

Relational Bayesian Inference for Musical Pleasure: An Experimental Protocol

Amaç Erdem

Independent researcher

Correspondence: amace@bu.edu · ORCID 0009-0000-1711-0036

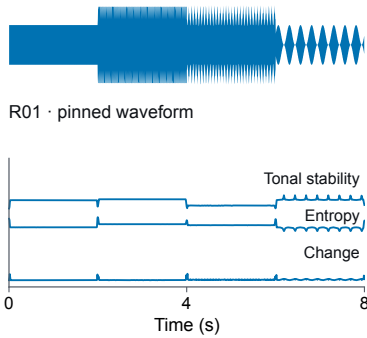
Abstract

Can a response about familiarity or expectation improve the prediction of later musical pleasure? This protocol tests that possibility in Musical Intelligence (MI), combining a fixed, inspectable audio representation with a proposed personal Bayesian state. Direct and auxiliary reports enter distinct likelihoods. Their value is assessed through predictions committed before each listening block, with controls for history, direct feedback and cross-task dependence. Analytic examples show both informative transfer and the harm caused by an incorrect relation. Source-checked audio calculations and an executable reduced simulation make these examples reproducible. A separate discovery procedure admits dimensions without a prior pleasure label, while preserving selection, replication and causal boundaries. The contribution is a testable route from named musical computations to listener-specific inference and new hypotheses. No participant outcome or full-protocol validation is claimed.

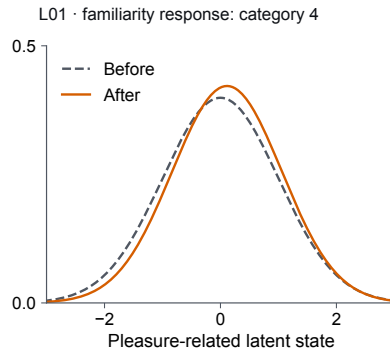
The recording stays fixed. Knowledge of the listener can change.

Overview | Audio fixture, analytic update and synthetic forecast; personal layer proposed

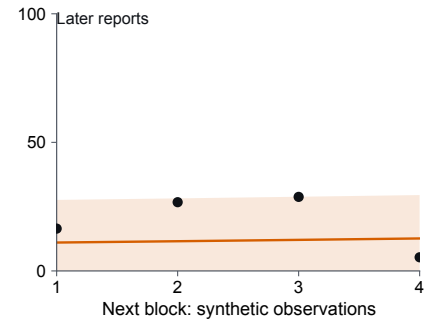
a Analyse a verified recording



b Update a listener distribution



c Predict the next block



FIXED AUDIO ENGINE

No feedback-trained audio encoder

PROPOSED PERSONAL INFERENCE

Completed responses only

LATER EMPIRICAL TEST

Does R beat direct feedback C?

Figure 1. From a recording to a future prediction. (a) Saved acoustic coordinates from an eight-second teaching signal. (b) Exact marginal update for the ordinal example in Figure 6. (c) A separate synthetic listener's block-2 predictive medians, central 80% intervals and later reports from the reduced simulation in Figure 13. These three examples explain the workflow; they are not an end-to-end participant demonstration. Blue denotes fixed computation, orange personal inference and black observations.

1. The question is whether another response helps

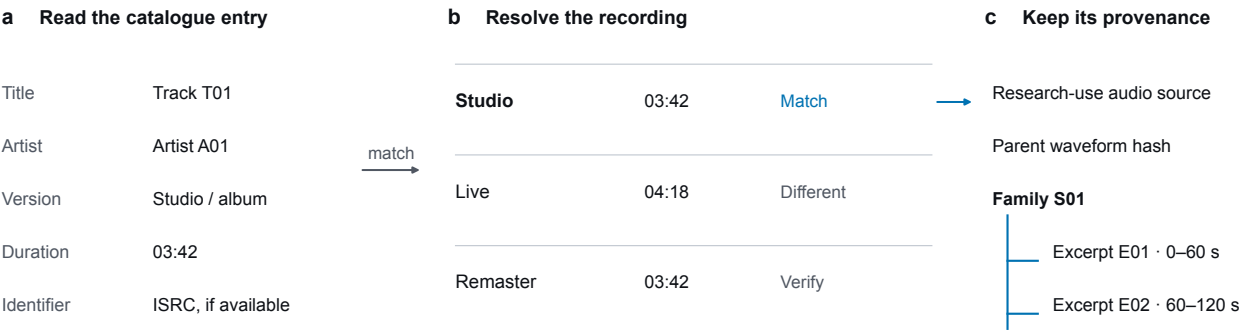
A pleasure report is direct evidence, but its interpretation is incomplete. Familiarity and expectation may constrain different explanations of that report. Their relevance is motivated by music research (van den Bosch et al., 2013; Cheung et al., 2024); neither is equivalent to pleasure, and perceptual and reward responses can dissociate (Mas-Herrero et al., 2014).

The proposed contribution is an inspectable experiment around those distinctions. Joint inference and cross-task transfer have established precedents (Turner et al., 2013; Bonilla et al., 2007). MI must demonstrate added value through its representation, controls and reproducible findings. The personal study layer remains proposed; source calculations, analytic examples and reduced-model simulations are labelled separately throughout.

2. From catalogue history to a verified signal

The same title can refer to different recordings.

Proposed recording workflow | Fictional identifiers; catalogue metadata is not the audio source



Playlist membership is history metadata, not proof of liking. Both excerpts remain in family S01.

Figure 2. Recording identity precedes analysis. Catalogue entries from Spotify or Apple Music identify candidate recordings; a separate research-use source supplies the waveform. Title alone cannot establish identity. Artist, version, duration and available identifiers support verification; the manifest retains source, permission, hashes, excerpt boundaries and family membership. Entries here are fictional. A playlist is not a measured preference.

An audible change becomes a measured coordinate.

Archived MI teaching fixture | 44.1 kHz; hop 256 samples; 97 saved channels; no human ratings

