk from unobserved response style)**.** For a joint distribution $q(\theta, s)$ with finite second moments, define

$$\Delta_s(q) = R(q) - \mathbb{E}_q[R(q \mid s)]. \tag{6}$$

Averaging style-conditional risk over q gives the expected error if style were observed. Thus, $\Delta_s(q)$ measures the reduction in minimum expected squared error from observing style, with answers and fitted model unchanged.

Proposition 1 (Risk from unknown response style)**.** *For any joint distribution $q(\theta, s)$ with finite second moments,*

$$\Delta_s(q) = \mathrm{tr}\,\mathrm{Cov}_q(\mathbb{E}_q[\theta \mid s])\,. \tag{7}$$

Unobserved style increases trait-estimation risk whenever the conditional trait mean varies with style (Appendix A.2). Such distortions can alter recovered trait structure and propagate to biased, less powerful downstream analyses (Laajaj et al., 2019; Schoeler et al., 2025). Because finite samples admit multiple calibrations that explain the same responses through different trait–style relations, a single-fit posterior understates this uncertainty.

A Gaussian mixture model (GMM) represents these alternatives with distinct joint means m_b, covariances V_b, and weights w_b. Each component retains uncertainty and trait–style dependence within one explanation. For M components indexed by $b = 0, \ldots, M - 1$,

$$g_b(z) = \mathcal{N}(z \mid m_b, V_b), \qquad G_w(z) = \sum_{b=0}^{M-1} w_b g_b(z). \tag{8}$$

Here $w \in \Delta_M = \{w \in \mathbb{R}^M : w_b \geq 0, \sum_b w_b = 1\}$. Superscripts select trait and style coordinates: $m_b^\theta \in \mathbb{R}^D$, $m_b^s \in \mathbb{R}$, and covariance blocks $V_b^{\theta\theta}$, $V_b^{\theta s}$, V_b^{ss}. For a Gaussian component with

$V_b^{ss} > 0$,

$$\mathbb{E}_{g_b}[\theta \mid s] = m_b^\theta + \frac{V_b^{\theta s}}{V_b^{ss}}(s - m_b^s),$$
$$\Delta_s(g_b) = \frac{\|V_b^{\theta s}\|_2^2}{V_b^{ss}}. \tag{9}$$

The ratio $V_b^{\theta s}/V_b^{ss}$ describes how the conditional trait mean changes with style; the second line relates this dependence to excess risk. A single Gaussian imposes a linear conditional mean and style-independent conditional covariance. In a mixture, conditional weights depend on s, permitting a nonlinear conditional mean and style-dependent covariance (Appendix A.3).

Style can be marginalized analytically. Let G_w^θ be the trait marginal, $\bar{m}_w^\theta = \sum_b w_b m_b^\theta$ its mean, and $\hat{\theta} \in \mathbb{R}^D$ a candidate report. Then

$$G_w^\theta(\theta) = \sum_b w_b \mathcal{N}(\theta \mid m_b^\theta, V_b^{\theta\theta}),$$
$$\min_{\hat{\theta}} \mathbb{E}_{G_w}\left[\|\theta - \hat{\theta}\|_2^2\right] = \sum_b w_b \operatorname{tr} V_b^{\theta\theta} + \sum_b w_b \|m_b^\theta - \bar{m}_w^\theta\|_2^2. \tag{10}$$

The two terms measure within-component uncertainty and disagreement between component means. Each component's risk is $\mathbb{E}_{g_b}\|\theta - \hat{\theta}\|_2^2 = \|\hat{\theta} - m_b^\theta\|_2^2 + \operatorname{tr} V_b^{\theta\theta}$. These calculations require only component moments and weights; retaining separate risks supports the minimax analysis below.

Calibration uncertainty is represented by the original fit and B whole-record bootstrap refits (Efron, 1979), yielding $M = B + 1$ members estimated by approximate generalized EM (Dempster et al., 1977). Resampling complete response records produces alternative item parameters and latent correlations. For a common history, each member defines a posterior q_b, with p_b denoting its prior and response probabilities. The moments $m_b = \mathbb{E}_{q_b}[z]$ and $V_b = \operatorname{Cov}_{q_b}(z)$ form the GMM components in Eq. (8). Moment matching preserves squared-error reporting risk while exposing calibration-dependent trait–style relations hidden by a single fit (Appendix A.3). Component differences capture calibration variation, whereas Δ_s isolates the effect of unobserved style within a member.

3.4 Shared-attention post-calibration

The measurement model assumes that responses are independent of preceding questions conditional on (θ, s), yet self-report judgments can shift with retrieved information and comparison standards (Tourangeau & Rasinski, 1988; Schwarz et al., 1991), producing documented order effects (Kwan et al., 2012; Bonning et al., 2025). This mismatch necessitates post-calibration from the accumulated item–answer context. A shared-attention network (Vaswani et al., 2017) applies self-attention to observed pairs and cross-attention between candidate items and history, with unobserved responses masked. It outputs a shared latent tilt $\xi = \xi_\psi(H)$ and item-specific residuals $r_j = r_{\psi,j}(H)$ to correct trait–style posteriors and response probabilities.

Let $f_b(z)$ contain the standardized linear and upper-triangular quadratic features of z, with $(D+1)(D+4)/2$ entries. The latent correction is

$$\widetilde{q}_b(z) = \frac{q_b(z)\exp\left(\xi^\top f_b(z)\right)}{\mathbb{E}_{q_b}[\exp(\xi^\top f_b(z))]}. \tag{11}$$

The tilt redistributes posterior mass, adjusting Gaussian means, dispersions, and trait–style dependence while retaining Gaussianity when the updated precision is positive definite (Appendix A.7). For the implemented finite-support distributions, normalization exists for finite logits.

Structural predictions $P_b(y \mid j) = \mathbb{E}_{\widetilde{q}_b}[p_b(y \mid j, z)]$ receive category residuals $r_{j,y}$:

$$\widetilde{P}_b(y \mid j) = \frac{P_b(y \mid j)\exp(r_{j,y})}{\sum_{c=1}^{K_j} P_b(c \mid j)\exp(r_{j,c})}. \tag{12}$$

Training uses withheld-response log loss, auxiliary structural-prediction loss, and a Kullback–Leibler (KL) penalty between corrected and base posteriors, with calibration parameters frozen.

The encoder represents the evolving answer set through permutation-invariant attention, capturing interactions among observed item–answer pairs. Appendix A.9 details this representation.

3.5 MINIMAX AGGREGATION

Corrected members can differ in trait means and uncertainty. With style marginalized, let $\mu_b = \mathbb{E}_{\widetilde{q}_b}[\theta]$ and $v_b = \operatorname{tr}\operatorname{Cov}_{\widetilde{q}_b}(\theta)$. Member risk $R_b(\hat{\theta}) = v_b + \|\mu_b - \hat{\theta}\|_2^2$ combines posterior uncertainty with disagreement between the report and member mean. Equal averaging minimizes average risk, whereas minimax aggregation minimizes the largest member risk and has no greater maximum risk than any fixed-weight average. Define $\bar{\mu}_w = \sum_b w_b \mu_b$.

Proposition 2 (Minimax risk and least-favorable mixture)**.** *For finitely many member distributions with finite second moments, the minimax risk and an attaining trait report satisfy*

$$\min_{\hat{\theta}\in\mathbb{R}^D} \max_b R_b(\hat{\theta}) = \max_{w\in\Delta_M} \sum_b w_b \left[v_b + \|\mu_b - \bar{\mu}_w\|_2^2\right], \tag{13}$$

$$\hat{\theta} = \bar{\mu}_{w^*} = \sum_b w_b^* \mu_b. \tag{14}$$

Here w^* maximizes the right-hand side of Eq. (13).

The weights optimize worst-case risk across retained calibrations (proof and procedure in Appendix A.6). They also aggregate corrected response probabilities, $\widehat{P} = \sum_b w_b^* \widetilde{P}_b$; category residuals do not enter trait-risk optimization.

3.6 INFORMATION LEAKAGE CONTROL FOR ITEM SELECTION

To balance trait risk against information conveyed by item selection, first form structural hypothetical updates $\widetilde{q}_b^{j,y}(z) \propto \widetilde{q}_b(z) p_b(y \mid j, z)$ and minimax reports $\hat{\theta}^{j,y}$. The one-step cost is

$$C(j) = \max_b \sum_{y=1}^{K_j} P_b(y \mid j) \mathbb{E}_{\widetilde{q}_b^{j,y}}\left[\|\theta - \hat{\theta}^{j,y}\|_2^2\right]. \tag{15}$$

This lookahead uses structural response probabilities and the current Transformer correction for all hypothetical answers.

Let $\mathcal{U}$ index response profiles sharing the observed-item mask: $u = 0$ is the actual profile, and the others are training-derived references. Standardizing each profile's costs over available items gives c_u. Policy Π^* approximately solves

$$\min_{\{\pi_u\}} \frac{1}{|\mathcal{U}|} \sum_{u\in\mathcal{U}} \left[\sum_{j\in\mathcal{Q}\setminus\mathcal{S}} \pi_u(j) c_u(j) + \kappa\ \mathrm{KL}(\pi_u \| \bar{\pi})\right], \tag{16}$$

where $\pi_u(j)$ is the selection probability for available item j, $\bar{\pi} = \frac{1}{|\mathcal{U}|}\sum_{u\in\mathcal{U}} \pi_u$, and $\kappa > 0$ is validation-selected. The mean KL penalty equals mutual information $I(U; J)$ for uniform profile index U and next item $J \mid U = u \sim \pi_u$. We use alternating Blahut–Arimoto updates (Blahut, 1972; Arimoto, 1972):

$$\begin{aligned} \pi_u(j) &\leftarrow \frac{\bar{\pi}(j) e^{-c_u(j)/\kappa}}{\sum_{\ell\in\mathcal{Q}\setminus\mathcal{S}} \bar{\pi}(\ell) e^{-c_u(\ell)/\kappa}}, \\ \bar{\pi}(j) &\leftarrow \frac{1}{|\mathcal{U}|} \sum_{u\in\mathcal{U}} \pi_u(j). \end{aligned} \tag{17}$$

Iterative updates yield a policy conditioned on the observed history, from which the next item is sampled. The penalty reduces distinguishability among next-item distributions across response profiles, controlling profile-dependent selection separately from minimax aggregation of calibration uncertainty.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETTING

Datasets. Held-out response prediction was evaluated on six questionnaire banks using five random train/validation/test splits. Main-text results cover DASS-42, Xinxiang-23, and Corona-23 (Open-Source Psychometrics Project, 2019; Su et al., 2024; Winter et al., 2026), with maximum budgets of 20, 15, and 15 items, respectively (Table 1). All methods used identical permanently hidden query items. Appendix results cover IPIP-50, PROMIS Pediatric Life Satisfaction, and OECD SSES 2019 (Table 3); split sizes and replay protocols appear in Appendix B.

Table 1: Filtered main datasets. Samples denote records, including repeated Corona visits; domains give item counts, and response shares use eligible responses. Scored categories are bold; split sizes appear in Table 4.

Dataset	Sample size	Items by domain	Response share (%)
DASS-42	32,495	Depression / anxiety / stress: 14/14/14	**0:** 28.0 / **1:** 31.5 / **2:** 20.1 / **3:** 20.5
Xinxiang-23	24,073	Depression / anxiety / insomnia: 9/7/7	**0:** 71.0 / **1:** 24.8 / **2:** 3.1 / **3:** 1.0 / **4:** 0.1
Corona-23	14,082	Depression / anxiety / insomnia: 9/7/7	**0:** 43.7 / **1:** 32.2 / **2:** 12.1 / **3:** 10.6 / **4:** 1.4

Comparators and selection. Comparators included Random + EAP, EAP-MFI, Transformer-U, and ordinal NCAT, UATS, and RECAT (Zhuang et al., 2022; Yu et al., 2024; Wang et al., 2025). Validation-selected configurations were fixed across test budgets (Appendix B). Appendix C.2 analyzes sensitivity to bootstrap count B and GMM size $M = B + 1$.

Metrics. Reported metrics were multiclass accuracy (ACC), area under the receiver operating characteristic curve (AUC), and normalized ranked probability score (RPS). ACC measures top-category accuracy; AUC and ACC are maximized, whereas RPS is minimized. For item j with K_j ordered categories, prediction p_j, and response y_j,

$$\mathrm{RPS}(p_j, y_j) = \frac{1}{K_j - 1} \sum_{k=1}^{K_j - 1} \left[\sum_{c=1}^{k} p_j(c) - \mathbf{1}\{y_j \le k\} \right]^2 . \tag{18}$$

RPS measures ordered cumulative-probability error. RPS and one-versus-rest AUC were averaged equally over eligible items, whereas ACC was weighted by query-response counts. Five-split means and sample standard deviations are reported; full and domain-level results appear in Appendices C and D.

4.2 MAIN RESULTS

Table 2 and Figure 3 report performance across budgets using seeds 2026–2030 with identical split-specific partitions and query masks.

PsyCAT achieved the best mean AUC at every tabulated budget. ACC and RPS also led, except at five items in Xinxiang-23, where ACC was 79.63% versus 79.73% for EAP-MFI. Both exceptions were concentrated in depression, with gains in anxiety and insomnia partly offsetting the shortfalls. By ten items, PsyCAT led whole-bank ACC and RPS. Appendix D.1 reports the domain decomposition and its relation to item allocation and response distributions.

4.3 COUNTERFACTUAL ANALYSIS

Paired-response DASS-42 experiments tested separation of response-style and depression-trait changes. For each synthetic respondent, pre/post answers shared the baseline state and random draws. Interventions changed style alone, $(\Delta s, 0)$, or style and depression jointly, $(\Delta s, \Delta\theta_{\mathrm{dep}})$. Asymmetric generators varied threshold gaps independently beyond the fitted uniform-scaling form. All methods processed the same 20 items; a common frozen decoder mapped pre/post response probabilities to trait–style changes, isolating estimation from selection. Figure 4 shows initial truth plus estimated change, with stars marking targets. PsyCAT separated style from depression under

Table 2: Response prediction at specified item budgets, averaged over five seeds (2026–2030). Multiclass ACC/AUC pairs are percentages; RPS is unscaled. Bold marks the best unrounded mean for each metric and budget. Standard deviations and all seven metrics appear in Appendix C.

(a) Results on DASS-42

Metric	Multiclass ACC/AUC (%) ↑				RPS ↓			
Method / Budget	5	10	15	20	5	10	15	20
Random + EAP	51.20/75.39	53.27/77.40	54.50/78.31	54.78/78.86	0.1381	0.1296	0.1256	0.1233
EAP-MFI	50.57/73.96	53.32/76.94	54.72/78.12	55.30/78.77	0.1391	0.1295	0.1254	0.1232
Transformer-U	50.97/75.83	53.98/78.45	54.68/79.55	55.61/80.07	0.1384	0.1294	0.1258	0.1241
NCAT	51.11/75.48	53.61/77.53	54.49/78.49	54.93/78.85	0.1368	0.1285	0.1251	0.1235
UATS	50.24/75.10	54.00/77.39	54.41/78.16	55.01/78.67	0.1379	0.1286	0.1252	0.1234
RECAT	51.76/75.71	53.92/77.47	54.68/78.20	55.30/78.77	0.1348	0.1282	0.1251	0.1232
PsyCAT	**53.01/77.11**	**55.23/79.20**	**56.35/80.20**	**57.00/80.77**	**0.1321**	**0.1252**	**0.1218**	**0.1198**

(b) Results on Xinxiang-23

Metric	Multiclass ACC/AUC (%) ↑			RPS ↓		
Method / Budget	5	10	15	5	10	15
Random + EAP	78.58/86.07	80.06/88.16	80.79/89.09	0.0515	0.0474	0.0455
EAP-MFI	**79.73**/87.50	80.63/88.82	80.98/89.31	**0.0485**	0.0459	0.0449
Transformer-U	78.57/86.86	80.39/89.17	80.91/90.10	0.0512	0.0466	0.0452
NCAT	78.90/86.49	80.24/88.40	80.91/89.11	0.0506	0.0472	0.0454
UATS	79.12/85.86	80.64/88.84	80.96/89.35	0.0521	0.0462	0.0449
RECAT	79.66/87.39	80.70/88.79	80.97/89.29	0.0487	0.0458	0.0449
PsyCAT	79.63/**87.69**	**80.80/89.73**	**81.32/90.33**	0.0490	**0.0453**	**0.0440**

(c) Results on Corona-23

Metric	Multiclass ACC/AUC (%) ↑			RPS ↓		
Method / Budget	5	10	15	5	10	15
Random + EAP	55.75/78.82	57.61/81.15	58.45/81.87	0.1136	0.1060	0.1031
EAP-MFI	57.56/80.13	58.47/81.48	59.03/82.05	0.1073	0.1037	0.1020
Transformer-U	57.66/81.01	60.86/84.18	62.80/85.56	0.1075	0.0975	0.0923
NCAT	56.40/79.75	58.24/81.11	58.50/81.92	0.1104	0.1054	0.1027
UATS	54.37/78.85	58.21/81.42	58.72/81.93	0.1160	0.1042	0.1025
RECAT	57.65/80.02	58.45/81.57	58.93/82.03	0.1081	0.1032	0.1020
PsyCAT	**59.75/82.21**	**63.34/84.60**	**64.58/85.86**	**0.1019**	**0.0925**	**0.0885**

style-only shifts and tracked both under joint shifts. Each panel contains 5,120 estimates; controls appear in Appendix E.

4.4 Expert evaluation of information-regularized selection

Two psychology experts evaluated ordered, 15-item Xinxiang-23 sequences from frozen PsyCAT and PsyCAT-MFI models. The models shared an estimator but used information-regularized and maximum Fisher information selection, respectively. Blinded to responses and methods, experts inferred four-level severity; binary low–high accuracy was primary and exact agreement secondary. Expert-stratified exact tests assumed case independence (Appendix F). MFI items lead to higher inference accuracy for both outcomes across experts which is information leakage (Figure 5).

