# DOES PREDICTIVE UNCERTAINTY GUIDE US TO THE MISSING INFORMATION? A CONTROLLED AUDIT UNDER PARTIAL EEG OBSERVATION

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Predictive uncertainty can identify difficult predictions, but can it also identify which unavailable measurements matter? We audit this stronger interpretation through controlled channel withholding in EEG: the history, future target, horizon, and retained channel count stay fixed while channel identity changes. A standard heteroscedastic predictor trained on 1,600 hours produces sample-dependent, error-associated uncertainty. Yet its measurement-level interpretation breaks down in a specific way. Teacher-state consequence rankings reproduce across disjoint subject halves with correlation approximately $0.68$, and uncertainty rankings with $0.81$–$0.96$ across three constructions, while their aggregate cross-half alignment is weak or negative. Both rankings therefore contain population structure, but that structure does not reliably align; many individual mask-pair orderings also remain unresolved. To characterize what the score tracks, we measure Local Predictive-State Dispersion, a model-defined structural probe that covaries positively with uncertainty. An analytic oracle shows that interpreting this probe as exact conditional ambiguity is regime-dependent. Together, these results establish a controlled boundary between recognizing prediction difficulty and identifying missing-measurement consequences, motivating validation against the specific semantic claim made about an uncertainty score.

## 1 INTRODUCTION

*When is prediction difficult?* and *what information is missing?* are different questions. Error association and predictive calibration can support using uncertainty to recognize difficult predictions (Kendall & Gal, 2017; Kuleshov et al., 2018). Interpreting the same score as evidence about which unavailable measurements matter requires an additional connection: its ordering of observation choices must agree with their predictive consequences.

Partial EEG observation lets us test that connection while controlling the prediction task. We withhold different channel subsets from the same stored history, keeping its future target, prediction horizon, and retained channel count fixed. Only the identity of observed channels changes. This separates measurement identity from measurement count and yields comparable outcomes for alternative masks under a shared reference task.

Our central finding is a **dissociation between reproducible mask rankings**. Teacher-state consequence rankings reproduce across disjoint held-out subject halves with correlation $0.679$; uncertainty-induced rankings are more reproducible, at $0.811$–$0.957$ across three constructions. Their aggregate cross-half alignment nevertheless remains weak or negative. The mismatch is therefore not explained solely by either ranking lacking population structure. It does not imply that every condition fails, that correlations equal zero, or that individual mask pairs can be precisely ordered.

This result raises a second question: what structure does predictive uncertainty capture? The score is associated with prediction error even in fixed-mask diagnostics, and it covaries with Local Predictive-State Dispersion (LPSD), computed from full-view teacher states retrieved through partial-view neighborhoods. LPSD supplies an explanatory probe of what uncertainty tracks, rather than a



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.

demonstrated causal mechanism for the mismatch. A separate analytic oracle then tests the probe itself and finds regime-dependent recovery of conditional ambiguity.

We make three contributions. First, we construct a **controlled semantic audit** of the inference from ordinary predictive uncertainty to missing-measurement relevance. Second, we establish a **structural–interventional dissociation**: reproducible uncertainty and consequence rankings need not positively align across subjects. Third, we provide an **audited structural interpretation**, combining positive uncertainty–LPSD associations with an oracle boundary on what LPSD can mean. The contribution is this evidence chain under partial observation; the predictor remains deliberately standard.

## 2 Related Work and Positioning

**Predictive uncertainty and reliability.** Heteroscedastic regression learns input-dependent variance (Kendall & Gal, 2017; Gal & Ghahramani, 2016; Lakshminarayanan et al., 2017), while calibration, selective prediction, and shift evaluation assess predictive distributions or decision reliability (Kuleshov et al., 2018; Guo et al., 2017; Geifman & El-Yaniv, 2019; Ovadia et al., 2019; Malinin & Gales, 2018; van Amersfoort et al., 2020). Optimization can also affect the learned mean and variance (Seitzer et al., 2022). These evaluations establish properties of predictions; they do not directly test whether a score ranks the consequences of alternative measurements.

**Missing-observation acquisition.** Methods such as EDDI evaluate additional observations under an explicit information objective (Ma et al., 2019; Gal et al., 2017; Kirsch et al., 2019). Our retrospective audit compares prediction errors under alternative masks of an already recorded signal. It neither integrates over unknown acquired values nor optimizes measurement costs or an acquisition policy. The question is whether ordinary predictive uncertainty supports a measurement-consequence interpretation, not whether a dedicated acquisition objective succeeds.

**Uncertainty attribution and explanation.** Information-theoretic Shapley values assign feature contributions to a specified uncertainty functional (Watson et al., 2023); CXPlain constructs explanations from prediction consequences (Schwab & Karlen, 2019; Sundararajan et al., 2017). Explaining such a functional differs from using raw predictive variance as evidence of missing-measurement utility. We do not assume that these attribution methods make the latter substitution.

**Semantic validity and our position.** Measurement research distinguishes operational proxies from the constructs attributed to them (Jacobs & Wallach, 2021; Kuhn et al., 2023). Our question concerns *inferential license, not estimator intent*: after an uncertainty score passes ordinary predictive tests, what stronger interpretation is empirically justified? EEG provides a controlled setting for this audit. Unlike representation studies such as BENDR and LaBraM (Kostas et al., 2021; Jiang et al., 2024; Yang et al., 2023), our objective is to characterize a semantic boundary rather than introduce a new EEG representation, calibration method, or explanation algorithm.

## 3 Problem Setting and Audit Questions

### 3.1 Partial EEG observation

Let $H$ denote a complete multichannel EEG history, $M \in \{0,1\}^{19}$ an observation mask, and $k = \sum_c M_c$ the retained channel count. The partial view $V = g(H, M)$ includes the mask and observed-channel reference canonicalization. A frozen descriptor maps the future segment to $Y = \Psi(F_{t+\Delta}) \in \mathbb{R}^{64}$ at horizon $\Delta$. The heteroscedastic predictor returns a mean and a diagonal variance:

$$q_\theta(Y \mid V, \Delta) = \mathcal{N}\big(\mu_\theta(V,\Delta), \operatorname{diag}\sigma^2_\theta(V,\Delta)\big), \quad U_\theta(V) = \frac{1}{d}\sum_{j=1}^{d}\sigma^2_{\theta,j}(V), \quad d = 64. \tag{1}$$

We suppress $\Delta$ where it is fixed. $U$ is an operational prediction score; its interpretation as ambiguity or measurement relevance requires evidence beyond its definition.

### 3.2 THREE LEVELS OF INTERPRETATION

**Q1: Response validity.** Does higher uncertainty accompany greater prediction error, including within fixed observation conditions? **Q2: Structural validity.** Does it accompany greater dispersion of full-view predictive states among similar partial views? **Q3: Interventional validity.** At fixed $H, Y, \Delta, k$, does the score order masks in the same direction as a specified predictive consequence $D$, so that higher $U(M)$ indicates greater $D(M)$? Q3 is a stronger use of the score than recognizing difficult predictions and needs its own validation target. Q2 characterizes a structural association; it does not logically imply Q3.

### 3.3 CONTROLLED AUDIT DESIGN

Figure 1 shows how the same prediction setting supports the three tests. Mask withholding changes model inputs, including the mask-dependent reference operation, while leaving the recorded physiology and future target unchanged. We establish Q1, test Q3 by checking each ranking's reproducibility before comparing them, and then use Q2 to characterize what uncertainty tracks. Figure 2 previews the central result: both mask rankings reproduce across subjects, but their aggregate alignment does not recover a consistently positive ordering. Section 5 defines the estimands and evaluates this mismatch; the final oracle audit tests the structural probe's construct interpretation.

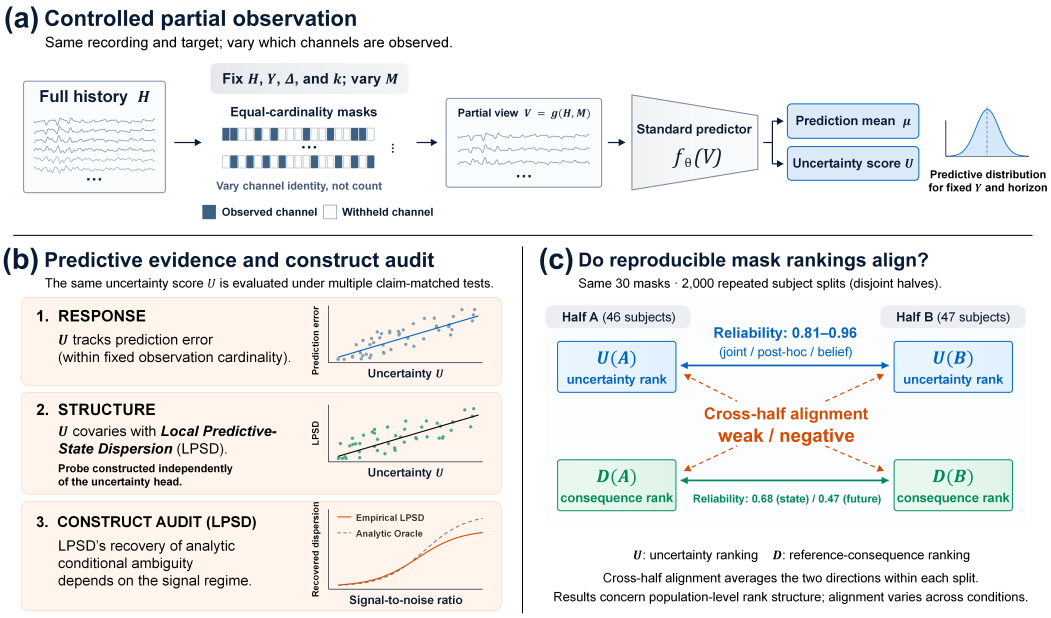


Figure 1: **Controlled semantic audit.** (a) Hold history $H$, target $Y$, horizon $\Delta$, and count $k$ fixed; vary the channel mask $M$. (b) Response and structural tests evaluate error association and LPSD; an oracle separately audits the probe. Waveforms, scatterplots, and curves are schematic, not empirical samples or fitted recovery curves. (c) The same 30 masks are ranked in disjoint subject halves. Horizontal comparisons measure each ranking's reproducibility; diagonal comparisons measure uncertainty–consequence alignment. Values summarize population rank structure, with condition-dependent alignment. Each semantic claim has its own validation target.

## 4 RESPONSE VALIDITY: IS THE UNCERTAINTY SIGNAL INFORMATIVE?

**Data and predictor.** The study uses 1,600 hours of TUEG from 1,918 training subjects and holds out 126 hours from 101 subjects (Neural Engineering Data Consortium, 2026). A roughly one-million-parameter encoder predicts a 64-dimensional future descriptor from ten seconds of EEG, with five seconds as the primary horizon. Full and degraded histories share the same training target. Appendix B specifies the data and descriptor; Appendix C details the architecture and optimization.

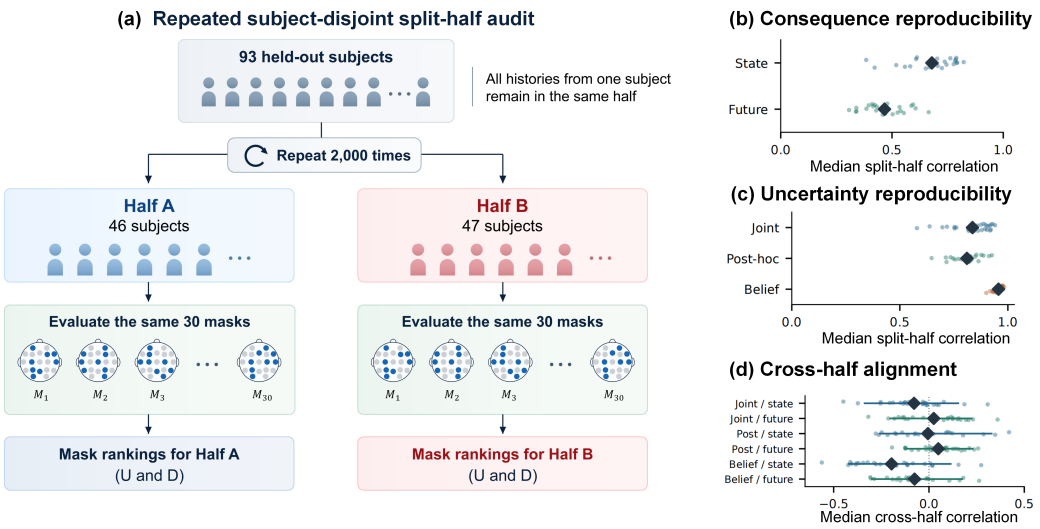


Figure 2: **Reproducible mask rankings do not reliably align.** (a) The 93 held-out subjects are partitioned into groups of 46 and 47, evaluating the same 30 masks in each half, over 2,000 splits. (b–c) Consequence and uncertainty rankings each reproduce across halves. (d) Cross-half alignment is weak or negative in aggregate, with both signs across conditions. Dots show 27 seed–bank–keep condition medians; diamonds show their equal-weight mean. Bars in (d) are the 5th–95th percentiles across condition medians, not confidence intervals. Reproducibility concerns population structure and does not establish precise pairwise ordering.

**Error association within observation conditions.** Uncertainty carries a nontrivial response signal. Figure 3a–b shows fixed-cardinality uncertainty–error correlations of 0.546–0.637 and a fixed-mask joint-score correlation of 0.591, with reported 95% subject-cluster bootstrap interval $[0.516, 0.657]$ over 400 replicates. The fixed-mask diagnostic also has nonconstant uncertainty (SD approximately 4.1). Its summary combines seeds and keep levels, so it supports a bounded sample-dependence diagnostic rather than a claim about every exact mask (Appendix E).

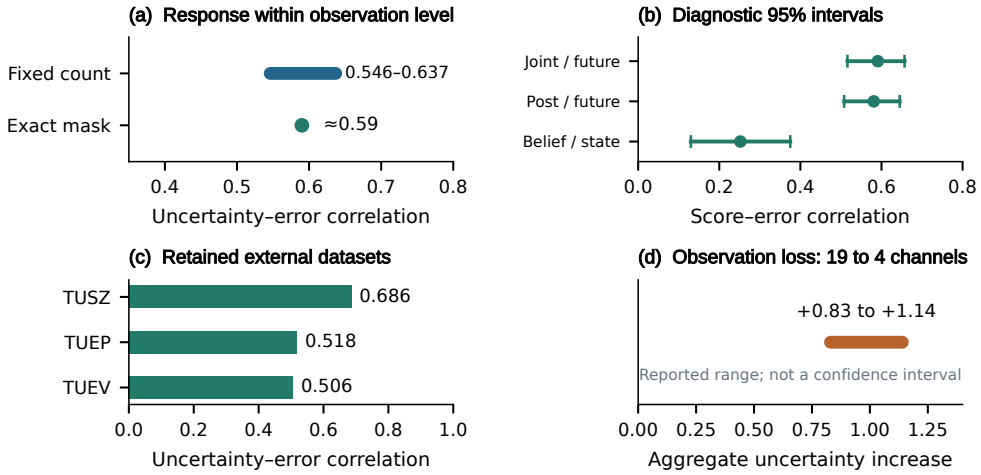


Figure 3: **Uncertainty retains predictive response information.** (a) Fixed-cardinality range and approximate fixed-mask association. (b) Diagnostic correlations and reported 95% percentile intervals from 400 subject-cluster replicates; belief uses teacher-state error, the other scores future error. Rows summarize seeds and keep levels together. (c) Subject-disjoint TUH-derived external evaluations. (d) Aggregate increase under 19-to-4-channel withholding. Panels (a), (c), and (d) show ranges or point summaries, not confidence intervals.

**Degradation and external robustness.** Withholding channels from 19 to 4 increases aggregate uncertainty by 0.83–1.14. Without predictive-model refitting, uncertainty–error correlations are 0.686, 0.518, and 0.506 on TUSZ, TUEP, and TUEV, respectively (Figure 3c–d). The identity audit finds the retained subsets subject-disjoint from each other and TUEG training/evaluation, but all share the TUH ecosystem. CHB-MIT is excluded because of person-overlap and bipolar-montage concerns (Appendix B).

These tests establish sample dependence, error association, and degradation sensitivity. They do not establish probabilistic superiority: the optimized homoscedastic comparator has better aggregate NLL on TUEG and TUSZ. Calibration, temperature rescaling, prediction-error comparisons, and common-reference behavior are reported in Appendices D–E. A useful response signal is the starting point for the mask audit, not evidence that the score identifies which measurements matter.

# 5 Interventional Validity: Do Reproducible Rankings Align?

## 5.1 Measurement consequences

The intervention pool contains 256 complete 19-channel histories from 93 held-out subjects. At fixed history, target, horizon, and $k \in \{4, 8, 12\}$, each bank supplies 30 alternative masks. A frozen full-view teacher defines $s_H = \mu_T(H)$. The partial-view belief mean $m_B$ reconstructs $s_H$, and a designated future predictor $\mu_0$ predicts $Y$:

$$D_{i,r}^{\text{state}} = \|m_B(H_i, M_r) - s_{H_i}\|_2^2, \tag{2}$$

$$D_{i,r}^{\text{future}} = \|Y_i - \mu_0(H_i, M_r)\|_2^2. \tag{3}$$

Teacher-state consequence is model-relative, not physiological ground truth. Both quantities are squared errors, without subtracting a full-view baseline. The three uncertainty constructions share these reference consequences; they are not necessarily each score's own predictive risk. Joint and post-hoc scores model future residual variance; the belief diagnostic models dispersion over teacher states (Appendix C).