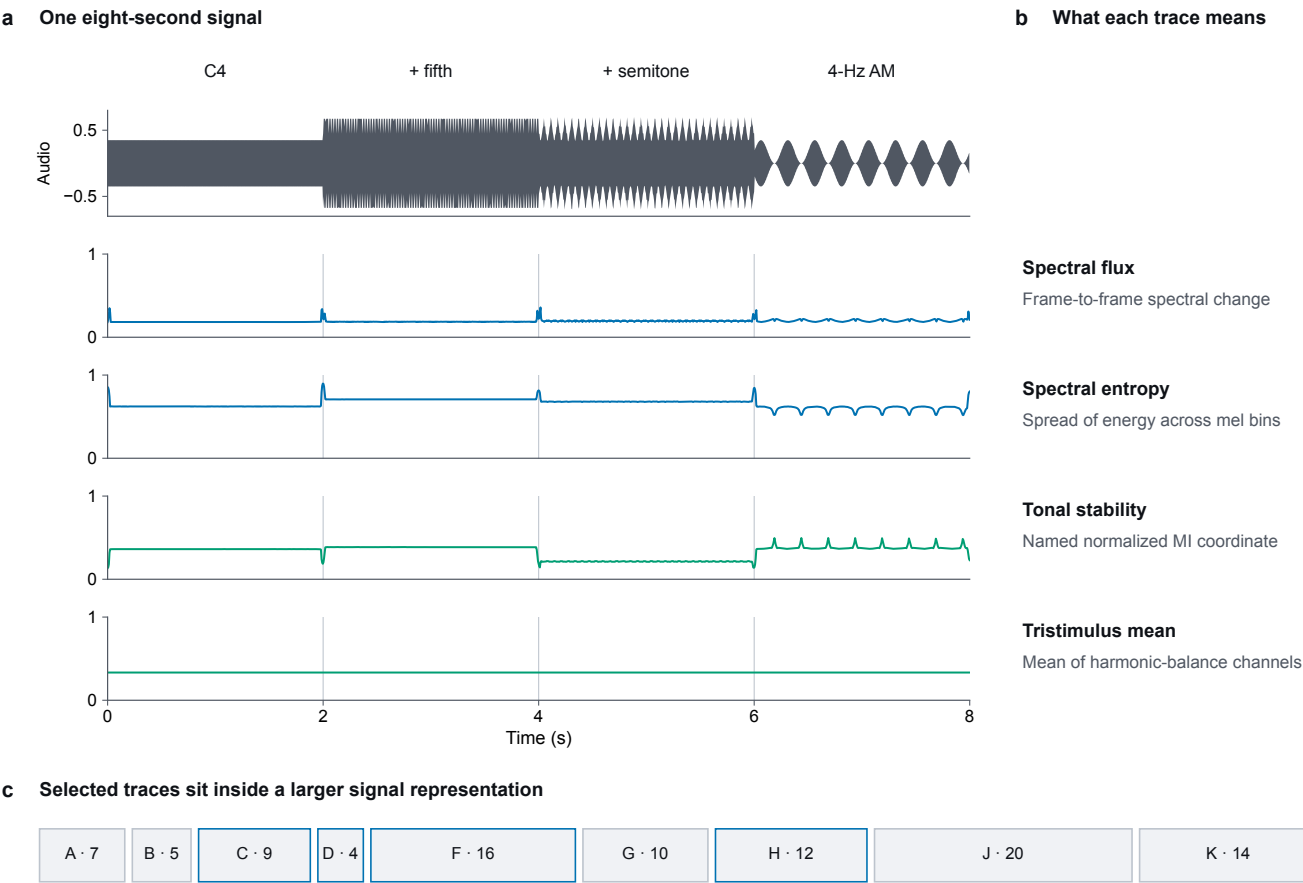


Figure 3. Read the signal before interpreting its labels. (a) An archived 44.1-kHz fixture adds a fifth, a semitone and amplitude modulation across four two-second segments. Saved R^3 coordinates share a 256-sample hop. (b) Trace labels describe operational quantities, not measured mental states. (c) The nine group counts sum to 97; Supplementary Figure S1 lists every coordinate. The fixture illustrates the engine and is not a proposed study stimulus.

The waveform, its identity and the listener’s history answer different questions. Catalogue metadata may help find a recording; permission to analyse audio and the identity of that recording require separate evidence. The analysis retains this provenance rather than joining tracks by title alone.

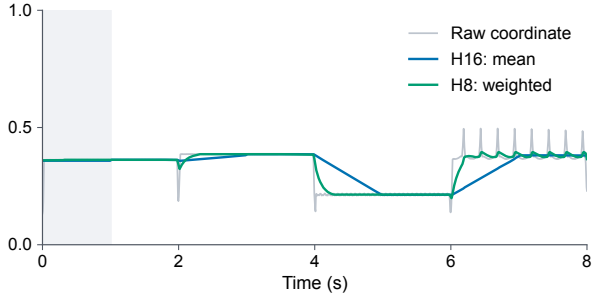
The fixed engine maps audio to named coordinates, $\mathbf{z}_{st} = F_{MI}(a_s)_t$. Coordinates, temporal operations and composite scores remain inspectable. They do not acquire psychological validity simply because their labels resemble mental processes.

3. Separate the audio computation from the listener model

Open one computation: HTP extraction output E0.

Executed current temporal/HTP code on the archived R3 fixture | Operational score, not neural latency

a Two summaries of tonal stability



b Inspect the weighted result

Tristimulus mean	0.333×0.40	
Tonal H16 mean	0.330×0.35	
Tonal H8 weighted	0.213×0.25	

At 4.301 s: sum $\rightarrow$ clip $\rightarrow$ E0 = 0.302

H16: 1,000 ms / 172 frames. H8: 300 ms / 52 frames. Initial shaded values use edge replication.

$$E_0(t) = \text{clip}_{[0,1]} \left[0.40 \bar{z}_{18:20}(t) + 0.35 H_{60,16,1,0}(t) + 0.25 H_{60,8,0,0}(t) \right]$$

Horizon durations follow the registry; an older variable name incorrectly says 500 ms for H8.

Figure 4. A source-checked computation. HTP means Hierarchical Temporal Prediction. Current temporal code and the HTP extraction function were executed on the archived R³ arrays; independent NumPy calculations reproduced E0 within 2×10^{-7} . (a) Temporal summaries of coordinate 60. (b) Their actual contributions at the marked numerical example. The tristimulus mean remains one third in this fixture. H16 and H8 use 172 and 52 frames. The first complete backward window is replicated at the boundary, so those initial outputs are not strictly causal. See Supplementary Methods for the registry/comment discrepancy.

Different responses can constrain the same personal state.

Proposed observation model | Filled marks: observed responses; hollow circles: latent components

a Three actual response scales

How much pleasure do you feel?

0 100

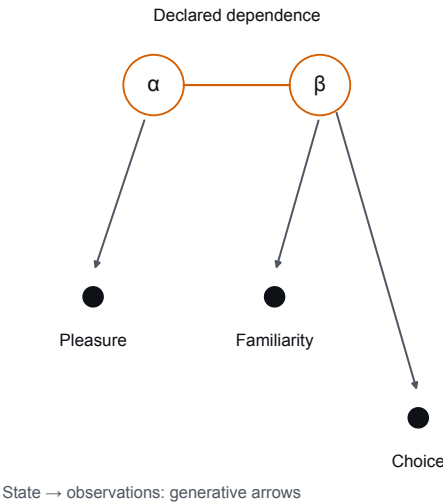
Did you know this excerpt before today?

☐ 1 ☐ 2 ☐ 3 ☒ 4 ☐ 5

Which continuation did you expect?

☐ Option A ☐ Option B

b A generative model



c Different lifetimes

Global

Links and priors
frozen after development

Personal

A distribution
updated between blocks

Transient

Playback, raw reports
and interruption events

$$r_k \sim \mathcal{L}_k(h_k(\mathbf{z}, \theta_u, \mathbf{c})), \quad p(\theta_u | D, r_k) \propto p(r_k | \theta_u, \mathbf{z}, \mathbf{c}) p(\theta_u | D)$$

Figure 5. Observed tasks and inferred state have different roles. (a) Proposed controls retain the pleasure, ordinal familiarity and binary continuation scales. (b) A simplified joint model links distinct latent components; the undirected line denotes statistical dependence, not causation. (c) Shared settings remain fixed during evaluation, while a personal distribution can update. Transient session state is separate from the immutable raw-event archive. The equation gives the general observation and posterior-update rules.

The audio calculation above is executed code. The personal state is an additional inference layer. In the inspected HCMC memory interface, eleven outputs depend on acoustic, temporal and relay inputs; participant responses and persistent listening history are not arguments. A personal memory therefore needs explicit observations, uncertainty, persistence and versioned updates.

This protocol uses no deep neural network and does not retrain an audio encoder from ratings. Bayesian adaptation can nevertheless fall under the broad term statistical learning; the precise distinction concerns what is fixed and what is updated.

4. Work through a response before generalizing the claim

Follow a familiarity answer through a complete Bayesian update.

Analytic ordinal teaching model | Fictional listener L01; latent units; no fitted participant

a Observe category 4

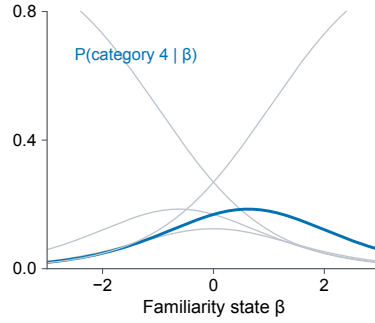
Known before this session?