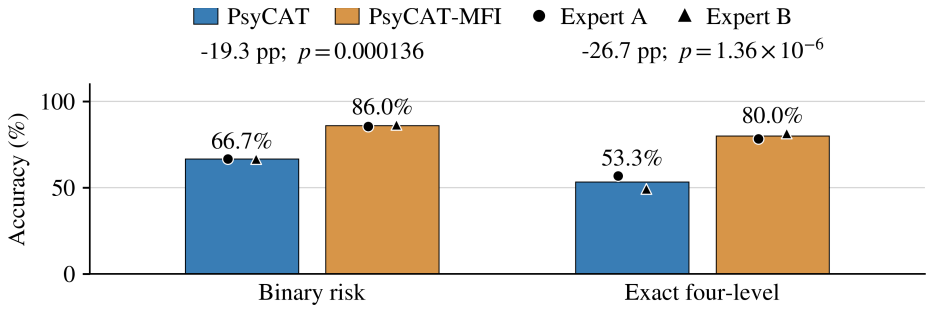

Figure 5: Expert inference accuracy. Markers denote experts; differences are PsyCAT minus PsyCAT-MFI. Exact tests are two-sided and expert-stratified.

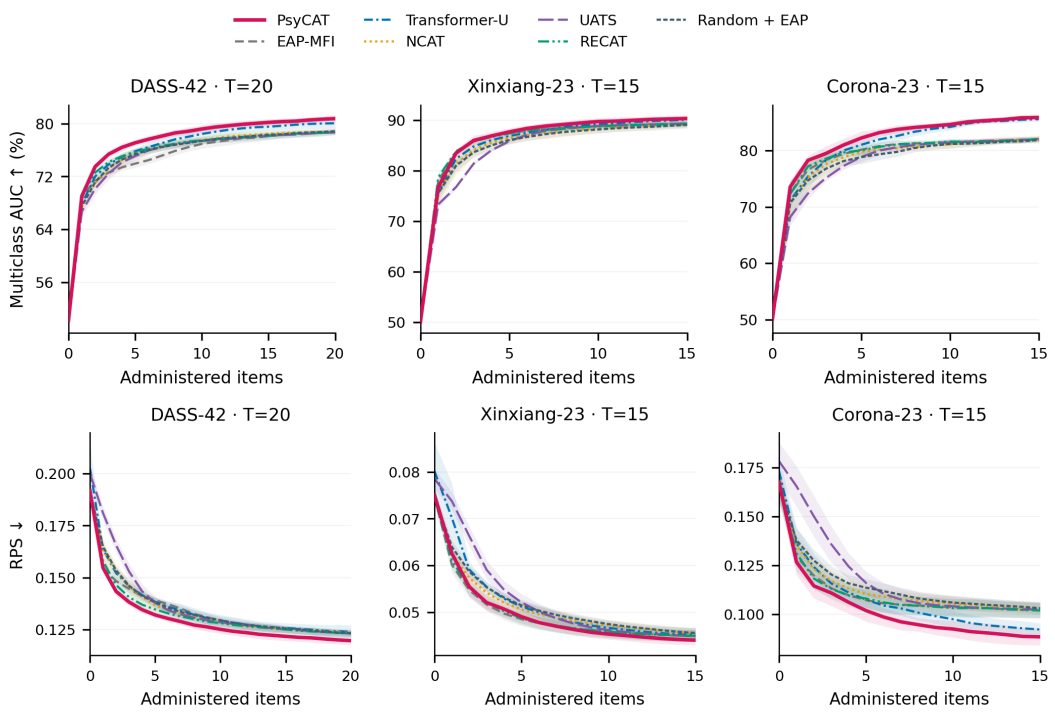

Figure 3: Multiclass AUC (top) and normalized RPS (bottom) across budgets. Curves and shading show five-seed means and sample standard deviations for identical hidden query items.

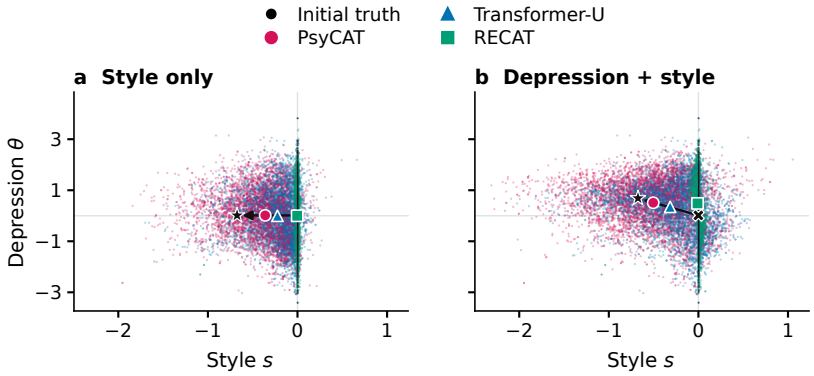

Figure 4: Asymmetric DASS-42 interventions: (a) style-only; (b) joint depression–style. Faint colored points show initial truth plus decoded change. Black asterisks mark targets; red circles, blue triangles, and green squares mark the PsyCAT, Transformer-U, and RECAT means. GRM-based CAT lacks an explicit style variable, while Transformer-U may absorb style shifts without identifying them; PsyCAT estimates style and trait separately.

5 Discussion

PsyCAT combines explicit response-style modelling, shared-attention post-calibration, Gaussian-mixture uncertainty, minimax aggregation, and information-regularized selection. Across six banks, it achieved the lowest mean endpoint RPS, tracked trait–style changes under controlled interventions, and reduced system-label inferability from selected questions. These results cover predictive performance, model-based latent recovery, and observer inference, but not behavioral effects during live administration. Prospective studies are therefore needed to test measurement reactivity, alongside broader inference grids and external measurement comparisons. Numerical diagnostics and sampling assumptions appear in Appendices B and F.

AI USE STATEMENT

Generative AI tools assisted with drafting, translation, and language polishing to improve clarity, concision, and terminology consistency. They supported searches for related work and baseline methods published in leading peer-reviewed conferences and journals, and checks of bibliographic details and citation relevance against source publications. AI assistance also covered the inspection of implementation code and experimental outputs, and analysis of expert-study results.

ETHICS STATEMENT

The study uses existing questionnaire data, controlled simulations, and expert evaluations. Two psychology experts reviewed question sequences derived from published data, with respondent answers and method identities withheld. The submission reports aggregate results and excludes participant-level records and expert identities.

REPRODUCIBILITY STATEMENT

The appendices provide model equations, proofs, training and selection settings, metric definitions, data preparation, baseline adaptations, and numerical diagnostics. The source package includes the manuscript, original displayed graphics, aggregate tables, and provenance and citation audits. Displayed results derive from frozen benchmark outputs and author-provided expert-study judgments. Expert-level aggregate counts, analysis assumptions, and an exact-test verification script accompany the study figure. The package supports verification of reported results and theoretical identities. Training reproduction requires the questionnaire data and training pipeline described in the appendices, subject to the source datasets' access conditions. All main and supplementary files are prepared without author-identifying affiliations or participant-level data.

REFERENCES

S. Arimoto. An algorithm for computing the capacity of arbitrary discrete memoryless channels. *IEEE Transactions on Information Theory*, 18(1):14–20, 1972. doi: 10.1109/tit.1972.1054753. URL https://doi.org/10.1109/tit.1972.1054753.

Nivedita Bhaktha, Henning Silber, and Clemens M. Lechner. A unified framework for characterizing response quality in surveys with multi-item scales. *Communications Psychology*, 4 (1):86, 2026. doi: 10.1038/s44271-026-00463-2. URL https://doi.org/10.1038/s44271-026-00463-2.

Christopher M. Bishop. Mixture Density Networks. Technical Report NCRG/94/004, Aston University, 1994. URL https://www.microsoft.com/en-us/research/publication/mixture-density-networks/.

R. Blahut. Computation of channel capacity and rate-distortion functions. *IEEE Transactions on Information Theory*, 18(4):460–473, 1972. doi: 10.1109/tit.1972.1054855. URL https://doi.org/10.1109/tit.1972.1054855.

R. Darrell Bock and Murray Aitkin. Marginal Maximum Likelihood Estimation of Item Parameters: Application of an EM Algorithm. *Psychometrika*, 46(4):443–459, 1981. doi: 10.1007/bf02293801. URL https://doi.org/10.1007/bf02293801.

Eva Bonning, Mevagh Sanson, Andrea Taylor, Rachel Zajac, and Maryanne Garry. The effects of symptom scale order on people's judgments of their psychological experiences. *Psychology of Consciousness: Theory, Research, and Practice*, 12(4):745–758, 2025. doi: 10.1037/cns0000430. URL https://doi.org/10.1037/cns0000430.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977. doi: 10.1111/j.2517-6161.1977.tb01600.x. URL https://doi.org/10.1111/j.2517-6161.1977.tb01600.x.

B. Efron. Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*, 7(1): 1–26, 1979. doi: 10.1214/aos/1176344552. URL https://doi.org/10.1214/aos/1176344552.

Christopher Forrest. PROMIS Pediatric Measure Evaluation, 2021. URL https://doi.org/10.7910/DVN/QOO7QX. Version 1.0; Life Satisfaction child-report developmental data; accessed September 25, 2026.

Robert D. Gibbons, David J. Weiss, Paul A. Pilkonis, Ellen Frank, Tara Moore, Jong Bae Kim, and David J. Kupfer. Development of a Computerized Adaptive Test for Depression. *Archives of General Psychiatry*, 69(11):1104–1112, 2012. doi: 10.1001/archgenpsychiatry.2012.14. URL https://doi.org/10.1001/archgenpsychiatry.2012.14.

Robert D. Gibbons, Diane S. Lauderdale, Robert S. Wilson, David A. Bennett, Tesnim Arar, and David A. Gallo. Adaptive measurement of cognitive function based on multidimensional item response theory. *Alzheimer's & Dementia: Translational Research & Clinical Interventions*, 10 (4):e70018, 2024. doi: 10.1002/trc2.70018. URL https://doi.org/10.1002/trc2.70018.

Virginia S. Y. Kwan, Sean P. Wojcik, Talya Miron-Shatz, Ashley M. Votruba, and Christopher Y. Olivola. Effects of Symptom Presentation Order on Perceived Disease Risk. *Psychological Science*, 23(4):381–385, 2012. doi: 10.1177/0956797611432177. URL https://doi.org/10.1177/0956797611432177.

Rachid Laajaj, Karen Macours, Daniel Alejandro Pinzon Hernandez, Omar Arias, Samuel D. Gosling, Jeff Potter, Marta Rubio-Codina, and Renos Vakis. Challenges to capture the big five personality traits in non-WEIRD populations. *Science Advances*, 5(7):eaaw5226, 2019. doi: 10.1126/sciadv.aaw5226. URL https://doi.org/10.1126/sciadv.aaw5226.

Dirk Lubbe and Christof Schuster. A Scaled Threshold Model for Measuring Extreme Response Style. *Journal of Educational and Behavioral Statistics*, 45(1):86–107, 2020. doi: 10.3102/1076998619859541. URL https://doi.org/10.3102/1076998619859541.

Lisa M. Miles, Angela M. Rodrigues, Falko F. Sniehotta, and David P. French. Asking questions changes health-related behavior: an updated systematic review and meta-analysis. *Journal of Clinical Epidemiology*, 123:59–68, 2020. doi: 10.1016/j.jclinepi.2020.03.014. URL https://doi.org/10.1016/j.jclinepi.2020.03.014.

OECD. OECD Survey on Social and Emotional Skills: Technical Report. Technical report, OECD, 2021. URL https://www.oecd.org/content/dam/oecd/en/about/programmes/edu/survey-on-social-and-emotional-skills/Technical%20Report%20SSES.pdf. SSES 2019; official database at https://www.oecd.org/en/data/datasets/SSES-Round-1-Database.html; accessed September 25, 2026.

Open-Source Psychometrics Project. Big Five Personality Test Responses (IPIP-FFM Data), 2018. URL https://openpsychometrics.org/_rawdata/. IPIP-FFM-data-8Nov2018 release; accessed September 25, 2026.

Open-Source Psychometrics Project. DASS Responses: Raw Data from Online Personality Tests, 2019. URL https://openpsychometrics.org/_rawdata/. DASS_data_21.02.19.zip release; accessed September 25, 2026.

Martin T. Orne. On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications. *American Psychologist*, 17(11):776–783, 1962. doi: 10.1037/h0043424. URL https://doi.org/10.1037/h0043424.

Yannick Rudolph, Kai Neubauer, and Ulf Brefeld. Self-improvement for Computerized Adaptive Testing. In *Machine Learning and Knowledge Discovery in Databases. Research Track (ECML PKDD 2025)*, volume 16014 of *Lecture Notes in Computer Science*, pp. 70–86. Springer, 2026. doi: 10.1007/978-3-032-05981-9_5. URL https://doi.org/10.1007/978-3-032-05981-9_5.

Fumiko Samejima. Estimation of Latent Ability Using a Response Pattern of Graded Scores. *Psychometrika*, 34(S1):1–97, 1969. doi: 10.1007/bf03372160. URL https://doi.org/10.1007/bf03372160.

Tabea Schoeler, Jean-Baptiste Pingault, and Zoltán Kutalik. The impact of self-report inaccuracy in the UK Biobank and its interplay with selective participation. *Nature Human Behaviour*, 9(3): 584–594, 2025. doi: 10.1038/s41562-024-02061-w. URL https://doi.org/10.1038/s41562-024-02061-w.

Norbert Schwarz. Self-reports: How the questions shape the answers. *American Psychologist*, 54 (2):93–105, 1999. doi: 10.1037/0003-066x.54.2.93. URL https://doi.org/10.1037/0003-066x.54.2.93.

Norbert Schwarz, Hans-J. Hippler, Brigitte Deutsch, and Fritz Strack. Response Scales: Effects of Category Range on Reported Behavior and Comparative Judgments. *Public Opinion Quarterly*, 49(3):388–395, 1985. doi: 10.1086/268936. URL https://doi.org/10.1086/268936.

Norbert Schwarz, Fritz Strack, and Hans-Peter Mai. Assimilation and Contrast Effects in Part-Whole Question Sequences: A Conversational Logic Analysis. *Public Opinion Quarterly*, 55(1): 3–23, 1991. doi: 10.1086/269239. URL https://doi.org/10.1086/269239.

Leonard J. Simms and Lee Anna Clark. Validation of a Computerized Adaptive Version of the Schedule for Nonadaptive and Adaptive Personality (SNAP). *Psychological Assessment*, 17 (1):28–43, 2005. doi: 10.1037/1040-3590.17.1.28. URL https://doi.org/10.1037/1040-3590.17.1.28.

Zhao Su, Rongxun Liu, Yange Wei, Ran Zhang, Xiao Xu, Yang Wang, Yue Zhu, Lifei Wang, Lijuan Liang, Fei Wang, and Xizhe Zhang. Temporal dynamics in psychological assessments: a novel dataset with scales and response times. *Scientific Data*, 11(1):1046, 2024. doi: 10.1038/s41597-024-03888-8. URL https://doi.org/10.1038/s41597-024-03888-8.

Roger Tourangeau and Kenneth A. Rasinski. Cognitive processes underlying context effects in attitude measurement. *Psychological Bulletin*, 103(3):299–314, 1988. doi: 10.1037/0033-2909.103.3.299. URL https://doi.org/10.1037/0033-2909.103.3.299.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Changqian Wang, Shangshang Yang, Siyu Song, Ziwen Wang, Haiping Ma, Xingyi Zhang, and Bo Jin. Explicit and Implicit Examinee-Question Relation Exploiting for Efficient Computerized Adaptive Testing. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(12):12685–12693, 2025. doi: 10.1609/aaai.v39i12.33383. URL https://doi.org/10.1609/aaai.v39i12.33383.

Michael Winter, Carsten Vogel, Johannes Schobel, Miriam Schlüter, Harald Baumeister, Yannik Terhorst, Winfried Schlee, Berthold Langguth, Peter Heuschmann, Caroline Cohrdes, and Rüdiger Pryss. Longitudinal mental health data collected via the Corona Health smartphone app during COVID-19. *Scientific Data*, 13(1):392, 2026. doi: 10.1038/s41597-026-07015-7. URL https://doi.org/10.1038/s41597-026-07015-7.

Junhao Yu, Yan Zhuang, Zhenya Huang, Qi Liu, Xin Li, Rui Li, and Enhong Chen. A Unified Adaptive Testing System Enabled by Hierarchical Structure Search. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 57803–57817. PMLR, 2024. URL https://proceedings.mlr.press/v235/yu24r.html.

Yan Zhuang, Qi Liu, Zhenya Huang, Zhi Li, Shuanghong Shen, and Haiping Ma. Fully Adaptive Framework: Neural Computerized Adaptive Testing for Online Education. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(4):4734–4742, 2022. doi: 10.1609/aaai.v36i4.20399. URL https://doi.org/10.1609/aaai.v36i4.20399.

A PROBABILITY-CLOUD REPRESENTATION AND DECISION THEORY

A.1 NOTATION AND MODULE INTERFACES

The number of bootstrap resamples is B; the ensemble has $M = B + 1$ members, including the model fitted to the original training records. The joint latent state is $z = (\theta, s) \in \mathbb{R}^{D+1}$. As in the main text, the current history H is fixed and time indices are suppressed. Sums over b run from 0 to B. The term *probability cloud* denotes a distribution over this joint state. The following interfaces distinguish the model's inputs and outputs.