## 5.2 Is the consequence ranking reproducible?

A mismatch is informative only if the target ordering has measurable structure. We repeatedly partition the 93 subjects into disjoint halves $A$ and $B$ of sizes 46 and 47, keeping all histories of each person together. Each half evaluates the same 30 masks. For each mask, we first average histories within a subject, then average subjects within a half. These subject-balanced means define

$$R_D^a = \text{Spearman}_r(\bar{D}_r^{a,A}, \bar{D}_r^{a,B}), \qquad R_U^e = \text{Spearman}_r(\bar{U}_r^{e,A}, \bar{U}_r^{e,B}), \tag{4}$$

where $a$ indexes consequence and $e$ estimator. For each seed–bank–keep condition, we take the median over 2,000 splits; reported aggregate values equally average all 27 condition medians. Banks are analyzed separately, not pooled into 90-mask rankings (Appendix H.8).

Consequence rankings reproduce in both spaces: aggregate $R_D^{\text{state}} = 0.679$ and $R_D^{\text{future}} = 0.466$, with positive medians in all 27 conditions for each target (Figure 2b). Future-error rankings have greater split variability. Teacher-state emphasis follows this reliability analysis; both spaces remain reported. Repeated partitions measure stability within this held-out pool, not 2,000 independent population replications.

## 5.3 Is the uncertainty ranking reproducible?

The score-induced ordering is also reproducible: $R_U = 0.836$ for joint, 0.811 for post-hoc, and 0.957 for belief uncertainty (Figure 2c). All 27 condition medians are positive for each estimator. Thus the subsequent comparison involves reproducible structure on both sides, rather than only a structured uncertainty score and an unstructured target.

### 5.4 Do the rankings align across subjects?

We compare uncertainty in one half with consequence in the other, averaging both directions *within each split*:

$$R_{UD}^{e,a,\text{cross}} = \tfrac{1}{2}\left[\text{Spearman}_r(\bar{U}_r^{e,A}, \bar{D}_r^{a,B}) + \text{Spearman}_r(\bar{U}_r^{e,B}, \bar{D}_r^{a,A})\right]. \tag{5}$$

The same median-then-mean reduction gives Table 1. Despite reproducibility on both sides, teacher-state alignment is weak or negative, and future-error alignment remains weak or negative in aggregate. Figure 2d displays the heterogeneity: every estimator–target combination has condition medians of both signs. Same-half comparisons yield the same qualitative pattern (Appendix H).

Table 1: **Rank reproducibility and cross-half alignment.** Equal-weight means of 27 seed–bank–keep condition medians over 2,000 subject splits; entries are descriptive coefficients, not confidence limits. Consequence reproducibility is 0.679 (state) and 0.466 (future). Table 15 and Tables 16–18 provide same-half comparisons and complete condition summaries.

| Estimator | Uncertainty reproducibility $R_U$ | Cross-half alignment $R_{UD}$ Teacher state | Future error |
|---|---|---|---|
| Joint | 0.836 | −0.078 | +0.025 |
| Post-hoc | 0.811 | −0.006 | +0.049 |
| Belief | 0.957 | −0.196 | −0.076 |

We call this pattern a **structural–interventional dissociation**: uncertainty and consequence each exhibit reproducible population-level mask structure, but their rankings do not reliably align. This weakens an explanation based solely on absent ranking signal. It does not remove all measurement error, establish equivalence to zero, or validate an acquisition policy based on reversing the score.

### 5.5 Population structure is not pairwise identification

Global reproducibility coexists with limited pairwise resolution. In the teacher-state bootstrap audit, 7,390 of 11,745 condition-specific mask pairs (62.9%) have ordering frequencies between 0.05 and 0.95. Only about 9% of 810 mask entries have narrow rank ranges; Appendix H gives the denominators and a discrepancy between stored widths and interval endpoints. Subject-balanced and window-weighted consequence rankings are highly concordant (0.931 state; 0.944 future), although they define different estimands. The supported finding is reproducible population structure without consistent positive cross-ranking recovery, not precise identification of every consequential subset.

## 6 Structural Validity: What Does Predictive Uncertainty Track?

### 6.1 Local Predictive-State Dispersion

The intervention audit leaves an informative score whose mask ordering does not recover shared measurement consequences. We now ask what structure it tracks. **LPSD is a model-defined structural probe, not a ground-truth ambiguity label.** For partial query $V_i$, retrieve $K$ similar partial views from reference subjects disjoint from the query, at the same keep level and horizon. The corresponding frozen full-view teacher states define

$$D_{\text{LPSD}}(V_i) = \text{Tr}\,\widehat{\text{Cov}}\{s_{H_j} : j \in \mathcal{N}_K(V_i)\}. \tag{6}$$

No uncertainty-head output enters retrieval or state construction. The primary analysis uses $K = 20$. Figure 4a shows the probe; Appendix F gives its computation. Conditional total covariance motivates predictive-state dispersion, but neither that identity nor Gaussian likelihood identifies the learned score with it (Appendix C).

### 6.2 Uncertainty–LPSD association

Uncertainty covaries positively with the probe across all three constructions. Table 2 reports mean correlations of 0.449–0.534; all 18 underlying seed coefficients are positive. Persistence under

post-hoc variance fitting weakens a dependence solely on joint mean–variance optimization. The finding characterizes predictive-state structure in the score, while leaving shared training distribution, teacher error, and representation geometry as possible dependencies.

Table 2: **Uncertainty–LPSD association.** Mean ± sample SD across three seeds for TUEG encoder-latent neighborhoods, $K = 20$, 1,024 queries from 93 held-out subjects. SD uses denominator $3 - 1$ and describes seed variation, not subject-sampling uncertainty. All input coefficients appear in Table 11.

| Estimator | $k = 8$ | $k = 12$ |
|---|---|---|
| Joint heteroscedastic | $+0.459 \pm 0.114$ | $+0.471 \pm 0.129$ |
| Post-hoc | $+0.449 \pm 0.121$ | $+0.469 \pm 0.130$ |
| Belief diagnostic | $+0.534 \pm 0.075$ | $+0.514 \pm 0.092$ |

### 6.3 Geometry and Negative Controls

The association extends across encoder-latent and handcrafted bandpower/temporal neighborhoods, with ranges approximately 0.30–0.61 and 0.21–0.49. Regional-STFT neighborhoods are less stable and include a sign reversal (Figure 4c). Sensitivity to $K \in \{10, 20, 50\}$ and alternative dispersion statistics is detailed in Appendix F; the results do not establish geometry-independent ambiguity recovery.

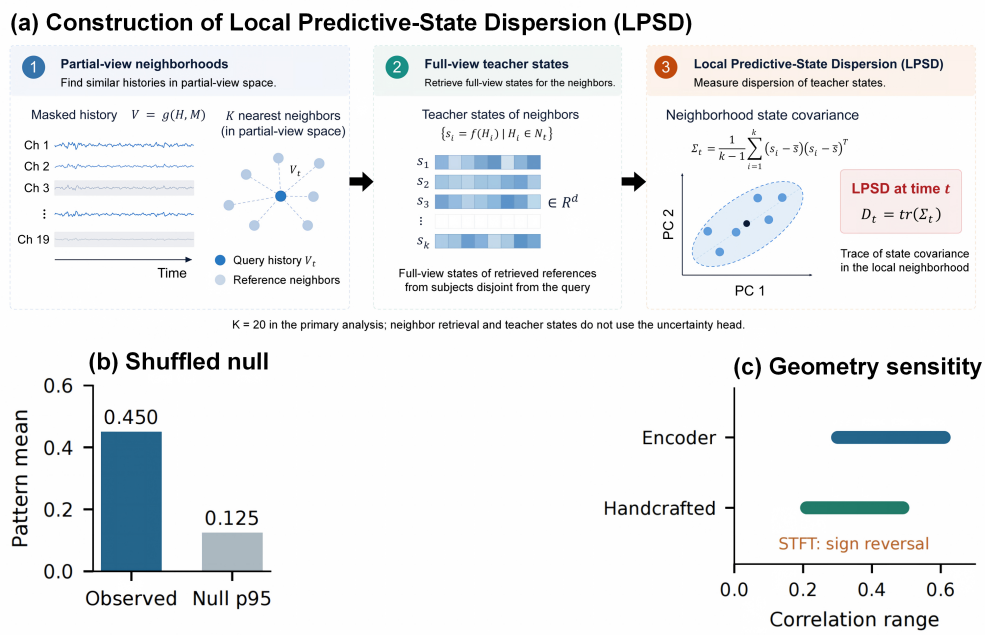


Figure 4: **Predictive-state structure and its controls.** (a) Partial-view neighborhoods retrieve full-view teacher states from disjoint reference subjects; LPSD is their covariance trace, with $K$ denoting neighborhood size and $k$ retained channels. The point cloud is schematic. (b) Mean observed correlations across 16 contrasts are compared with empirical null 95th percentiles from 20 state-assignment permutations; the averaged threshold is descriptive rather than a pooled significance test. (c) Correlation ranges across encoder and handcrafted geometries are shown; regional STFT includes a sign reversal. Ranges are not confidence intervals.

Shuffling teacher-state assignments while preserving queries and neighborhood geometry supplies a targeted negative control. Fifteen of 16 observed contrasts exceed their empirical null 95th percentile across 20 permutations; the average observed statistic is 0.450, compared with an average threshold of 0.125 (Figure 4b). The exception is TUSZ, $k = 8$, PCA-64: 0.126 versus 0.148. This is a descriptive threshold pattern, not 15 multiplicity-adjusted discoveries (Appendix G). The score therefore carries

structural information under specified probes, even though its reproducible mask ordering does not consistently recover measurement consequence.

## 7 Construct Validity: How Far Can LPSD Be Interpreted?

### 7.1 Analytic Oracle

Association with LPSD establishes Q2 only for the defined probe. To test the stronger conditional-ambiguity interpretation, we use an eight-class oracle with uniform prior, centers $\mu_c \sim \mathcal{N}(0, 3^2 I)$, and independent Gaussian noise:

$$Y \mid C \sim \mathcal{N}(\mu_C, \sigma_0^2 I), \qquad V = \mu_C^{(1:d_V)} + \sigma_v \epsilon, \qquad \sigma_0 = 1. \tag{7}$$

The known posterior satisfies $p(c \mid V) \propto \exp[-\|V - \mu_c^{(1:d_V)}\|^2/(2\sigma_v^2)]$. Writing $\bar{\mu}(V) = \sum_c p(c \mid V)\mu_c$, the conditional between-state covariance is

$$B^*(V) = \sum_c p(c \mid V)(\mu_c - \bar{\mu}(V))(\mu_c - \bar{\mu}(V))^\top, \quad b^*(V) = \operatorname{Tr} B^*(V). \tag{8}$$

Total predictive covariance additionally contains $\sigma_0^2 I$. We compare retrieved class-mean dispersion with $b^*$ using 50,000 references, 5,000 queries, and $K = 20$ (Appendix I).

### 7.2 Regime-Dependent Recovery

Figure 5 shows strong recovery at $d_V = 2, 4$ with $\sigma_v = 1$ $(0.91, 0.75)$, but approximately zero correlation at $d_V = 8, 16$. At the latter dimensions, increased observation noise yields intermediate recovery $(0.44, 0.43)$ and stronger recovery in non-degenerate regimes $(0.83, 0.73)$. Increasing the reference bank from 20,000 to 100,000 does not rescue the reported degenerate regimes. Dimension alone is therefore insufficient to explain the pattern. Dependence on signal regime and target dynamic range is a plausible interpretation, but noise also changes retrieval geometry and mixture overlap.

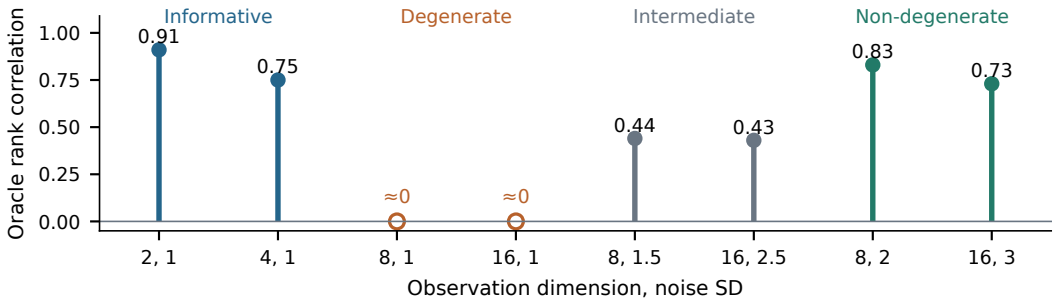


Figure 5: **LPSD's conditional-ambiguity interpretation depends on regime.** Rank recovery across eight oracle settings with uniform class prior, centers drawn from $\mathcal{N}(0, 3^2 I)$, and $\sigma_0 = 1$. Points are reported regime summaries, not confidence intervals. Hollow markers denote approximately-zero summaries; Spearman correlation would be undefined for an exactly constant target.

### 7.3 Construct Boundary

LPSD is not uniformly validated as exact conditional ambiguity. The oracle establishes a boundary for this finite-bank probe, not an impossibility result for conditional-moment estimation or a validation of latent EEG ambiguity. Matched representation controls also show dataset- and statistic-dependent ordering reversals (Appendix J). A useful structural association therefore supports a narrower interpretation than exact recovery of a latent conditional quantity.

## 8 Discussion and Scope

**Claim-to-evidence synthesis.** Table 3 summarizes the evidence ladder. Response tests establish that uncertainty is informative; the rank audit establishes that informative, reproducible uncertainty

need not recover a separately reproducible consequence ordering. LPSD characterizes a structure with which the score covaries, and the oracle bounds the probe's interpretation. These are distinct estimands, so their correlation magnitudes should not be subtracted or treated as interchangeable measures of validity.

Table 3: **Semantic claims and their evidential scope.** Each interpretation is evaluated against its own target. The structural result concerns a defined probe; the intervention result is an aggregate finding with condition heterogeneity.

| Semantic claim | Evidence | Conclusion |
|---|---|---|
| Prediction difficulty | Error association and degradation | Supported in evaluated settings |
| Predictive-state structure | LPSD association and controls | Supported within the defined probe |
| Missing-information ranking | Reproducible mask rankings; cross-half comparison | Shared consequence ordering not reliably recovered |
| Exact conditional ambiguity | Analytic oracle | Regime-dependent recovery |

**Interpretive scope.** The findings preserve the predictive usefulness of uncertainty. They do not establish absent electrode information, physiological electrode importance, clinical utility, or failure of dedicated acquisition methods. The target is a reference-model consequence under retrospective withholding, not the expected value of acquiring a new measurement. Global rank reproducibility also does not provide precise ordering of every mask pair or a general impossibility theorem.

**Generality and limitations.** The study uses one EEG ecosystem, a fixed montage and future descriptor, learned reference models, finite banks, and a 93-subject intervention pool. Future-error ranking is less reproducible than teacher-state ranking, and the latter's emphasis is audit-informed. Both signs occur across condition-specific alignment estimates; the aggregate does not imply universal conditionwise failure. External evaluations remain within TUH. Exact-mask aggregation and several implementation details limit full experimental reproduction (Appendix M). Whether the same dissociation generalizes beyond partial EEG observation remains an empirical question.

**Implications.** For uncertainty estimation and explanation, response accuracy should be paired with a validation target for the intended semantic use. For acquisition and multimodal learning, future audits should distinguish score reproducibility from agreement with an explicit acquisition objective, including measurement costs and unknown values. The principle is specific: *validate a score against the semantic claim for which it is used*.

## 9 Conclusion

Predictive uncertainty identifies difficult predictions in the evaluated EEG setting. Yet reproducible uncertainty rankings do not consistently recover independently reproducible measurement-consequence rankings. Their association with LPSD supports a bounded structural interpretation, not unrestricted measurement utility or exact ambiguity. Knowing when the model is uncertain is not the same as knowing what information it needs.

## Reproducibility Statement

Appendices B–L specify the data, prediction models, consequence operators, statistical estimands, and derivations. Complete condition tables, nine CSV summaries, the rank-aggregation rule, and a deterministic recomputation script accompany the manuscript. Figures preserve the original designs or are generated from stated summary values; schematic elements are identified in their captions. These materials support reproduction of aggregate arithmetic. Raw EEG experiments, model training, individual split coefficients, and all mask/configuration artifacts are outside the accomanying materials; Appendix M details this scope.

## Ethics Statement

This study uses EEG data to examine predictive uncertainty under controlled measurement withholding. The results do not validate diagnosis, seizure prediction, electrode placement, clinical decision-making, or a measurement-acquisition policy, and channel-mask rankings should not be interpreted as physiological importance. All datasets remain subject to the access, privacy, and usage requirements of their original sources. No new prospective human-subject experiment was conducted, and we do not assert any new institutional approval, exemption, or consent status beyond that associated with the original data resources.

## AI Use Statement

In this work, we used generative AI tools for English drafting and revision, LaTeX preparation, and visualization of results. We have not used generative AI tools to train the EEG models or independently rerun the underlying experiments, and other experimental data-collection or model-training tasks are not applicable to this use. We reviewed all AI-assisted work. AI-assisted text, mathematical expressions, LaTeX, visualizations, and recomputed summaries were checked against the manuscript, source materials, and reported aggregate results for consistency and correctness. Generative AI outputs were not treated as independent experimental evidence or as a substitute for verification of the underlying analyses. We take responsibility for the final content of this work, including text, claims, figures, and other artifacts produced with the aid of generative AI.

## References

Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 1050–1059, 2016. URL https://proceedings.mlr.press/v48/gal16.html.

Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep Bayesian active learning with image data. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1183–1192, 2017. URL https://proceedings.mlr.press/v70/gal17a.html.

Yonatan Geifman and Ran El-Yaniv. SelectiveNet: A deep neural network with an integrated reject option. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2151–2159, 2019. URL https://proceedings.mlr.press/v97/geifman19a.html.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330, 2017. URL https://proceedings.mlr.press/v70/guo17a.html.

John Guttag. CHB-MIT Scalp EEG Database. *PhysioNet*, 2010. doi: 10.13026/C2K01R. URL https://physionet.org/content/chbmit/1.0.0/. Version 1.0.0.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979. URL https://www.jstor.org/stable/4615733.

Abigail Z. Jacobs and Hanna Wallach. Measurement and fairness. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*, 2021. doi: 10.1145/3442188.3445901. URL https://arxiv.org/abs/1912.05511.