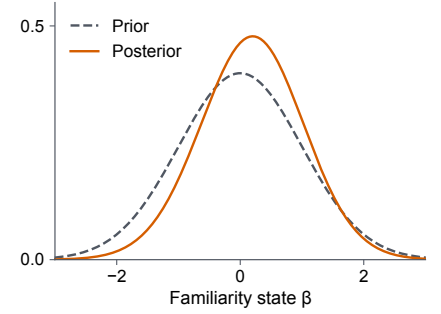
Response $r = 4$

Possible familiarity states are not all equally likely to produce this answer.

b Select its likelihood



c Reweight and normalize



$$p(\alpha, \beta | r = 4) \propto \underbrace{[\sigma(1 - \beta) - \sigma(0.25 - \beta)]}_{\text{likelihood of category 4}} p(\alpha, \beta)$$

d Transfer requires the specified relation

Connected prior: correlation 0.6

α : mean 0 $\rightarrow$ 0.110; variance 1 $\rightarrow$ 0.896

Independent prior: correlation 0

α : mean 0 $\rightarrow$ 0; variance 1 $\rightarrow$ 1

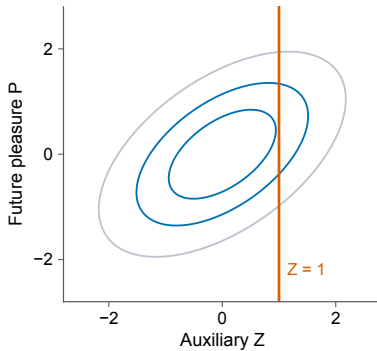
The answer updates familiarity directly. Pleasure-related uncertainty changes only through the joint model.

Figure 6. An ordinal answer updates a distribution. The joint standard-normal prior has correlation 0.6. Category 4 uses ordered-logit cutpoints 0.25 and 1, with σ the logistic function; other category likelihoods appear in grey. The response probability is 0.14524. Numerical quadrature gives posterior mean 0.11046 and variance 0.89562 for the pleasure-related component α . Setting the prior correlation to zero leaves its marginal unchanged. These latent units are not pleasure-scale points.

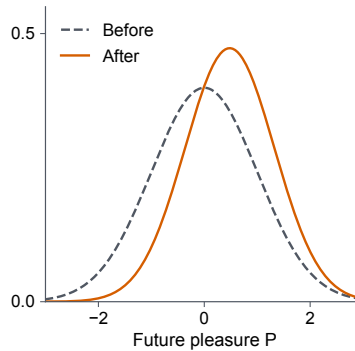
An indirect observation selects a conditional prediction.

Exact Gaussian example | Continuous standardized Z; distinct from the ordinal answer in Figure 6

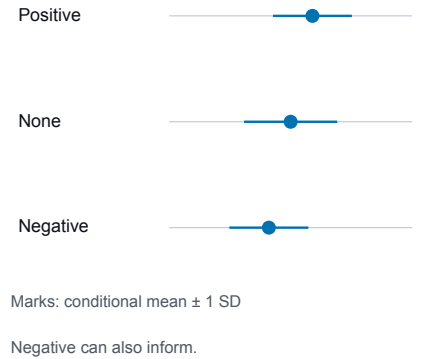
a Specify a joint distribution



b Condition on Z = 1



c Compare directions



$$\mathbb{E}[P | Z = 1] = 0.48, \quad \text{Var}(P | Z = 1) = 1 - \frac{0.60^2}{1.25} = 0.712$$

Assumptions: $\text{Var}(P) = 1$; $\text{Var}(Z) = 1.25$ including noise; $\text{Cov}(P, Z) = 0.60$; prior means zero.

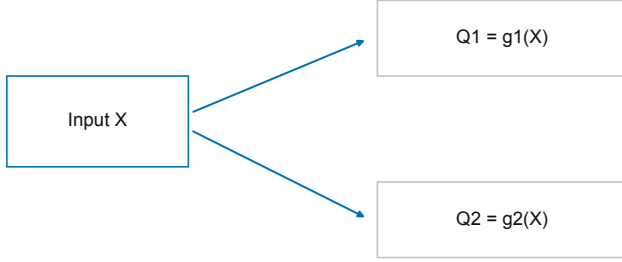
Figure 7. Conditioning in a separate Gaussian example. The joint distribution has zero means, $\text{Var}(P) = 1$, $\text{Var}(Z) = 1.25$ and covariance 0.60. The slice $Z = 1$ produces conditional mean 0.48 and variance 0.712. (c) Reversing the covariance reverses the mean shift while preserving the uncertainty reduction; a zero covariance transfers no information. This continuous observation is distinct from the ordinal response in Figure 6.

5. More labels do not necessarily provide more evidence

A second observation can add a constraint. A copied score cannot.

Analytic examples | Conditional information and structural identifiability are different from finite-sample recovery

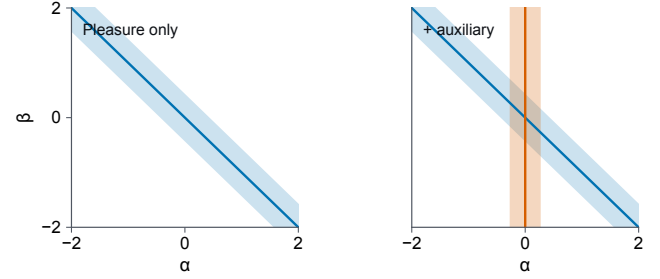
a Same input, two derived labels



Already conditioned on X: no new evidence

$$Q = g(X) \implies I(P; Q | X, D) = 0$$

b A new measurement resolves a direction



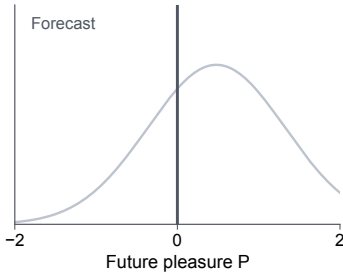
$$P = \alpha + \beta + \epsilon_P, \quad Z = \alpha + \epsilon_Z$$

Figure 8. Information requires an additional constraint. (a) A deterministic score contributes no conditional information once its input is already known, although it can help a restricted predictor represent that input. (b) Pleasure sensitive only to $\alpha + \beta$ leaves an unresolved direction; a separate noisy observation of α can constrain it. Bands are geometric teaching sketches, not fitted posteriors. Structural identifiability does not establish finite-sample recoverability.

An incorrect relation can increase prediction error.

Exact Gaussian risk | The true covariance is +0.60 in every column; only the forecasting rule changes

a Ignore Z

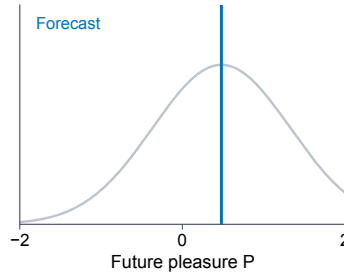


Gain: +0.00



Expected squared error: 1.000

b Correct relation

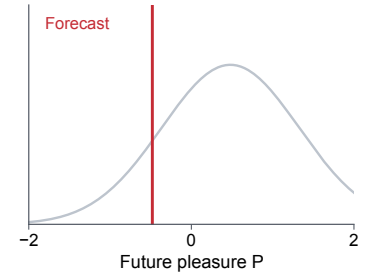


Gain: +0.48



Expected squared error: 0.712

c Wrong sign



Gain: -0.48



Expected squared error: 1.864

$$R(a) = v_P - 2ac + a^2v_Z$$

Grey curve: true $P | Z = 1$. Bars: average risk over Z .

Figure 9. The same world can punish a wrong update. The grey conditional density is identical in all three panels. Lines indicate forecasts after $Z = 1$; bars show expected squared error averaged over the joint distribution of P and Z . The true covariance remains +0.60. Using gains 0, +0.48 and -0.48 gives risks 1, 0.712 and 1.864, respectively. A narrower assumed distribution does not guarantee more accurate prediction.

Conditional dependence determines whether transfer is possible. In the Gaussian example, uncertainty decreases by c_{PZ}^2/v_Z under the specified model (Rasmussen and Williams, 2006). Both positive and negative relations can help. A duplicated transformation contributes no new information once its input is known, although it can help a restricted predictor use that input.