Module	Input	Output
Style GRM calibration	Whole response records; domain labels	$\widehat{\Phi}_b$, including thresholds, discrimination, style sensitivity, and correlation matrix
Member inference	$\widehat{\Phi}_b$ and observed item–response pairs	Base cloud q_b over (θ, s)
Transformer	Observed item, domain, and response embeddings	Shared tilt $\xi = \xi_\psi(H)$ and category residual $r_j = r_{\psi,j}(H)$
Cloud correction	q_b and shared tilt	Corrected cloud $\widetilde{q}_b$, structural predictive P_b, and trait moments (μ_b, v_b)
Categorical correction	P_b and category residual	$\widetilde{P}_b$
Minimax aggregation	Corrected trait moments of all M members	Decision weights w, trait report $\sum_b w_b \mu_b$, and predictive mixture $\sum_b w_b \widetilde{P}_b$
Item policy	Structural one-step risks for actual and prototype histories; available-item mask	A distribution over the next item

Trait and style inference. The response model represents traits and threshold-spacing style jointly through a fitted correlation matrix. The Transformer tilt includes trait–style cross-products, preserving their dependence. For any report $\hat{\theta}$, the squared-error risk satisfies

$$\begin{aligned} \widetilde{q}_b^{\theta}(\theta) &= \int \widetilde{q}_b(\theta, s)\, ds, \\ \mathbb{E}_{\widetilde{q}_b}\left[\|\theta - \hat{\theta}\|_2^2\right] &= \mathbb{E}_{\widetilde{q}_b^{\theta}}\left[\|\theta - \hat{\theta}\|_2^2\right]. \end{aligned} \tag{19}$$

The implementation evaluates these expectations by summing joint-particle weights against trait coordinates. This integrates style uncertainty into the trait report. The same weights also provide a style marginal.

A.2 STYLE-ATTRIBUTION RISK

The distribution $q(\theta, s)$ conditions on a fixed set of observed answers and a specified measurement model. Equations (4) and (5) define the minimum posterior squared-error risk $R(q)$ and the style-conditional risk $R(q \mid s)$, respectively. The latter depends on the observed style value; its average under the style marginal of q enters Definition 2. Each scalar-variance sum equals the trace of the corresponding covariance matrix.

Proof of Proposition 1. **Step 1: express the risks as covariance traces.** Under squared loss, the optimal reports are $\hat{\theta}_q = \mathbb{E}_q[\theta]$ without style and $\hat{\theta}_q(s) = \mathbb{E}_q[\theta \mid s]$ with style. By the definition of conditional covariance,

$$\mathrm{Cov}_q(\theta \mid s) = \mathbb{E}_q\Big[(\theta - \hat{\theta}_q(s))(\theta - \hat{\theta}_q(s))^\top \mid s\Big]. \tag{20}$$

The trace of a vector's outer product is its squared Euclidean norm. Taking the trace and interchanging the finite diagonal sum with expectation therefore gives

$$\begin{aligned} R(q \mid s) &= \mathbb{E}_q\Big[\|\theta - \hat{\theta}_q(s)\|_2^2 \mid s\Big] \\ &= \sum_{d=1}^{D} \mathrm{Var}_q(\theta_d \mid s) = \mathrm{tr}\,\mathrm{Cov}_q(\theta \mid s). \end{aligned} \tag{21}$$

The same argument without conditioning gives $R(q) = \mathrm{tr}\,\mathrm{Cov}_q(\theta)$.

Step 2: substitute into the excess-risk definition. The average $\mathbb{E}_q[R(q \mid s)]$ is taken over the style marginal of q. It describes the expected error that would remain if style were revealed and the estimate were adjusted to that value. Substituting the two trace expressions into Eq. (6) gives

$$\begin{aligned} \Delta_s(q) &= \mathrm{tr}\,\mathrm{Cov}_q(\theta) - \mathbb{E}_q[\mathrm{tr}\,\mathrm{Cov}_q(\theta \mid s)] \\ &= \mathrm{tr}(\mathrm{Cov}_q(\theta) - \mathbb{E}_q[\mathrm{Cov}_q(\theta \mid s)])\,. \end{aligned} \tag{22}$$

The second equality uses the linearity of trace and expectation.

Step 3: decompose the covariance. Separate the total estimation error into the error within a given style and the difference between its conditional mean and the overall mean:

$$\theta - \hat{\theta}_q = (\theta - \hat{\theta}_q(s)) + (\hat{\theta}_q(s) - \hat{\theta}_q). \tag{23}$$

For a fixed s, the first term has conditional mean zero, whereas the second term is fixed. Hence their conditional cross-product satisfies

$$\begin{aligned} &\mathbb{E}_q\Big[(\theta - \hat{\theta}_q(s))(\hat{\theta}_q(s) - \hat{\theta}_q)^\top \mid s\Big] \\ &\quad = \mathbb{E}_q[\theta - \hat{\theta}_q(s) \mid s]\,(\hat{\theta}_q(s) - \hat{\theta}_q)^\top = 0. \end{aligned} \tag{24}$$

Averaging over s makes this cross-product zero unconditionally; its transpose is also zero. Expanding the outer product of the total error and taking expectations therefore leaves

$$\begin{aligned} \mathrm{Cov}_q(\theta) &= \mathbb{E}_q\Big[(\theta - \hat{\theta}_q(s))(\theta - \hat{\theta}_q(s))^\top\Big] \\ &\quad + \mathbb{E}_q\Big[(\hat{\theta}_q(s) - \hat{\theta}_q)(\hat{\theta}_q(s) - \hat{\theta}_q)^\top\Big] \\ &= \mathbb{E}_q[\mathrm{Cov}_q(\theta \mid s)] + \mathrm{Cov}_q(\hat{\theta}_q(s)). \end{aligned} \tag{25}$$

The first term averages the conditional covariance over style. The second is the covariance of the conditional mean because $\mathbb{E}_q[\hat{\theta}_q(s)] = \hat{\theta}_q$. This derives the law of total covariance directly from the estimation errors.

Step 4: subtract and take the trace. Substituting the decomposition from Step 3 into Step 2 yields

$$\begin{aligned} \Delta_s(q) &= \mathrm{tr}\,\mathrm{Cov}_q(\hat{\theta}_q(s)) \\ &= \mathrm{tr}\,\mathrm{Cov}_q(\mathbb{E}_q[\theta \mid s]) \geq 0. \end{aligned} \tag{26}$$

A covariance matrix is positive semidefinite, so its trace is nonnegative. It is zero exactly when every coordinate of the conditional mean has zero variance, or equivalently when $\hat{\theta}_q(s) = \hat{\theta}_q$ almost surely under q. □

The excess risk compares trait estimation before and after observing style within a fixed model. It vanishes when the conditional trait mean is constant, including posterior independence of θ and s; conditional variances may still depend on style. The identity applies to any joint posterior with finite second moments.

A.3 Gaussian components, conditioning, and reporting risk

The moment-matched component $g_b = \mathcal{N}(m_b, V_b)$ is a continuous representation of member q_b. The superscripts θ and s denote coordinate blocks: $V_b^{\theta\theta} \in \mathbb{R}^{D\times D}$, $V_b^{\theta s} \in \mathbb{R}^{D\times 1}$, $V_b^{s\theta} = (V_b^{\theta s})^\top$,

and $V_b^{ss} \in \mathbb{R}$. When $V_b^{ss} > 0$, Gaussian conditioning gives the mean in Eq. (9) and

$$\mathrm{Cov}_{g_b}(\theta \mid s) = V_b^{\theta\theta} - \frac{V_b^{\theta s} V_b^{s\theta}}{V_b^{ss}}. \tag{27}$$

The difference between marginal and conditional covariance has trace $\|V_b^{\theta s}\|_2^2 / V_b^{ss}$, proving the second line of Eq. (9). If the Gaussian style coordinate is deterministic, revealing it adds no information and its excess risk is zero.

For mixture components with positive style variance, let $\gamma_b(s)$ be the conditional probability of component b given style:

$$\begin{aligned} \gamma_b(s) &= \frac{w_b \mathcal{N}(s \mid m_b^s, V_b^{ss})}{\sum_{c=0}^{B} w_c \mathcal{N}(s \mid m_c^s, V_c^{ss})}, \\ \mathbb{E}_{G_w}[\theta \mid s] &= \sum_b \gamma_b(s) \left[m_b^\theta + \frac{V_b^{\theta s}}{V_b^{ss}} (s - m_b^s) \right]. \end{aligned} \tag{28}$$

For components with positive style variance, each conditional mean is affine, while the conditional weights change with style. The mixture can therefore express a nonlinear conditional trait mean. Components index fitted measurement models. Their weights specify each calibration's contribution; the reporting procedure optimizes these weights under the minimax loss.

Marginalizing a Gaussian selects its trait mean and covariance blocks. Moreover, moment matching preserves every member's squared-error reporting risk:

$$\begin{aligned} \mathbb{E}_{q_b}[\|\theta - \hat{\theta}\|_2^2] &= \mathrm{tr}\, V_b^{\theta\theta} + \|m_b^\theta - \hat{\theta}\|_2^2 \\ &= \mathbb{E}_{g_b}[\|\theta - \hat{\theta}\|_2^2]. \end{aligned} \tag{29}$$

Averaging this identity over members and minimizing at $\hat{\theta} = \bar{m}_w^\theta$ gives Eq. (10). This is the computational role of the Gaussian moment representation: marginalization and quadratic reporting risk use stored means and covariances. Updating the ordinal-response posteriors and predicting categorical answers still require numerical integration.

Moment matching preserves reporting risk, while conditional means and Δ_s can differ between q_b and g_b. Equation (9) concerns g_b; Proposition 1 applies directly to q_b. Numerical integration approximates joint expectations. Conditional-risk identities concern the underlying distribution: conditioning on a unique particle coordinate would isolate one particle and yield artificial zero conditional variance.

The between-member term in Eq. (10) measures the spread of trait means across complete fitted parameter sets. It is distinct from $\Delta_s(q)$: ordinary item-parameter variation also contributes when all of Φ is refitted. A decomposition into the average within-component style penalty plus between-component mean variation would instead compare against an oracle that reveals both style and component identity. Each oracle therefore defines a different information gain.

A.4 Within- and between-model covariance

Proposition 3 (Within- and between-calibration uncertainty)**.** *For any member distributions with finite second moments and weights $w \in \Delta_M$, let $q_w = \sum_b w_b q_b$ and $\bar{m}_w = \sum_b w_b m_b$. Then*

$$\mathrm{Cov}_{q_w}(z) = \sum_b w_b V_b + \sum_b w_b (m_b - \bar{m}_w)(m_b - \bar{m}_w)^\top. \tag{30}$$

Proof. Let L be a component index with $\Pr(L = b) = w_b$, and draw $z \mid L = b$ from q_b. Then $\mathbb{E}[z] = \bar{m}_w$. Writing $z - \bar{m}_w = (z - m_L) + (m_L - \bar{m}_w)$, the cross terms have zero expectation because $\mathbb{E}[z - m_L \mid L] = 0$. Therefore

$$\begin{aligned} \mathbb{E}[(z - \bar{m}_w)(z - \bar{m}_w)^\top] &= \sum_b w_b \mathbb{E}_{q_b}[(z - m_b)(z - m_b)^\top] \\ &\quad + \sum_b w_b (m_b - \bar{m}_w)(m_b - \bar{m}_w)^\top. \end{aligned} \tag{31}$$

The second term is positive semidefinite. Its trace is $\sum_b w_b \|m_b - \bar{m}_w\|_2^2$, which is zero exactly when every positive-weight component has mean $\bar{m}_w$. Replacing q_b by a Gaussian with the same first two moments leaves the identity unchanged. This applies the law of total covariance to calibration uncertainty.

□

Interpretation and resampling. The disagreement term measures variation in posterior means across retained calibrations for a fixed history and weights. It supplements the weighted average of component covariances. Its scope is disagreement within the ensemble; structural misspecification shared by all members remains outside this measure.

An independent-sample bootstrap interpretation requires independent sampling units. We resample response rows; Corona's repeated visits are resampled separately within person-disjoint training, validation, and test partitions. Preserving within-person longitudinal dependence would require a person-cluster bootstrap. Reporting uses optimized decision weights, distinct from empirical bootstrap frequencies or posterior model probabilities.

A.5 Why retaining components matters for minimax reporting

For any distribution Q with finite second moments,

$$\mathbb{E}_Q\left[\|\theta - \hat{\theta}\|_2^2\right] = \operatorname{tr} \operatorname{Cov}_Q(\theta) + \|\mathbb{E}_Q[\theta] - \hat{\theta}\|_2^2. \tag{32}$$

Thus, a moment-matched Gaussian has exactly the same expected squared loss for every report $\hat{\theta}$. This holds both for an individual member and for a mixture with fixed weights. A single Gaussian matched to the complete mixture therefore preserves its averaged reporting risk.

Minimax reporting additionally requires member-specific risks, which pooled moments do not identify. Consider scalar traits and two ensembles whose equally weighted mixtures both have mean 0 and variance 2:

$$\begin{aligned} \mathcal{P}_A &= \{\mathcal{N}(-1, \tfrac{1}{2}), \mathcal{N}(1, \tfrac{3}{2})\}, \\ \mathcal{P}_B &= \{\mathcal{N}(0, 2), \mathcal{N}(0, 2)\}. \end{aligned} \tag{33}$$

For $\mathcal{P}_A$, the risks $\frac{1}{2} + (\hat{\theta}+1)^2$ and $\frac{3}{2} + (\hat{\theta}-1)^2$ intersect at $\hat{\theta} = \frac{1}{4}$, which minimizes their maximum with value $33/16$. For $\mathcal{P}_B$, the optimal report is $\hat{\theta} = 0$ with value 2. Both ensembles have the same pooled Gaussian $\mathcal{N}(0, 2)$, but their minimax reports and risks differ. Hence, pooled first and second moments do not determine a minimax decision across members. The result motivates retaining member-specific moments for minimax reporting.

A.6 Proof of Proposition 2

Let $R_b(\hat{\theta}) = v_b + \|\mu_b - \hat{\theta}\|_2^2$. The primal problem is

$$\min_{\hat{\theta}, r} \; r \qquad \text{s.t.} \quad R_b(\hat{\theta}) \le r, \qquad b = 0, \dots, B. \tag{34}$$

It is convex and strictly feasible: for any finite $\hat{\theta}$, choose $r > \max_b R_b(\hat{\theta})$. Its Lagrangian is

$$\mathcal{L}(\hat{\theta}, r, w) = r\left(1 - \sum_b w_b\right) + \sum_b w_b R_b(\hat{\theta}), \qquad w_b \ge 0. \tag{35}$$

The infimum over r is finite only when $w \in \Delta_M$. For such weights, minimizing over $\hat{\theta}$ gives $\hat{\theta} = \bar{\mu}_w$. Substitution yields

$$\begin{aligned} F(w) &= \sum_b w_b\big(v_b + \|\mu_b\|_2^2\big) - \|\bar{\mu}_w\|_2^2 \\ &= \sum_b w_b \left[v_b + \|\mu_b - \bar{\mu}_w\|_2^2\right]. \end{aligned} \tag{36}$$

Strong duality gives the asserted equality. Compactness of the simplex gives a dual optimum. The primal risk is coercive and strongly convex, so it has a unique minimizer. For any dual optimum, complementary slackness and minimization of the Lagrangian yield $\hat{\theta} = \bar{\mu}_{w^*}$.

The same maximum risk can be written as a supremum over the convex hull of the member distributions because expectations are linear in mixtures. This gives a distributionally robust interpretation over the convex hull of the retained members.

Implemented numerical approximation. The code initializes $w_b = 1/M$ and applies exponentiated-gradient ascent to the concave dual:

$$\begin{aligned} w_b^{(k+1)} &\propto w_b^{(k)} \exp\left\{\eta \left[v_b + \|\mu_b\|_2^2 - 2\mu_b^\top \bar{\mu}_{w^{(k)}}\right]\right\}, \\ \eta &= \frac{0.3}{\max_b v_b + \max_b \|\mu_b\|_2^2 + 10^{-3}}. \end{aligned} \tag{37}$$

There are 35 updates for reporting and 20 for hypothetical-answer reporting during selection. For any computed w, optimality can be assessed through the directly evaluable gap

$$0 \le \min_{\hat{\theta}} \max_b R_b(\hat{\theta}) - F(w) \le \max_b R_b(\bar{\mu}_w) - F(w). \tag{38}$$

The right side certifies optimization error for a given history. The benchmark implementation check compares cached and original iterations for numerical equivalence.

A.7 Quadratic tilting of a Gaussian cloud

Suppose the base distribution is $\mathcal{N}(m, V)$ with $V \succ 0$, and an unnormalized correction is $\exp\{h^\top z + z^\top A z\}$ with symmetric A. Collecting quadratic and linear terms gives

$$-\tfrac{1}{2} z^\top (V^{-1} - 2A) z + z^\top (V^{-1} m + h) + \text{const.} \tag{39}$$

If $\widetilde{V}^{-1} = V^{-1} - 2A \succ 0$, completing the square yields

$$\begin{aligned} \widetilde{V} &= (V^{-1} - 2A)^{-1}, \\ \widetilde{m} &= \widetilde{V}(V^{-1} m + h). \end{aligned} \tag{40}$$

Thus, linear features alter the natural mean parameter, and quadratic features alter precision, including trait–style cross terms. Positive-definite updated precision is required for normalization on continuous support.

The implemented feature vector contains z_i and $z_i z_j$ for $i \le j$, centered and scaled using each member's integration grid. For an off-diagonal term with coefficient α_{ij}, the symmetric matrix has $A_{ij} = A_{ji} = \alpha_{ij}/2$; a diagonal coefficient is $A_{ii} = \alpha_{ii}$. Standardization contributes a normalization constant and rescales coefficients. A shared network output therefore induces member-dependent (h_b, A_b) through member-specific feature scales.