Wei-Bang Jiang, Li-Ming Zhao, and Bao-Liang Lu. Large brain model for learning generic representations with tremendous EEG data in BCI. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2405.18765.

Alex Kendall and Yarin Gal. What uncertainties do we need in Bayesian deep learning for computer vision? In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/hash/2650d6089a6d640c5e85b2b88265dc2b-Abstract.html.

Andreas Kirsch, Joost van Amersfoort, and Yarin Gal. BatchBALD: Efficient and diverse batch acquisition for deep Bayesian active learning. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://arxiv.org/abs/1906.08158.

Demetres Kostas, Stephane Aroca-Ouellette, and Frank Rudzicz. BENDR: Using transformers and a contrastive self-supervised learning task to learn from massive amounts of EEG data. *arXiv preprint arXiv:2101.12037*, 2021. URL https://arxiv.org/abs/2101.12037.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2302.09664.

Volodymyr Kuleshov, Nathan Fenner, and Stefano Ermon. Accurate uncertainties for deep learning using calibrated regression. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2796–2804, 2018. URL https://proceedings.mlr.press/v80/kuleshov18a.html.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://arxiv.org/abs/1612.01474.

Chao Ma, Sebastian Tschiatschek, Konstantina Palla, José Miguel Hernández-Lobato, Sebastian Nowozin, and Cheng Zhang. EDDI: Efficient dynamic discovery of high-value information with partial VAE. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 4234–4243, 2019. URL https://proceedings.mlr.press/v97/ma19c.html.

Andrey Malinin and Mark Gales. Predictive uncertainty estimation via prior networks. In *Advances in Neural Information Processing Systems*, volume 31, 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/hash/3ea2db50e62ceefceaf70a9d9a56a6f4-Abstract.html.

Neural Engineering Data Consortium. The Temple University Hospital EEG corpus: Corpora and downloads. Temple University, 2026. URL https://isip.piconepress.com/projects/nedc/html/tuh_eeg/. Accessed September 5, 2026; project resource, not a statement of the study's inventory size.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, D. Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/hash/8558cb408c1d76621371888657d2eb1d-Abstract.html.

Tom Pollard, Benjamin E. Moody, Li wei H. Lehman, Brian J. Gow, Chrystinne Fernandes, Chen Xie, Alistair Johnson, Roger G. Mark, and Thomas Heldt. PhysioNet as a global platform for biomedical research. *Nature Health*, 1(8):792–795, 2026. doi: 10.1038/s44360-026-00096-z. URL https://doi.org/10.1038/s44360-026-00096-z.

Patrick Schwab and Walter Karlen. CXPlain: Causal explanations for model interpretation under uncertainty. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/hash/3ab6be46e1d6b21d59a3c3a0b9d0f6ef-Abstract.html.

Maximilian Seitzer, Arash Tavakoli, Dimitrije Antic, and Georg Martius. On the pitfalls of heteroscedastic uncertainty estimation with probabilistic neural networks. In *International Conference on Learning Representations*, 2022. URL https://arxiv.org/abs/2203.09168.

Ali Shoeb. *Application of Machine Learning to Epileptic Seizure Onset Detection and Treatment*. PhD thesis, Massachusetts Institute of Technology, September 2009.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 3319–3328, 2017. URL https://proceedings.mlr.press/v70/sundararajan17a.html.

Joost van Amersfoort, Lewis Smith, Yee Whye Teh, and Yarin Gal. Uncertainty estimation using a single deep deterministic neural network. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 9690–9700, 2020. URL https://proceedings.mlr.press/v119/van-amersfoort20a.html.

David Watson, Joshua O'Hara, Niek Tax, Richard Mudd, and Ido Guy. Explaining predictive uncertainty with information theoretic Shapley values. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/16e4be78e61a3897665fa01504e9f452-Abstract-Conference.html.

Chaoqi Yang, M. Brandon Westover, and Jimeng Sun. BIOT: Biosignal transformer for cross-data learning in the wild. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://arxiv.org/abs/2305.10351.

# A Appendix

The appendices provide the evidence and implementation details supporting the nine main-text sections. Appendices B–C specify the setting and estimators; Appendices D–E expand the response tests in Section 4. Appendix H gives the complete rank audit for Section 5, including condition-level results and pairwise denominators. Appendices F–G support the structural analysis in Section 6; Appendices I–J bound its construct interpretation in Section 7. Appendices K–M collect scope checks, statistical definitions, and reproducibility limits.

# B Datasets, Canonicalization, and Preprocessing

## B.1 Study inventory and evaluation roles

The study-specific TUEG inventory, filtering, and split counts appear in Table 4. They describe the available source/cache, not the entire public TUH release (Neural Engineering Data Consortium, 2026). Rounded hours are retained at the precision of the study summaries. TUEG supplies training data and the primary held-out population; the three retained external datasets are used without refitting the predictive model. Table 5 separates final evaluation sizes from source inventories.

Table 4: **Source inventory and primary split.** External inventory rows are not the final evaluation subsets.

| Inventory | Subjects | Recordings | Hours |
|---|---|---|---|
| TUEG source | 2,042 | 6,820 | 2,533 |
| TUEG filtered | 2,019 | 6,558 | 1,726 |
| TUEG training | 1,918 | 6,271 | 1,600 |
| TUEG held out | 101 | 287 | 126 |
| TUSZ source/cache | 120 | 1,124 | 258 |
| TUEP source/cache | 31 | 400 | 102 |
| TUEV source/cache | 127 | 127 | 37 |

Table 5: **Retained frozen evaluation subsets.** The identity audit reports distinct persons across the TUH-derived evaluation subsets and TUEG training/evaluation.

| Evaluation | Subjects | Recordings | Hours |
|---|---|---|---|
| TUEG | 101 | 287 | 126 |
| TUSZ | 18 | 111 | 44 |
| TUEP | 9 | 209 | 47 |
| TUEV | 38 | 38 | 12 |

## B.2 Identity and split audit

Table 6 specifies the nine comparisons underlying the subject-disjoint external evaluations. The supplementary subject/recording manifest contains 8,873 entries. Its overlap audit reports the TUSZ test, TUEP held-out, and TUEV held-out subsets as pairwise subject-disjoint and disjoint from TUEG training and evaluation. This addresses shared-person leakage in the stated partitions. It does not establish independent acquisition institutions, nor does a subject audit by itself certify every possible duplicate, preprocessing, or target-selection pathway. The retained claim is subject-disjoint cross-dataset robustness within the TUH-derived ecosystem.

Source training and held-out persons are separated. LPSD query and reference banks are also specified as subject-disjoint, and PCA is fitted on the training pool. Repeated recordings and windows from one person remain within the same person-level partition. A window count therefore does not equal the number of independent units for uncertainty estimation. The main intervention pool contains 256 histories but 93 persons.

Table 6: **Exported overlap audit.** Source labels are retained. All nine reported comparisons have zero shared subjects and recordings; these are the study's evaluation partitions, not a claim about every public-corpus record.

| Partition A | Partition B | Subjects | Recordings |
|---|---|---|---|
| tueg_train | tusz_eval | 0 | 0 |
| tueg_train | tuep_eval | 0 | 0 |
| tueg_train | tuev_eval | 0 | 0 |
| tueg_eval | tusz_eval | 0 | 0 |
| tueg_eval | tuep_eval | 0 | 0 |
| tueg_eval | tuev_eval | 0 | 0 |
| tusz | tuep | 0 | 0 |
| tusz | tuev | 0 | 0 |
| tuep | tuev | 0 | 0 |

### B.3 Canonical channels and windows

The canonical 19-channel order is

FP1, FP2, F7, F3, FZ, F4, F8, T7, C3, CZ, C4, T8, P7, P3, PZ, P4, P8, O1, O2.

Legacy aliases are T3→T7, T5→P7, T4→T8, and T6→P8. Missing channels have mask bit zero and produce no valid token; they are not treated as measured zero-valued EEG. Resampling targets 200 Hz. The history duration is 10 s, the future segment duration 4 s, and prediction horizons are $\{1, 5, 15, 30\}$ s, with 5 s primary. Windows never cross recording boundaries. Sampling is balanced hierarchically by subject, recording, and anchor.

Quality filtering rejects NaNs, flatlines, clipping, and source observations with more than 50% missing channels. This source-quality check precedes deliberate mask degradation. Otherwise a deliberately generated four-channel input would be confused with an invalid source record. The audited intervention pool is stricter: all 256 histories have 19 genuinely available canonical channels before masking.

### B.4 CHB-MIT exclusion

CHB-MIT distinguishes cases from persons, including repeated recordings of the same individual across chb01 and chb21 (Guttag, 2010; Shoeb, 2009). The supplementary audit identifies these cases in nominal training and held-out partitions respectively and identifies a 16-lead bipolar input montage. A bipolar lead is a voltage difference, not a single canonical electrode potential. Directly renaming FP1–F7 as FP1 would therefore change the observation model. A defensible conversion requires an explicit measurement operator and reference constraint, with attention to connected components and identifiability.

We exclude CHB-MIT from the external response/calibration results and downstream numerical comparisons. This does not imply that the TUEG-trained predictor was trained on CHB-MIT; the identity finding concerns the dataset-specific partition labels, while the montage issue concerns input comparability. No corrected person-disjoint and measurement-consistent evaluation is included. The platform and resource are documented in (Guttag, 2010; Pollard et al., 2026); excluded results are not used to enlarge the evidence for the main claim.

### B.5 Fixed future descriptor

The descriptor uses an STFT with FFT length 128, window length 128, hop 32, and a 1–70 Hz retained range. At 200 Hz, these are 0.64 s windows and 0.16 s hops. Features summarize nine anatomical regions using power means/standard deviations and phase-coupling or mean-resultant statistics. Masked median/IQR standardization and training-pool PCA produce a 64-dimensional target.

Writing $\phi$ for the feature map, a generic expression is

$$\Psi(F) = P_{64}^{\top}\left(\frac{\phi(F) - m_{\text{train}}}{s_{\text{train}}}\right). \tag{9}$$

The expression specifies the functional stages, without assuming an undocumented whitening option. Exact region memberships, phase features, quality thresholds, horizon anchoring, and fitted transforms are required for numerical reproduction and are not fully specified in the accompanying summaries (Appendix M). Across history-mask interventions, the future segment, transform, and resulting target are held fixed.

# C Predictive Model and Estimator Variants

## C.1 Observed-channel reference canonicalization

For a nonempty mask,

$$\bar{x}_t(M) = \frac{\sum_c M_c x_{t,c}}{\sum_c M_c}, \qquad \tilde{x}_{t,c} = x_{t,c} - \bar{x}_t(M), \quad M_c = 1. \tag{10}$$

If all observed channels receive the same reference shift $a_t$, then $\bar{x}'_t(M) = \bar{x}_t(M) + a_t$ and $\tilde{x}'_{t,c} = \tilde{x}_{t,c}$. This is an algebraic invariance for a fixed mask. Changing $M$ changes the observed-channel mean: re-referencing is part of the implemented measurement intervention, not a separate physiological perturbation. The numerical reference experiment tests this common-shift setting rather than arbitrary montage reconstruction.

## C.2 Encoder and heads

A shared Conv1d uses kernel length 50 and stride 25, giving 250 ms patches and 125 ms spacing at 200 Hz. Channel and time embeddings identify tokens. A four-layer Transformer has model width 128, four attention heads, and feed-forward width 512. Masked mean pooling produces $h \in \mathbb{R}^{128}$; the appended horizon representation gives $h' \in \mathbb{R}^{160}$. Total size is approximately one million parameters.

The mean head maps $h'$ to $\mu \in \mathbb{R}^{64}$. The variance head receives $[h', M, \bar{k}]$, including the observation mask and keep-level feature. Its variance output is

$$\sigma_j^2 = \text{clip}_{[10^{-4}, 10^2]}\left(\text{softplus}(r_j) + 10^{-4}\right). \tag{11}$$

Receiving the mask explicitly is why a fixed-cardinality test alone cannot exclude exact-mask dependence. The supplementary exact-mask diagnostic tests variation beyond a particular fixed mask; it is not an architectural guarantee.

## C.3 Likelihood and variance target

The diagonal Gaussian density is

$$q(Y \mid V) = \prod_{j=1}^{d} \frac{1}{\sqrt{2\pi\sigma_j^2(V)}} \exp\left[-\frac{(Y_j - \mu_j(V))^2}{2\sigma_j^2(V)}\right]. \tag{12}$$

The negative log likelihood is

$$-\log q(Y \mid V) = \frac{1}{2}\sum_j \left[\frac{(Y_j - \mu_j(V))^2}{\sigma_j^2(V)} + \log\sigma_j^2(V) + \log(2\pi)\right]. \tag{13}$$

Removing the additive constant and normalizing by dimension gives

$$\mathcal{L}_{\text{NLL}} = \frac{1}{2d}\sum_j \left[\frac{(Y_j - \mu_j)^2}{\sigma_j^2} + \log\sigma_j^2\right]. \tag{14}$$

Full and degraded views share the same true future target. The reported keep ratios are $\alpha \in \{1, .75, .5, .25\}$.

For fixed $\mu_j$ and positive variance $v$, let $r(V) = \mathbb{E}[(Y_j - \mu_j(V))^2 \mid V]$. Conditional expected loss, up to an additive constant, is $\frac{1}{2}(r/v + \log v)$. Differentiation yields $v^*(V) = r(V)$ when $r > 0$. Expanding the conditional square gives

$$v_j^*(V) = \mathrm{Var}(Y_j \mid V) + \big(\mathbb{E}[Y_j \mid V] - \mu_j(V)\big)^2. \tag{15}$$

If the mean is not the conditional mean, its squared bias contributes to optimal variance. Finite samples, shared parameterization, optimization, and clipping further separate a trained score from this unconstrained optimum. None of these calculations identifies the empirical cause of the observed dissociation.

### C.4 Conditional covariance derivation

Assume finite second moments and a fixed deterministic observation map $V = g_M(H)$. With $m(H) = \mathbb{E}[Y \mid H]$,

$$Y - \mathbb{E}[Y \mid V] = \big(Y - m(H)\big) + \big(m(H) - \mathbb{E}[m(H) \mid V]\big). \tag{16}$$

The conditional cross terms vanish by iterated expectation because $\mathbb{E}[Y - m(H) \mid H] = 0$. Conditional outer products give

$$\mathrm{Cov}(Y \mid V) = \mathbb{E}[\mathrm{Cov}(Y \mid H) \mid V] + \mathrm{Cov}(m(H) \mid V). \tag{17}$$

For randomized masks, the same argument is conditional on $M$. Dropping $M$ from full-history quantities requires $Y \perp M \mid H$, appropriate to exogenous withholding that leaves the stored future unchanged.

For nested information sets $\mathcal{G}_1 \subseteq \mathcal{G}_2$,

$$\mathbb{E}\,\mathrm{Tr}\,\mathrm{Cov}(Y \mid \mathcal{G}_1) - \mathbb{E}\,\mathrm{Tr}\,\mathrm{Cov}(Y \mid \mathcal{G}_2) = \mathbb{E}\|\mathbb{E}[Y \mid \mathcal{G}_2] - \mathbb{E}[Y \mid \mathcal{G}_1]\|^2 \geq 0. \tag{18}$$

This expectation result neither orders every realized conditional variance nor compares arbitrary nonnested equal-cardinality masks. A learned variance score also need not satisfy the population identities exactly.

### C.5 Optimization and estimator constructions

Table 7 gives the common optimization settings for the estimators in Section 5.

Table 7: **Primary optimization settings.** Settings not specified here require the original model configuration.

| Setting | Value |
|---|---|
| Optimizer | Adam |
| Reported learning rate | $3 \times 10^{-4}$ |
| Warmup | 1,000 steps |
| Schedule | Cosine decay |
| Batch size | 64 |
| Training length | 30,000 steps |
| Gradient clipping | 1.0 |
| Training seeds | 3 |

**Joint heteroscedastic:** encoder, mean, and variance are optimized jointly for future-descriptor prediction. **Post-hoc:** a deterministic predictor is frozen, and a variance model is fitted to its training-pool residuals; final held-out subjects do not supply fitting targets. **Belief diagnostic:** a predictor models frozen full-view teacher states using 25% of training-pool windows and 12,000 steps. Its mean is denoted $m_B$ in the teacher-state consequence. This is a diagnostic distribution over model-defined states, not a physiological posterior. The accompanying summaries do not fully specify the belief distribution parameterization or scalar score reduction; exact reproduction requires the estimator configuration."

# D CALIBRATION AND PROBABILISTIC BASELINES

## D.1 SCALAR RESCALING AND RANKING

For a fixed positive variance multiplier $T$, $\tilde{\sigma}_j^2 = T\sigma_j^2$ gives $\tilde{U} = TU$. Within a group sharing $T$, rankings are unchanged in exact arithmetic. Different temperatures across datasets or horizons need not preserve a subsequently pooled ranking. Scalar calibration therefore cannot create independent evidence for LPSD or measurement-utility semantics.

## D.2 HOMOSCEDASTIC OPTIMUM

For frozen residuals $e_{ij} = Y_{ij} - \hat{\mu}_{ij}$ on a designated fitting set,

$$\mathcal{L}_j(v_j) = \frac{1}{2N} \sum_i \left( \frac{e_{ij}^2}{v_j} + \log v_j \right) \tag{19}$$

has unconstrained optimum

$$\hat{v}_j = \frac{1}{N} \sum_i e_{ij}^2, \qquad \hat{v}_{j,\Delta} = \frac{1}{N_\Delta} \sum_{i:\Delta_i=\Delta} e_{ij}^2 \tag{20}$$

for global and per-horizon fits. A numerical variance bound clips this optimum. For a fixed heteroscedastic predictor, the analogous NLL-optimal scalar variance multiplier is

$$\hat{T}_{\mathrm{NLL}} = \frac{1}{Nd} \sum_{i,j} \frac{e_{ij}^2}{\sigma_{ij}^2}. \tag{21}$$

This formula applies when NLL is the fitting criterion; it is not a claim that every reported temperature used that criterion. The fitting objective and partition define the interpretation of a calibrated comparison.