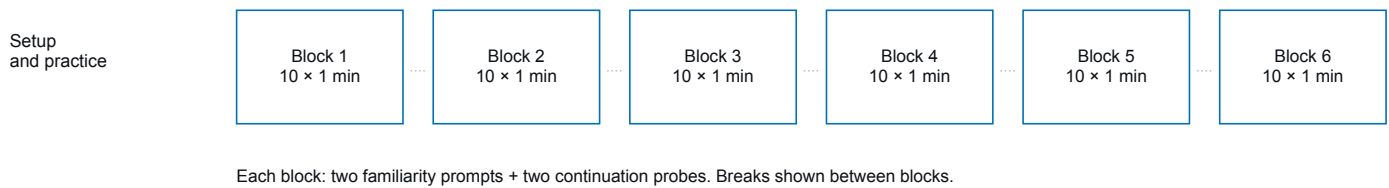
The corresponding general statements are expectations over possible observations. They do not guarantee a narrower interval after every response, nor accuracy under a wrong likelihood. Supplementary Figure S3 states these limits explicitly. Predictive evaluation is necessary even when every update is mathematically well defined.

6. Collect a common listening trajectory

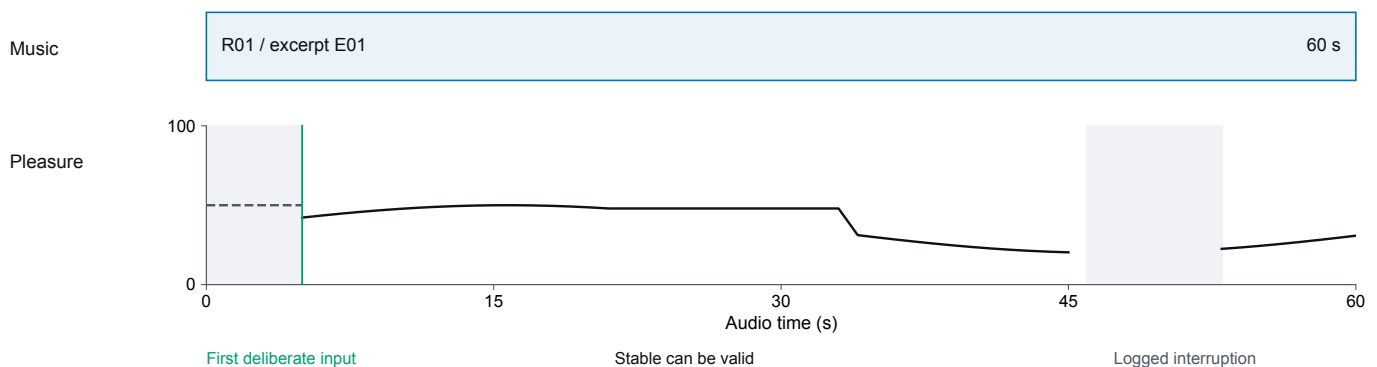
One listener, one trajectory, three kinds of response.

Proposed protocol; illustrative response trace | 60 min scored music, plus practice, probes and breaks

a See the whole session



b Open one scored excerpt: participant actions and event logging



c After selected excerpts: auxiliary probes outside scored playback



Probe allowance $\approx$ 8 min is provisional. Pilot duration, instructions and coverage before recruitment.

Figure 10. The participant's schedule and the analysis clock. (a) Six blocks contain ten one-minute scored excerpts each. Two familiarity reports and two continuation choices follow selected excerpts in each block. (b) The schematic pleasure trace distinguishes untouched initialization, deliberate reporting, stable responses and logged interruption. (c) Auxiliary probes occur outside scored playback. Durations, category anchors, probe matching and breaks require piloting; the eight-minute probe allowance is provisional.

The initial study fixes the stimulus schedule and task exposure. Every virtual model receives the same listening trajectory and is assessed against the same future pleasure reports. Sixty one-minute excerpts are arranged in six blocks. Scored playback is separated from auxiliary probes, practice and breaks. The proposed probe count balances repeated observations with burden; piloting must establish that participants can complete it reliably.

Familiarity refers to knowledge before the session. Category anchors should distinguish recognition from confident identification without assuming equal spacing. The continuation task measures

which of two alternatives the listener expected, not which they liked more. Alternatives require matched duration, loudness and plausibility, with order randomized and recorded.

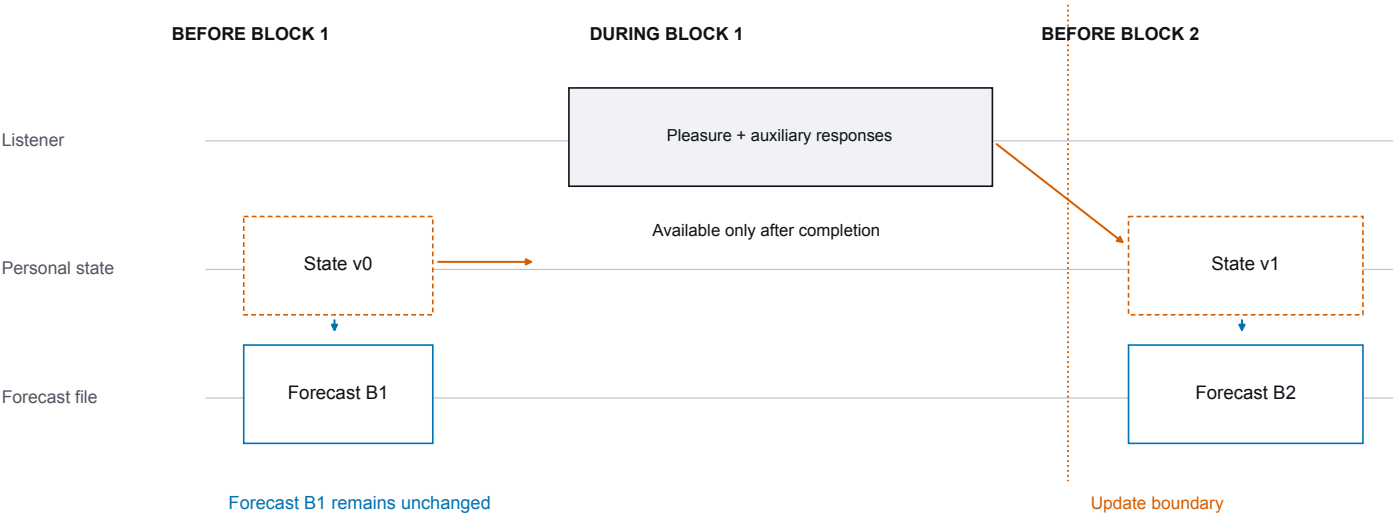
Raw timestamps use the audio clock and record pause, seek, focus and device interruptions. An untouched slider is not a report; a stable slider after deliberate engagement need not be missing. Scoring rules are finalized from pilot data and applied identically to all models. Reporting and probing may alter listening, so the resulting inference is about this explicit measurement setting.

7. Commit predictions and control information access

Save the prediction before collecting the response.

Proposed persistent memory | Forecast files are immutable for scoring; responses update only later blocks

a Follow the block boundary



b Saved record State version · uncertainty · observation IDs · data cutoff · code / configuration hash

Figure 11. A personal memory must record what was known and when. Before a block, the model saves immutable predictions linked to its state and configuration. Only completed observations can enter the next update. Raw event records remain available for audit; updating state does not rewrite an earlier forecast. This persistence layer is proposed. The inspected memory-labelled engine interface does not itself accept or store a listener’s response history.

The experiment changes information access, not the listening trajectory.

Six virtual models | Same listener, music, reports and task burden; primary comparison R versus C

a One event record feeds every model



b Information access and the link into pleasure prediction

Model	History	Past pleasure	Auxiliary data	Cross-task link
A Population	No	No	No	No
B History, fixed	Own	No	No	No
C Direct feedback	Own	Own	No	No
C-perp Separate tasks	Own	Own	Own	No
R Relational	Own	Own	Own	Yes
R-yoked Control	Own	Own	Matched other	Yes

Match initial pleasure predictions and fitting budgets. Yoking must preserve compatible probe conditions.