The implementation evaluates the GRM posterior and normalizes its correction on a finite integration grid. Equation (40) provides the Gaussian special case. Tilting precedes moment calculation; reversing this order generally changes the result for a non-Gaussian posterior. Category residuals in Eq. (12) separately modify response probabilities.

A.8 Calibration, integration, and correction training

Each member starts from constructor defaults. Calibration uses 2,048 fixed quasi-Monte Carlo points and 65 approximate generalized-EM rounds. The E step computes normalized latent-grid weights for training respondents. The M step takes four Adam updates at learning rate 0.025 on the expected complete-data log likelihood with parameter penalties. The discrete prior weights include an importance correction relative to the fixed proposal distribution.

For inference, each calibrated member uses 4,096 quasi-Monte Carlo points. Let $z_{b\ell}$ and $\omega_{b\ell}$ be its grid and prior weights. The base and corrected clouds are represented by

$$\begin{aligned} q_{b\ell} &\propto \omega_{b\ell} \prod_{(j,y)\in H} p_b(y \mid j, z_{b\ell}), \\ \widetilde{q}_{b\ell} &\propto q_{b\ell} e^{\xi^\top f_b(z_{b\ell})}. \end{aligned} \tag{41}$$

These are normalized discrete approximations to conditional latent distributions. Corrected trait means, second moments, and category probabilities are weighted sums over the grid.

The Transformer uses embedding dimension 128, two encoder layers, two cross-attention decoder blocks, eight attention heads, and dropout 0.2. Its output heads are initialized to zero for correction training, so the initial tilt and residual leave base member distributions unchanged. The pretrained item/domain/response representations initialize the backbone. For a training history H, withheld-item set $\mathcal{V}$, and a cyclically selected member b, the objective is

$$\begin{aligned} \mathcal{L}(\psi) &= \ell_{\mathcal{V}}(\widetilde{P}_b) + \tfrac{1}{4}\ell_{\mathcal{V}}(P_b) + \beta\ \mathrm{KL}(\widetilde{q}_b \| q_b), \\ \ell_{\mathcal{V}}(P) &= -\frac{1}{|\mathcal{V}|}\sum_{j\in\mathcal{V}} \log P(y_j \mid j). \end{aligned} \tag{42}$$

Here y_j is the observed withheld response; predictions are conditional on H. The implementation averages over each person's observed withheld entries, then over the batch. The base calibration parameters remain frozen. Member b is cycled by epoch and minibatch index; correction training holds minimax weights outside the gradient computation. All retained members participate in inference. Appendix B reports member coverage at the selected checkpoints. Histories initially use randomly sampled observed-item masks. After trajectory pools are constructed at epochs 9, 17, and 25, alternating minibatches use collected adaptive histories. Correction training uses AdamW, learning rate 3×10^{-4}, weight decay 10^{-4}, minibatches of 128, and at most 32 epochs with validation-based checkpoint selection. The candidates $\beta \in \{0.05, 0.5\}$ are chosen by validation negative log likelihood. Selection strength $\kappa \in \{0.05, 0.2, 1\}$ is subsequently chosen by validation endpoint ranked probability score.

Style model scope and coordinates. The fitted latent correlation matrix has unit diagonal; positive discriminations and fixed item-to-domain assignments establish a common coordinate convention across members. This coordinate convention makes member means comparable. Threshold spacings remain ordered because the multiplier $e^{\lambda_j s}$ is positive. For numerical evaluation, $\lambda_j s$ is clipped to $[-15, 15]$. For binary items, the single threshold satisfies $\tau_{j1} = \bar{\tau}_j$ and remains unchanged by spacing adjustment. The style coordinate specifically models category-spacing preferences; location-shift response styles require a separate parameterization.

A.9 Answer-set conditioning

The input comprises a response array and an observed-item mask. Tokens sum item, domain, and response embeddings. This representation encodes the observed answer set without position or timing features.

In evaluation mode, self-attention is equivariant to permutation of observed tokens. The fixed start-token output is invariant to jointly permuted observed keys and values. Cross-attention from a fixed item query is likewise invariant to a common permutation of memory keys and values. Thus, histories containing the same observed pairs yield the same tilt and category residual, regardless of administration order. The architecture models answer-set context; temporal extensions would require additional input features.

A.10 Information-regularized item selection

The local objective in Eq. (16) combines expected standardized one-step trait risk with a profile-information penalty. Trait risk integrates out response style within each calibration member. The implementation selects κ by validation performance and approximates policy optimization with a fixed number of updates.

For a candidate item j, hypothetical answer y, and member b, form the structural update $\widetilde{q}_b^{j,y}$ defined in the main text. The report $\hat{\theta}^{j,y}$ approximately minimizes the largest conditional member risk for that answer. The candidate cost subsequently averages each member's conditional risks using that member's structural answer probabilities, then takes the maximum over members. Thus, the order is a maximum *after* each member's answer expectation. The reports are computed separately for each hypothetical answer before evaluating candidate costs.

Hypothetical updates begin from the current corrected clouds, holding the Transformer correction fixed. Lookahead uses structural probabilities P_b; final prediction additionally applies categorical residuals to obtain $\widetilde{P}_b$.

Prototype histories. In the experiments, four training respondents are chosen as prototypes, giving $\mathcal{U} = \{0, \dots, 4\}$ after including the actual profile. First, complete, unique training records are summarized by their domain-average responses normalized by the category ranges. These summaries are standardized, clustered by K-means with four clusters, and the respondent nearest each cluster center supplies a prototype response vector. At a decision step, each prototype supplies answers only at the actual observed-item positions. All profiles share the actual available-item mask. They define a finite reference set for profile-dependent selection. The derivation below applies to any finite profile collection $\mathcal{U}$.

Let $\bar{C}_u$ and σ_u be the mean and population standard deviation of $C_u(j)$ over $j \in \mathcal{Q} \setminus \mathcal{S}$. For each profile, costs are standardized as

$$c_u(j) = \frac{C_u(j) - \bar{C}_u}{\max(\sigma_u, 10^{-6})}. \tag{43}$$

Let U be uniform on $\mathcal{U}$, and let $J \mid U = u$ follow π_u. With $\bar{\pi} = \frac{1}{|\mathcal{U}|} \sum_{u \in \mathcal{U}} \pi_u$,

$$I(U; J) = \frac{1}{|\mathcal{U}|} \sum_{u \in \mathcal{U}} \mathrm{KL}(\pi_u \| \bar{\pi}). \tag{44}$$

The policy objective is therefore an expected standardized cost plus average divergence from the marginal policy. Introducing an auxiliary distribution ρ over $\mathcal{Q} \setminus \mathcal{S}$ gives

$$\min_{\{\pi_u\}, \rho} \frac{1}{|\mathcal{U}|} \sum_{u \in \mathcal{U}} \left[\sum_{j \in \mathcal{Q} \setminus \mathcal{S}} \pi_u(j) c_u(j) + \kappa \, \mathrm{KL}(\pi_u \| \rho) \right]. \tag{45}$$

For fixed $\{\pi_u\}$, the minimizing ρ is $\bar{\pi}$. For fixed ρ, Lagrange multipliers for normalization give $\pi_u(j) \propto \rho(j) \exp[-c_u(j)/\kappa]$. Alternating these operations yields Eq. (17). The implementation starts ρ uniformly over available items, performs 60 iterations, and samples the actual-profile policy with Gumbel-max noise.

Policy information and empirical evaluation. The information quantity $I(U; J)$ measures dependence between the reference-profile index and next-item choice, conditional on the current candidate mask. The expert study evaluates the related empirical endpoint of system-label inference from selected sequences (Section 4.4). These are distinct mathematical and observational measures of selection informativeness.

B Experimental protocol and data provenance

Data and splits. Table 4 describes each bank and the split sizes. DASS responses were obtained from OpenPsychometrics (Open-Source Psychometrics Project, 2019); Xinxiang and Corona provide depression, anxiety and insomnia items (Su et al., 2024; Winter et al., 2026). The additional banks cover personality, pediatric life satisfaction and social–emotional skills (Open-Source Psychometrics Project, 2018; Forrest, 2021; OECD, 2021). Each bank was calibrated and trained on its own training split. Within each seed, training, validation and test IDs were disjoint. Across seeds, the same underlying population was resplit, so individuals may recur in different roles.

IPIP-50 covers extraversion, neuroticism, agreeableness, conscientiousness and openness (10 items each). SSES covers assertiveness, cooperation, creativity, curiosity, emotional control, empathy,

Table 3: Additional datasets reported in the appendix after eligibility filtering. Each item has five ordered response categories. PROMIS uses the released 56-item developmental Pediatric Life Satisfaction bank; SSES uses 120 direct self-report items. Eligible pools differ from the subsets used in each split (Table 4).

Dataset	Eligible sample	Items by domain	Options
IPIP-50	603,322 records	Five personality domains: 10 each	5
PROMIS	1,673 children	Life satisfaction: 56	5
OECD SSES 2019	37,584 students	Fifteen social–emotional skills: 8 each	5

energy, optimism, persistence, responsibility, self-control, sociability, stress resistance, tolerance and trust (eight items each). All three additional banks use five response options per item. DASS uses four options throughout; Xinxiang and Corona use four options for depression and anxiety, and five for insomnia. Scored implementation categories start at zero; the manuscript uses the order-equivalent labels starting at one. Table 1 reports eligible main-bank pools. DASS partitions are disjoint by source record ID; stable person identities are unavailable for person-level deduplication.

Table 4: Bank sizes and split-specific evaluation protocol. D excludes the additional style coordinate. Training counts are response records. Corona training records vary across seeds and originate from 1,886 distinct people; its validation and test sets contain one retained record per person.

Bank	Items	D	Train	Validation	Test	Endpoint
DASS-42	42	3	6,000	500	500	20
Xinxiang-23	23	3	6,000	3,611	3,611	15
Corona-23	23	3	9,874–10,229	404	405	15
IPIP-50	50	5	6,000	500	500	25
PROMIS-56	56	1	1,171	251	251	28
SSES-120	120	15	6,000	500	500	60

The main-bank eligibility rules and query masks reproduce the preceding benchmark construction. DASS permanently hides three items per domain (nine total); Xinxiang and Corona hide one per domain (three total). The external banks hide approximately 20% of items per domain, fixed per respondent. Query items are excluded from the administration pool $\mathcal{Q}$ for each respondent, and are withheld from observed histories. Maximum administered budgets are bank-specific, and all benchmark replay episodes and correction/policy training respect those caps. The reported budget counts items administered across the entire bank.

For IPIP, eligibility required all 50 responses in categories 1–5 and the source's single-response-per-IP indicator, retaining 603,322 of 1,015,341 records. Source-row IDs define IPIP partitions; longitudinal person identities are unavailable. PROMIS used complete child-report records after duplicate-child-ID exclusion, retaining 1,673 of 1,993 records. The evaluated bank is the released 56-item developmental version. Negative items were already reverse-coded in that release. SSES retained 37,584 complete student records from 61,010 source records, with 120 valid direct self-report items and unique nonmissing student IDs. The official codebook supplied reverse-scoring flags. Complete-case selection retained eight of ten sites. Evaluation uses unweighted, within-bank response prediction with item responses as inputs.

Calibration and fitting. PsyCAT calibrated one original-data member and B resampled-data members from constructor defaults, using 65 generalized EM rounds and a fixed 2,048-point quasi-Monte Carlo (QMC) calibration grid. Inference used 4,096 fixed QMC particles per member. Resampling operated on training records, including individual Corona visits. Ordinary GRM baselines were recalibrated using all training records with 180 variational-inference epochs. Main-bank Transformer initializers for seeds 2026–2028 reused historical train-split checkpoints; seeds 2029–2030 reproduced the original pretraining recipe on their own splits. Main-bank Transformer pretraining used histories of up to 19 items and a 10-item validation budget. External Transformers were pretrained within each bank's training split.

PsyCAT correction fitting used at most 32 epochs and 4,096 training records per epoch. External Transformer pretraining and learned baseline policies used at most 40 epochs and 512 training people per epoch. For Corona policy trajectories, repeated records were deduplicated to unique training IDs. Transformer-U loaded the original pretrained checkpoint without additional fitting.

Comparisons use the method-specific training budgets described above. NCAT, UATS and RECAT use documented ordinal adaptations (Zhuang et al., 2022; Yu et al., 2024; Wang et al., 2025). Random selected uniformly without replacement and used the same ordinary-GRM posterior-predictive readout as EAP-MFI.

EAP-MFI selects the item with maximum Fisher information at the current EAP trait estimate. Transformer-U predicts response probabilities and selects the item with largest categorical entropy normalized by $\log K_j$. NCAT learns attention-based action values for item selection. UATS learns item-feature targeting with differentiable latent adaptation and evaluates GRM probabilities at the adapted trait estimate. RECAT uses learned retrieval proposals to construct a candidate set and then selects by Fisher information; its retrieval count specifies the number of neighbors returned by each proposal generator. NCAT and RECAT use ordinary-GRM posterior-predictive readouts. These descriptions distinguish the baseline selection policies from their response estimators.

The bootstrap grid was $B \in \{1, 2, 3, 4, 5, 10, 50, 100\}$, giving $M = B + 1$ calibration components in the GMM representation. This varies the number of fitted models, while the posterior integration grid remains fixed at 4,096 particles per member. For each B, the correction coefficient $\beta \in \{0.05, 0.5\}$ and checkpoint were selected by endpoint validation negative log likelihood. The selection coefficient $\kappa \in \{0.05, 0.2, 1\}$ was selected by endpoint validation macro-item RPS on a fixed subset of up to 256 validation respondents. Then B was locked by the same split's endpoint validation RPS, with ties resolved toward smaller B. NCAT, UATS and RECAT checkpoints used endpoint validation negative log likelihood; the RECAT retrieval-count grid was $\{2, 4, 6\}$. All selection uses validation metrics; the selected configuration is reused at earlier budgets. Bootstrap sensitivity profiles independently retune β and κ at each B, measuring performance under validation-selected configurations.

Table 5: Selected hyperparameters in seed order 2026–2030. The number of inference members is $B + 1$.

Bank	Items	Endpoint	B (seeds 2026–2030)	β	κ
DASS-42	42	20	100/50/2/4/4	0.05/0.05/0.05/0.5/0.05	1/0.05/0.05/1/0.2
Xinxiang-23	23	15	1/4/4/50/2	0.5/0.05/0.05/0.5/0.05	0.2/0.2/1/1/0.05
Corona-23	23	15	3/1/1/3/100	0.5/0.05/0.5/0.05/0.05	1/1/0.05/1/1
IPIP-50	50	25	5/10/100/4/5	0.05/0.05/0.05/0.05/0.05	1/1/0.05/1/0.05
PROMIS-56	56	28	100/10/100/100/10	0.5/0.05/0.05/0.05/0.05	1/1/0.05/0.05/0.05
SSES-120	120	60	50/100/50/50/50	0.05/0.05/0.5/0.05/0.05	1/1/1/0.05/1

Metric definitions. The manuscript labels categories $1, \ldots, K_j$; implementation arrays use $0, \ldots, K_j - 1$, related by subtracting one. Equation (18) defines the normalized RPS. Let $\mathcal{I}_j$ contain test respondents for whom item j is a hidden query item. The item score is $r_j = |\mathcal{I}_j|^{-1} \sum_{i \in \mathcal{I}_j} \mathrm{RPS}(p_{ij}, y_{ij})$. For AUC, let $\mathcal{C}_j$ contain categories with both positive and negative observations among $\mathcal{I}_j$. The item score is $a_j = |\mathcal{C}_j|^{-1} \sum_{c \in \mathcal{C}_j} \mathrm{AUC}(\{p_{ij}(c), \mathbf{1}[y_{ij} = c]\}_{i \in \mathcal{I}_j})$. AUC compares respondents within each item and class. Items without query responses are excluded from RPS, and items without any eligible classes are excluded from AUC.

For metric $m \in \{\mathrm{RPS}, \mathrm{AUC}\}$, let $\mathcal{J}_{d,m}$ be the eligible items in domain d, and let m_d be their equally weighted mean. Whole-bank aggregation is

$$m_{\mathrm{bank}} = \frac{\sum_{d=1}^{D} |\mathcal{J}_{d,m}| \, m_d}{\sum_{d=1}^{D} |\mathcal{J}_{d,m}|}. \tag{46}$$

Domains with no eligible items are omitted. Each domain contributes according to its eligible item count. Main AUC averages eligible class-specific ROC areas within items, then averages across eligible items.

ACC and log loss are weighted by the number of query responses. Binary metrics define the event $y_j \geq \lceil (K_j+1)/2 \rceil$, using the sum of predicted category probabilities above that threshold. These are response-category events. Machine-readable per-item results retain eligible-class and query-support denominators. Displayed ACC/AUC values and their SDs are multiplied by 100. Sample SD uses denominator $5 - 1$ and describes split/training variation within the resampled source population.

Numerical diagnostics and training coverage. For eight fixed SSES validation respondents, median endpoint posterior effective sample sizes were 6.04, 6.27, 2.55, 4.48 and 4.21 out of 4,096

particles across seeds 2026–2030, pooled within each seed across respondents and selected members. These eight-person diagnostics quantify weight concentration; quadrature convergence requires larger-grid comparisons, which were not evaluated. Direct correction training used a rotating member schedule. For some large ensembles, direct training at the selected checkpoint covered a subset of members; inference included all $B + 1$ members. The source tables distinguish members seen by the selected checkpoint from members seen by the completed training run.

