Learning, expectation and musical pleasure: an integrated protocol for evaluating Musical Intelligence

Amaç Erdem | amace@bu.edu | ORCID 0009-0000-1711-0036

Abstract. This study has two coequal aims: to investigate how listeners update expectations and experience musical pleasure, and to test the independent predictive contribution of a first, literature-anchored version of Musical Intelligence (MI). Personal calibration samples previously known music, model-recommended music and unfamiliar music, recording familiarity and predicted preference as crossed attributes. A mobile interface collects one response dimension at a time. Controlled exposure tests listener updating; new musical passages test personal prediction using continuous pleasure and excerpt-mean pleasure. Acoustic and MI learners receive the same admissible history, outcomes and fitting resources. A practically meaningful MI contribution requires an increase in correlation greater than 0.10, assessed separately for within-passage trajectories and between-excerpt means. Nine candidate mechanisms are examined individually and jointly without treating their names as biological measurements. A feasibility pilot qualifies measurement, model recovery and sample size before a frozen confirmatory study. Measurement engineering and detailed analysis contracts accompany this protocol; an independent LLM-analysis benchmark remains a future study. No participant results are reported.

Author protocol · 12 October 2026 · Supersedes the study and measurement specifications dated 11 October 2026 where listed in Supplement S1. Not a completed preregistration or empirical report.

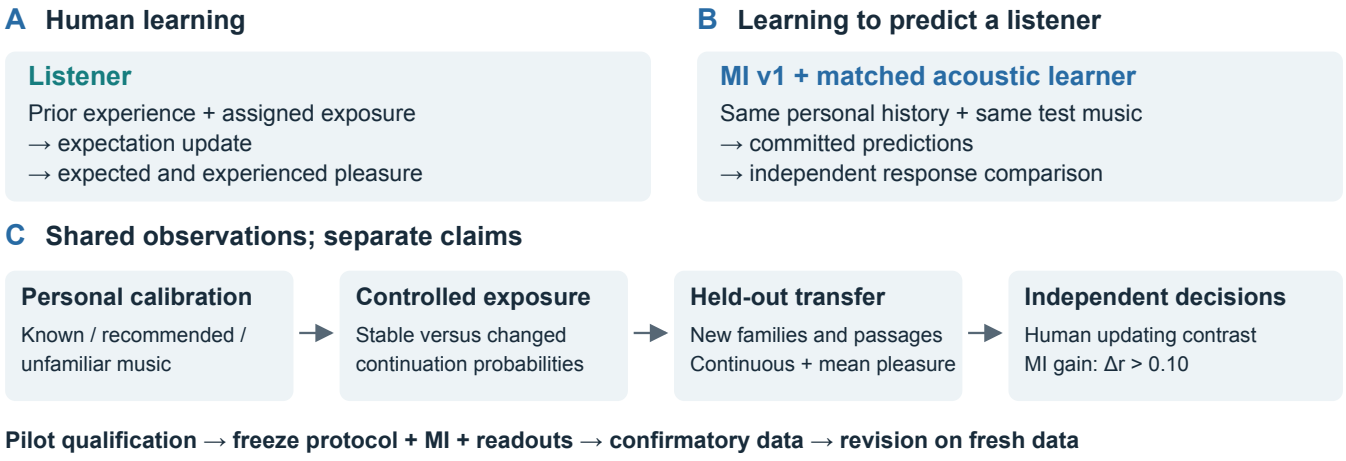


Fig. 1. Two aims, one experimental programme. Human learning and MI personalization have different update rules and separate evidence. Personal calibration precedes controlled exposure and held-out transfer. Each predicted response is committed before its target is revealed. All panels are design schematics.

1. Scientific rationale and claims

Learning and prediction offer a framework for relating musical structure, prior experience and pleasure[1, 2, 3, 4]. Their coordinated study requires distinguishing a listener who changes an expectation from an algorithm that learns to predict that listener. This protocol gives both questions equal scientific standing. Evidence of human updating does not require MI to outperform alternatives; successful response prediction does not uniquely identify a psychological or neural mechanism.