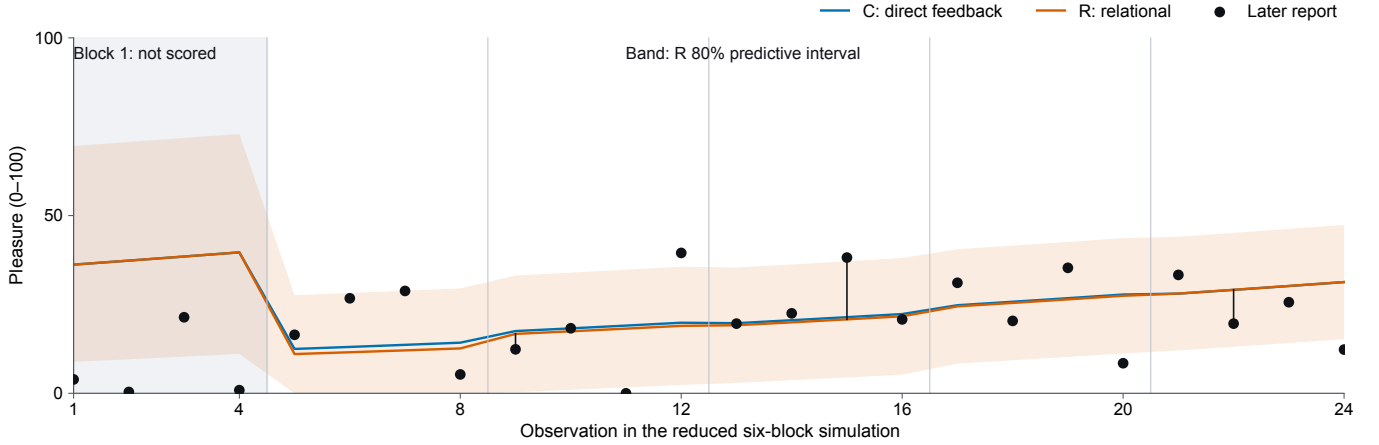
Figure 12. Information controls share one listening record. A is population-only; B uses initial history without updating; C uses history and direct pleasure; C_⊥ models tasks separately; R permits the specified cross-task relation; R_{yoked} substitutes compatible auxiliary reports from another listener. B, C, C_⊥ and R share the initial pleasure marginal. Capacity and development fitting budgets are matched. These are virtual models evaluated on the same responses, not six participant groups.

8. Define success on observations that were still in the future

Future prediction error is the primary test.

Synthetic reduced-model illustration | Predictive medians; real study uses a frozen common time mask

a Compare committed medians with later reports

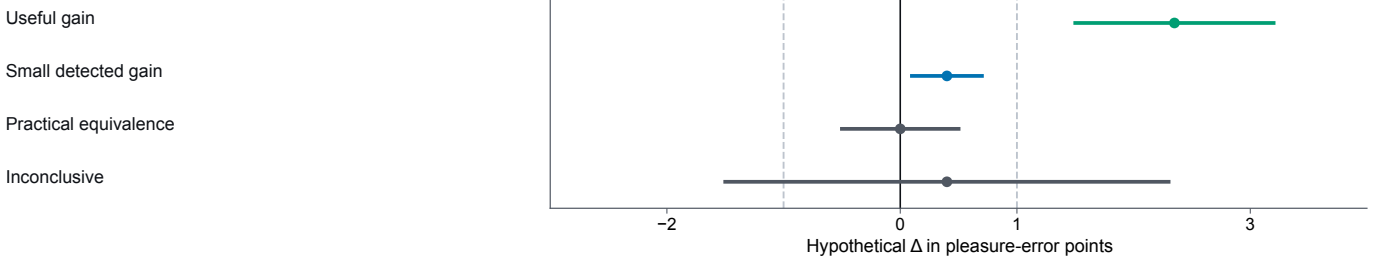


b Aggregate without treating time points as independent people



$$e_{um} = \frac{1}{|\mathcal{B}_u|} \sum_{b \in \mathcal{B}_u} \frac{1}{|V_{ub}|} \sum_{t \in V_{ub}} |y_{ut} - \hat{y}_{utm}|, \quad \Delta = \frac{1}{N} \sum_u (e_{uC} - e_{uR})$$

c Interpret the interval, not only its sign



Interval examples are not results. The illustrated practical threshold $\delta^* = 1$ is not a preregistered choice.

Figure 13. Score a forecast only against later data. (a) The first synthetic listener, selected by index before plotting, supplies six blocks of four observations. C and R nearly coincide; no separation is exaggerated. The shaded band is R's central 80% predictive interval. (b) The planned study uses common valid times, equal block weights and equal listener weights. (c) Hypothetical confidence intervals separate useful gain, small gain, equivalence and uncertainty. The illustrative $\delta^* = 1$ is not a chosen study threshold.

The primary contrast is R versus C: does auxiliary evidence add accuracy beyond direct pleasure feedback? Positive Δ means lower absolute error for R. Predictive medians are scored because absolute loss is minimized by a median. Blocks 2–6 form the evaluation interval; block 1 supplies initial observations.

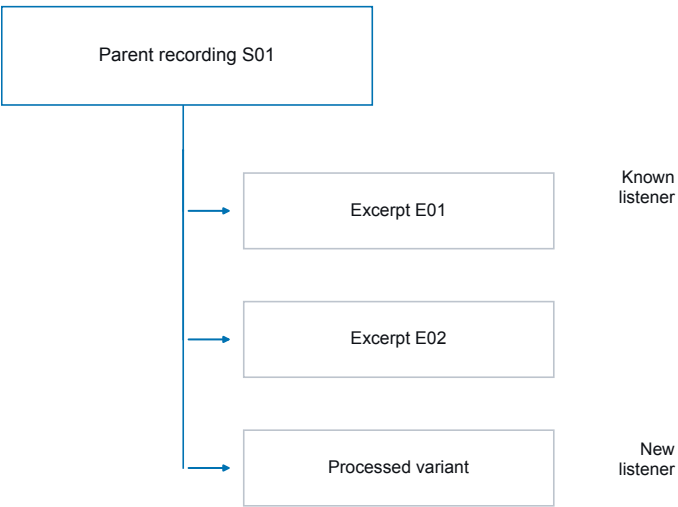
A participant-resampled confidence interval preserves each person's paired errors and internal time dependence. Practical usefulness additionally requires a prespecified δ^* in pleasure-error units. Secondary checks assess predictive coverage, width and a proper distributional score; lower median error cannot excuse unjustified confidence.

9. Protect the comparison from shared music and repeated testing

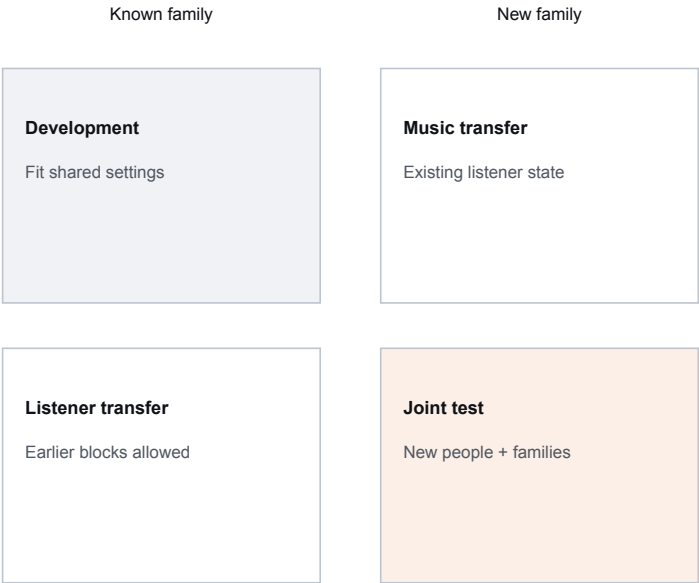
A new listener and a new recording are separate generalization tests.