C COMPLETE RESPONSE-PREDICTION RESULTS

Tables 6–9 report endpoint means and sample SD across all five splits. PsyCAT achieved the lowest mean endpoint RPS in all six banks. Transformer-U led multiclass AUC on PROMIS and binary AUC on Corona. Table 2 reports the three main banks at every specified budget. On PROMIS, endpoint RPS was 0.06649 for PsyCAT and 0.06663 for Transformer-U. The original PDF and the full multi-budget mean/SD PDF are retained in the accompanying source package, together with all per-seed numerical outputs.

Table 6: Main-bank endpoint multiclass metrics and RPS (mean ± sample SD, five splits). AUC and ACC are percentages. Bold identifies the best unrounded mean.

Method	AUC ↑ (%)	ACC ↑ (%)	LogLoss ↓	RPS ↓
DASS-42 (T=20)				
PsyCAT (tuned)	$\mathbf{80.77 \pm 0.49}$	$\mathbf{57.00 \pm 0.65}$	$\mathbf{0.9653 \pm 0.0080}$	$\mathbf{0.1198 \pm 0.0022}$
EAP-MFI	78.77 ± 0.47	55.30 ± 0.43	1.0012 ± 0.0068	0.1232 ± 0.0015
Transformer-U (non-refit)	80.07 ± 0.78	55.61 ± 0.65	0.9968 ± 0.0147	0.1241 ± 0.0030
NCAT-GRM	78.85 ± 0.34	54.93 ± 0.84	1.0033 ± 0.0059	0.1235 ± 0.0012
UATS-CAT-GRM	78.67 ± 0.32	55.01 ± 0.44	1.0043 ± 0.0072	0.1234 ± 0.0014
RECAT-GRM	78.77 ± 0.47	55.30 ± 0.52	1.0013 ± 0.0070	0.1232 ± 0.0015
Random + EAP	78.86 ± 0.29	54.78 ± 0.54	1.0025 ± 0.0066	0.1233 ± 0.0014
Xinxiang-23 (T=15)				
PsyCAT (tuned)	$\mathbf{90.33 \pm 0.62}$	$\mathbf{81.32 \pm 0.81}$	$\mathbf{0.4541 \pm 0.0100}$	$\mathbf{0.0440 \pm 0.0011}$
EAP-MFI	89.31 ± 0.58	80.98 ± 0.64	0.4666 ± 0.0115	0.0449 ± 0.0012
Transformer-U (non-refit)	90.10 ± 0.53	80.91 ± 0.56	0.4655 ± 0.0128	0.0452 ± 0.0015
NCAT-GRM	89.11 ± 0.69	80.91 ± 0.59	0.4715 ± 0.0103	0.0454 ± 0.0012
UATS-CAT-GRM	89.35 ± 0.53	80.96 ± 0.69	0.4676 ± 0.0120	0.0449 ± 0.0013
RECAT-GRM	89.29 ± 0.56	80.97 ± 0.66	0.4668 ± 0.0117	0.0449 ± 0.0012
Random + EAP	89.09 ± 0.66	80.79 ± 0.55	0.4726 ± 0.0102	0.0455 ± 0.0011
Corona-23 (T=15)				
PsyCAT (tuned)	$\mathbf{85.86 \pm 0.35}$	$\mathbf{64.58 \pm 1.40}$	$\mathbf{0.8265 \pm 0.0176}$	$\mathbf{0.0885 \pm 0.0043}$
EAP-MFI	82.05 ± 0.50	59.03 ± 1.70	0.9285 ± 0.0178	0.1020 ± 0.0039
Transformer-U (non-refit)	85.56 ± 0.41	62.80 ± 1.55	0.8564 ± 0.0120	0.0923 ± 0.0033
NCAT-GRM	81.92 ± 0.69	58.50 ± 1.35	0.9335 ± 0.0121	0.1027 ± 0.0031
UATS-CAT-GRM	81.93 ± 0.48	58.72 ± 1.09	0.9356 ± 0.0184	0.1025 ± 0.0042
RECAT-GRM	82.03 ± 0.50	58.93 ± 1.69	0.9286 ± 0.0177	0.1020 ± 0.0040
Random + EAP	81.87 ± 0.24	58.45 ± 1.43	0.9371 ± 0.0129	0.1031 ± 0.0031

Table 7: Main-bank endpoint binary response metrics (mean ± sample SD, five splits). The binary event is defined by a response-category threshold.

Method	AUC ↑ (%)	ACC ↑ (%)	LogLoss ↓
DASS-42 (T=20)			
PsyCAT (tuned)	$\mathbf{86.55 \pm 0.64}$	$\mathbf{80.09 \pm 0.63}$	$\mathbf{0.4225 \pm 0.0059}$
EAP-MFI	85.54 ± 0.53	79.56 ± 0.62	0.4346 ± 0.0057
Transformer-U (non-refit)	86.08 ± 0.90	79.56 ± 0.86	0.4364 ± 0.0111
NCAT-GRM	85.67 ± 0.38	79.48 ± 0.59	0.4360 ± 0.0046
UATS-CAT-GRM	85.66 ± 0.42	79.65 ± 0.27	0.4351 ± 0.0061
RECAT-GRM	85.54 ± 0.52	79.60 ± 0.64	0.4347 ± 0.0055
Random + EAP	85.78 ± 0.40	79.59 ± 0.51	0.4340 ± 0.0059
Xinxiang-23 (T=15)			
PsyCAT (tuned)	$\mathbf{93.25 \pm 1.00}$	96.46 ± 0.08	$\mathbf{0.0989 \pm 0.0038}$
EAP-MFI	91.95 ± 0.90	96.45 ± 0.16	0.1036 ± 0.0053
Transformer-U (non-refit)	93.20 ± 0.76	96.38 ± 0.15	0.1033 ± 0.0067
NCAT-GRM	91.91 ± 0.86	96.39 ± 0.19	0.1049 ± 0.0055
UATS-CAT-GRM	92.10 ± 0.86	$\mathbf{96.50 \pm 0.13}$	0.1028 ± 0.0058
RECAT-GRM	91.96 ± 0.93	96.44 ± 0.14	0.1037 ± 0.0053
Random + EAP	91.88 ± 1.03	96.41 ± 0.09	0.1043 ± 0.0052
Corona-23 (T=15)			
PsyCAT (tuned)	91.49 ± 0.66	$\mathbf{86.17 \pm 1.39}$	$\mathbf{0.3118 \pm 0.0175}$
EAP-MFI	89.22 ± 0.80	83.51 ± 1.46	0.3553 ± 0.0193
Transformer-U (non-refit)	$\mathbf{91.52 \pm 0.84}$	85.53 ± 1.05	0.3255 ± 0.0144
NCAT-GRM	89.01 ± 0.61	83.41 ± 0.68	0.3581 ± 0.0111
UATS-CAT-GRM	89.01 ± 0.83	83.47 ± 1.51	0.3573 ± 0.0187
RECAT-GRM	89.21 ± 0.79	83.49 ± 1.51	0.3559 ± 0.0195
Random + EAP	89.10 ± 0.84	83.41 ± 1.32	0.3587 ± 0.0169

Table 8: Additional-bank endpoint multiclass metrics and RPS (mean ± sample SD, five splits). Each bank was trained separately.

Method	AUC ↑ (%)	ACC ↑ (%)	LogLoss ↓	RPS ↓
IPIP Big Five (T=25)				
PsyCAT (tuned)	$\mathbf{77.94 \pm 0.82}$	$\mathbf{49.58 \pm 0.87}$	$\mathbf{1.1792 \pm 0.0179}$	$\mathbf{0.1189 \pm 0.0023}$
EAP-MFI	71.96 ± 0.44	44.72 ± 1.04	1.2729 ± 0.0126	0.1270 ± 0.0021
Transformer-U (non-refit)	76.61 ± 0.67	47.46 ± 0.53	1.2186 ± 0.0102	0.1239 ± 0.0017
NCAT-GRM	71.88 ± 0.49	44.86 ± 0.99	1.2763 ± 0.0105	0.1279 ± 0.0019
UATS-CAT-GRM	72.20 ± 0.38	45.16 ± 1.20	1.2695 ± 0.0133	0.1263 ± 0.0020
RECAT-GRM	71.94 ± 0.42	44.63 ± 0.97	1.2737 ± 0.0122	0.1272 ± 0.0020
Random + EAP	71.77 ± 0.49	44.64 ± 1.36	1.2785 ± 0.0150	0.1282 ± 0.0026
PROMIS Life Satisfaction (T=28)				
PsyCAT (tuned)	89.46 ± 0.35	$\mathbf{70.17 \pm 0.94}$	$\mathbf{0.7272 \pm 0.0198}$	$\mathbf{0.0665 \pm 0.0023}$
EAP-MFI	85.85 ± 0.91	68.55 ± 1.17	0.8178 ± 0.0233	0.0725 ± 0.0036
Transformer-U (non-refit)	$\mathbf{89.46 \pm 0.32}$	70.03 ± 0.91	0.7388 ± 0.0176	0.0666 ± 0.0023
NCAT-GRM	85.86 ± 0.85	67.82 ± 1.33	0.8205 ± 0.0221	0.0728 ± 0.0034
UATS-CAT-GRM	85.76 ± 0.91	68.42 ± 1.33	0.8204 ± 0.0229	0.0728 ± 0.0035
RECAT-GRM	85.86 ± 0.89	68.53 ± 1.19	0.8180 ± 0.0230	0.0726 ± 0.0036
Random + EAP	85.80 ± 0.81	68.08 ± 1.24	0.8211 ± 0.0234	0.0730 ± 0.0036
OECD SSES 2019 (T=60)				
PsyCAT (tuned)	$\mathbf{79.77 \pm 0.79}$	$\mathbf{53.98 \pm 0.92}$	$\mathbf{1.0856 \pm 0.0188}$	$\mathbf{0.1069 \pm 0.0029}$
EAP-MFI	70.99 ± 0.95	47.61 ± 0.80	1.2289 ± 0.0190	0.1181 ± 0.0031
Transformer-U (non-refit)	78.90 ± 0.90	52.72 ± 0.86	1.1163 ± 0.0215	0.1104 ± 0.0031
NCAT-GRM	70.91 ± 0.86	47.39 ± 0.93	1.2296 ± 0.0172	0.1183 ± 0.0030
UATS-CAT-GRM	72.76 ± 0.88	49.79 ± 0.78	1.1963 ± 0.0196	0.1134 ± 0.0031
RECAT-GRM	71.00 ± 0.91	47.66 ± 0.81	1.2287 ± 0.0191	0.1181 ± 0.0031
Random + EAP	70.93 ± 0.91	47.24 ± 0.90	1.2293 ± 0.0182	0.1182 ± 0.0031

Table 9: Additional-bank endpoint binary response metrics (mean ± sample SD, five splits).

Method	AUC ↑ (%)	ACC ↑ (%)	LogLoss ↓
IPIP Big Five (T=25)			
PsyCAT (tuned)	$\mathbf{81.70 \pm 0.85}$	$\mathbf{82.48 \pm 0.31}$	$\mathbf{0.3912 \pm 0.0031}$
EAP-MFI	77.98 ± 0.69	81.89 ± 0.39	0.4111 ± 0.0049
Transformer-U (non-refit)	79.92 ± 0.77	82.15 ± 0.11	0.4052 ± 0.0016
NCAT-GRM	77.70 ± 0.96	81.44 ± 0.20	0.4165 ± 0.0021
UATS-CAT-GRM	78.17 ± 0.57	81.82 ± 0.37	0.4118 ± 0.0045
RECAT-GRM	77.88 ± 0.66	81.84 ± 0.36	0.4120 ± 0.0044
Random + EAP	77.56 ± 0.44	81.42 ± 0.47	0.4174 ± 0.0045
PROMIS Life Satisfaction (T=28)			
PsyCAT (tuned)	93.01 ± 0.81	93.75 ± 0.32	$\mathbf{0.1639 \pm 0.0119}$
EAP-MFI	88.31 ± 1.18	93.56 ± 0.40	0.1925 ± 0.0094
Transformer-U (non-refit)	$\mathbf{93.13 \pm 1.03}$	$\mathbf{93.79 \pm 0.51}$	0.1662 ± 0.0114
NCAT-GRM	88.58 ± 1.17	93.50 ± 0.38	0.1911 ± 0.0092
UATS-CAT-GRM	88.25 ± 1.25	93.50 ± 0.44	0.1931 ± 0.0098
RECAT-GRM	88.37 ± 1.19	93.52 ± 0.37	0.1924 ± 0.0096
Random + EAP	88.39 ± 1.03	93.55 ± 0.43	0.1928 ± 0.0098
OECD SSES 2019 (T=60)			
PsyCAT (tuned)	$\mathbf{81.22 \pm 0.43}$	$\mathbf{86.61 \pm 0.97}$	$\mathbf{0.3199 \pm 0.0155}$
EAP-MFI	73.80 ± 0.95	85.61 ± 1.08	0.3572 ± 0.0176
Transformer-U (non-refit)	80.54 ± 0.73	86.40 ± 1.00	0.3282 ± 0.0156
NCAT-GRM	73.70 ± 0.74	85.47 ± 1.10	0.3578 ± 0.0175
UATS-CAT-GRM	75.15 ± 0.71	86.25 ± 1.07	0.3477 ± 0.0175
RECAT-GRM	73.80 ± 0.95	85.61 ± 1.07	0.3573 ± 0.0176
Random + EAP	73.70 ± 0.88	85.53 ± 1.13	0.3569 ± 0.0177

C.1 REMAINING BUDGET CURVES

The following plots show the frozen benchmark outputs. Each band is one sample SD across five split-specific runs. Figure 6 groups the three binary metrics across all six banks. Figures 7 and 8 report the remaining main-bank multiclass metrics and all additional-bank multiclass metrics, respectively. Main-bank multiclass AUC and RPS appear in the main text.

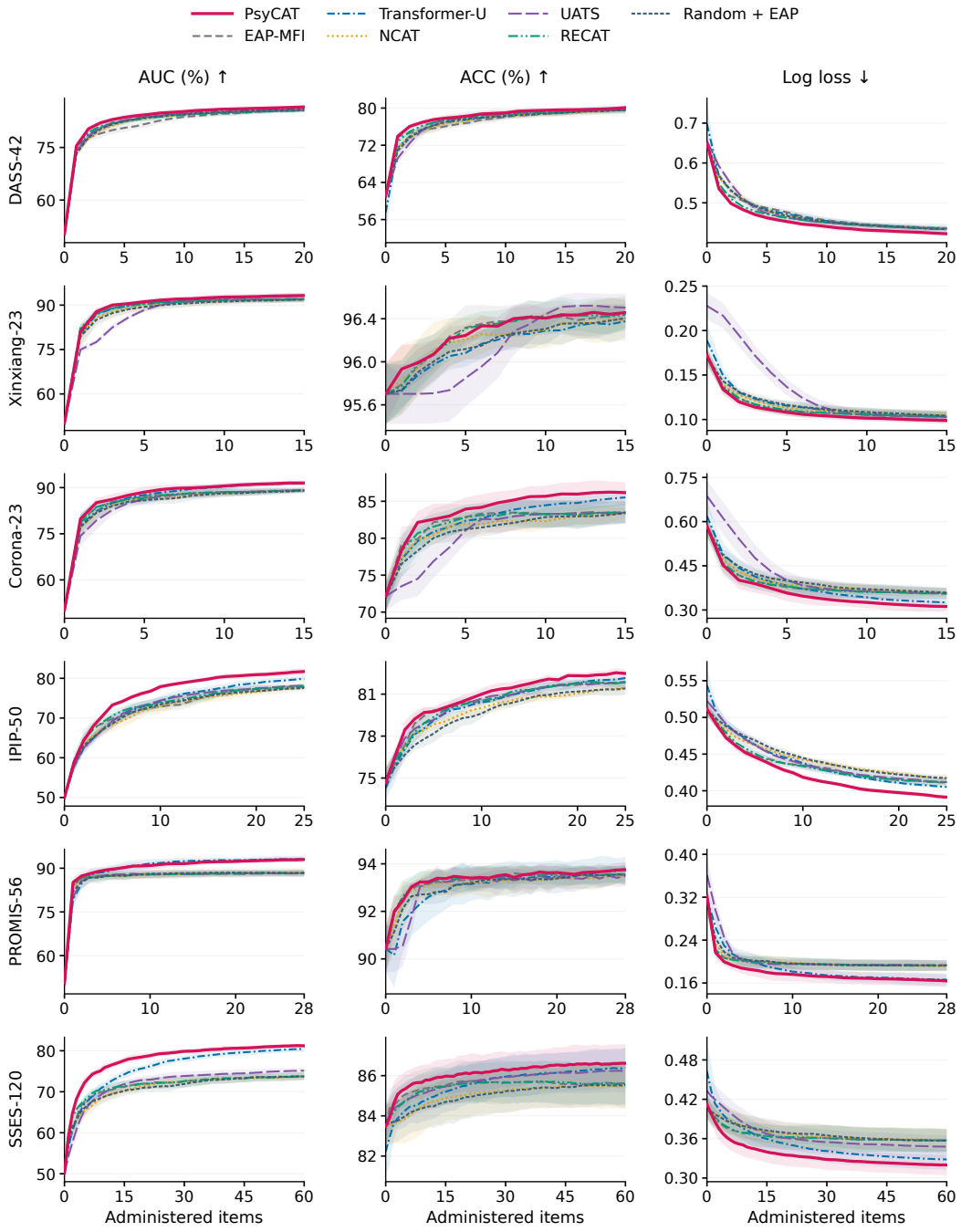

Figure 6: Binary response-prediction budget curves across all six banks. Rows identify banks; columns show AUC, ACC and log loss. Curves and shading show means and sample SD over five splits. The endpoint-selected configuration is reused at all budgets. AUC and ACC are percentages; log loss retains its original scale.