The human aim concerns experimentally induced updating and its relationship to expected and experienced pleasure. The computational aim concerns the added information supplied by MI's fixed transformations beyond an equally informed acoustic learner. Nine modules provide testable operational proposals, combining literature motivations with MI-specific input choices, weights, temporal summaries and interactions. Citation coverage is not a percentage of copied equations or independently established novelty.

Two confirmatory families are proposed: a randomized exposure-dependent updating contrast, and a joint MI incremental-prediction claim across two outcome scales. Prespecified secondary analyses relate expectations to pleasure, quantify personalization and evaluate mechanism contributions. An entropy–surprise interaction requires a separate, sufficiently rich expectation assay; the low-burden core does not claim to identify that interaction from pleasure alone.

This is a first-version evaluation. Model revision follows the primary test and requires new independent evaluation. Methods, apparatus and computational provenance are specified in the accompanying Supplement, which also preserves the earlier figures and records every substantive amendment.

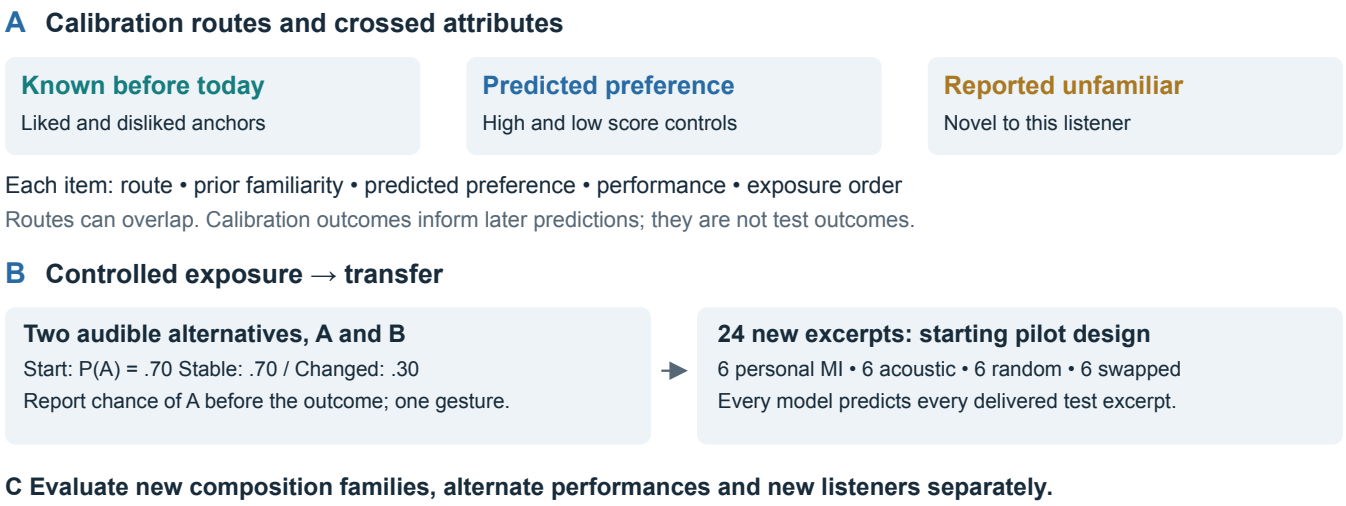


Fig. 2. Crossed calibration and transfer. A, Three recruitment routes populate the calibration pool; knownness and predicted liking remain separate attributes. A recommended track can be unfamiliar. **B,** Controlled exposure tests updating; new composition families test transfer. **C,** Distinct holdouts answer different generalization questions. Repeated excerpts from one family never cross a new-family split.

2. Participants, calibration and listening

Cohort. The planning cap remains 1,000 distinct eligible registrations, including an initial 60-registration feasibility pilot excluded from confirmation. The cap is not an achieved or powered sample. Adults aged at least 18 provide study-specific consent; a phone, compatible browser and headphones must pass delivery and comprehension checks[5, 6]. Record training, listening background, prior recognition, hearing/device information and missingness without inferring cognitive traits from demographics. Start with a rights-cleared classical catalogue; broader repertoire requires explicit sampling and validation. Recruitment, compensation, institutional responsibility and approvals precede invitations.