## D.3 COVERAGE AND LIKELIHOOD SUMMARIES

Table 8 separates the TUEG rescaling evaluation from same-set external scale diagnostics.

Table 8: **Retained scale diagnostics.** Coverage values use the original aggregate coverage statistic. They are not relabeled as a particular nominal or simultaneous coverage guarantee. External rows are same-set post-hoc diagnostics.

| Dataset | Raw | Rescaled | $T^*$ | Scale-fitting relation |
|---|---|---|---|---|
| TUEG | .939 | .960–.968 | 1.23–1.42 | Outside test |
| TUSZ | .902 | .960 | 1.88 | Same evaluation set |
| TUEP | .911 | .957 | 1.66 | Same evaluation set |
| TUEV | .928 | .955 | 1.28 | Same evaluation set |

The NLL point summaries favor the optimized homoscedastic comparator on TUEG (2.007 versus 2.056–2.086) and TUSZ (2.149 versus 2.282). These are aggregate fit comparisons on the study's reporting scale. Comparisons require common target transforms, normalization, and additive-constant conventions. They are not interpreted as evidence that one score has superior sample-dependent semantics. The nominal coverage level, precise reduction, temperature objective, and baseline fitting details are among the implementation metadata distinguished in Appendix M; no unreported calibration guarantee is inferred.

## D.4 COUNT AND EXACT-MASK CONTROLS

A comparator depending only on keep level and horizon has reported error association $-0.02$ in its aggregate evaluation. Within one fixed $(k, \Delta)$ cell such a comparator is constant, so this number

cannot be a meaningful within-cell Spearman coefficient. It is used only as the original aggregate count/horizon comparison.

The fixed-mask diagnostic reports $\mathrm{SD}(U) \approx 4.1$ and joint-score/future-error correlation $0.591$. The three exported summary intervals are positive; their score/error pairings and 400-replicate intervals appear in Table 10. These rows combine seeds and keep levels rather than providing separate seed estimates. Nonconstant scores at fixed mask reject the narrow explanation that the evaluated outputs are a deterministic function of that mask alone; positive error association additionally characterizes predictive relevance in the diagnostic. Neither result demonstrates that every mask was tested or that an independently fitted mask-only model is inferior.

A separate mask-only residual-risk comparator would require a fitting set disjoint from the final evaluation subjects, sufficient repeated mask support, and explicit handling of unseen masks. Weak performance of an inadequately fitted comparator would not exclude all mask dependence. No such comparator result is inferred from the exact-mask diagnostic.

# E Response-Validity Results and Scope

Table 9 collects the response evidence supporting Section 4.

Table 9: **Response evidence for Section 4.** Ranges describe the evaluated conditions and are not confidence intervals.

| Quantity | Summary |
|---|---|
| Mean prediction | Heteroscedastic error 55.9; deterministic 53.5; approximately 4.5% increase |
| Fixed-cardinality error association | 0.546–0.637 |
| Fixed-exact-mask diagnostic | Uncertainty SD approximately 4.1; error association approximately $+0.59$ |
| Count/horizon-only comparison | Aggregate association $-0.02$ |
| Degradation | $U(4\mathrm{ch}) - U(19\mathrm{ch}) = +0.83$ to $+1.14$ |
| Common-reference discrepancy | Reported normalized value approximately $10^{-9}$ |
| External error association | TUSZ .686; TUEP .518; TUEV .506 |

The exact-mask diagnostic strengthens the response interpretation within its evaluated setting. Its score standard deviation is scale-dependent and is not an uncertainty interval on the correlation. The fixed-cardinality range describes the original comparisons, while the exact-mask result is a supplementary diagnostic; they are not assumed to share identical sample supports.

The degradation differences characterize aggregate response to substantial observation loss. No strict within-history monotonicity or ordering of arbitrary masks follows. Reference sensitivity is restricted to the evaluated common-reference operation. Cross-dataset ranking is distinct from absolute uncertainty scale: external temperature values do not establish physiological population differences or held-out calibration transfer.

All retained numerical summaries are reported at the precision available in the evaluation summaries. The complete keep-by-horizon response arrays are not included in the accompanying summaries; reported ranges therefore do not specify individual condition values.

## E.1 Fixed-mask diagnostic: exported support and intervals

The diagnostic export specifies a 5 s horizon, 93 subjects, 256 histories, seeds 0/1/2, and keep levels 4/8/12. It identifies fixed bank masks by reference to the mask-bank manifest rather than listing their bits. Table 10 preserves the three exported summary rows and their target definitions.

The diagnostic specifies the confidence level, percentile rule, and sample support, but not the exact mask bits or the reduction across mask strata, keep levels, and seeds. The table retains the reported intervals. In particular, their nominal level does not by itself validate a pooled or seed-averaged estimand. The diagnostic supports the evaluated within-mask response claim; it does not establish

Table 10: **Fixed-mask diagnostic summaries.** Reported correlations and 95% percentile bootstrap intervals (400 subject-cluster replicates; 2.5th/97.5th percentile endpoints). Each row summarizes seeds and keep levels together; it is not a single-seed or single-mask result.

| Score | Error target | Reported $\rho$ | Reported 95% interval |
|---|---|---|---|
| Joint | Future | .591 | $[.516, .657]$ |
| Post-hoc | Future | .581 | $[.508, .645]$ |
| Belief | Teacher state | .252 | $[.130, .375]$ |

positive association for every exact mask or superiority to an independently trained mask-only predictor. Its 400 replicates remain distinct from the rank audit's 2,000-replicate bootstrap.

# F LPSD Construction and Sensitivity

## F.1 Definition and retrieval unit

For each query, retrieve $K$ references in the specified partial-view representation at matched cardinality and horizon, excluding query-subject overlap. A frozen full-view teacher supplies states $s_j$. With $\bar{s}_i = K^{-1} \sum_{j \in \mathcal{N}_K(V_i)} s_j$, the conventional unbiased sample-covariance form is

$$D_{\mathrm{LPSD}}(V_i) = \frac{1}{K-1} \sum_{j \in \mathcal{N}_K(V_i)} \|s_j - \bar{s}_i\|^2. \tag{22}$$

A $1/K$ convention changes this by a constant at fixed $K$, leaving ranks unchanged. The covariance convention nevertheless matters for absolute values. The primary $K$ is 20; $\epsilon = 10^{-6}$ is used with the reported log transform. A strictly increasing $\log(x + \epsilon)$ leaves Spearman ranks unchanged in exact arithmetic.

The retrieval specification includes the representation transform, metric, reference sampling, and whether exact masks are also matched. Matching cardinality alone is not equivalent to matching masks. No uncertainty output enters the neighborhood or teacher-state construction. The full implementation specification, rather than the marginal correlation, defines what local similarity means.

## F.2 Seed-level structural association

The primary export uses TUEG, 93 query subjects, 1,024 queries, $K = 20$, and frozen encoder-latent neighborhoods. Query subjects are held out and reference subjects come from the training pool. Table 11 records the four-decimal coefficients, including all 18 nonconstant-score results. Six homoscedastic-control entries are NaN because the score is constant; they are not zero correlations.

Table 11: **Structural association by seed.** Exported coefficients and recomputed mean $\pm$ sample SD ($n - 1$ denominator); four-decimal input precision is retained.

| Estimator | $k$ | Seed 0 | Seed 1 | Seed 2 | Mean $\pm$ sample SD |
|---|---|---|---|---|---|
| Joint | 8 | 0.5848 | 0.3619 | 0.4313 | $0.459 \pm 0.114$ |
| Joint | 12 | 0.6002 | 0.3429 | 0.4704 | $0.471 \pm 0.129$ |
| Post-hoc | 8 | 0.5872 | 0.3635 | 0.3967 | $0.449 \pm 0.121$ |
| Post-hoc | 12 | 0.6103 | 0.3541 | 0.4434 | $0.469 \pm 0.130$ |
| Belief | 8 | 0.6165 | 0.4691 | 0.5155 | $0.534 \pm 0.075$ |
| Belief | 12 | 0.6181 | 0.4475 | 0.4753 | $0.514 \pm 0.092$ |

## F.3 Reproducible computation

1. Fix query/reference persons, observation condition, teacher, representation, and distance rule.

2. Encode partial views without consulting query future targets or uncertainty scores for neighbor selection.
3. Retrieve eligible reference indices, maintaining the declared subject exclusion and mask-matching policy.
4. Compute dispersion of the corresponding full-view teacher states.
5. Pair query-level dispersion with uncertainty, compute the declared rank statistic, and preserve subject identity for inference.

Finite neighborhoods approximate conditioning rather than implement exact conditional distributions. Teacher error, reference density, embedding geometry, bank composition, and neighborhood size can all affect the probe. Independence from the uncertainty head addresses direct reuse of that head; it does not remove model dependence.

### F.4 Geometry, neighborhood scale, and alternative spread

Table 12 summarizes the representation, neighborhood, and statistic dependence of the structural claim in Section 6.

Table 12: **Structural sensitivity summaries.** The regional-STFT and statistic-sensitive exceptions are retained.

| Control | Result or tested setting |
|---|---|
| Encoder-latent space | Correlations approximately $+0.30$ to $+0.61$ |
| Handcrafted bandpower/temporal | Approximately $+0.21$ to $+0.49$ |
| Regional STFT | Unstable, including a sign reversal |
| Neighbor count | $K \in \{10, 20, 50\}$ in the audited sweep; primary 20 |
| Median pairwise statistic | 13/16 contrasts preserve sign; TUSZ $k = 8$ is a boundary |

The alternative dispersion is

$$D_{\text{pair}}(V_i) = \text{median}_{p<q;\ p,q\in\mathcal{N}_K(V_i)} \|s_p - s_q\|^2. \tag{23}$$

It is not an algebraic copy of covariance trace. By contrast, the *mean* pairwise squared distance equals twice the unbiased covariance trace because $\sum_{p<q} \|s_p - s_q\|^2 = K \sum_p \|s_p - \bar{s}\|^2$. Median-based differences can therefore reveal sensitivity to a different aspect of spread. The results establish association across several specified constructions, not universally representation-invariant ambiguity measurement. The independent-neighborhood sweep uses $K = 10, 20, 50$. The aggregate ranges do not specify separate geometry-by-$K$ effects.

## G Shuffled Structural Reference

### G.1 Construction

Queries, their uncertainty outputs, reference representations, and neighbor lists remain fixed. A permutation $\pi_b$ reassigns reference teacher states:

$$D_i^{(b)} = \text{Tr}\,\widehat{\text{Cov}}\{s_{\pi_b(j)} : j \in \mathcal{N}_K(V_i)\}, \qquad \rho^{(b)} = \text{Spearman}_i(U_i, D_i^{(b)}). \tag{24}$$

This disrupts the specific reference-observation/state pairing while retaining retrieval geometry. The condition summary covers 16 contrasts at model seed 0, each using 20 permutations. The negative control addresses this assignment, not all possible geometry or model-dependence explanations.

### G.2 Finite-permutation scope

Table 13 reports every observed statistic and its corresponding shuffled-reference threshold. Fifteen of 16 observed statistics exceed the corresponding empirical 95th percentile. The observed pattern

mean is 0.450 and the mean null threshold 0.125. TUSZ at $k = 8$ in PCA-64 space is the exception: observed 0.126 versus null threshold 0.148. Means of thresholds do not form a pooled null distribution.

Under a valid exchangeable permutation scheme, a conventional one-sided Monte Carlo calculation is

$$p_{\mathrm{MC}} = \frac{1 + \sum_{b=1}^{B} \mathbf{1}\{\rho^{(b)} \geq \rho_{\mathrm{obs}}\}}{B + 1}. \tag{25}$$

For $B = 20$, its minimum is $1/21 \approx 0.0476$. We therefore report an empirical-threshold pattern, not 15 multiplicity-adjusted discoveries. Formal inference additionally depends on permutation units and restrictions compatible with dependent EEG windows; increasing $B$ alone would not fix an invalid exchangeability assumption.

Table 13: **All shuffled-reference contrasts.** Model seed 0 and 20 permutations per row. The threshold is the empirical null 95th percentile, not a multiplicity-adjusted test.

| Dataset | $k$ | Representation | Observed | Null $q_{.95}$ | Exceeds |
|---|---|---|---|---|---|
| TUEG | 12 | PCA64 | 0.3087 | 0.0529 | Yes |
| TUEG | 12 | PRED64 | 0.5524 | 0.0873 | Yes |
| TUEG | 12 | RP64 | 0.5670 | 0.0691 | Yes |
| TUEG | 12 | VAR64 | 0.5930 | 0.0709 | Yes |
| TUEG | 8 | PCA64 | 0.2374 | 0.1161 | Yes |
| TUEG | 8 | PRED64 | 0.5919 | 0.0704 | Yes |
| TUEG | 8 | RP64 | 0.5918 | 0.1007 | Yes |
| TUEG | 8 | VAR64 | 0.6391 | 0.0810 | Yes |
| TUSZ | 12 | PCA64 | 0.2660 | 0.1574 | Yes |
| TUSZ | 12 | PRED64 | 0.6608 | 0.1572 | Yes |
| TUSZ | 12 | RP64 | 0.5555 | 0.1904 | Yes |
| TUSZ | 12 | VAR64 | 0.5609 | 0.1949 | Yes |
| TUSZ | 8 | PCA64 | 0.1258 | 0.1478 | No |
| TUSZ | 8 | PRED64 | 0.2665 | 0.1358 | Yes |
| TUSZ | 8 | RP64 | 0.3454 | 0.1900 | Yes |
| TUSZ | 8 | VAR64 | 0.3388 | 0.1747 | Yes |

# H Controlled Channel-Mask Intervention

## H.1 Audited pool and effective cardinality

The intervention pool consists of 256 histories from 93 held-out subjects, each with all 19 canonical channels available. If $M_{\mathrm{native}}$ denotes native availability, the effective mask is $M_{\mathrm{effective}} = M_{\mathrm{planned}} \odot M_{\mathrm{native}}$. Here $M_{\mathrm{native}} = \mathbf{1}$ for every history, and the reconciliation reports equality of planned and effective keep levels. Thus a planned eight-channel mask is not silently reduced by native missingness in this pool.

Each comparison fixes history, future descriptor, horizon, and keep level. Observed-channel reference canonicalization is recomputed using the chosen mask. A distinct sweep uses 25 masks per history; population analyses use three banks with 30 masks per keep level. Masks select measurements of the same stored signal rather than generate new EEG. Three trained seeds and three estimator constructions are evaluated.

## H.2 Reference consequence operators

Define $s_H = \mu_T(H)$ for the frozen full-view teacher. The belief mean $m_B(H, M)$ predicts that state from a partial view. The designated future mean $\mu_0(H, M)$ predicts the real descriptor $Y$. Per-history consequences are

$$D^{\mathrm{state}}(H, M) = \|m_B(H, M) - s_H\|_2^2, \tag{26}$$

$$D^{\mathrm{future}}(H, M, Y) = \|Y - \mu_0(H, M)\|_2^2. \tag{27}$$

The corresponding population operators are their history expectations. These are summed squared distances in the 64-dimensional target spaces, while $U$ is the mean of predicted variances over those dimensions. Multiplying a consequence by a fixed positive $1/64$ would not change its rank coefficient. Nonlinear clipping, an absolute-value transformation, or changing the comparator would generally change the estimand and is not part of the equations above.

The teacher-state error measures how accurately the partial-view belief mean reconstructs the full-view teacher state. Future-space error measures realized prediction error relative to the true descriptor. Neither subtracts a full-view baseline, and neither is simply the displacement between partial and full predictions. The same consequence operator is compared with each score construction. Because these are reference-model consequences, a negative result does not establish that each score fails to rank the error of its own mean model.

### H.3 Within-history statistic

For fixed history $i$, horizon, keep level, estimator, and seed,

$$\rho_i^a = \mathrm{Spearman}_r(U_{i,r}, D_{i,r}^a). \tag{28}$$

The reported within-history results are centered near zero. Only a qualitative center summary is available for this sweep; it does not support distributional or equivalence claims. A constant score or consequence makes Spearman undefined; this requires an explicit exclusion/flag policy rather than zero imputation.

### H.4 Population statistic and run counts

For the common $N = 256$ histories, use window means

$$\bar{U}_r = N^{-1} \sum_i U_{i,r}, \qquad \bar{D}_r^a = N^{-1} \sum_i D_{i,r}^a. \tag{29}$$

These means define $\rho_{\mathrm{pop}}^a = \mathrm{Spearman}_r(\bar{U}_r, \bar{D}_r^a)$ across masks. Averaging before ranking differs from averaging within-history coefficients, and neither is a pooled history–mask correlation. The resulting estimand is window-weighted over the fixed pool. Subject-cluster resampling accounts for within-person dependence but does not automatically change it into a subject-balanced estimand.

The complete factorial design has 3 estimators $\times 3$ keep levels $\times 2$ consequence spaces $\times 3$ seeds $\times 3$ banks, or 162 run-level cells. Each table condition summarizes nine seed-by-bank cells; those cells reuse subjects and are not nine independent biological replicates. Table 14 reports the means and run SDs over these nine cells. The supplementary record identifies the SD as a sample SD; without the individual nine values, this normalization cannot be independently checked. No SD is converted into a subject-sampling confidence interval.