Proposed partition contract | Familiarity reported by a listener is not the same as development-set membership

a Keep related recordings together



b Cross listener and source-family identity



Different file hashes; one declared family.

c Freeze the complete pipeline before evaluation

Coordinates	Normalization	Lags	Likelihoods	Shared coefficients	Scoring rules
-------------	---------------	------	-------------	---------------------	---------------

No test-set arrow returns to model selection. Covers and versions follow a documented family policy.

Figure 14. Generalization has two axes. (a) Excerpts, transformations and near-duplicates remain with their recording family. (b) The four cells distinguish shared-parameter development, music transfer, listener transfer and their joint test. New listeners may update from their own earlier blocks under the same schedule. (c) All preprocessing and shared choices are frozen before evaluation. Listener familiarity with a song and a recording's presence in development data are different attributes.

The primary participant-resampled interval describes the chosen music panel. Transfer to new music requires a source-family analysis and an independent test. A file hash is an identity check, not a rule for statistical independence: excerpts, processed copies and near-duplicate versions can have different hashes while sharing a parent recording.

All comparisons use the same eligibility mask, lag treatment, normalization and held-out partitions. A test set becomes development data once its results influence model choices (Dwork et al.,

2015). Partition creation, access dates and subsequent decisions belong in the study record.

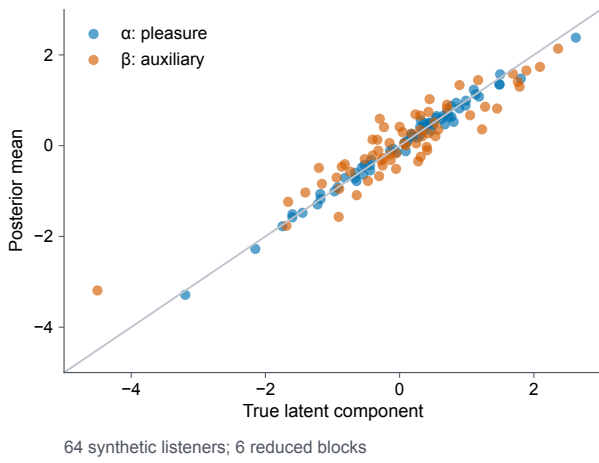
R versus C isolates the value of auxiliary information within the chosen representation. R^3 -only, simple history/running-mean and capacity-matched generic joint models are also needed to establish the specific value of MI's composites. Relational improvement alone would not settle that question.

10. Make failures visible before recruiting listeners

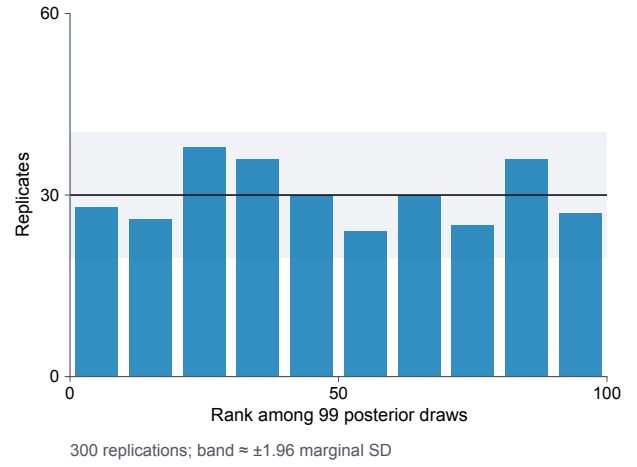
Numerical correctness, recoverability and predictive adequacy differ.

Executed reduced-model simulations | 2 latent states; independent synthetic responses; not full-protocol validation

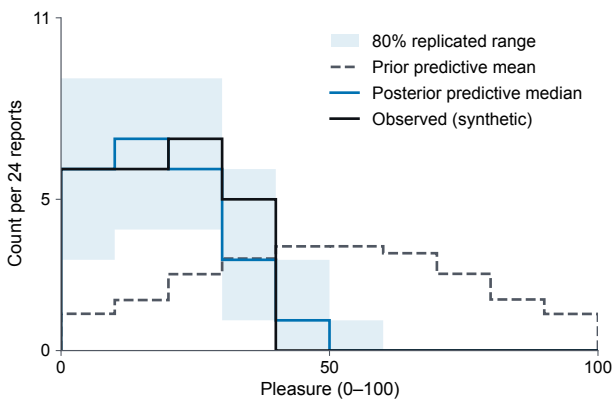
a Recover known personal states



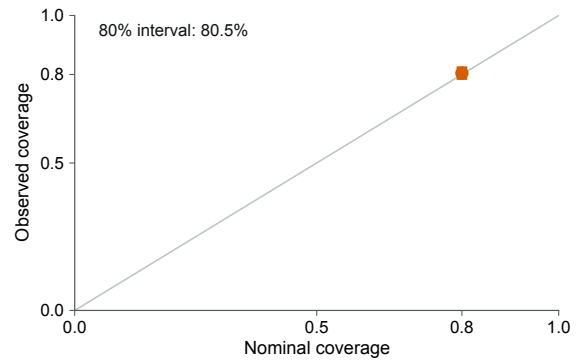
b Check posterior computation



c Check prior and posterior predictions



d Check future predictive coverage



Passing these small-model checks does not establish calibrated inference for real continuous listening responses.

Figure 15. Executed checks of a reduced model. (a) Posterior means versus generating states for 64 synthetic listeners. (b) Ranks of α among 99 posterior draws across 300 fresh simulations; the grey band is an approximate pointwise count reference, not a simultaneous acceptance region. (c) Prior and posterior predictive counts for the first synthetic listener; shading is the pointwise central 80% range across 1,000 replicated datasets. This diagnostic conditions on all observed blocks. (d) Coverage of 80% intervals is 80.5%; the error bar is ± 1.96 listener-level standard errors. All tasks use independent conditional errors; this does not validate the complete continuous-response protocol.

The executable two-state model is a numerical demonstration, not a miniature human experiment. Its six blocks contain 24 pleasure observations in total. With fixed global parameters, direct and relational predictors differ only in access to the auxiliary likelihoods. Across 64 synthetic listeners, mean Δ is 0.035 pleasure-error points: a small model-dependent difference, not an estimate of an MI benefit.

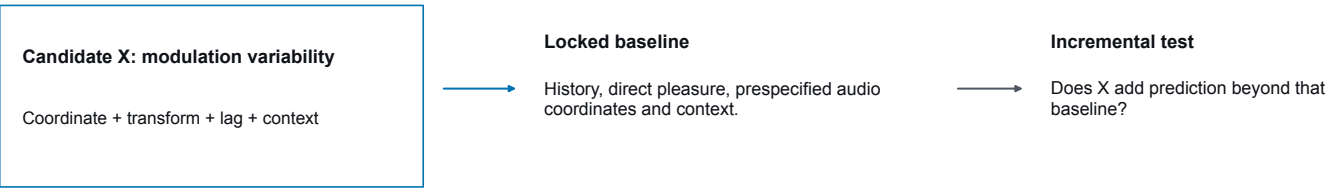
Recovery, rank diagnostics and predictive checks address distinct questions (Wilson and Collins, 2019; Gabry et al., 2019). Supplementary Figure S4 includes null and reversed-relation conditions. Full planning still requires serial dependence, reporting delay, uncertain history, missingness and participant-recording variation; no final sample size follows from these reduced simulations.

11. Admit unexpected dimensions without lowering the evidence standard

An unexpected coordinate becomes a finding only through a test.