Personal calibration. Three routes supply candidates: music known before today, music recommended from the information available so far, and music reported as unfamiliar. Include known disliked works and high/low predicted-preference controls. Route, prior familiarity, performance recognition and model score are distinct fields. A pilot compares 12, 18 and 24-reference designs; the final dose and minimum usable ratings are frozen before confirmation. Within each dose, seed references precede recommendation; predictions for subsequent calibration items use only earlier information. Recommended and unfamiliar routes may overlap, but each exposure is counted once. A short preview is an exposure, never evidence of prior familiarity.

References initialize a regularized personal readout; their outcomes are not test outcomes. Keep selection probabilities, search effort, repetitions and incomplete intake. Population priors provide a fallback when a dimension has little personal information. Calibration coverage, rather than a requirement to know exactly 20 works, determines readiness under the pilot-frozen rule. Musical and demographic coverage must be reported; a larger dataset does not automatically repair selective recruitment.

Controlled learning. Listeners hear two demonstrated continuation alternatives, A and B, in initially stable contexts. A single vertical gesture records the predicted chance of A before the outcome; the complementary probability is defined as one minus that response. Initial probability schedules are (.70,.30). Random assignment either retains this schedule or changes it to (.30,.70), with alternative identity counterbalanced. The starting pilot schedule has 16 training exposures and 32 probes, with change after probe 16; shorter schedules may replace it only through prespecified feasibility and recovery criteria. Expected enjoyment and experienced pleasure are sampled in designated probes, one dimension per screen. No continuous pleasure task runs simultaneously with probability reporting.

Transfer. The starting design retains 24 previously unpresented 25-second excerpts, excluding calibration composition families. Six are selected by each of four policies: personal MI, an equally calibrated acoustic model, matched random selection and MI with another person's matched profile. All prediction models score the same delivered excerpts. Mask policy labels and interleave order; model scores never appear during listening. Record outside-study familiarity after playback. Four paired-performance comparisons remain secondary. Counts and duration are pilot-qualified; they are planning specifications, not completed observations.

Pilot and sample planning

Pilot decisions. The initial 60-registration pilot qualifies the mobile gesture, the three-source calibration, probability comprehension, reporting delay, trial timing, data retention and the ability to recover the specified effects. It compares calibration doses and candidate shorter controlled-task schedules using completion, reliability and recovery criteria, not whichever produces the most favourable MI result. The pilot determines the final number of references, usable-reference rule, task counts, response masks, lag treatment, readout basis, regularization and confirmatory sample. These choices are recorded together in a dated amendment before recruitment of the confirmatory cohort. If a required endpoint cannot be measured or powered, redesign and re-pilot it; do not silently promote a convenient substitute.

The former 55–75-minute session estimate and one-point absolute-error target are superseded as primary planning commitments. Session duration will be measured for the mobile design. Retain the 1,000-registration cap as an operational constraint unless separately amended. Simulate candidate analysable cohorts of 300, 500, 700 and 900 with crossed family coverage, temporal dependence, bounded ratings, calibration burden and missingness. Require adequate error control and target at least 90% power for a prespecified alternative exceeding the 0.10 boundary by a justified margin; 90% power cannot be required at the null boundary itself. The pilot-exit report must state that alternative and achievable precision, not infer adequacy from the cap.