Table 14: **Population mask-ranking correlations.** Reported mean $\pm$ descriptive sample SD over nine seed-by-bank cells (three seeds, three banks), not nine independent biological replicates. Consequences are Equations (2)–(3). Run dispersion is not a confidence interval or a test of equivalence to zero.

| Consequence | Score | $k = 4$ | $k = 8$ | $k = 12$ |
|---|---|---|---|---|
| Teacher-state error | Joint | $+.02 \pm .19$ | $-.11 \pm .20$ | $-.03 \pm .18$ |
| | Post-hoc | $+.07 \pm .18$ | $-.02 \pm .14$ | $+.10 \pm .26$ |
| | Belief | $-.18 \pm .12$ | $-.04 \pm .23$ | $-.23 \pm .28$ |
| Future-prediction error | Joint | $+.05 \pm .22$ | $-.05 \pm .23$ | $+.02 \pm .29$ |
| | Post-hoc | $+.06 \pm .17$ | $+.01 \pm .09$ | $+.07 \pm .28$ |
| | Belief | $-.07 \pm .09$ | $+.08 \pm .20$ | $-.11 \pm .28$ |

### H.5 Why pooled association is insufficient

For fixed keep and horizon, the ordinary covariance decomposition over histories is

$$\mathrm{Cov}(U, D) = \mathbb{E}_H[\mathrm{Cov}(U, D \mid H)] + \mathrm{Cov}_H(\mathbb{E}[U \mid H], \mathbb{E}[D \mid H]). \tag{30}$$

A positive between-history term can coexist with a weak within-history term. This explains why error association pooled across histories cannot replace a mask-ranking test. The identity concerns ordinary covariance, not an additive decomposition of Spearman coefficients, and does not numerically decompose the LPSD correlation. The structural and intervention tests target different relationships; their point coefficients should not be subtracted as if they estimate a common parameter.

### H.6 Repeated subject-disjoint split-half audit

For each of 2,000 partitions $t$, the 93 held-out subjects are split into disjoint groups $A_t$ and $B_t$ of sizes 46 and 47. Each person's histories stay together. At fixed seed, bank, and keep level, both halves evaluate the same 30 masks. Banks have identifiers 7000, 8000, and 9000 and are analyzed separately; their masks are never concatenated into a 90-mask ranking.

The audit uses *subject-balanced* half means for both uncertainty and consequence. For $Z \in \{U^e, D^a\}$ and subject $s$ with $n_s$ histories,

$$\bar{Z}_r^{A_t} = \frac{1}{|A_t|} \sum_{s \in A_t} \left( \frac{1}{n_s} \sum_{i:\,\mathrm{subject}(i)=s} Z_{i,r} \right), \tag{31}$$

and analogously for $B_t$. This gives equal weight to subjects, unlike the window-weighted full-pool statistic above. The distinction is part of the estimand, not merely an uncertainty-estimation choice.

At each split, $R_D$ and $R_U$ correlate the corresponding mask means across halves (Equation (4)). The frozen models, target, horizon, and bank are shared. Thus the audit estimates reproducibility under subject partitioning, not across independently retrained models or independent data-collection institutions. Repeated partitions reuse the same 93 subjects.

### H.7 Cross-half and same-half definitions

For estimator $e$ and consequence $a$, the two directional cross coefficients are

$$r^{e,a}_{A_t \to B_t} = \mathrm{Spearman}_r(\bar{U}_r^{e,A_t}, \bar{D}_r^{a,B_t}), \tag{32}$$

$$r^{e,a}_{B_t \to A_t} = \mathrm{Spearman}_r(\bar{U}_r^{e,B_t}, \bar{D}_r^{a,A_t}). \tag{33}$$

The audit averages directions *inside each split*, before taking any split quantile:

$$R^{e,a,\mathrm{cross}}_{UD,c,t} = \tfrac{1}{2}(r^{e,a}_{A_t \to B_t} + r^{e,a}_{B_t \to A_t}), \tag{34}$$

$$R^{e,a,\mathrm{same}}_{UD,c,t} = \tfrac{1}{2}\left[\mathrm{Spearman}_r(\bar{U}_r^{e,A_t}, \bar{D}_r^{a,A_t}) + \mathrm{Spearman}_r(\bar{U}_r^{e,B_t}, \bar{D}_r^{a,B_t})\right]. \tag{35}$$

Cross-half estimation separates subjects contributing to uncertainty and consequence. It does not separate mask identity, model construction, or the source ecosystem. The same-half comparison is a diagnostic, not a test of equality between the two alignment statistics.

### H.8 Aggregation of split-half reliability summaries

Let $c = (s, b, k)$ index seed $s \in \{0, 1, 2\}$, bank $b \in \{7000, 8000, 9000\}$, and keep $k \in \{4, 8, 12\}$. For each fixed quantity, estimator, and consequence as applicable, the complete design contains 27 conditions. The aggregation specification gives

$$\widetilde{R}_c = \mathrm{median}_{t \in \mathcal{T}_c} R_{c,t}, \qquad R_{\mathrm{headline}} = \frac{1}{27} \sum_{c=1}^{27} \widetilde{R}_c, \tag{36}$$

where $\mathcal{T}_c$ contains valid splits. Conditions are equally weighted, so seeds, banks, and keep levels receive equal total weight. Estimators and consequences are never pooled into one headline. The cross-direction average precedes the median and generally cannot be interchanged with it.

Ties use average ranks (midranks). A constant-vector Spearman coefficient is NaN, not zero, and NaN split coefficients are excluded from quantile calculation. Every one of the 162 exported estimator–consequence–condition rows records 2,000 valid and zero invalid splits. All expected rows are present and unique; no condition is excluded. The consequence fields agree exactly across their estimator

copies, and uncertainty fields agree across consequence copies. Deduplication therefore yields 54 unique consequence conditions and 81 unique uncertainty conditions, rather than treating repeated fields as extra replication.

The paper recomputes headlines from the exported three-decimal condition medians. All eleven recomputed headlines differ from the separately reported headline values by at most $0.000519$; the largest difference is post-hoc $R_U$, displayed here as $0.811$ rather than $0.810$. Full-precision split medians are not available to resolve that final-digit difference. For aggregate graphics, 5th–95th percentiles across the 27 condition medians use linear interpolation. These *between-condition* quantiles differ from each condition's exported *split-distribution* quantiles. Neither is a confidence interval for a population parameter.

### H.9 Reproducibility, alignment, and condition heterogeneity

Table 15 compares cross-half and same-half aggregates and quantifies positive condition-level alignment. Consequence headline reproducibility is 0.679 in teacher-state space and 0.466 in future space. All 27 condition medians are positive for each target: ranges are $[0.384, 0.822]$ and $[0.307, 0.665]$, respectively. The within-condition split 5th percentile is positive in 25/27 teacher-state conditions and 9/27 future conditions. These counts describe partition distributions rather than significant discoveries, and show the weaker resolution of future-error rankings.

All 27 uncertainty medians and all 27 split 5th percentiles are positive for each estimator. Headline reproducibility is 0.836 joint, 0.811 post-hoc, and 0.957 belief. Thus both uncertainty and consequence exhibit reproducible population mask structure. However, cross-half condition medians have both signs for every estimator–target pairing. The weak or negative aggregate alignment is not a claim that every condition fails or has exactly zero correlation.

Table 15: **Recomputed alignment headlines.** Equal-weight means of the 27 condition medians. The last columns count positive cross-half medians and positive within-condition split 5th percentiles; these are descriptive, not multiplicity-adjusted tests.

| Target | Score | Cross-half | Same-half | Median $> 0$ | $q_{.05} > 0$ |
|---|---|---|---|---|---|
| State | Joint | $-0.078$ | $-0.124$ | 6/27 | 1/27 |
| State | Post-hoc | $-0.006$ | $-0.064$ | 14/27 | 3/27 |
| State | Belief | $-0.196$ | $-0.226$ | 5/27 | 1/27 |
| Future | Joint | $+0.025$ | $+0.030$ | 17/27 | 4/27 |
| Future | Post-hoc | $+0.049$ | $+0.049$ | 19/27 | 1/27 |
| Future | Belief | $-0.076$ | $-0.089$ | 5/27 | 0/27 |

Teacher-state error receives primary interpretive emphasis because its ordering is more reproducible in this audit; future error remains complementary. This is audit-informed emphasis, not a preregistered outcome designation. Positive rank reproducibility weakens an explanation based solely on absent signal, but does not eliminate attenuation or validate an inverse-score acquisition policy from negative coefficients. No correction dividing by $\sqrt{R_U R_D}$ is used.

### H.10 Complete condition-level results

Tables 16–18 report all unique conditions. The source exports split medians and 5th/95th percentiles; within-condition split minima and maxima were not included and are not inferred from these quantiles. Same-half summaries are retained alongside cross-half results. Bank identifiers specify independent 30-mask banks, not 90 jointly ranked masks.

Table 16: **Consequence reliability by condition.** All 54 unique target–seed–bank–keep rows. Each has 2,000 valid splits and zero invalid splits. Quantiles describe the split distribution; they are not confidence intervals.

| Target | Seed | Bank | $k$ | Median $R_D$ | $q_{.05}$ | $q_{.95}$ |
|---|---|---|---|---|---|---|
| State | 0 | 7000 | 4 | $+0.652$ | $+0.274$ | $+0.938$ |
| State | 0 | 7000 | 8 | $+0.724$ | $+0.454$ | $+0.915$ |
| State | 0 | 7000 | 12 | $+0.422$ | $-0.103$ | $+0.925$ |
| State | 0 | 8000 | 4 | $+0.669$ | $+0.403$ | $+0.921$ |
| State | 0 | 8000 | 8 | $+0.558$ | $+0.060$ | $+0.946$ |
| State | 0 | 8000 | 12 | $+0.582$ | $+0.169$ | $+0.927$ |
| State | 0 | 9000 | 4 | $+0.705$ | $+0.213$ | $+0.969$ |
| State | 0 | 9000 | 8 | $+0.686$ | $+0.421$ | $+0.899$ |
| State | 0 | 9000 | 12 | $+0.519$ | $+0.157$ | $+0.949$ |
| State | 1 | 7000 | 4 | $+0.745$ | $+0.460$ | $+0.950$ |
| State | 1 | 7000 | 8 | $+0.777$ | $+0.351$ | $+0.959$ |
| State | 1 | 7000 | 12 | $+0.667$ | $+0.019$ | $+0.945$ |
| State | 1 | 8000 | 4 | $+0.702$ | $+0.320$ | $+0.935$ |
| State | 1 | 8000 | 8 | $+0.791$ | $+0.313$ | $+0.962$ |
| State | 1 | 8000 | 12 | $+0.793$ | $+0.340$ | $+0.960$ |
| State | 1 | 9000 | 4 | $+0.822$ | $+0.554$ | $+0.960$ |
| State | 1 | 9000 | 8 | $+0.737$ | $+0.250$ | $+0.944$ |
| State | 1 | 9000 | 12 | $+0.674$ | $+0.282$ | $+0.926$ |
| State | 2 | 7000 | 4 | $+0.783$ | $+0.605$ | $+0.963$ |
| State | 2 | 7000 | 8 | $+0.683$ | $+0.398$ | $+0.956$ |
| State | 2 | 7000 | 12 | $+0.384$ | $-0.253$ | $+0.948$ |
| State | 2 | 8000 | 4 | $+0.787$ | $+0.595$ | $+0.967$ |
| State | 2 | 8000 | 8 | $+0.608$ | $+0.250$ | $+0.963$ |
| State | 2 | 8000 | 12 | $+0.703$ | $+0.359$ | $+0.968$ |
| State | 2 | 9000 | 4 | $+0.762$ | $+0.409$ | $+0.970$ |
| State | 2 | 9000 | 8 | $+0.789$ | $+0.535$ | $+0.958$ |
| State | 2 | 9000 | 12 | $+0.612$ | $+0.255$ | $+0.931$ |
| Future | 0 | 7000 | 4 | $+0.458$ | $-0.031$ | $+0.980$ |
| Future | 0 | 7000 | 8 | $+0.503$ | $+0.030$ | $+0.986$ |
| Future | 0 | 7000 | 12 | $+0.467$ | $-0.014$ | $+0.858$ |
| Future | 0 | 8000 | 4 | $+0.470$ | $-0.013$ | $+0.978$ |
| Future | 0 | 8000 | 8 | $+0.553$ | $-0.059$ | $+0.960$ |
| Future | 0 | 8000 | 12 | $+0.525$ | $-0.105$ | $+0.951$ |
| Future | 0 | 9000 | 4 | $+0.665$ | $-0.025$ | $+0.989$ |
| Future | 0 | 9000 | 8 | $+0.338$ | $-0.248$ | $+0.970$ |
| Future | 0 | 9000 | 12 | $+0.340$ | $-0.190$ | $+0.915$ |
| Future | 1 | 7000 | 4 | $+0.539$ | $+0.143$ | $+0.870$ |
| Future | 1 | 7000 | 8 | $+0.429$ | $-0.029$ | $+0.899$ |
| Future | 1 | 7000 | 12 | $+0.307$ | $-0.288$ | $+0.846$ |
| Future | 1 | 8000 | 4 | $+0.467$ | $+0.052$ | $+0.902$ |
| Future | 1 | 8000 | 8 | $+0.455$ | $-0.158$ | $+0.895$ |
| Future | 1 | 8000 | 12 | $+0.385$ | $-0.091$ | $+0.874$ |
| Future | 1 | 9000 | 4 | $+0.606$ | $+0.143$ | $+0.937$ |
| Future | 1 | 9000 | 8 | $+0.340$ | $-0.191$ | $+0.880$ |
| Future | 1 | 9000 | 12 | $+0.473$ | $+0.005$ | $+0.837$ |
| Future | 2 | 7000 | 4 | $+0.534$ | $-0.011$ | $+0.981$ |
| Future | 2 | 7000 | 8 | $+0.432$ | $-0.068$ | $+0.979$ |
| Future | 2 | 7000 | 12 | $+0.415$ | $-0.450$ | $+0.983$ |
| Future | 2 | 8000 | 4 | $+0.587$ | $+0.062$ | $+0.984$ |
| Future | 2 | 8000 | 8 | $+0.395$ | $-0.116$ | $+0.985$ |
| Future | 2 | 8000 | 12 | $+0.423$ | $-0.105$ | $+0.980$ |
| Future | 2 | 9000 | 4 | $+0.595$ | $+0.354$ | $+0.984$ |
| Future | 2 | 9000 | 8 | $+0.411$ | $+0.054$ | $+0.979$ |
| Future | 2 | 9000 | 12 | $+0.479$ | $+0.055$ | $+0.993$ |

Table 17: **Uncertainty reliability by condition.** All 81 unique estimator–seed–bank–keep rows, with 2,000 valid splits and zero invalid splits each. Duplicate entries across consequence targets are counted once. Quantiles are split-distribution quantiles.

| Score | Seed | Bank | $k$ | Median $R_U$ | $q_{.05}$ | $q_{.95}$ |
|---|---|---|---|---|---|---|
| Joint | 0 | 7000 | 4 | +0.842 | +0.677 | +0.929 |
| Joint | 0 | 7000 | 8 | +0.820 | +0.654 | +0.917 |
| Joint | 0 | 7000 | 12 | +0.861 | +0.728 | +0.938 |
| Joint | 0 | 8000 | 4 | +0.918 | +0.827 | +0.964 |
| Joint | 0 | 8000 | 8 | +0.879 | +0.755 | +0.947 |
| Joint | 0 | 8000 | 12 | +0.903 | +0.803 | +0.956 |
| Joint | 0 | 9000 | 4 | +0.921 | +0.839 | +0.965 |
| Joint | 0 | 9000 | 8 | +0.816 | +0.635 | +0.919 |
| Joint | 0 | 9000 | 12 | +0.894 | +0.766 | +0.956 |
| Joint | 1 | 7000 | 4 | +0.848 | +0.715 | +0.935 |
| Joint | 1 | 7000 | 8 | +0.753 | +0.527 | +0.911 |
| Joint | 1 | 7000 | 12 | +0.579 | +0.104 | +0.887 |
| Joint | 1 | 8000 | 4 | +0.838 | +0.665 | +0.927 |
| Joint | 1 | 8000 | 8 | +0.762 | +0.522 | +0.916 |
| Joint | 1 | 8000 | 12 | +0.697 | +0.402 | +0.918 |
| Joint | 1 | 9000 | 4 | +0.753 | +0.526 | +0.893 |
| Joint | 1 | 9000 | 8 | +0.718 | +0.408 | +0.912 |
| Joint | 1 | 9000 | 12 | +0.638 | +0.273 | +0.875 |
| Joint | 2 | 7000 | 4 | +0.906 | +0.806 | +0.959 |
| Joint | 2 | 7000 | 8 | +0.915 | +0.817 | +0.965 |
| Joint | 2 | 7000 | 12 | +0.859 | +0.697 | +0.942 |
| Joint | 2 | 8000 | 4 | +0.941 | +0.881 | +0.976 |
| Joint | 2 | 8000 | 8 | +0.898 | +0.793 | +0.960 |
| Joint | 2 | 8000 | 12 | +0.876 | +0.758 | +0.953 |
| Joint | 2 | 9000 | 4 | +0.922 | +0.852 | +0.966 |
| Joint | 2 | 9000 | 8 | +0.930 | +0.850 | +0.971 |
| Joint | 2 | 9000 | 12 | +0.888 | +0.767 | +0.954 |
| Post-hoc | 0 | 7000 | 4 | +0.716 | +0.451 | +0.881 |
| Post-hoc | 0 | 7000 | 8 | +0.820 | +0.632 | +0.920 |
| Post-hoc | 0 | 7000 | 12 | +0.822 | +0.636 | +0.923 |
| Post-hoc | 0 | 8000 | 4 | +0.864 | +0.713 | +0.941 |
| Post-hoc | 0 | 8000 | 8 | +0.819 | +0.642 | +0.924 |
| Post-hoc | 0 | 8000 | 12 | +0.819 | +0.642 | +0.919 |
| Post-hoc | 0 | 9000 | 4 | +0.825 | +0.650 | +0.918 |
| Post-hoc | 0 | 9000 | 8 | +0.819 | +0.616 | +0.920 |
| Post-hoc | 0 | 9000 | 12 | +0.831 | +0.652 | +0.923 |
| Post-hoc | 1 | 7000 | 4 | +0.755 | +0.473 | +0.897 |
| Post-hoc | 1 | 7000 | 8 | +0.786 | +0.568 | +0.921 |
| Post-hoc | 1 | 7000 | 12 | +0.771 | +0.386 | +0.947 |
| Post-hoc | 1 | 8000 | 4 | +0.833 | +0.658 | +0.926 |
| Post-hoc | 1 | 8000 | 8 | +0.747 | +0.480 | +0.912 |
| Post-hoc | 1 | 8000 | 12 | +0.647 | +0.286 | +0.936 |
| Post-hoc | 1 | 9000 | 4 | +0.799 | +0.605 | +0.911 |
| Post-hoc | 1 | 9000 | 8 | +0.713 | +0.372 | +0.901 |
| Post-hoc | 1 | 9000 | 12 | +0.735 | +0.450 | +0.935 |
| Post-hoc | 2 | 7000 | 4 | +0.826 | +0.652 | +0.925 |
| Post-hoc | 2 | 7000 | 8 | +0.849 | +0.700 | +0.934 |
| Post-hoc | 2 | 7000 | 12 | +0.827 | +0.640 | +0.931 |
| Post-hoc | 2 | 8000 | 4 | +0.913 | +0.832 | +0.961 |
| Post-hoc | 2 | 8000 | 8 | +0.804 | +0.629 | +0.915 |
| Post-hoc | 2 | 8000 | 12 | +0.887 | +0.777 | +0.950 |
| Post-hoc | 2 | 9000 | 4 | +0.926 | +0.845 | +0.966 |
| Post-hoc | 2 | 9000 | 8 | +0.877 | +0.760 | +0.944 |
| Post-hoc | 2 | 9000 | 12 | +0.854 | +0.677 | +0.940 |
| Belief | 0 | 7000 | 4 | +0.900 | +0.788 | +0.962 |
| Belief | 0 | 7000 | 8 | +0.945 | +0.883 | +0.978 |
| Belief | 0 | 7000 | 12 | +0.961 | +0.925 | +0.984 |