Proposed discovery workflow | X is a candidate, not an observed MI-pleasure association

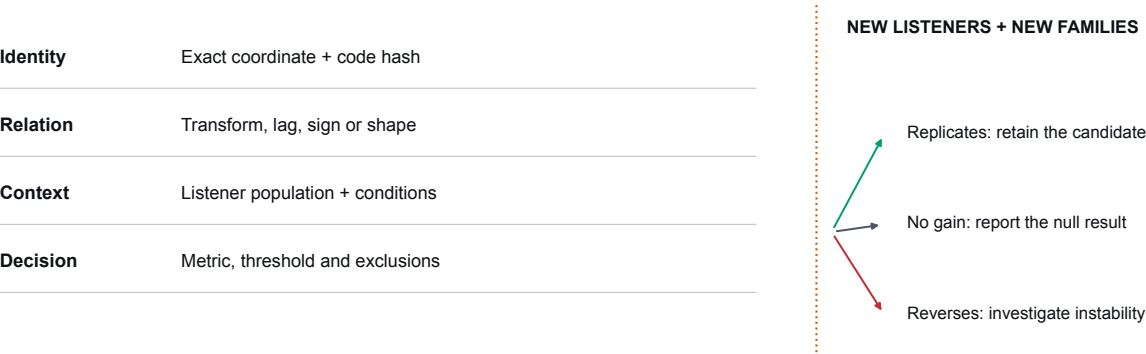
a Define exactly what is being searched



b Select only inside development folds



c Lock the candidate before independent replication



The search is hypothesis-generating. Replication supports association; a targeted intervention tests causality.

Figure 16. An unexpected dimension needs its own route to evidence. (a) Modulation variability is an example candidate, not an observed finding. Its incremental value is tested against a locked baseline. (b) Feature construction and tuning are repeated inside the training portion of every outer partition. (c) The selected definition, relation, context and decision rule are frozen before testing new listeners and new source families. Reversal and failure to replicate remain reportable outcomes.

A coordinate need not have a pleasure-related name to become scientifically useful. Its contribution must survive adjustment for the baseline, its own selection process and the context in which it was found. Large feature libraries already support systematic discovery (Fulcher et al., 2013; Fulcher and Jones, 2017; Denny et al., 2013). MI’s contribution would be the particular relation, its provenance and the experiment it motivates.

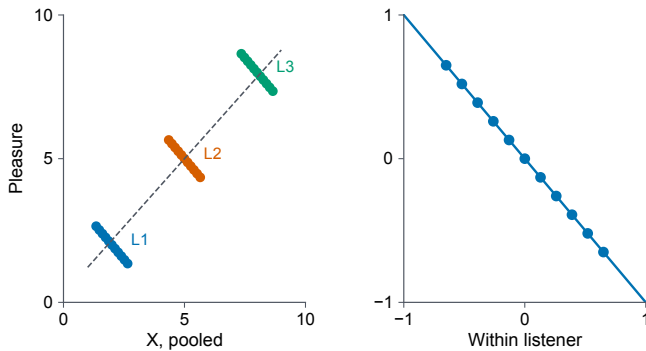
Selection includes transformations, lags, interactions and sub-group choices. Reusing the selected data for confirmation is circular (Kriegeskorte et al., 2009). Nested development, clustered stability checks and an independently frozen replication specification provide separate safeguards. A relation that reverses or fails must remain in the record.

12. Ask what an association actually describes

A plausible association can disappear under a better question.

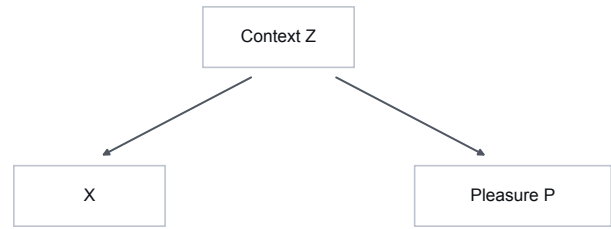
Synthetic / analytic counterexamples | These are not measured MI discoveries

a Whose relation is it?



Positive pooled; negative within each listener.

b Does a common context explain it?

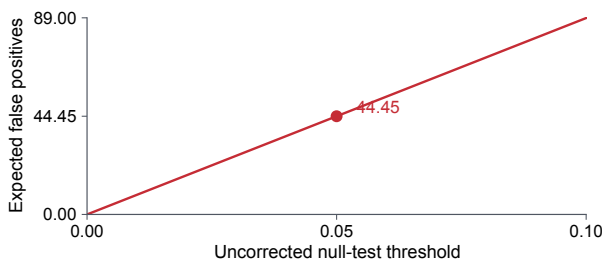


Marginal correlation: 0.50

After conditioning on Z: covariance 0

$$X = Z + \epsilon_X, \quad P = Z + \epsilon_P$$

c How many opportunities were searched?



Bayesian evidence also needs its prior.

Prior inclusion = 0.01; Bayes factor = 100

Posterior inclusion ≈ 0.503

889 outputs are not 889 independent constructs.

Figure 17. Three diagnostic counterexamples. (a) Synthetic listeners show a positive pooled trend and a negative within-listener relation. (b) Independent unit-variance context and noise terms give marginal correlation 0.5 but zero residual covariance after conditioning on context. (c) If 889 valid null tests each have size 0.05, their expected false-positive count is 44.45, without requiring independence. A Bayes factor of 100 with prior inclusion 0.01 yields posterior inclusion 0.5025. Neither calculation is a participant result.

A pooled association is not automatically a personal update rule (Fisher et al., 2018). Separate within-listener deviations from between-listener differences, and distinguish a candidate's incremental information from a shared acoustic or contextual cause. Nonlinearity also matters: zero Pearson correlation does not establish independence.

Multiplicity applies to the complete search. Sparse priors expose its assumptions; Bayesian analysis does not automatically correct

every adaptive search (Scott and Berger, 2010). Frequentist confirmation must use valid sampling units and a declared error criterion (Benjamini and Hochberg, 1995; Benjamini and Yekutieli, 2001). Stability selection can guide follow-up, but does not replace replication (Meinshausen and Bühlmann, 2010).

Even a replicated predictor remains an association. A targeted manipulation must assess the candidate alongside plausible alternatives and off-target changes before a causal interpretation is warranted (Peters et al., 2016).

13. Interpret prediction, discovery and learning separately

Different scientific claims require different evidence.

Interpretation map | Proposed protocol; no participant results or established benefit

a What the present design can test

Evidence obtained	Permitted conclusion	Still required
R beats C on later reports with appropriate calibration	Auxiliary information improved prediction under this panel.	Capacity controls; replication.
A frozen unexpected relation replicates on new people/music	A reproducible conditional association was identified.	Targeted manipulation for a causal claim.
No meaningful gain, poor calibration or worse error	The tested relation or model is unsupported or harmful.	Revise on development data; retain the negative result.

b Two distinct subsequent experiments

Does the listener learn?	Which question should be asked next?
Manipulate exposure or transition histories; measure expectation and pleasure separately.	Randomize static versus adaptive probe policies at an equal burden. Asking can change listening.

Model updating is not evidence of listener learning. Predictive association is not a demonstrated mechanism.

Figure 18. Keep the conclusion at the level of the evidence. Improved future prediction, a replicated association and a causal account require different tests. An absent or harmful transfer is informative about the specified model. Listener learning would require manipulated exposure or transition histories. Adaptive questioning would require randomized policies and matched burden because the act of asking can change subsequent listening. Neither follows from a personal posterior changing.

The central hypothesis is modest in scope but testable: an observation about another aspect of listening may improve a later pleasure prediction. MI makes the proposed path available for inspection. A researcher can trace its acoustic inputs, remove a relation or identify which response changed a personal posterior. Whether this organization improves prediction or discovery remains an empirical question.

Listener learning is a different target. Manipulated exposure or transition histories would test whether expectation changes and how that change relates to pleasure; computational memory models offer hypotheses for such experiments (Harrison et al., 2020).