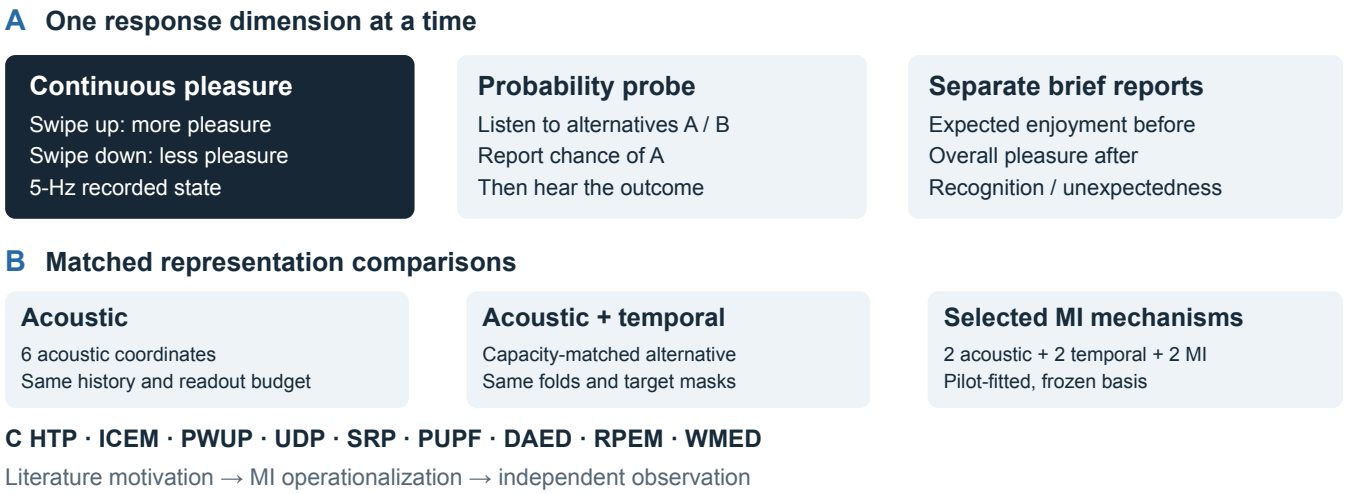


Fig. 3. Measurement and inspectable predictions. A, One response meaning per screen; continuous pleasure and probability probes occupy separate blocks. A dark interface remains active. **B,** Acoustic, temporal and mechanism information feed matched readouts with the same personal history. **C,** Literature motivation, implemented operation and observed outcome remain separate records. Nine modules are evaluated; no causal connections are implied by their grouping.

3. Measurement and model specification

Low-demand reporting. During transfer, upward/downward finger movement increases/decreases a 0–100 pleasure value. The active screen can be visually dark; an operating-system-locked screen is not the collection mode. Freeze gesture sensitivity, anchors, initial-value handling and practice requirements in the pilot. Record the current state at 5 Hz, synchronized to delivered audio; a stationary value is a valid response, whereas a suspended page or lost measurement interval is not. Retain raw events and masks. Timing resolution does not establish physical accuracy[6].

Two primary computational targets are the aligned within-excerpt pleasure trajectory and the mean of its valid, prespecified time samples. A separate immediate overall pleasure rating is secondary: it is not assumed equal to a trace mean or peak. Recognition and brief unexpectedness reports follow playback. The pilot estimates reporting delay, repeat reliability, scale use, completion burden and reactivity through counterbalanced reporting-versus-passive listening. The confirmatory response-lag rule is common to all models and never selected on their test correlations. The task measures listening while reporting; it does not observe an unconstrained internal state.

Human updating. Let $q_{it}(A)$ be the pre-outcome probability report. For the initial dominant alternative, compare the mean of probes 25–32 with probes 9–16 under the starting schedule. The primary human effect is the changed-minus-stable difference in this change; a negative value supports updating. Use participant and motif-family effects, realized exposure and counterbalanced identity. If the pilot shortens the task, define corresponding non-overlapping before/after windows in the amendment. Expected enjoyment V and experienced pleasure Y remain distinct reports. Their associations and later changes are secondary, with acoustic/context and exposure controls; $Y-V$ is a signed value discrepancy, not a neural reward-prediction-error measurement[7, 8].

MI v1. Freeze the engine, preprocessing, catalogue, module outputs and global learning specification before confirmation. HTP, ICEM, PWUP, UDP, SRP, PUPF, DAED, RPEM and WMED constitute the question-specific module set[9]. The local audit covers 103 output coordinates across these nine modules, not 103 independent psychological constructs. Acoustic novelty, spectral entropy and named uncertainty channels are not automatically a listener's probability-defined uncertainty. Known-recording response prediction may use prepared full audio; any time-local anticipation claim requires past-only support throughout all dependencies.