Continued on next page

Table 17 continued

| Score | Seed | Bank | $k$ | Median $R_U$ | $q_{.05}$ | $q_{.95}$ |
|---|---|---|---|---|---|---|
| Belief | 0 | 8000 | 4 | +0.930 | +0.847 | +0.973 |
| Belief | 0 | 8000 | 8 | +0.935 | +0.862 | +0.972 |
| Belief | 0 | 8000 | 12 | +0.954 | +0.902 | +0.982 |
| Belief | 0 | 9000 | 4 | +0.918 | +0.822 | +0.967 |
| Belief | 0 | 9000 | 8 | +0.951 | +0.891 | +0.980 |
| Belief | 0 | 9000 | 12 | +0.931 | +0.861 | +0.975 |
| Belief | 1 | 7000 | 4 | +0.964 | +0.918 | +0.986 |
| Belief | 1 | 7000 | 8 | +0.971 | +0.939 | +0.987 |
| Belief | 1 | 7000 | 12 | +0.956 | +0.891 | +0.984 |
| Belief | 1 | 8000 | 4 | +0.973 | +0.948 | +0.989 |
| Belief | 1 | 8000 | 8 | +0.959 | +0.915 | +0.984 |
| Belief | 1 | 8000 | 12 | +0.960 | +0.900 | +0.985 |
| Belief | 1 | 9000 | 4 | +0.963 | +0.922 | +0.984 |
| Belief | 1 | 9000 | 8 | +0.947 | +0.882 | +0.978 |
| Belief | 1 | 9000 | 12 | +0.955 | +0.912 | +0.981 |
| Belief | 2 | 7000 | 4 | +0.965 | +0.922 | +0.985 |
| Belief | 2 | 7000 | 8 | +0.969 | +0.931 | +0.987 |
| Belief | 2 | 7000 | 12 | +0.983 | +0.963 | +0.992 |
| Belief | 2 | 8000 | 4 | +0.975 | +0.948 | +0.989 |
| Belief | 2 | 8000 | 8 | +0.969 | +0.936 | +0.987 |
| Belief | 2 | 8000 | 12 | +0.977 | +0.946 | +0.991 |
| Belief | 2 | 9000 | 4 | +0.967 | +0.924 | +0.986 |
| Belief | 2 | 9000 | 8 | +0.976 | +0.952 | +0.990 |
| Belief | 2 | 9000 | 12 | +0.974 | +0.948 | +0.989 |

Table 18: **Cross-half and same-half alignment by condition.** All 162 estimator–target–seed–bank–keep rows. Entries are split medians followed by $[q_{.05}, q_{.95}]$ of the split distribution, not confidence intervals. Every row has 2,000 valid splits and zero invalid splits.

| Score | Target | Seed | Bank | $k$ | Cross: median $[q_{.05}, q_{.95}]$ | Same: median $[q_{.05}, q_{.95}]$ |
|---|---|---|---|---|---|---|
| Joint | State | 0 | 7000 | 4 | $-0.009$ $[-0.208, +0.179]$ | $-0.051$ $[-0.233, +0.145]$ |
| Joint | State | 0 | 7000 | 8 | $+0.051$ $[-0.115, +0.226]$ | $+0.051$ $[-0.120, +0.227]$ |
| Joint | State | 0 | 7000 | 12 | $-0.014$ $[-0.181, +0.189]$ | $-0.005$ $[-0.178, +0.203]$ |
| Joint | State | 0 | 8000 | 4 | $+0.025$ $[-0.124, +0.148]$ | $-0.011$ $[-0.155, +0.123]$ |
| Joint | State | 0 | 8000 | 8 | $+0.041$ $[-0.106, +0.186]$ | $+0.023$ $[-0.113, +0.163]$ |
| Joint | State | 0 | 8000 | 12 | $+0.308$ $[+0.161, +0.455]$ | $+0.312$ $[+0.155, +0.454]$ |
| Joint | State | 0 | 9000 | 4 | $+0.186$ $[-0.015, +0.343]$ | $+0.144$ $[-0.038, +0.333]$ |
| Joint | State | 0 | 9000 | 8 | $-0.033$ $[-0.200, +0.150]$ | $-0.030$ $[-0.202, +0.141]$ |
| Joint | State | 0 | 9000 | 12 | $-0.068$ $[-0.237, +0.130]$ | $-0.067$ $[-0.238, +0.133]$ |
| Joint | State | 1 | 7000 | 4 | $-0.114$ $[-0.336, +0.144]$ | $-0.168$ $[-0.370, +0.083]$ |
| Joint | State | 1 | 7000 | 8 | $-0.134$ $[-0.457, +0.105]$ | $-0.210$ $[-0.464, +0.081]$ |
| Joint | State | 1 | 7000 | 12 | $-0.031$ $[-0.538, +0.254]$ | $-0.207$ $[-0.550, +0.197]$ |
| Joint | State | 1 | 8000 | 4 | $-0.255$ $[-0.474, -0.018]$ | $-0.319$ $[-0.504, -0.099]$ |
| Joint | State | 1 | 8000 | 8 | $-0.373$ $[-0.677, -0.092]$ | $-0.445$ $[-0.689, -0.149]$ |
| Joint | State | 1 | 8000 | 12 | $-0.023$ $[-0.467, +0.268]$ | $-0.132$ $[-0.474, +0.249]$ |
| Joint | State | 1 | 9000 | 4 | $-0.015$ $[-0.273, +0.201]$ | $-0.078$ $[-0.309, +0.154]$ |
| Joint | State | 1 | 9000 | 8 | $-0.206$ $[-0.558, +0.103]$ | $-0.288$ $[-0.563, +0.065]$ |
| Joint | State | 1 | 9000 | 12 | $-0.149$ $[-0.481, +0.156]$ | $-0.238$ $[-0.505, +0.089]$ |
| Joint | State | 2 | 7000 | 4 | $-0.189$ $[-0.363, -0.012]$ | $-0.233$ $[-0.382, -0.051]$ |
| Joint | State | 2 | 7000 | 8 | $-0.128$ $[-0.269, +0.024]$ | $-0.159$ $[-0.292, -0.012]$ |
| Joint | State | 2 | 7000 | 12 | $-0.125$ $[-0.353, +0.124]$ | $-0.169$ $[-0.388, +0.073]$ |
| Joint | State | 2 | 8000 | 4 | $-0.108$ $[-0.211, -0.012]$ | $-0.146$ $[-0.237, -0.051]$ |
| Joint | State | 2 | 8000 | 8 | $-0.253$ $[-0.495, -0.057]$ | $-0.321$ $[-0.504, -0.077]$ |
| Joint | State | 2 | 8000 | 12 | $+0.078$ $[-0.060, +0.246]$ | $+0.057$ $[-0.076, +0.210]$ |
| Joint | State | 2 | 9000 | 4 | $-0.031$ $[-0.143, +0.081]$ | $-0.073$ $[-0.165, +0.027]$ |
| Joint | State | 2 | 9000 | 8 | $-0.449$ $[-0.567, -0.271]$ | $-0.466$ $[-0.577, -0.294]$ |
| Joint | State | 2 | 9000 | 12 | $-0.096$ $[-0.311, +0.093]$ | $-0.115$ $[-0.320, +0.081]$ |

Continued on next page

Table 18 continued

| Score | Target | Seed | Bank | $k$ | Cross: median $[q_{.05}, q_{.95}]$ | Same: median $[q_{.05}, q_{.95}]$ |
|---|---|---|---|---|---|---|
| Joint | Future | 0 | 7000 | 4 | $+0.019$ $[-0.212, +0.205]$ | $-0.038$ $[-0.256, +0.174]$ |
| Joint | Future | 0 | 7000 | 8 | $+0.077$ $[-0.123, +0.249]$ | $+0.056$ $[-0.152, +0.241]$ |
| Joint | Future | 0 | 7000 | 12 | $+0.208$ $[+0.002, +0.388]$ | $+0.197$ $[-0.026, +0.374]$ |
| Joint | Future | 0 | 8000 | 4 | $+0.194$ $[-0.016, +0.351]$ | $+0.156$ $[-0.051, +0.337]$ |
| Joint | Future | 0 | 8000 | 8 | $+0.166$ $[+0.008, +0.328]$ | $+0.136$ $[-0.025, +0.278]$ |
| Joint | Future | 0 | 8000 | 12 | $+0.360$ $[+0.193, +0.531]$ | $+0.346$ $[+0.179, +0.520]$ |
| Joint | Future | 0 | 9000 | 4 | $+0.231$ $[+0.013, +0.361]$ | $+0.198$ $[-0.022, +0.354]$ |
| Joint | Future | 0 | 9000 | 8 | $+0.148$ $[-0.079, +0.316]$ | $+0.123$ $[-0.102, +0.307]$ |
| Joint | Future | 0 | 9000 | 12 | $+0.124$ $[-0.091, +0.326]$ | $+0.114$ $[-0.100, +0.309]$ |
| Joint | Future | 1 | 7000 | 4 | $-0.109$ $[-0.297, +0.156]$ | $-0.108$ $[-0.314, +0.108]$ |
| Joint | Future | 1 | 7000 | 8 | $-0.125$ $[-0.331, +0.169]$ | $-0.057$ $[-0.266, +0.195]$ |
| Joint | Future | 1 | 7000 | 12 | $+0.099$ $[-0.273, +0.464]$ | $+0.252$ $[-0.099, +0.536]$ |
| Joint | Future | 1 | 8000 | 4 | $-0.181$ $[-0.376, +0.083]$ | $-0.182$ $[-0.377, +0.049]$ |
| Joint | Future | 1 | 8000 | 8 | $-0.213$ $[-0.436, +0.168]$ | $-0.170$ $[-0.371, +0.193]$ |
| Joint | Future | 1 | 8000 | 12 | $+0.156$ $[-0.117, +0.401]$ | $+0.255$ $[+0.038, +0.479]$ |
| Joint | Future | 1 | 9000 | 4 | $+0.009$ $[-0.194, +0.210]$ | $+0.001$ $[-0.214, +0.178]$ |
| Joint | Future | 1 | 9000 | 8 | $+0.125$ $[-0.164, +0.490]$ | $+0.226$ $[-0.066, +0.537]$ |
| Joint | Future | 1 | 9000 | 12 | $+0.100$ $[-0.170, +0.380]$ | $+0.224$ $[-0.038, +0.467]$ |
| Joint | Future | 2 | 7000 | 4 | $-0.029$ $[-0.169, +0.161]$ | $-0.065$ $[-0.184, +0.132]$ |
| Joint | Future | 2 | 7000 | 8 | $+0.020$ $[-0.141, +0.201]$ | $-0.003$ $[-0.157, +0.180]$ |
| Joint | Future | 2 | 7000 | 12 | $-0.140$ $[-0.452, +0.219]$ | $-0.168$ $[-0.467, +0.193]$ |
| Joint | Future | 2 | 8000 | 4 | $-0.085$ $[-0.200, +0.023]$ | $-0.108$ $[-0.220, -0.001]$ |
| Joint | Future | 2 | 8000 | 8 | $-0.131$ $[-0.435, +0.040]$ | $-0.172$ $[-0.440, +0.012]$ |
| Joint | Future | 2 | 8000 | 12 | $-0.160$ $[-0.414, +0.162]$ | $-0.179$ $[-0.420, +0.127]$ |
| Joint | Future | 2 | 9000 | 4 | $+0.116$ $[-0.048, +0.237]$ | $+0.084$ $[-0.084, +0.220]$ |
| Joint | Future | 2 | 9000 | 8 | $-0.318$ $[-0.471, -0.012]$ | $-0.331$ $[-0.483, -0.033]$ |
| Joint | Future | 2 | 9000 | 12 | $+0.018$ $[-0.170, +0.276]$ | $+0.012$ $[-0.170, +0.261]$ |
| Post-hoc | State | 0 | 7000 | 4 | $+0.179$ $[-0.068, +0.385]$ | $+0.121$ $[-0.119, +0.349]$ |
| Post-hoc | State | 0 | 7000 | 8 | $+0.002$ $[-0.198, +0.198]$ | $-0.027$ $[-0.224, +0.167]$ |
| Post-hoc | State | 0 | 7000 | 12 | $-0.112$ $[-0.467, +0.240]$ | $-0.122$ $[-0.470, +0.235]$ |
| Post-hoc | State | 0 | 8000 | 4 | $-0.217$ $[-0.420, -0.063]$ | $-0.287$ $[-0.451, -0.113]$ |
| Post-hoc | State | 0 | 8000 | 8 | $-0.263$ $[-0.466, -0.048]$ | $-0.304$ $[-0.489, -0.082]$ |
| Post-hoc | State | 0 | 8000 | 12 | $+0.347$ $[+0.136, +0.555]$ | $+0.332$ $[+0.124, +0.530]$ |
| Post-hoc | State | 0 | 9000 | 4 | $-0.020$ $[-0.203, +0.144]$ | $-0.096$ $[-0.255, +0.091]$ |
| Post-hoc | State | 0 | 9000 | 8 | $-0.056$ $[-0.259, +0.161]$ | $-0.065$ $[-0.280, +0.139]$ |
| Post-hoc | State | 0 | 9000 | 12 | $+0.286$ $[+0.107, +0.443]$ | $+0.267$ $[+0.097, +0.423]$ |
| Post-hoc | State | 1 | 7000 | 4 | $-0.075$ $[-0.326, +0.174]$ | $-0.162$ $[-0.368, +0.097]$ |
| Post-hoc | State | 1 | 7000 | 8 | $-0.172$ $[-0.463, +0.063]$ | $-0.258$ $[-0.491, +0.019]$ |
| Post-hoc | State | 1 | 7000 | 12 | $+0.024$ $[-0.415, +0.258]$ | $-0.135$ $[-0.429, +0.193]$ |
| Post-hoc | State | 1 | 8000 | 4 | $+0.009$ $[-0.209, +0.209]$ | $-0.079$ $[-0.280, +0.135]$ |
| Post-hoc | State | 1 | 8000 | 8 | $-0.279$ $[-0.621, +0.016]$ | $-0.380$ $[-0.636, -0.064]$ |
| Post-hoc | State | 1 | 8000 | 12 | $-0.071$ $[-0.545, +0.240]$ | $-0.202$ $[-0.560, +0.195]$ |
| Post-hoc | State | 1 | 9000 | 4 | $+0.122$ $[-0.137, +0.303]$ | $+0.049$ $[-0.166, +0.254]$ |
| Post-hoc | State | 1 | 9000 | 8 | $-0.178$ $[-0.554, +0.079]$ | $-0.269$ $[-0.580, +0.032]$ |
| Post-hoc | State | 1 | 9000 | 12 | $+0.009$ $[-0.358, +0.315]$ | $-0.098$ $[-0.393, +0.255]$ |
| Post-hoc | State | 2 | 7000 | 4 | $+0.002$ $[-0.132, +0.115]$ | $-0.056$ $[-0.181, +0.056]$ |
| Post-hoc | State | 2 | 7000 | 8 | $+0.089$ $[-0.055, +0.227]$ | $+0.060$ $[-0.083, +0.193]$ |
| Post-hoc | State | 2 | 7000 | 12 | $+0.109$ $[-0.234, +0.448]$ | $+0.078$ $[-0.267, +0.420]$ |
| Post-hoc | State | 2 | 8000 | 4 | $-0.141$ $[-0.263, -0.025]$ | $-0.174$ $[-0.298, -0.071]$ |
| Post-hoc | State | 2 | 8000 | 8 | $-0.248$ $[-0.467, -0.061]$ | $-0.312$ $[-0.480, -0.093]$ |
| Post-hoc | State | 2 | 8000 | 12 | $+0.420$ $[+0.294, +0.571]$ | $+0.401$ $[+0.283, +0.543]$ |
| Post-hoc | State | 2 | 9000 | 4 | $-0.023$ $[-0.123, +0.099]$ | $-0.062$ $[-0.154, +0.040]$ |
| Post-hoc | State | 2 | 9000 | 8 | $+0.098$ $[-0.071, +0.285]$ | $+0.080$ $[-0.084, +0.267]$ |
| Post-hoc | State | 2 | 9000 | 12 | $+0.008$ $[-0.156, +0.178]$ | $-0.018$ $[-0.183, +0.160]$ |
| Post-hoc | Future | 0 | 7000 | 4 | $+0.170$ $[-0.030, +0.347]$ | $+0.119$ $[-0.100, +0.295]$ |
| Post-hoc | Future | 0 | 7000 | 8 | $+0.096$ $[-0.130, +0.306]$ | $+0.073$ $[-0.177, +0.275]$ |
| Post-hoc | Future | 0 | 7000 | 12 | $+0.010$ $[-0.242, +0.233]$ | $-0.010$ $[-0.253, +0.219]$ |
| Post-hoc | Future | 0 | 8000 | 4 | $+0.021$ $[-0.331, +0.290]$ | $-0.054$ $[-0.379, +0.272]$ |
| Post-hoc | Future | 0 | 8000 | 8 | $-0.090$ $[-0.301, +0.135]$ | $-0.142$ $[-0.334, +0.077]$ |
| Post-hoc | Future | 0 | 8000 | 12 | $+0.188$ $[-0.001, +0.410]$ | $+0.170$ $[-0.018, +0.384]$ |
| Post-hoc | Future | 0 | 9000 | 4 | $-0.055$ $[-0.245, +0.110]$ | $-0.126$ $[-0.291, +0.040]$ |