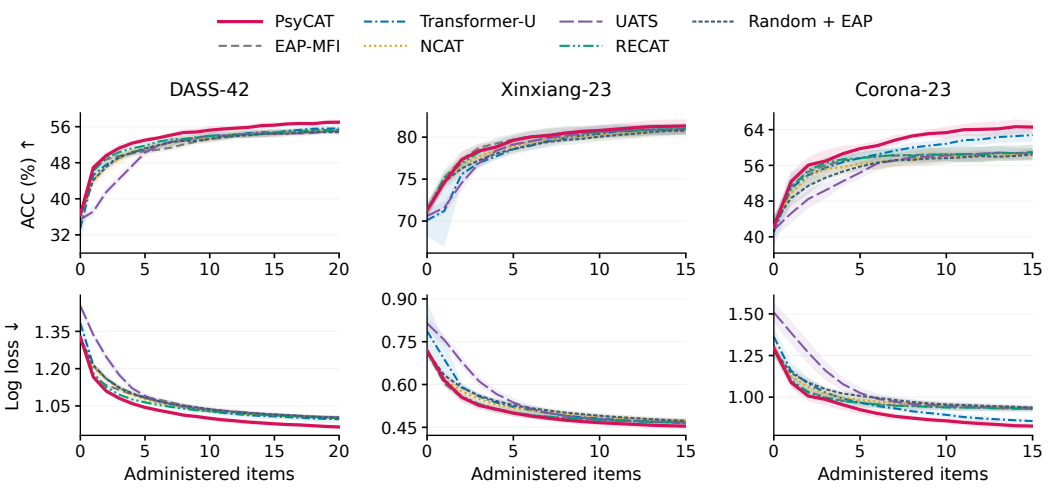

Figure 7: Remaining main-bank multiclass budget curves. Columns identify banks; rows show ACC (percent) and log loss. Curves and shading show means and sample SD over five splits. The endpoint-selected configuration is reused at all budgets. Main-bank multiclass AUC and RPS appear in Figure 3.

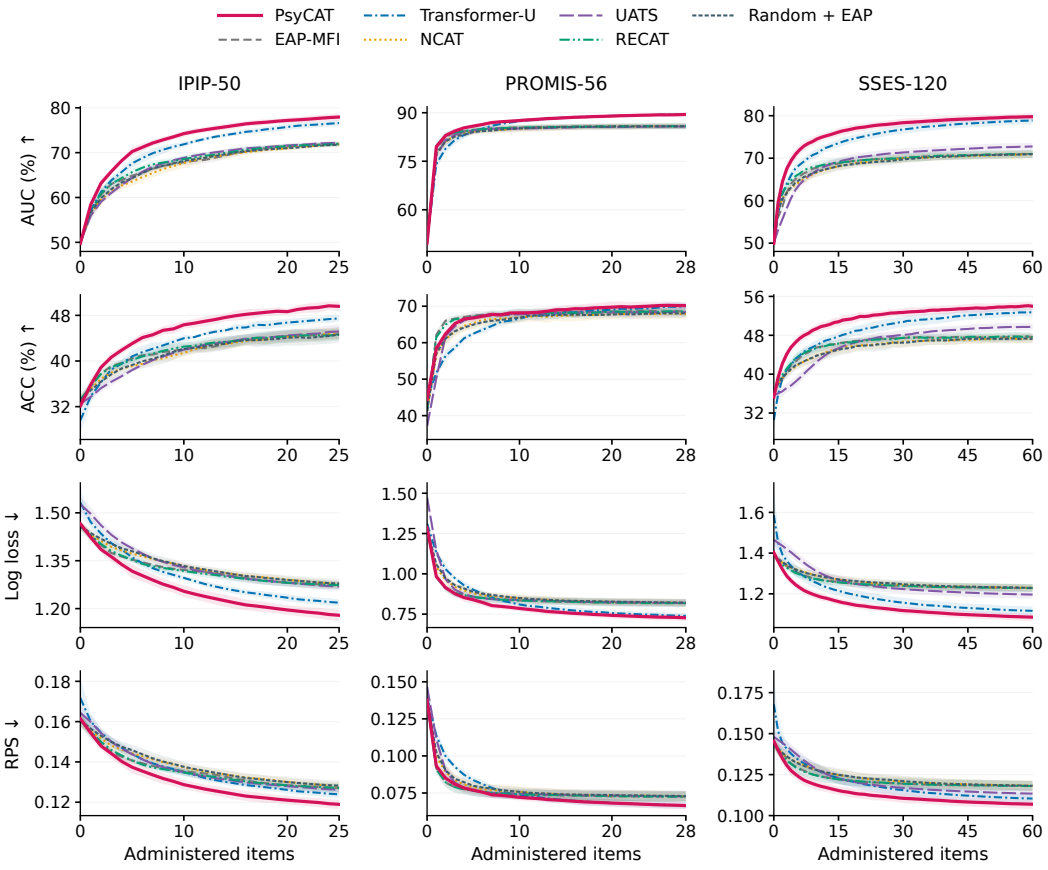

Figure 8: Additional-bank multiclass budget curves and ranked probability score. Columns identify banks; rows show AUC, ACC, log loss and RPS. Curves and shading show means and sample SD over five splits. The endpoint-selected configuration is reused at all budgets. AUC and ACC are percentages; losses retain their original scales.

C.2 BOOTSTRAP SENSITIVITY

Figures 9 and 10 group multiclass and binary metrics, respectively, for every evaluated main-bank bootstrap count, $B \in \{1, 2, 3, 4, 5, 10, 50, 100\}$. Hollow points identify validation-selected counts and their corresponding test values. Each B has separately selected correction and policy coefficients, giving a validation-profiled sensitivity analysis across ensemble sizes.

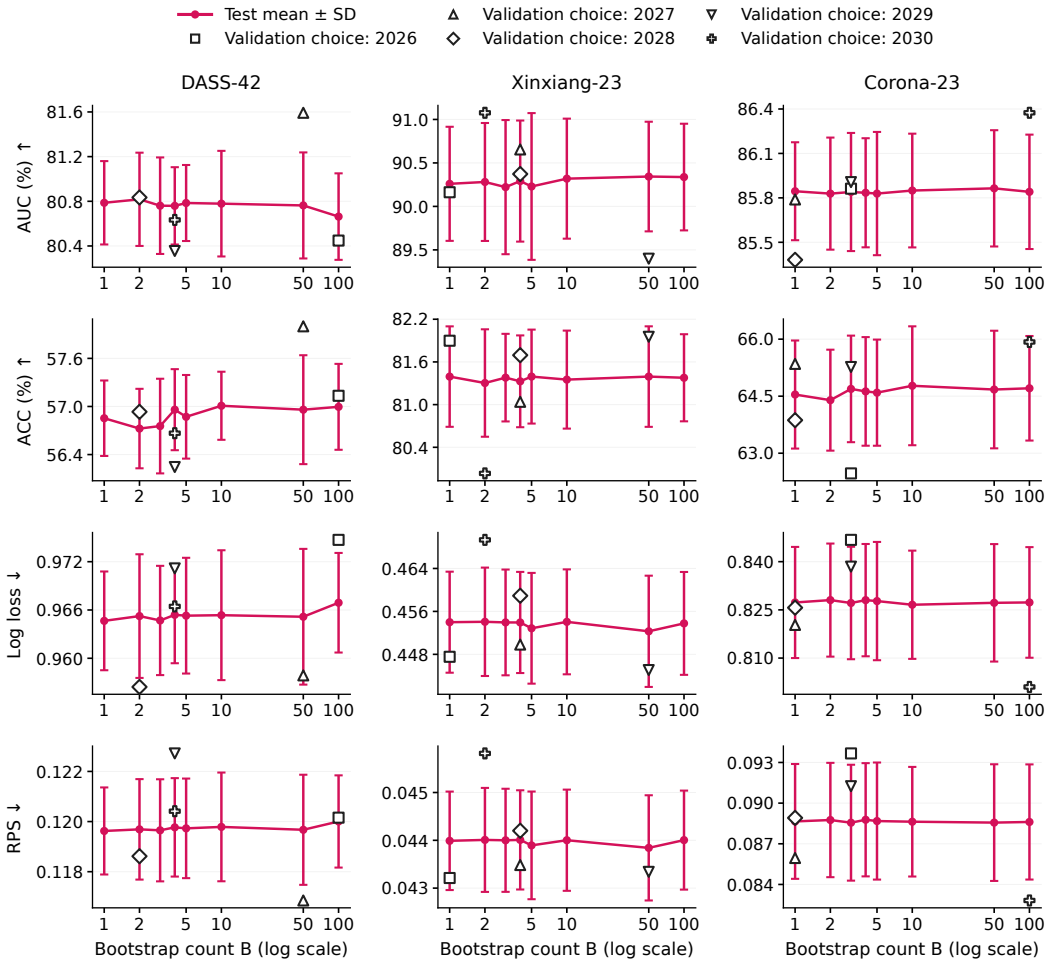

Figure 9: Validation-profiled bootstrap sensitivity for multiclass prediction. Columns identify banks; rows show AUC (percent), ACC (percent), log loss and RPS. Error bars show means and sample SD over five splits. Hollow markers show each split's validation-selected count and corresponding test value. Correction and policy coefficients are retuned at each count; the ensemble contains $B + 1$ members.

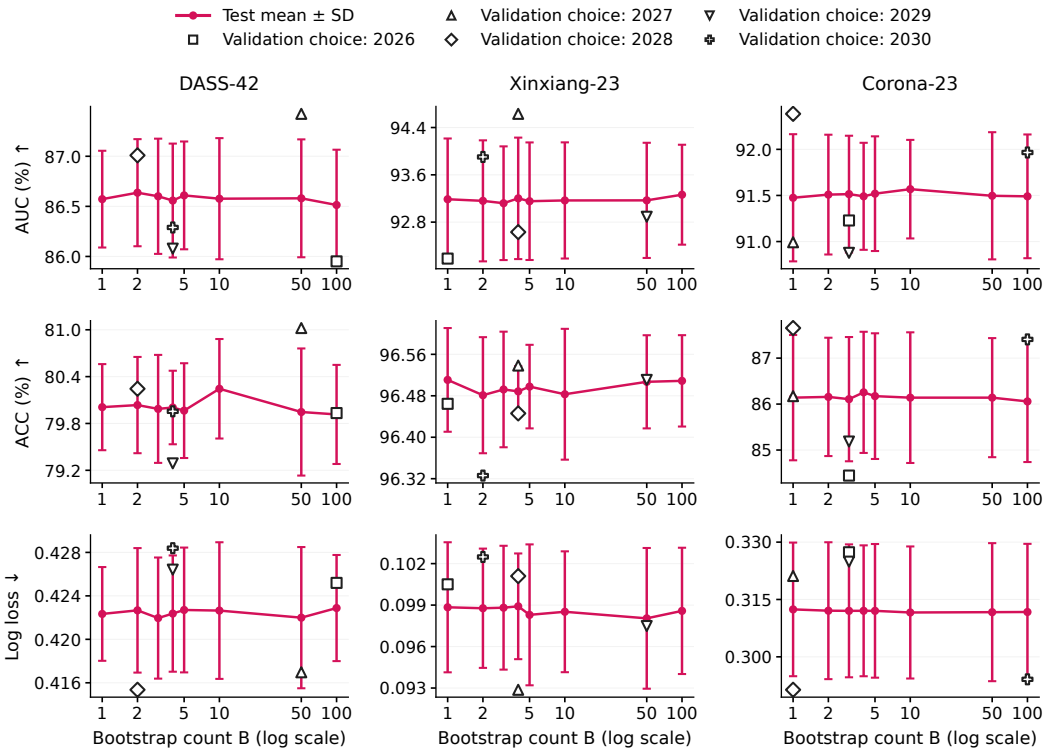

Figure 10: Validation-profiled bootstrap sensitivity for binary prediction. Columns identify banks; rows show AUC (percent), ACC (percent) and log loss. Error bars show means and sample SD over five splits. Hollow markers show each split's validation-selected count and corresponding test value. Correction and policy coefficients are retuned at each count; the ensemble contains $B + 1$ members.

D DOMAIN-LEVEL EVALUATION

Domain results use the same selected model, administered trajectory and fixed query mask as the corresponding whole-bank result. Each domain uses its bank-level selected configuration. Weighting the domain metrics by their original response/item denominators reconstructs whole-bank metrics at every budget. Table 11 compares PsyCAT with the lowest mean-RPS baseline for each of the nine domains in DASS-42, Xinxiang-23 and Corona-23. The per-domain comparator is selected descriptively from the reported test means. Tables 12–14 report all seven methods' domain-level RPS at the reported budgets. The packaged `tables/domains/` directory contains every method's seven metrics, all budgets, per-seed values and support denominators.

D.1 XINXIANG AT A FIVE-ITEM BUDGET

The five-item whole-bank RPS exception is attributable to the depression domain in the item-weighted comparison. Relative to EAP-MFI, PsyCAT's mean RPS difference was $+0.002264$ for depression, -0.000597 for anxiety and -0.000557 for insomnia. Weighting these by 9, 7 and 7 items reconstructs the whole-bank difference of $+0.000535$. The five-item depression difference had the same sign in all five seeds; by ten items, PsyCAT achieved lower mean depression RPS than EAP-MFI.

ACC showed the same domain pattern (Table 10). PsyCAT minus EAP-MFI differences were -0.7533 percentage points for depression, $+0.0997$ for anxiety and $+0.3545$ for insomnia. Each domain contributed 3,611 query responses per split, so their equally weighted mean reconstructs the whole-bank ACC difference of -0.0997 points. Depression ACC was lower in four seeds and tied in one; the whole-bank difference was positive in three seeds and negative in two. By ten items, mean depression ACC was higher by 0.3046 points, and whole-bank ACC reached 80.80% versus 80.63% for EAP-MFI. ACC evaluates the highest-probability category, whereas AUC evaluates classwise probability rankings; higher AUC can therefore coexist with lower ACC.

Table 10: Xinxiang domain context at five administered items. P/E denotes PsyCAT/EAP-MFI. Allocations, ACC and RPS are five-seed means; ACC is in percent. Lowest-category percentages use all eligible source-bank records. Item allocations sum to five before rounding.

Domain	Items	Lowest category (%)	First-five items (P/E)	ACC (%) (P/E)	RPS (P/E)
Depression	9	68.7	1.54/2.01	78.10/78.85	0.05634/0.05408
Anxiety	7	76.0	1.44/1.00	83.89/83.79	0.04174/0.04234
Insomnia	7	69.0	2.02/1.99	76.89/76.54	0.04695/0.04750

PsyCAT allocated fewer of its first five items to depression than EAP-MFI (1.54 versus 2.01), and more to anxiety (1.44 versus 1.00). Xinxiang also had 71.0% of all eligible item responses in the lowest category, compared with 28.0% for DASS and 43.7% for Corona. Among the actually administered first five Xinxiang responses, the lowest-category shares were 64.1% for PsyCAT and 60.8% for EAP-MFI. These descriptive associations relate early depression performance to item allocation and concentration in the lowest response category.

Table 1 uses questionnaire records as its sample-size unit: 32,495 for DASS, 24,073 for Xinxiang and 14,082 visits for Corona. These contain 1,364,790, 553,679 and 323,886 item responses, respectively. Category percentages pool these eligible item responses once, without multiplying records across experimental seeds; Corona's repeated visits contribute separately. Scores are zero-based, and the fifth category in Xinxiang and Corona occurs only in the seven insomnia items. Rounded percentages may not sum to exactly 100%. The accompanying aggregate tables provide category counts and denominators for each bank and domain.

D.2 MAIN-BANK DOMAIN RESULTS

Table 11: Domain-level endpoint RPS (mean ± sample SD over five splits). "Best baseline" denotes the minimum unrounded mean among the six baselines for that domain.

Bank	Domain	PsyCAT RPS	Best baseline RPS	Baseline
DASS-42	anxiety	0.1243 ± 0.0036	0.1276 ± 0.0028	Transformer-U
DASS-42	depression	0.1089 ± 0.0034	0.1096 ± 0.0036	EAP-MFI
DASS-42	stress	0.1260 ± 0.0058	0.1290 ± 0.0042	Random + EAP
Xinxiang-23	anxiety	0.0364 ± 0.0012	0.0369 ± 0.0010	UATS
Xinxiang-23	depression	0.0496 ± 0.0022	0.0508 ± 0.0029	Transformer-U
Xinxiang-23	insomnia	0.0445 ± 0.0009	0.0450 ± 0.0010	UATS
Corona-23	anxiety	0.0982 ± 0.0086	0.0997 ± 0.0069	Transformer-U
Corona-23	depression	0.0824 ± 0.0084	0.0863 ± 0.0089	Transformer-U
Corona-23	insomnia	0.0867 ± 0.0066	0.0925 ± 0.0074	Transformer-U

Table 12: DASS-42 domain-level RPS at the reported budgets. Entries are means ± sample SD over five splits. T counts all administered bank items. Bold identifies the lowest unrounded mean among all seven methods within each domain and budget. All domains use the bank-level selected configuration.