Matched learners. Compare population, calibration-only and earlier-feedback-updated readouts for both acoustic and MI representations. Every comparator receives identical allowed responses, background, trial order, outer splits and tuning resources. Retain the initial six-coordinate ridge specification: two pilot-fitted components from each acoustic/temporal/mechanism group versus six acoustic components, with shrinkage in $\{.1, 1, 10, 100\}$. Separate temporal and mean readouts use their appropriate targets. Capacity-matched temporal and nonlinear alternatives qualify representation claims. Personal-mean, order/repetition and swapped-profile controls test specificity. All transforms are trained inside eligible folds; new-listener test participants supply only their permitted personal calibration and earlier responses.

Acquisition and quality

Prepared acquisition. Compute each approved audio representation once per version. A personal preparation stage fixes calibration state, selection policy, trial schedule and initial predictions; the phone presents prepared stimuli and appends observations. No live MI, GPU or LLM inference is required for collection. Store stimulus family, performance, excerpt and exposure separately, with audio and software hashes, timing support, units, versioned consent and local event sequences. Separate audio-context time, browser-monotonic time and server receipt: upload latency is not response latency.

An acknowledged event is durably committed; identical retries return its receipt and conflicting duplicates are audited. Local queues retain unacknowledged events within fixed capacity. A replayed sound creates a new exposure; it never overwrites the first response. Visibility loss, stalled audio, device changes and interrupted intervals remain explicit. Chronological analysis commits a prediction using only eligible earlier outcomes, reveals the target and records the update. Offline replay is labelled as such and is not represented as a previously issued online forecast. Full numerical programs, transforms, masks and fold assignments are saved so that reproduction reruns the same calculation.

Quality and missingness. Starting transfer eligibility is at least 18 of 24 excerpts with at least 80% valid samples in the common scoring window; pilot changes require an explicit amendment. A constant human response remains valid for means and error metrics, although its within-excerpt Pearson correlation is undefined. The same target-based coverage rule applies to both models; a constant prediction on a variable target receives the prespecified zero-correlation performance convention and is reported as a prediction failure. Do not exclude people for low pleasure or disagreement with MI. Report endpoint-specific denominators, all-valid-observation sensitivity analyses, missingness by calibration route/background and complete participant flow. Do not impute primary pleasure outcomes or infer unobserved reports from an upload failure.