Adaptive questioning is also subsequent work. Bayesian elicitation and goal-oriented design provide precedents (Guo and Saner, 2010; Attia et al., 2018). Equal-budget randomized policies are required because asking a different question can change the experience. Recommendation outcomes (Nonnemaker et al., 2025) do not by themselves establish a mechanism of pleasure.

14. Methods and implementation status

Source and numerical provenance

This edition pins the prior manuscript, reference typography, inspected engine and experiment contract by SHA-256. The selected graph audit covers 52 source files. The broader inventory comprises 84 C³ mechanisms and 889 outputs; these are computation counts, not independent constructs. Figure 3 uses archived R³ outputs, while Figure 4 executes the current temporal and HTP code on those arrays. It does not silently substitute archived output for a current end-to-end audio run.

The ordinal calculation uses numerical quadrature. Gaussian examples use closed-form moments and risk. Reduced simulations use seed 20261009, a joint grid of 81×81 points over $[-5, 5]^2$ and log-normalized likelihood updates. Posterior predictive quantiles integrate over the state distribution, including censoring at 0 and 100. An 81 versus 161 point grid check on the displayed listener changes posterior means by less than 2×10^{-10} ; this is a local numerical check, not universal convergence evidence. Supplementary Methods specify all likelihoods, designs and counts.

History and personal state

Where available, a consented history profile may shift the initial mean, $\theta_u \sim \mathcal{N}(B\mathbf{h}_u, \Sigma)$. History is incomplete evidence; a saved track, repeated exposure or playlist position is not a pleasure rating. History encoding, missing-history treatment and uncertainty are developed and frozen before evaluation. The initial state dimension and its observation links must be small enough to recover with the task budget. The two-state demonstration is not presented as the finalized study model.

Comparators and scoring

B, C, C_⊥ and R share the initial pleasure marginal. In C_⊥, auxiliary likelihoods update their own task states without changing pleasure predictions. Yoked controls preserve compatible item and probe conditions while breaking listener-specific auxiliary information. Identical development fitting budgets constrain capacity differences. Their definitive specifications belong in the frozen analysis package.

Eligible response times are defined before error comparison. Interruption masks and minimum block/participant coverage are

fixed during piloting. Pleasure targets are not imputed. Reporting delay, smoothing, normalization and feature selection are estimated on development data. Secondary inferential contrasts receive a prespecified Holm adjustment; descriptive analyses remain labelled as such. Participant resampling is primary, with source-family and crossed-effects sensitivity analyses for music generalization.

Development and recruitment

Posterior predictive checks must include scale boundaries, response persistence, lag and missingness, not only the appearance of a density. Simulation-based calibration concerns computation under an assumed model; it is distinct from fit to human data ([Stan User's Guide, v2.37](#)). The full protocol is not yet implemented or validated by the reduced model. Its state dimension, likelihood dependencies, recovery criteria and inference procedure remain to be finalized before preregistration.

Sample-size planning follows a smallest useful effect and a precision target ([Lakens, 2022](#)). Pilot ranges must feed a generative simulation of the actual schedule, including null and adverse relations, failures and attrition. Freeze sample size, stopping rule, exclusions, δ^* , partitions, masks, contrasts and simulation criteria before confirmatory recruitment. No ethics approval, preregistration or completed human collection is claimed in this draft.

Data and code availability

This edition includes editable vector figures, typeset equations, manuscript source, numerical inputs, simulation outputs and hash-based provenance. It contains no participant dataset. Future releases should retain permitted de-identified responses, stimulus manifests, partition identities, committed predictions and independent scoring code. Waveform permissions and private listening histories require separate access arrangements. The full specification and limitations accompany the technical supplement.

Author statements

Amaç Erdem designed the proposed study and is responsible for its scientific content. No external funding supported this manuscript. Human recruitment requires ethics review and informed consent, including withdrawal and listening-history protections. Claude Code and Codex assisted code development, drafting, document preparation and computational checks; they are not authors and do not replace author review.

References

- Attia A, Alexanderian A, Saibaba AK (2018). Goal-oriented optimal design of experiments for large-scale Bayesian linear inverse problems. *Inverse Problems* 34, 095009. [Source](#)
- Benjamini Y, Hochberg Y (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J R Stat Soc B* 57, 289-300. [Source](#)
- Benjamini Y, Yekutieli D (2001). The control of the false discovery rate in multiple testing under dependency. *Ann Stat* 29, 1165-1188. [Source](#)
- Bonilla EV, Chai KMA, Williams CKI (2007). Multi-task Gaussian Process Prediction. *NeurIPS* 20. [Source](#)
- Cheung VKM et al. (2024; online 2023). Cognitive and sensory expectations independently shape musical expectancy and pleasure. *Philos Trans R Soc B* 379, 20220420. [Source](#)
- Denny JC et al. (2013). Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nat Biotechnol* 31, 1102-1111. [Source](#)
- Dwork C et al. (2015). The reusable holdout: Preserving validity in adaptive data analysis. *Science* 349, 636-638. [Source](#)
- Fisher AJ, Medaglia JD, Jeronimus BF (2018). Lack of group-to-individual generalizability is a threat to human subjects research. *PNAS* 115, E6106-E6115. [Source](#)
- Fulcher BD, Jones NS (2017). htcsa: A Computational Framework for Automated Time-Series Phenotyping Using Massive Feature Extraction. *Cell Syst* 5, 527-531.e3. [Source](#)
- Fulcher BD, Little MA, Jones NS (2013). Highly comparative time-series analysis: the empirical structure of time series and their methods. *J R Soc Interface* 10, 20130048. [Source](#)
- Gabry J et al. (2019). Visualization in Bayesian workflow. *J R Stat Soc A* 182, 389-402. [Source](#)
- Guo S, Sanner S (2010). Real-time Multiattribute Bayesian Preference Elicitation with Pairwise Comparison Queries. *AISTATS, PMLR* 9, 289-296. [Source](#)
- Harrison PMC et al. (2020). PPM-Decay: A computational model of auditory prediction with memory decay. *PLoS Comput Biol* 16, e1008304. [Source](#)
- Janata P (2009). The neural architecture of music-evoked autobiographical memories. *Cereb Cortex* 19, 2579-2594. [Source](#)
- Kriegeskorte N et al. (2009). Circular analysis in systems neuroscience: the dangers of double dipping. *Nat Neurosci* 12, 535-540. [Source](#)
- Lakens D (2022). Sample Size Justification. *Collabra: Psychology* 8, 33267. [Source](#)
- Mas-Herrero E et al. (2014). Dissociation between musical and monetary reward responses in specific musical anhedonia. *Curr Biol* 24, 699-704. [Source](#)
- Meinshausen N, Bühlmann P (2010). Stability selection. *J R Stat Soc B* 72, 417-473. [Source](#)
- Nonnemaker A et al. (2025). Feedback-Driven Gradual Discovery for Expanding Musical Preferences. *RecSys*, 605-609. [Source](#)
- Peters J, Bühlmann P, Meinshausen N (2016). Causal inference by using invariant prediction: identification and confidence intervals. *J R Stat Soc B* 78, 947-1012. [Source](#)
- Rasmussen CE, Williams CKI (2006). *Gaussian Processes for Machine Learning*. MIT Press. Chapter 2. [Source](#)
- Scott JG, Berger JO (2010). Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem. *Ann Stat* 38, 2587-2619. [Source](#)
- Stan Development Team (v2.37). *Stan User's Guide 2.37: Posterior and prior predictive checks; simulation-based calibration*. [Source](#)
- Turner BM et al. (2013). A Bayesian framework for simultaneously modeling neural and behavioral data. *NeuroImage* 72, 193-206. [Source](#)
- van den Bosch I, Salimpoor VN, Zatorre RJ (2013). Familiarity mediates the relationship between emotional arousal and pleasure during music listening. *Front Hum Neurosci* 7, 534. [Source](#)
- Wilson RC, Collins AGE (2019). Ten simple rules for the computational modeling of behavioral data. *eLife* 8, e49547. [Source](#)