Method	$T = 5$	$T = 10$	$T = 15$	$T = 20$
Anxiety				
PsyCAT	$\mathbf{0.1372 \pm 0.0056}$	$\mathbf{0.1300 \pm 0.0039}$	$\mathbf{0.1263 \pm 0.0034}$	$\mathbf{0.1243 \pm 0.0036}$
EAP-MFI	0.1502 ± 0.0039	0.1388 ± 0.0036	0.1331 ± 0.0031	0.1303 ± 0.0030
Transformer-U	0.1382 ± 0.0053	0.1312 ± 0.0045	0.1289 ± 0.0035	0.1276 ± 0.0028
NCAT	0.1400 ± 0.0038	0.1327 ± 0.0054	0.1294 ± 0.0044	0.1279 ± 0.0046
UATS	0.1435 ± 0.0038	0.1343 ± 0.0035	0.1312 ± 0.0033	0.1292 ± 0.0035
RECAT	0.1427 ± 0.0046	0.1354 ± 0.0038	0.1322 ± 0.0033	0.1300 ± 0.0033
Random + EAP	0.1396 ± 0.0046	0.1333 ± 0.0048	0.1301 ± 0.0042	0.1279 ± 0.0040
Depression				
PsyCAT	0.1213 ± 0.0043	0.1143 ± 0.0038	0.1111 ± 0.0036	$\mathbf{0.1089 \pm 0.0034}$
EAP-MFI	$\mathbf{0.1150 \pm 0.0031}$	$\mathbf{0.1117 \pm 0.0037}$	$\mathbf{0.1103 \pm 0.0034}$	0.1096 ± 0.0036
Transformer-U	0.1332 ± 0.0038	0.1212 ± 0.0041	0.1167 ± 0.0036	0.1144 ± 0.0045
NCAT	0.1305 ± 0.0087	0.1190 ± 0.0042	0.1150 ± 0.0036	0.1135 ± 0.0036
UATS	0.1240 ± 0.0047	0.1165 ± 0.0048	0.1125 ± 0.0041	0.1111 ± 0.0047
RECAT	0.1175 ± 0.0043	0.1127 ± 0.0040	0.1106 ± 0.0035	0.1098 ± 0.0037
Random + EAP	0.1331 ± 0.0042	0.1208 ± 0.0027	0.1157 ± 0.0035	0.1130 ± 0.0033
Stress				
PsyCAT	$\mathbf{0.1378 \pm 0.0046}$	$\mathbf{0.1312 \pm 0.0056}$	$\mathbf{0.1281 \pm 0.0063}$	$\mathbf{0.1260 \pm 0.0058}$
EAP-MFI	0.1522 ± 0.0045	0.1380 ± 0.0052	0.1327 ± 0.0049	0.1296 ± 0.0047
Transformer-U	0.1439 ± 0.0052	0.1356 ± 0.0056	0.1318 ± 0.0054	0.1301 ± 0.0062
NCAT	0.1400 ± 0.0051	0.1338 ± 0.0050	0.1308 ± 0.0049	0.1292 ± 0.0050
UATS	0.1462 ± 0.0042	0.1350 ± 0.0026	0.1320 ± 0.0029	0.1299 ± 0.0033
RECAT	0.1440 ± 0.0058	0.1365 ± 0.0055	0.1323 ± 0.0056	0.1298 ± 0.0047
Random + EAP	0.1415 ± 0.0042	0.1347 ± 0.0038	0.1311 ± 0.0037	0.1290 ± 0.0042

Table 13: Xinxiang-23 domain-level RPS at the reported budgets. Entries are means ± sample SD over five splits. T counts all administered bank items. Bold identifies the lowest unrounded mean among all seven methods within each domain and budget. All domains use the bank-level selected configuration.

Method	$T = 5$	$T = 10$	$T = 15$
Anxiety			
PsyCAT	$\mathbf{0.0417 \pm 0.0020}$	0.0379 ± 0.0012	$\mathbf{0.0364 \pm 0.0012}$
EAP-MFI	0.0423 ± 0.0010	$\mathbf{0.0378 \pm 0.0009}$	0.0369 ± 0.0009
Transformer-U	0.0474 ± 0.0017	0.0400 ± 0.0011	0.0377 ± 0.0014
NCAT	0.0441 ± 0.0027	0.0394 ± 0.0015	0.0375 ± 0.0007
UATS	0.0445 ± 0.0005	0.0380 ± 0.0010	0.0369 ± 0.0010
RECAT	0.0424 ± 0.0012	0.0378 ± 0.0008	0.0370 ± 0.0008
Random + EAP	0.0436 ± 0.0010	0.0393 ± 0.0009	0.0374 ± 0.0008
Depression			
PsyCAT	0.0563 ± 0.0019	$\mathbf{0.0513 \pm 0.0021}$	$\mathbf{0.0496 \pm 0.0022}$
EAP-MFI	$\mathbf{0.0541 \pm 0.0026}$	0.0522 ± 0.0026	0.0510 ± 0.0025
Transformer-U	0.0551 ± 0.0030	0.0517 ± 0.0028	0.0508 ± 0.0029
NCAT	0.0570 ± 0.0038	0.0535 ± 0.0031	0.0516 ± 0.0028
UATS	0.0591 ± 0.0033	0.0529 ± 0.0027	0.0510 ± 0.0024
RECAT	0.0542 ± 0.0024	0.0521 ± 0.0026	0.0510 ± 0.0026
Random + EAP	0.0572 ± 0.0025	0.0534 ± 0.0023	0.0517 ± 0.0027
Insomnia			
PsyCAT	$\mathbf{0.0469 \pm 0.0008}$	$\mathbf{0.0450 \pm 0.0009}$	$\mathbf{0.0445 \pm 0.0009}$
EAP-MFI	0.0475 ± 0.0011	0.0458 ± 0.0012	0.0451 ± 0.0010
Transformer-U	0.0499 ± 0.0008	0.0468 ± 0.0012	0.0453 ± 0.0010
NCAT	0.0487 ± 0.0013	0.0468 ± 0.0013	0.0455 ± 0.0010
UATS	0.0508 ± 0.0015	0.0456 ± 0.0011	0.0450 ± 0.0010
RECAT	0.0479 ± 0.0013	0.0458 ± 0.0012	0.0451 ± 0.0010
Random + EAP	0.0519 ± 0.0012	0.0480 ± 0.0008	0.0458 ± 0.0011

Table 14: Corona-23 domain-level RPS at the reported budgets. Entries are means ± sample SD over five splits. T counts all administered bank items. Bold identifies the lowest unrounded mean among all seven methods within each domain and budget. All domains use the bank-level selected configuration.

Method	$T = 5$	$T = 10$	$T = 15$
Anxiety			
PsyCAT	$\mathbf{0.1142 \pm 0.0093}$	$\mathbf{0.1023 \pm 0.0092}$	$\mathbf{0.0982 \pm 0.0086}$
EAP-MFI	0.1234 ± 0.0078	0.1180 ± 0.0071	0.1167 ± 0.0081
Transformer-U	0.1164 ± 0.0062	0.1057 ± 0.0069	0.0997 ± 0.0069
NCAT	0.1239 ± 0.0083	0.1188 ± 0.0076	0.1167 ± 0.0072
UATS	0.1339 ± 0.0079	0.1190 ± 0.0083	0.1170 ± 0.0073
RECAT	0.1237 ± 0.0070	0.1174 ± 0.0077	0.1167 ± 0.0082
Random + EAP	0.1257 ± 0.0067	0.1198 ± 0.0060	0.1177 ± 0.0065
Depression			
PsyCAT	$\mathbf{0.0948 \pm 0.0081}$	$\mathbf{0.0859 \pm 0.0077}$	$\mathbf{0.0824 \pm 0.0084}$
EAP-MFI	0.1016 ± 0.0091	0.0984 ± 0.0069	0.0959 ± 0.0059
Transformer-U	0.0999 ± 0.0106	0.0908 ± 0.0106	0.0863 ± 0.0089
NCAT	0.1069 ± 0.0099	0.1014 ± 0.0077	0.0972 ± 0.0064
UATS	0.1106 ± 0.0085	0.0979 ± 0.0078	0.0963 ± 0.0065
RECAT	0.1033 ± 0.0086	0.0975 ± 0.0068	0.0959 ± 0.0059
Random + EAP	0.1072 ± 0.0091	0.0996 ± 0.0066	0.0966 ± 0.0058
Insomnia			
PsyCAT	0.0987 ± 0.0055	$\mathbf{0.0913 \pm 0.0069}$	$\mathbf{0.0867 \pm 0.0066}$
EAP-MFI	$\mathbf{0.0985 \pm 0.0065}$	0.0963 ± 0.0062	0.0950 ± 0.0062
Transformer-U	0.1085 ± 0.0058	0.0980 ± 0.0061	0.0925 ± 0.0074
NCAT	0.1013 ± 0.0063	0.0974 ± 0.0068	0.0957 ± 0.0067
UATS	0.1052 ± 0.0069	0.0973 ± 0.0063	0.0959 ± 0.0060
RECAT	0.0987 ± 0.0066	0.0963 ± 0.0061	0.0951 ± 0.0061
Random + EAP	0.1099 ± 0.0053	0.1005 ± 0.0062	0.0969 ± 0.0068

E CONTROLLED LATENT-CHANGE EXPERIMENTS

Each seed generated 512 paired synthetic respondents under four frozen generators: warm/cold source calibrations crossed with matched/asymmetric response generators. Matched generators use the fitted threshold-scaling response model. Asymmetric generators permit distinct threshold-gap changes derived from a second fitted member, constituting a model-based stress test. Both generator families remain constructed from fitted model banks. Shared random uniforms couple reference and intervention responses; fixed balanced histories isolate the response readout from selection-policy differences. Each family averages two source generators within each seed before summarizing across five seeds.

Interventions changed style by -0.67449, the target trait by $+0.67449$, or both. Other traits were held fixed. In the reported simulations, the target was depression for the three main banks, extraversion for IPIP, and assertiveness for SSES. The common probability-to-latent decoder was frozen before method comparisons. The first three main-bank seeds preserved the original synthetic worlds and decoders; added seeds and external banks generated independent worlds and decoders using the same dimension-generalized recipe. Latent-change targets are defined on each bank's fitted coordinate scale.

Decoder and interpretation. The shared decoder maps probability-derived features to latent coordinates. It was calibrated on independent synthetic respondents and fixed before evaluation, with one decoder per dataset and seed shared across methods. Decoded change is the intervention estimate minus the reference estimate; Figure 4 adds it to the known initial latent value. RECAT contributes its ordinary-GRM response estimator under the shared history. Both generator families derive from fitted response models; asymmetric generators stress-test the measurement form through individual threshold-gap perturbations.

Under matched generators, PsyCAT's mean decoded changes were $(-0.3658, 0.0013)$ for style-only and $(-0.5096, 0.4961)$ for joint interventions. Individual change errors are summarized by MSE in Table 15. Under asymmetric depression-only intervention, PsyCAT's mean decoded style change was -0.1364.

No-Transformer control and native estimates. The control disabled both posterior tilt and output-probability residual while retaining the calibrated ensemble and minimax readout. This is an inference-time module removal under fixed histories, retaining the original training configuration. Table 15 reports all three DASS interventions and both generator families, including changes and change MSE. Under asymmetric style-only intervention, full PsyCAT had style-change MSE 0.2031 versus 0.1715 for the no-Transformer control, and trait-change MSE 0.0173 versus 0.0136. This control comparison measures paired latent-change estimation under the specified generators.

Native PsyCAT posterior changes are separately exported. For asymmetric style-only interventions, native mean changes were $\Delta s = -0.3834$ and $\Delta\theta_{\text{dep}} = -0.0004$, compared with decoded changes -0.3604 and 0.0010. Native latent estimates are exported where the method defines them; Transformer-U latent estimates and ordinary-GRM style estimates are recorded as unavailable. Evaluation excludes replay placeholders. Independent decoder diagnostics, source hashes, frozen-weight checks and unrevealed-answer invariance checks accompany the extension artifacts. The additional-bank figures retain the same initial-truth-plus-decoded-change convention as Figure 4.

Table 15: DASS common-decoder changes and change mean squared errors (mean ± sample SD over five seeds after averaging the two source generators per family). True changes are $(-0.67449, 0)$ for style-only, $(0, +0.67449)$ for trait-only, and $(-0.67449, +0.67449)$ for joint interventions.

Condition	Method	Δs	$\Delta\theta_1$	Style MSE	Trait MSE
Matched generator family					
Style only	PsyCAT	-0.3658 ± 0.0396	0.0013 ± 0.0116	0.2006 ± 0.0091	0.0175 ± 0.0061
Style only	PsyCAT w/o Transformer	-0.3878 ± 0.0129	-0.0011 ± 0.0050	0.1687 ± 0.0074	0.0138 ± 0.0017
Style only	Transformer-U	-0.2269 ± 0.0694	-0.0155 ± 0.0252	0.2592 ± 0.0375	0.0084 ± 0.0027
Style only	RECAT-GRM	-0.0037 ± 0.0020	-0.0152 ± 0.0084	0.4501 ± 0.0027	0.0236 ± 0.0015
Trait only	PsyCAT	-0.1364 ± 0.0313	0.4905 ± 0.0321	0.1263 ± 0.0384	0.1037 ± 0.0110
Trait only	PsyCAT w/o Transformer	-0.0861 ± 0.0226	0.4424 ± 0.0270	0.0853 ± 0.0062	0.0976 ± 0.0071
Trait only	Transformer-U	-0.0914 ± 0.0212	0.3354 ± 0.0411	0.0676 ± 0.0397	0.1668 ± 0.0331
Trait only	RECAT-GRM	-0.0062 ± 0.0062	0.4507 ± 0.0201	0.0009 ± 0.0004	0.0940 ± 0.0056
Joint	PsyCAT	-0.5096 ± 0.0801	0.4961 ± 0.0433	0.2367 ± 0.0466	0.1167 ± 0.0200
Joint	PsyCAT w/o Transformer	-0.4742 ± 0.0276	0.4499 ± 0.0350	0.1882 ± 0.0108	0.1013 ± 0.0096
Joint	Transformer-U	-0.3181 ± 0.0806	0.3349 ± 0.0669	0.2493 ± 0.0445	0.1744 ± 0.0465
Joint	RECAT-GRM	-0.0098 ± 0.0071	0.4711 ± 0.0303	0.4431 ± 0.0096	0.0968 ± 0.0076
Asymmetric generator family					
Style only	PsyCAT	-0.3604 ± 0.0358	0.0010 ± 0.0113	0.2031 ± 0.0080	0.0173 ± 0.0064
Style only	PsyCAT w/o Transformer	-0.3820 ± 0.0084	-0.0011 ± 0.0050	0.1715 ± 0.0068	0.0136 ± 0.0015
Style only	Transformer-U	-0.2236 ± 0.0662	-0.0153 ± 0.0247	0.2611 ± 0.0368	0.0083 ± 0.0026
Style only	RECAT-GRM	-0.0037 ± 0.0019	-0.0152 ± 0.0082	0.4502 ± 0.0025	0.0230 ± 0.0014
Trait only	PsyCAT	-0.1364 ± 0.0313	0.4905 ± 0.0321	0.1263 ± 0.0384	0.1037 ± 0.0110
Trait only	PsyCAT w/o Transformer	-0.0861 ± 0.0226	0.4424 ± 0.0270	0.0853 ± 0.0062	0.0976 ± 0.0071
Trait only	Transformer-U	-0.0914 ± 0.0212	0.3354 ± 0.0411	0.0676 ± 0.0397	0.1668 ± 0.0331
Trait only	RECAT-GRM	-0.0062 ± 0.0062	0.4507 ± 0.0201	0.0009 ± 0.0004	0.0940 ± 0.0056
Joint	PsyCAT	-0.5044 ± 0.0757	0.4959 ± 0.0432	0.2376 ± 0.0466	0.1165 ± 0.0199
Joint	PsyCAT w/o Transformer	-0.4686 ± 0.0252	0.4495 ± 0.0340	0.1899 ± 0.0111	0.1010 ± 0.0093
Joint	Transformer-U	-0.3154 ± 0.0783	0.3347 ± 0.0660	0.2509 ± 0.0450	0.1742 ± 0.0460
Joint	RECAT-GRM	-0.0097 ± 0.0072	0.4705 ± 0.0292	0.4431 ± 0.0097	0.0965 ± 0.0073

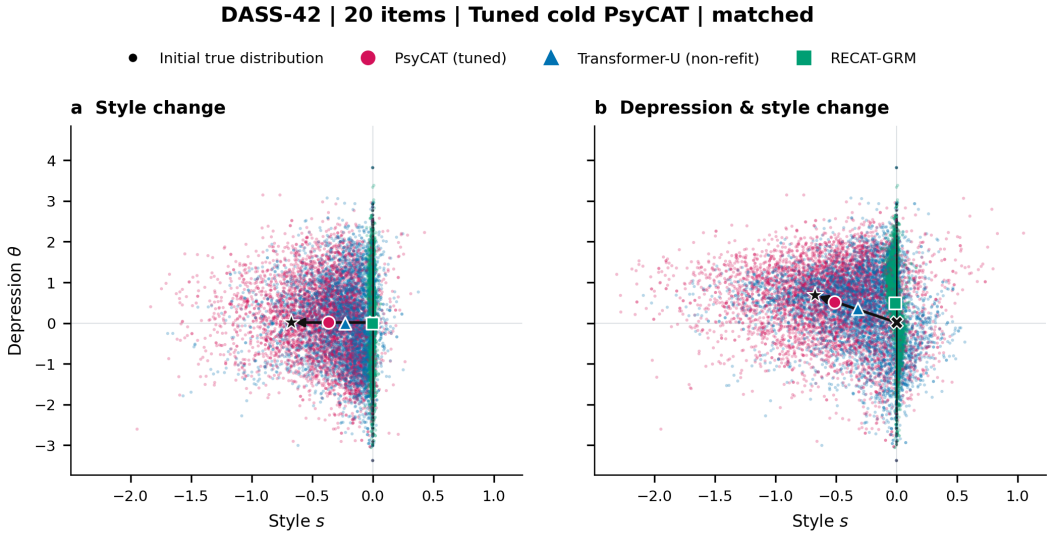

Figure 11: DASS-42 paired-change simulations under matched generators. Panels show style-only and joint depression–style interventions. Each method contributes 5,120 points per panel from 512 respondents, five seeds, and two source calibrations. Coordinates equal initial truth plus common-decoder paired changes. The corresponding asymmetric-generator comparison appears in Figure 4.