Continued on next page

Table 18 continued

| Score | Target | Seed | Bank | $k$ | Cross: median $[q_{.05}, q_{.95}]$ | Same: median $[q_{.05}, q_{.95}]$ |
|---|---|---|---|---|---|---|
| Post-hoc | Future | 0 | 9000 | 8 | $-0.015$ $[-0.263, +0.187]$ | $-0.034$ $[-0.285, +0.170]$ |
| Post-hoc | Future | 0 | 9000 | 12 | $+0.237$ $[-0.076, +0.499]$ | $+0.216$ $[-0.110, +0.492]$ |
| Post-hoc | Future | 1 | 7000 | 4 | $+0.083$ $[-0.134, +0.335]$ | $+0.090$ $[-0.130, +0.313]$ |
| Post-hoc | Future | 1 | 7000 | 8 | $+0.035$ $[-0.193, +0.280]$ | $+0.094$ $[-0.111, +0.296]$ |
| Post-hoc | Future | 1 | 7000 | 12 | $+0.153$ $[-0.229, +0.484]$ | $+0.267$ $[-0.081, +0.540]$ |
| Post-hoc | Future | 1 | 8000 | 4 | $+0.125$ $[-0.125, +0.307]$ | $+0.114$ $[-0.117, +0.298]$ |
| Post-hoc | Future | 1 | 8000 | 8 | $-0.124$ $[-0.342, +0.194]$ | $-0.059$ $[-0.271, +0.200]$ |
| Post-hoc | Future | 1 | 8000 | 12 | $+0.139$ $[-0.169, +0.421]$ | $+0.252$ $[+0.001, +0.473]$ |
| Post-hoc | Future | 1 | 9000 | 4 | $+0.258$ $[+0.063, +0.437]$ | $+0.252$ $[+0.032, +0.389]$ |
| Post-hoc | Future | 1 | 9000 | 8 | $+0.084$ $[-0.171, +0.398]$ | $+0.196$ $[-0.046, +0.433]$ |
| Post-hoc | Future | 1 | 9000 | 12 | $+0.212$ $[-0.043, +0.486]$ | $+0.336$ $[+0.115, +0.546]$ |
| Post-hoc | Future | 2 | 7000 | 4 | $+0.016$ $[-0.123, +0.152]$ | $-0.028$ $[-0.157, +0.106]$ |
| Post-hoc | Future | 2 | 7000 | 8 | $+0.085$ $[-0.077, +0.248]$ | $+0.061$ $[-0.081, +0.223]$ |
| Post-hoc | Future | 2 | 7000 | 12 | $-0.123$ $[-0.448, +0.241]$ | $-0.135$ $[-0.452, +0.226]$ |
| Post-hoc | Future | 2 | 8000 | 4 | $-0.125$ $[-0.274, +0.009]$ | $-0.150$ $[-0.289, -0.022]$ |
| Post-hoc | Future | 2 | 8000 | 8 | $-0.194$ $[-0.434, -0.005]$ | $-0.249$ $[-0.459, -0.029]$ |
| Post-hoc | Future | 2 | 8000 | 12 | $+0.093$ $[-0.145, +0.361]$ | $+0.076$ $[-0.150, +0.334]$ |
| Post-hoc | Future | 2 | 9000 | 4 | $-0.017$ $[-0.157, +0.092]$ | $-0.051$ $[-0.191, +0.055]$ |
| Post-hoc | Future | 2 | 9000 | 8 | $+0.004$ $[-0.159, +0.215]$ | $-0.008$ $[-0.165, +0.196]$ |
| Post-hoc | Future | 2 | 9000 | 12 | $+0.055$ $[-0.131, +0.285]$ | $+0.045$ $[-0.140, +0.272]$ |
| Belief | State | 0 | 7000 | 4 | $-0.171$ $[-0.351, +0.003]$ | $-0.210$ $[-0.390, -0.045]$ |
| Belief | State | 0 | 7000 | 8 | $+0.018$ $[-0.174, +0.205]$ | $-0.020$ $[-0.212, +0.174]$ |
| Belief | State | 0 | 7000 | 12 | $+0.018$ $[-0.479, +0.483]$ | $-0.005$ $[-0.495, +0.470]$ |
| Belief | State | 0 | 8000 | 4 | $-0.112$ $[-0.285, +0.048]$ | $-0.156$ $[-0.324, +0.007]$ |
| Belief | State | 0 | 8000 | 8 | $+0.274$ $[+0.018, +0.436]$ | $+0.229$ $[-0.030, +0.396]$ |
| Belief | State | 0 | 8000 | 12 | $-0.424$ $[-0.629, -0.189]$ | $-0.448$ $[-0.651, -0.207]$ |
| Belief | State | 0 | 9000 | 4 | $-0.173$ $[-0.315, -0.024]$ | $-0.221$ $[-0.355, -0.066]$ |
| Belief | State | 0 | 9000 | 8 | $+0.155$ $[-0.083, +0.372]$ | $+0.121$ $[-0.110, +0.342]$ |
| Belief | State | 0 | 9000 | 12 | $-0.114$ $[-0.322, +0.073]$ | $-0.156$ $[-0.369, +0.035]$ |
| Belief | State | 1 | 7000 | 4 | $-0.384$ $[-0.515, -0.240]$ | $-0.416$ $[-0.539, -0.283]$ |
| Belief | State | 1 | 7000 | 8 | $-0.255$ $[-0.386, -0.083]$ | $-0.288$ $[-0.412, -0.129]$ |
| Belief | State | 1 | 7000 | 12 | $-0.278$ $[-0.448, -0.076]$ | $-0.330$ $[-0.495, -0.157]$ |
| Belief | State | 1 | 8000 | 4 | $-0.220$ $[-0.378, -0.055]$ | $-0.249$ $[-0.417, -0.087]$ |
| Belief | State | 1 | 8000 | 8 | $-0.022$ $[-0.147, +0.147]$ | $-0.052$ $[-0.172, +0.095]$ |
| Belief | State | 1 | 8000 | 12 | $-0.407$ $[-0.557, -0.210]$ | $-0.449$ $[-0.578, -0.271]$ |
| Belief | State | 1 | 9000 | 4 | $-0.118$ $[-0.257, +0.014]$ | $-0.152$ $[-0.287, -0.014]$ |
| Belief | State | 1 | 9000 | 8 | $+0.008$ $[-0.192, +0.172]$ | $-0.036$ $[-0.226, +0.132]$ |
| Belief | State | 1 | 9000 | 12 | $-0.084$ $[-0.239, +0.100]$ | $-0.119$ $[-0.264, +0.062]$ |
| Belief | State | 2 | 7000 | 4 | $-0.369$ $[-0.521, -0.219]$ | $-0.394$ $[-0.531, -0.248]$ |
| Belief | State | 2 | 7000 | 8 | $-0.379$ $[-0.524, -0.249]$ | $-0.395$ $[-0.541, -0.268]$ |
| Belief | State | 2 | 7000 | 12 | $-0.288$ $[-0.693, +0.087]$ | $-0.306$ $[-0.703, +0.073]$ |
| Belief | State | 2 | 8000 | 4 | $-0.345$ $[-0.457, -0.254]$ | $-0.364$ $[-0.471, -0.268]$ |
| Belief | State | 2 | 8000 | 8 | $-0.130$ $[-0.389, +0.017]$ | $-0.144$ $[-0.415, +0.003]$ |
| Belief | State | 2 | 8000 | 12 | $-0.562$ $[-0.809, -0.355]$ | $-0.569$ $[-0.825, -0.367]$ |
| Belief | State | 2 | 9000 | 4 | $-0.322$ $[-0.455, -0.166]$ | $-0.345$ $[-0.465, -0.197]$ |
| Belief | State | 2 | 9000 | 8 | $-0.203$ $[-0.395, -0.046]$ | $-0.210$ $[-0.404, -0.053]$ |
| Belief | State | 2 | 9000 | 12 | $-0.413$ $[-0.652, -0.178]$ | $-0.421$ $[-0.663, -0.184]$ |
| Belief | Future | 0 | 7000 | 4 | $-0.073$ $[-0.268, +0.085]$ | $-0.104$ $[-0.303, +0.058]$ |
| Belief | Future | 0 | 7000 | 8 | $+0.263$ $[-0.022, +0.421]$ | $+0.240$ $[-0.047, +0.408]$ |
| Belief | Future | 0 | 7000 | 12 | $+0.107$ $[-0.157, +0.315]$ | $+0.101$ $[-0.166, +0.311]$ |
| Belief | Future | 0 | 8000 | 4 | $-0.150$ $[-0.299, -0.027]$ | $-0.173$ $[-0.338, -0.051]$ |
| Belief | Future | 0 | 8000 | 8 | $+0.161$ $[-0.031, +0.301]$ | $+0.125$ $[-0.081, +0.280]$ |
| Belief | Future | 0 | 8000 | 12 | $-0.079$ $[-0.290, +0.075]$ | $-0.093$ $[-0.300, +0.061]$ |
| Belief | Future | 0 | 9000 | 4 | $-0.109$ $[-0.257, +0.055]$ | $-0.146$ $[-0.282, +0.000]$ |
| Belief | Future | 0 | 9000 | 8 | $+0.020$ $[-0.125, +0.246]$ | $+0.000$ $[-0.144, +0.225]$ |
| Belief | Future | 0 | 9000 | 12 | $-0.068$ $[-0.362, +0.334]$ | $-0.085$ $[-0.365, +0.299]$ |
| Belief | Future | 1 | 7000 | 4 | $-0.170$ $[-0.347, +0.032]$ | $-0.173$ $[-0.351, +0.008]$ |
| Belief | Future | 1 | 7000 | 8 | $-0.051$ $[-0.266, +0.136]$ | $-0.052$ $[-0.283, +0.125]$ |
| Belief | Future | 1 | 7000 | 12 | $-0.191$ $[-0.528, +0.120]$ | $-0.204$ $[-0.527, +0.107]$ |
| Belief | Future | 1 | 8000 | 4 | $-0.066$ $[-0.302, +0.103]$ | $-0.071$ $[-0.306, +0.091]$ |
| Belief | Future | 1 | 8000 | 8 | $-0.002$ $[-0.140, +0.125]$ | $-0.005$ $[-0.148, +0.113]$ |

Continued on next page

Table 18 continued

| Score | Target | Seed | Bank | $k$ | Cross: median $[q_{.05}, q_{.95}]$ | Same: median $[q_{.05}, q_{.95}]$ |
|---|---|---|---|---|---|---|
| Belief | Future | 1 | 8000 | 12 | $-0.287$ $[-0.490, -0.023]$ | $-0.308$ $[-0.499, -0.039]$ |
| Belief | Future | 1 | 9000 | 4 | $-0.015$ $[-0.178, +0.113]$ | $-0.029$ $[-0.189, +0.091]$ |
| Belief | Future | 1 | 9000 | 8 | $-0.110$ $[-0.349, +0.082]$ | $-0.107$ $[-0.342, +0.072]$ |
| Belief | Future | 1 | 9000 | 12 | $-0.304$ $[-0.530, -0.079]$ | $-0.308$ $[-0.521, -0.084]$ |
| Belief | Future | 2 | 7000 | 4 | $-0.101$ $[-0.291, +0.032]$ | $-0.117$ $[-0.305, +0.013]$ |
| Belief | Future | 2 | 7000 | 8 | $-0.080$ $[-0.429, +0.235]$ | $-0.089$ $[-0.436, +0.233]$ |
| Belief | Future | 2 | 7000 | 12 | $-0.009$ $[-0.579, +0.492]$ | $-0.025$ $[-0.575, +0.492]$ |
| Belief | Future | 2 | 8000 | 4 | $-0.158$ $[-0.372, +0.006]$ | $-0.173$ $[-0.380, +0.002]$ |
| Belief | Future | 2 | 8000 | 8 | $+0.180$ $[-0.134, +0.424]$ | $+0.164$ $[-0.142, +0.421]$ |
| Belief | Future | 2 | 8000 | 12 | $-0.220$ $[-0.558, +0.099]$ | $-0.224$ $[-0.563, +0.097]$ |
| Belief | Future | 2 | 9000 | 4 | $-0.104$ $[-0.264, +0.069]$ | $-0.122$ $[-0.271, +0.041]$ |
| Belief | Future | 2 | 9000 | 8 | $-0.119$ $[-0.350, +0.014]$ | $-0.123$ $[-0.349, +0.010]$ |
| Belief | Future | 2 | 9000 | 12 | $-0.308$ $[-0.481, -0.189]$ | $-0.300$ $[-0.477, -0.182]$ |

### H.11 Teacher-state bootstrap rank resolution

The rank and pairwise exports cover *teacher-state consequence only*: 3 seeds $\times$ 3 banks $\times$ 3 keep levels gives 27 conditions, each with 30 unique mask entries. Thus the rank file contains exactly $27 \times 30 = 810$ indexed entries. These are repeated condition-specific mask entries, not 810 independent people or necessarily globally distinct masks. Future-space pairwise precision is not characterized by this export.

The bootstrap specification uses 2,000 subject-cluster resamples of 93 subjects with replacement, retaining all histories and repeated-subject multiplicity. The file contains full-sample ranks, bootstrap median ranks, and bootstrap 5th/95th percentile ranks. Population rank reproducibility can coexist with uncertain ordering of nearby masks.

Using the exported endpoints, 72/810 entries (8.9%) have $q_{.95}(\text{rank}) - q_{.05}(\text{rank}) \leq 6$. The stored `rank_width` column instead yields 73/810 (9.0%); 19 rows disagree with endpoint subtraction, each by one rank unit. Both counts are reported rather than silently reconciling them. Endpoint/width rounding or conversion cannot be resolved without the underlying bootstrap ranks and export implementation. Both definitions imply that only about 9% have narrow rank ranges.

### H.12 Pairwise denominators and ordering frequencies

Each condition has all $\binom{30}{2} = 435$ unordered pairs, indexed by `mask_r < mask_s`. The file therefore contains exactly $27 \times 435 = 11{,}745$ unique condition-pair entries, with no missing, duplicate, or non-finite probability fields. For the reported bootstrap ordering frequency $\hat{P}_{rs} = \Pr_{\text{boot}}(\bar{D}_r > \bar{D}_s)$, the exported classes agree with the thresholds in every row:

$$\hat{P}_{rs} > .95: \quad 2{,}129/11{,}745 = 18.13\%, \tag{37}$$

$$\hat{P}_{rs} < .05: \quad 2{,}226/11{,}745 = 18.95\%, \tag{38}$$

$$.05 \leq \hat{P}_{rs} \leq .95: \quad 7{,}390/11{,}745 = 62.92\%. \tag{39}$$

Thus 4,355/11,745 pairs (37.08%) lie beyond the two thresholds. These are pooled counts over the 27 teacher-state conditions. Because every condition has 435 valid pairs, pooling and equal-condition averaging give exactly the same percentages. The category orientation follows the exported $P(D_r > D_s)$ field; the underlying bootstrap tie treatment is not reconstructed from a probability summary. Probability values are compatible with increments of $1/2000$, but that arithmetic check is not a rerun of the bootstrap.

The frequencies are not Bayesian posterior probabilities or familywise-controlled significant pairwise comparisons. Pair entries share masks, subjects, and models and are not independent samples. Table 19 provides every condition's integer numerator and denominator for rank-width and pairwise summaries.

Table 19: **Teacher-state bootstrap denominators and counts.** Each of the 27 conditions contains 30 mask entries and 435 unordered pair entries. Width columns count entries with width at most six, using exported endpoints (EP) or the stored width field (SW). Pair columns use strict probability thresholds.

| Seed | Bank | $k$ | EP $\leq 6$ | SW $\leq 6$ | $P > .95$ | $P < .05$ | Ambiguous |
|---|---|---|---|---|---|---|---|
| 0 | 7000 | 4 | 1 | 1 | 83 | 72 | 280 |
| 0 | 7000 | 8 | 2 | 2 | 99 | 62 | 274 |
| 0 | 7000 | 12 | 1 | 1 | 36 | 51 | 348 |
| 0 | 8000 | 4 | 2 | 3 | 83 | 78 | 274 |
| 0 | 8000 | 8 | 2 | 2 | 41 | 74 | 320 |
| 0 | 8000 | 12 | 2 | 2 | 78 | 54 | 303 |
| 0 | 9000 | 4 | 2 | 2 | 71 | 100 | 264 |
| 0 | 9000 | 8 | 4 | 4 | 55 | 89 | 291 |
| 0 | 9000 | 12 | 1 | 1 | 35 | 80 | 320 |
| 1 | 7000 | 4 | 2 | 2 | 114 | 65 | 256 |
| 1 | 7000 | 8 | 4 | 4 | 126 | 60 | 249 |
| 1 | 7000 | 12 | 1 | 1 | 46 | 75 | 314 |
| 1 | 8000 | 4 | 2 | 2 | 84 | 76 | 275 |
| 1 | 8000 | 8 | 4 | 4 | 87 | 95 | 253 |
| 1 | 8000 | 12 | 3 | 3 | 113 | 76 | 246 |
| 1 | 9000 | 4 | 4 | 4 | 94 | 124 | 217 |
| 1 | 9000 | 8 | 1 | 1 | 58 | 102 | 275 |
| 1 | 9000 | 12 | 3 | 3 | 59 | 92 | 284 |
| 2 | 7000 | 4 | 8 | 8 | 147 | 79 | 209 |
| 2 | 7000 | 8 | 4 | 4 | 107 | 61 | 267 |
| 2 | 7000 | 12 | 0 | 0 | 30 | 40 | 365 |
| 2 | 8000 | 4 | 7 | 7 | 115 | 113 | 207 |
| 2 | 8000 | 8 | 2 | 2 | 44 | 99 | 292 |
| 2 | 8000 | 12 | 2 | 2 | 113 | 79 | 243 |
| 2 | 9000 | 4 | 5 | 5 | 95 | 102 | 238 |
| 2 | 9000 | 8 | 2 | 2 | 76 | 140 | 219 |
| 2 | 9000 | 12 | 1 | 1 | 40 | 88 | 307 |

### H.13 Sensitivity to subject weighting

Subject-balanced and window-weighted means are

$$\bar{D}^{a}_{r,\mathrm{subj}} = \frac{1}{93}\sum_{s=1}^{93}\frac{1}{n_s}\sum_{i:\,\mathrm{subject}(i)=s} D^{a}_{i,r}, \qquad \bar{D}^{a}_{r,\mathrm{win}} = \frac{1}{256}\sum_{i} D^{a}_{i,r}. \tag{40}$$

The 54-row weighting export covers all 27 conditions for each target. Their mean rank concordance is 0.931 for teacher-state error (minimum 0.842) and 0.944 for future error (minimum 0.815). The full CSV accompanies the paper. These values support consequence-order robustness to the weighting choice within this pool. They do not make the estimands identical or supply unreported subject-balanced full-pool uncertainty–consequence correlations.