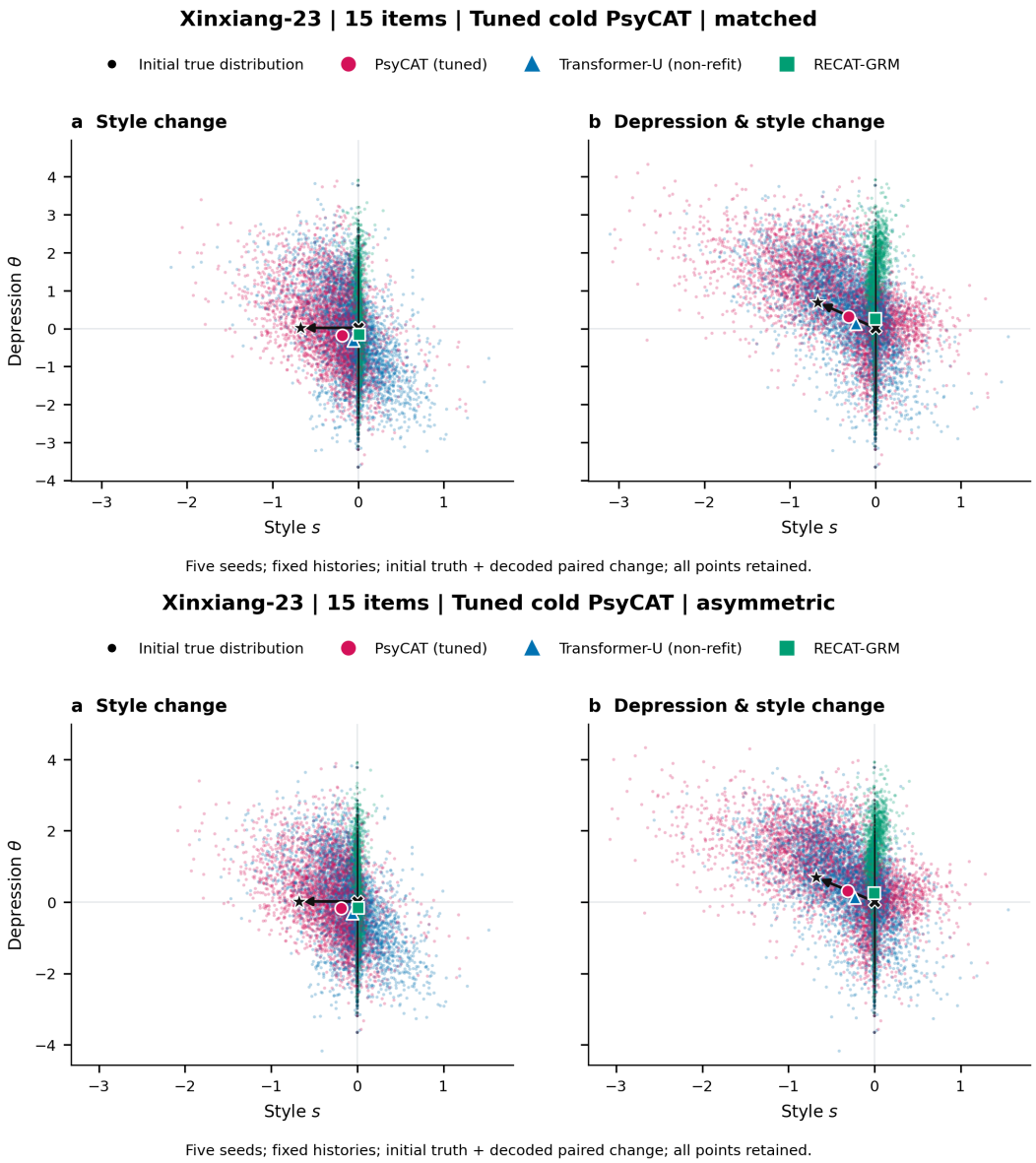

Figure 12: Xinxiang-23 paired-change simulations under matched (top) and asymmetric (bottom) generators. Each plot contrasts style-only and joint changes. Coordinates equal initial truth plus common-decoder paired changes. There are 512 synthetic respondents per seed, five seeds and two source calibrations per family. Shared fixed histories isolate the response estimators.

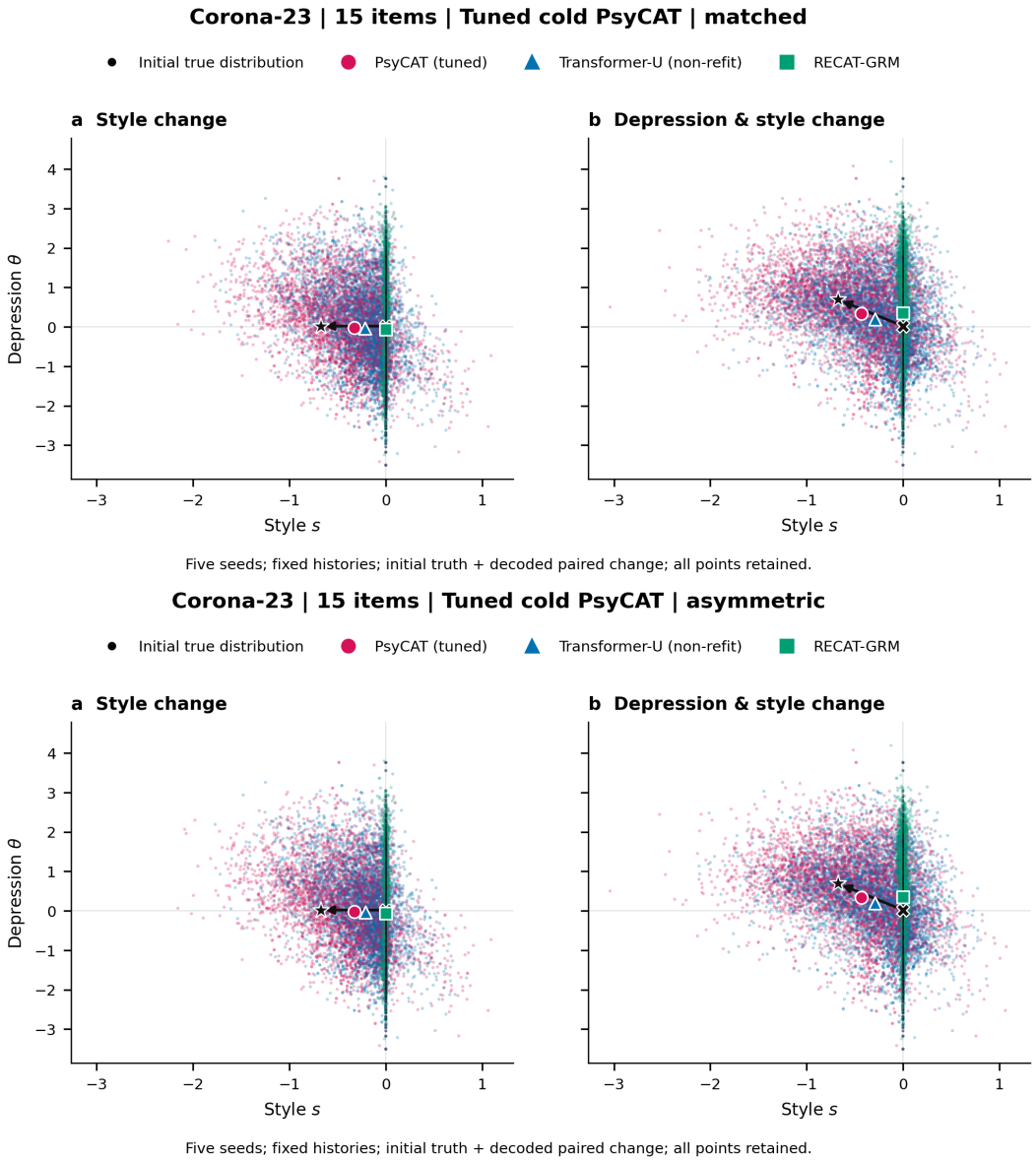

Figure 13: Corona-23 paired-change simulations under matched (top) and asymmetric (bottom) generators. Each plot contrasts style-only and joint changes. Coordinates equal initial truth plus common-decoder paired changes. There are 512 synthetic respondents per seed, five seeds and two source calibrations per family. Shared fixed histories isolate the response estimators.

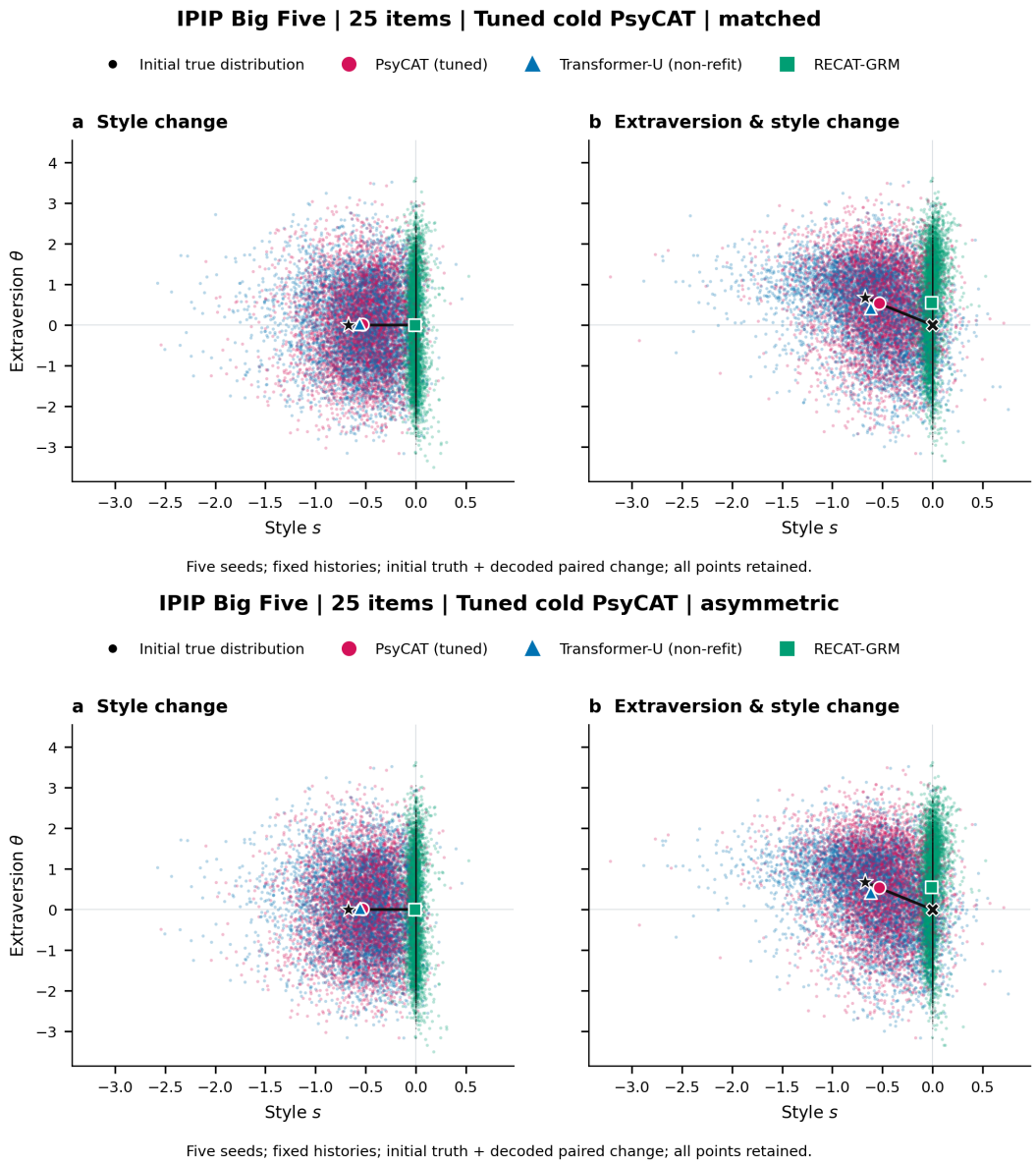

Figure 14: IPIP-50 paired-change simulations under matched (top) and asymmetric (bottom) generators. Each plot contrasts style-only and joint changes. Coordinates equal initial truth plus common-decoder paired changes. There are 512 synthetic respondents per seed, five seeds and two source calibrations per family. Shared fixed histories isolate the response estimators.

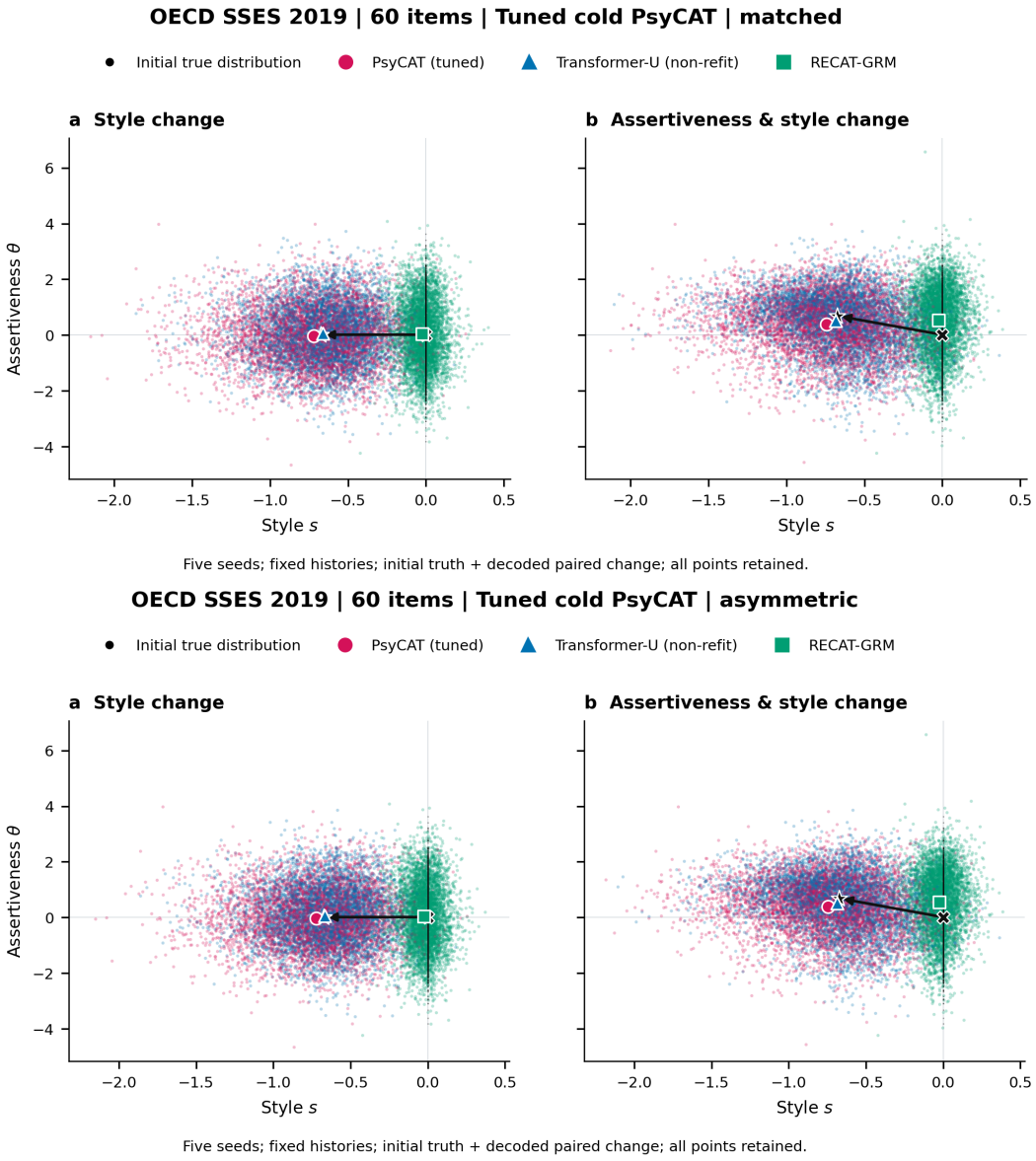

Figure 15: SSES-120 paired-change simulations under matched (top) and asymmetric (bottom) generators. Each plot contrasts style-only and joint changes. Coordinates equal initial truth plus common-decoder paired changes. There are 512 synthetic respondents per seed, five seeds and two source calibrations per family. Shared fixed histories isolate the response estimators.

F EXPERT INFERENCE STUDY: ALLOCATION AND ANALYSIS

All 300 recorded judgments were analyzed. Expert A assessed 81 PsyCAT and 69 PsyCAT-MFI sequences; expert B assessed 69 and 81, respectively. Binary correct counts were 54/81 and 59/69 for expert A, and 46/69 and 70/81 for expert B. Corresponding four-level counts were 46/81 and 54/69, and 34/69 and 66/81. PsyCAT-MFI shared PsyCAT's estimation components and differed in selection; it was distinct from the EAP-MFI benchmark baseline. Severity labels were the system-defined study scoring categories.

The null hypothesis was equal correctness probabilities between methods within each expert. Conditional on each expert's method counts and total correct judgments, the number correct under PsyCAT followed a hypergeometric distribution. We convolved the two distributions and computed twice the smaller inclusive tail probability, capped at one. Tests were two-sided and unadjusted for multiplicity, with binary accuracy primary and four-level agreement secondary.

Inference treated experts as fixed and assumed independent, exchangeable case judgments under the null within each stratum. The analysis file lacked case identifiers, preventing assessment of pairing or reuse. The estimand is the between-method inference-accuracy difference for these two experts and the presented materials. Aggregate counts and an exact-test verification script accompany the manuscript.