### H.14 Interpretive boundary

The audit establishes positive condition-median reproducibility on each side, together with weak or negative *aggregate* cross-half alignment and substantial condition heterogeneity. Teacher-state pairwise ambiguity further limits precise measurement-subset ordering. The target is a learned reference consequence, not physiological ground truth; an acquisition policy would additionally integrate unknown measurements and costs. Neither exact-zero equivalence nor failure of a dedicated acquisition objective follows.

# I Oracle Construct-Validity Study

## I.1 Generator and known conditional quantity

The oracle has eight classes with uniform prior, centers drawn from $\mu_c \sim \mathcal{N}(0, 3.0^2 I_{d_Y})$, and irreducible future-noise scale $\sigma_0 = 1.0$. For $d_Y \geq d_V$,

$$C \sim \text{Uniform}\{1, \ldots, 8\}, \qquad Y = \mu_C + \sigma_0 \eta, \qquad V = \mu_C^{(1:d_V)} + \sigma_v \epsilon, \tag{41}$$

where $\eta$ and $\epsilon$ are independent standard Gaussian noise vectors. The audited state is the class mean $\mu_C$, not a noisy future realization. The center distribution and future-noise scale are specified independently of the recovery correlations. The realized centers, target dimension, and random seeds are not specified in the accompanying summaries and are required to reproduce the generator exactly. Adding the fixed $\sigma_0^2 I$ shifts the total covariance trace by $d_Y \sigma_0^2$ and therefore does not change its rank within a regime.

## I.2 Posterior derivation

The Gaussian likelihood is

$$p(V \mid c) = (2\pi\sigma_v^2)^{-d_V/2} \exp\left[ -\frac{\|V - \mu_c^{(1:d_V)}\|^2}{2\sigma_v^2} \right]. \tag{42}$$

The common prior and normalizing factors cancel, giving

$$p(c \mid V) = \frac{\exp[-\|V - \mu_c^{(1:d_V)}\|^2/(2\sigma_v^2)]}{\sum_{c'} \exp[-\|V - \mu_{c'}^{(1:d_V)}\|^2/(2\sigma_v^2)]}. \tag{43}$$

Then

$$\bar{\mu}(V) = \sum_c p(c \mid V) \mu_c, \tag{44}$$

$$B^*(V) = \sum_c p(c \mid V) \mu_c \mu_c^\top - \bar{\mu}(V)\bar{\mu}(V)^\top, \tag{45}$$

$$b^*(V) = \sum_c p(c \mid V) \|\mu_c\|^2 - \|\bar{\mu}(V)\|^2, \tag{46}$$

$$\text{Cov}(Y \mid V) = \sigma_0^2 I + B^*(V). \tag{47}$$

The centered form in Equation (8) is algebraically equivalent and avoids subtracting large second moments when posterior concentration is high. $B^*$ measures ambiguity over latent class means. Equating it to an EEG full-history conditional quantity would require assumptions about what the full EEG history reveals.

## I.3 Retrieval estimator and regime results

An independent reference bank contains $(V_j, \mu_{C_j})$. Its retrieved dispersion is

$$\hat{b}(V) = \text{Tr}\, \widehat{\text{Cov}}\{\mu_{C_j} : j \in \mathcal{N}_K(V)\}. \tag{48}$$

The primary experiment uses 50,000 references, 5,000 queries, and $K = 20$. Table 20 retains the available regime summaries.

The reported 20,000-to-100,000-reference checks do not rescue degenerate regimes. They do not establish performance for arbitrary bank sizes. A low-dynamic-range $b^*$ can make ranks sensitive to finite-neighborhood fluctuations and numerical ties. Exactly constant $b^*$ has undefined Spearman correlation. Accordingly, approximate-zero summaries support no conclusion about a mathematically defined correlation with an exactly constant target.

Table 20: **Oracle recovery.** Uniform class prior, centers from $\mathcal{N}(0, 3.0^2 I)$, and $\sigma_0 = 1$. Approximately-zero summaries are not assigned additional precision.

| $d_V$ | $\sigma_v$ | Regime | Spearman $(\hat{b}, b^*)$ |
|---|---|---|---|
| 2 | 1.0 | Informative | $+0.91$ |
| 4 | 1.0 | Informative | $+0.75$ |
| 8 | 1.0 | Degenerate | $\approx 0$ |
| 16 | 1.0 | Degenerate | $\approx 0$ |
| 8 | 1.5 | Intermediate | $+0.44$ |
| 16 | 2.5 | Intermediate | $+0.43$ |
| 8 | 2.0 | Non-degenerate | $+0.83$ |
| 16 | 3.0 | Non-degenerate | $+0.73$ |

### I.4 What the oracle establishes

Recovery at $d_V = 8, 16$ improves in different noise regimes, so dimension alone does not explain the pattern. Noise changes both conditional ambiguity and local observation geometry; the analysis does not isolate either as the sole mechanism. The finite-bank probe is not uniformly validated as exact ambiguity. Suitable nonparametric estimators may still consistently recover conditional moments under regularity assumptions and an appropriate joint growth of sample size and neighborhood size. Fixed $K = 20$ is not such a guarantee.

## J Matched-Representation Control

### J.1 Matched factors and reporting target

Table 21 compares prediction error with the separate split-reliability statistic under matched representation settings. The control compares PCA-64, RP-64, VAR-64, and PRED-64 using the same frozen source, 64-dimensional size, normalization, $N_Q = 2{,}048$, 20 subject splits, ridge-readout setup, and LPSD estimator. The two reported quantities are prediction MSE and median split reliability. Median split reliability is distinct from the uncertainty–LPSD association in Table 2. Its complete splitwise definition and representation construction are not specified in the accompanying matched-control summaries.

Table 21: **Matched representation summaries.** Entries are prediction MSE / median split reliability, with the shared settings above.

| Dataset | Keep | PCA-64 | RP-64 | VAR-64 | PRED-64 |
|---|---|---|---|---|---|
| TUEG | 8 | 8.8/.489 | 9.2/.526 | 8.8/.531 | 6.7/.597 |
| TUEG | 12 | 4.9/.492 | 5.2/.542 | 5.1/.547 | 2.5/.564 |
| TUSZ | 8 | 43.3/.507 | 41.8/.310 | 42.1/.291 | 37.9/.193 |
| TUSZ | 12 | 26.0/.477 | 25.5/.610 | 25.7/.618 | 20.2/.660 |

### J.2 Limited conclusion

PRED-64 has the smallest prediction-error entry in each row. Its reliability is highest in three rows but lowest in TUSZ at $k = 8$. In that condition the primary statistic favors PCA-64 over PRED-64 ($0.507 > 0.193$), whereas the median-pairwise statistic favors PRED-64 over PCA-64 ($0.609 > 0.518$). This is a reversal of representation ordering, distinct from a sign reversal of an uncertainty–LPSD coefficient. The shuffled-null exception ($0.126$ versus $0.148$) is another statistic and must not be substituted for either reliability comparison. This does not support a universal prediction–reliability trade-off. It also does not causally explain the main dissociation. The control serves to prevent an unsupported broader claim rather than supply an additional headline contribution.

# K Boundary Analyses

## K.1 Ranking stability

Across 16 evaluated conditions and 20 subject splits, the reported top-model flip rate is 0%. This indicates no observed top-model change under that control's reference ordering and ranking rule. It is not a population zero-risk guarantee or a result for every estimator family. Split variation, target reliability, and estimator ranking are distinct concepts; the main conclusion does not rely on instability of estimator ranking.

## K.2 Exploratory downstream probes

Table 22 shows that adding variance features provides no observed improvement in the reported exploratory probes.

Table 22: **Retained exploratory probe accuracy.** These point summaries support no downstream-benefit claim. CHB-MIT is excluded following Appendix B.

| Dataset | Mean features $\mu$ | Augmented $[\mu, \log \sigma^2]$ |
|---|---|---|
| TUEP | .50 | .50 |
| TUSZ | .35 | .18 |

Adding variance features does not improve the reported accuracies. Without a full account of task labels, class balance, readout, and sampling variability, their magnitude should not be generalized. They remain exploratory scope checks and do not establish a clinically useful uncertainty representation. No validated held-out seizure-onset signal is reported; the future-descriptor task is not presented as seizure prediction.

# L Statistical Reporting and Inference

## L.1 Estimand and unit

Spearman correlation is Pearson correlation of ranks, conventionally using midranks for ties. A constant vector gives undefined correlation. Pooled-window, subject-balanced, per-subject-averaged, within-history, and population-mask statistics are different estimands. The aggregation order and weights determine which is reported.

Subjects define biological resampling clusters. Training seeds describe optimization variation. Mask banks describe variation in the evaluated interventions. Reference banks affect the structural probe. Reusing persons across seeds or banks does not create new independent biological observations. The 256-history intervention pool contains 93 such subjects.

## L.2 Subject-cluster bootstrap

The primary statistical protocol specifies 2,000 subject-cluster replicates. The exact-mask diagnostic uses 400 replicates (Table 10). For a fixed analysis condition:

1. Sample the original number of subjects with replacement.
2. Include each sampled occurrence's full cluster of histories, preserving paired masks and outcomes and repeated sampling multiplicity.
3. Recompute the declared statistic, including mask means before population rank correlation.
4. Obtain percentile limits at the stated confidence level and record invalid replicates rather than silently assigning zeros.

A 95% percentile interval, when that level is specified, uses the 2.5th and 97.5th percentiles. The exact-mask export specifies a 95% percentile interval and 400 replicates; its cross-mask/seed reduction

remains distinct from the fully specified rank-reliability reduction. Query-only resampling conditions on a frozen LPSD reference bank; it does not account for new reference-subject sampling. Likewise, fixed-bank analysis does not quantify variability over new mask banks. Resampling at the subject level preserves dependence but leaves the original window/subject weighting choice intact.

### L.3 Point summaries, seed variation, and intervals

For three training-seed estimates,

$$\bar{\rho} = \frac{1}{3}\sum_{s=1}^{3}\hat{\rho}_s, \qquad \mathrm{SD}_{\mathrm{seed}} = \left[\frac{1}{2}\sum_{s}(\hat{\rho}_s - \bar{\rho})^2\right]^{1/2}. \tag{49}$$

These describe seed variability and are not confidence limits for subject sampling. An interval for a fixed three-model average can be computed by resampling subjects, recomputing each model's statistic, and applying the specified model aggregation within each replicate. That conditional interval does not by itself generalize to all future training seeds.

Table 2 uses the four-decimal seed coefficients and the $n-1$ sample-SD convention above. The mean and sample SD are recomputed directly, with the complete input coefficients in Table 11. For example, the joint $k = 8$ coefficients $(.5848, .3619, .4313)$ give mean $.459333$ and sample SD $.114064$. Seed SD is not interpreted as a subject-bootstrap interval.

For Table 14, the source identifies the spread as sample SD across nine seed-by-bank cells:

$$\bar{\rho}_{\mathrm{run}} = \frac{1}{9}\sum_{s=1}^{3}\sum_{b=1}^{3}\hat{\rho}_{s,b}, \qquad s_{\mathrm{run}} = \left[\frac{1}{8}\sum_{s,b}(\hat{\rho}_{s,b} - \bar{\rho}_{\mathrm{run}})^2\right]^{1/2}. \tag{50}$$

The summaries are retained as reported; individual run values are not available to verify the SD denominator independently. Reusing persons and crossing seeds with banks makes $s_{\mathrm{run}}$ descriptive dispersion, not a subject-level standard error. Dividing it by $\sqrt{9}$ would not resolve that dependence.

The 162-row rank-reliability export supplies split summaries, not the full-pool $\rho_{\mathrm{pop}}$ runs underlying Table 14. These two arrays share design factors but measure different statistics and cannot be substituted for each other. Rank headlines are independently reconstructed from the condition-median export; Table 14 retains its source-reported run SD. Exact-mask intervals in Table 10 are also retained as reported. No new pooled interval or practical-equivalence test is inferred.

### L.4 Split distributions, bootstrap ranks, and independent units

The 2,000 split-half partitions in Appendix H, the primary 2,000-replicate subject bootstrap protocol, and the exact-mask diagnostic's 400 bootstrap replicates are different procedures. Split-half 5th–95th percentile summaries, when available, describe a distribution over partitions; they are not automatically confidence intervals for a population parameter. Bootstrap rank ranges describe individual mask-rank variation and should not be substituted for uncertainty on $R_D$, $R_U$, or cross-half alignment. Tables 16–18 report the within-condition split quantiles. Figure 2 distinguishes those from the recomputed spread across condition medians.

Repeated splits, correlated mask pairs, and indexed mask entries do not increase the number of independently sampled subjects. The complete export shows positive medians in every consequence and uncertainty condition; its split quantiles retain substantial variation, especially for future error. The 810 rank entries and 11,745 pair entries are explicitly restricted to teacher-state error across all 27 conditions. The same-half/cross-half comparison remains descriptive; the split summaries do not support a paired-difference test without paired split-level values.

### L.5 Multiple testing and effect boundaries

The shuffled-reference comparison is reported descriptively. A formal multiple-testing claim would require a valid null scheme, the actual test family and all corresponding $p$-values. Holm's sequential rule compares ordered $p_{(j)}$ with $\alpha/(m-j+1)$ until the first non-rejection (Holm, 1979); the count

15/16 above empirical thresholds is not that analysis. Supplementary diagnostics are described as such, without retrospectively labeling them as pre-registered hypotheses.

Near-zero or sign-changing point coefficients do not prove practical equivalence to zero. Such a conclusion would require a justified smallest effect of interest and an appropriate interval or test against it. Distinct damage ranks alone do not establish precise ordering; the split-half audit supplies additional population-level reproducibility evidence while retaining pairwise uncertainty. The paper therefore concludes that positively reproducible consequence and uncertainty rankings have weak or negative cross-half alignment in the evaluated protocols, not that their underlying association is exactly zero.

# M Claim-to-Evidence Map and Reproducibility Scope

## M.1 Claim boundaries

Table 23 connects the main claims to their supporting tests and interpretive limits.

Table 23: **Claim-to-evidence map.** The consequence definitions, mask pool, and statistical limits determine the scope of the main finding.

| Claim | Evidence | Maximum supported interpretation |
|---|---|---|
| Non-trivial response | Error/degradation and exact-mask diagnostic | Sample-dependent, error-associated score in evaluated settings |
| Structural information | LPSD; geometries; shuffled reference | Covaries with a model-defined probe, with representation boundaries |
| Mask consequence ranking | Split-half $R_D > 0$, $R_U > 0$; cross-half alignment | Reproducible population rankings have weak or negative alignment |
| Individual mask ordering | Bootstrap rank ranges; about 63% ambiguous pairs | Population reproducibility does not establish precise pairwise ordering |
| Conditional ambiguity | Uniform-prior oracle regimes | LPSD not uniformly validated as exact ambiguity |
| External robustness | Three person-disjoint TUH-derived subsets | Cross-dataset response without predictive-model refitting; shared source ecosystem |
| Absolute scale transfer | External same-set rescaling | Post-hoc diagnostic only |
| Physiological/clinical benefit | No direct validation | Not claimed |

## M.2 Available materials and numerical reproduction

The accompanying source includes nine CSV summaries, the exact rank-aggregation rule, complete condition tables, figure assets, and a deterministic recomputation script. The script checks design completeness and duplicated fields; recomputes rank headlines, structural seed statistics, and bootstrap counts; and records source hashes. The condition summaries contain 162 estimator–consequence rows, yielding 54 unique consequence and 81 uncertainty records. Teacher-state bootstrap summaries contain 810 rank entries and 11,745 pairs. Additional files cover weighting sensitivity, fixed-mask diagnostics, LPSD coefficients, shuffled controls, and subject overlap.

These materials support summary-level reproduction, not a rerun of the underlying EEG experiments. Raw EEG, model weights, individual split coefficients, bootstrap ranks, exact mask bits, and complete implementation configurations are not included. Exact-mask reduction across seeds and masks, the 19 rank-width discrepancies, bootstrap tie/rounding rules, and full-pool run SDs require the corresponding underlying artifacts. Dataset access follows the original resource terms. Split quantiles, condition-median quantiles, seed/run SDs, and bootstrap intervals retain their distinct meanings; none establishes uniform conditionwise failure, precise pairwise identification, or exact latent ambiguity.