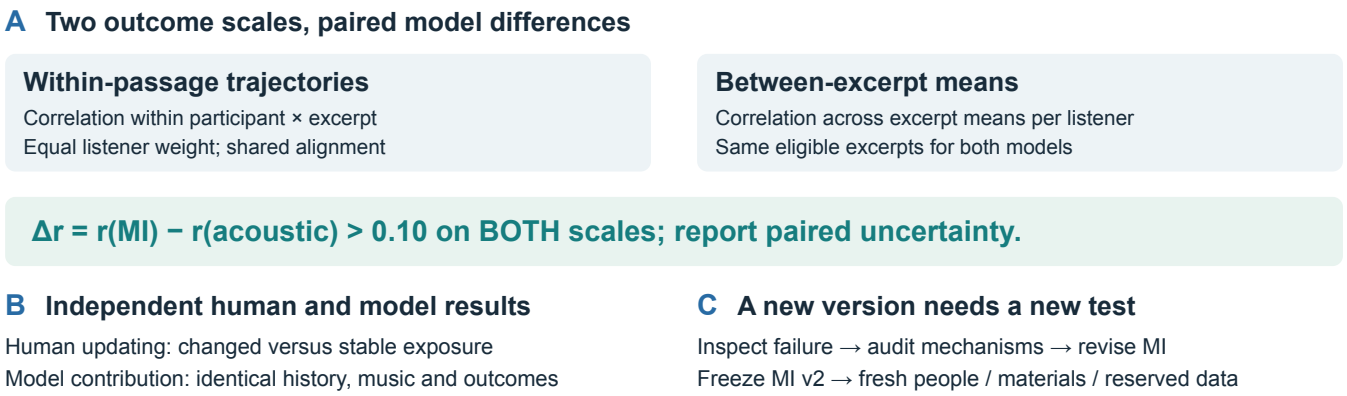


Fig. 4. Evaluation and revision. A, MI and its comparator predict the same targets. The practical threshold is a correlation difference of 0.10 for each outcome scale. **B,** Crossed participant/composition uncertainty accompanies effect estimates. Human and model findings are reported independently. **C,** Diagnostic ablations inform a revised version, which requires fresh evaluation. No performance values are shown.

4. Independent evaluation and decisions

Two correlation scales. For each eligible participant–excerpt pair, compute Pearson correlation between aligned continuous reports and predictions. Aggregate through Fisher transforms with equal participant weight and equal eligible-excerpt weight within participant; back-transform to obtain r_{time}. Separately compute each participant's correlation across excerpt means, then aggregate with equal participant weight to obtain r_{mean}. The Supplement defines masks, stable-response cases and minimum coverage. Pooling all frames or listeners into one correlation is not the primary analysis.

$$\Delta r_k = r_k(\text{MI}) - r_k(\text{acoustic}), \qquad k \in \{\text{time}, \text{mean}\}.$$

The primary MI contrast uses the same earlier-feedback-updated personal learner on both sides. A practically meaningful contribution requires $\Delta r_k > 0.10$ on both scales. Report point estimates and paired intervals even when this joint claim fails. A point estimate above the threshold is promising; the confirmatory claim requires threshold-exceeding evidence under the prespecified test. Gain on only one scale is a scale-specific result, not joint success. Mean absolute error, calibration, retrospective overall ratings and policy pleasure remain visible secondary outcomes.

Inference. Use 5,000 crossed participant-by-composition bootstrap replicates preserving observations and matching both models. Obtain one-sided tests against $\Delta r = 0.10$ through the prespecified centered bootstrap procedure, verified by simulation. The maximum of the two scale-specific p values tests the joint MI claim. Apply Holm control at $\alpha = .05$ across this claim and the human updating contrast. Report two-sided 95% intervals as descriptive uncertainty; use multiplicity-adjusted decisions. Secondary mechanism and personalization families receive separate false-discovery-rate control; exploratory outcomes require subsequent confirmation.

Generalization and model recovery. The primary transfer target is new composition families in new confirmatory participants after their permitted calibration. Report earlier-response prediction, new-family transfer, alternate-performance transfer and new-person generalization separately. Freeze outer participant/family splits before fitting; variants, movements, performances and excerpts of one family remain grouped. Shared acoustic sources and temporal dependence must survive resampling. Model recovery checks include acoustic-only signal, response-scale differences, smooth autocorrelation, no human updating, plausible updating, MI increments below/at/above threshold, missingness and attrition[10]. A mechanism name or a large frame count cannot substitute for identification or power.

Mechanisms. Compare acoustic, acoustic-plus-temporal and full selected-mechanism models, and leave-one-module-out variants with matched refitting resources. Shared upstream computations make ablation an information-contribution test, not unique biological attribution. Related mechanisms are also tested as groups; full-engine exploratory outputs cannot quietly enter the primary readout. Failure prompts checks of delivery, measurement, implementation, target alignment and source correspondence before mechanism revision. A null for an MI operationalization alone does not refute its motivating literature.

Interpretation and scope

Prespecified interpretation. The programme can find human updating with little MI increment, an MI increment with weak measured updating, both, or neither. Each outcome changes a different claim. A predictive failure can reveal a weak operational bridge, inadequate temporal support, an unsuitable target or insufficient precision. Only a sensitive test of a faithful operationalization can constrain the motivating theoretical claim. Preserve nulls, disagreements and viable alternatives. Changes to weights, inputs or mechanisms define a new MI version, evaluated on fresh people/materials or an untouched reserved set.

Publication and independent extensions. This integrated protocol absorbs the relevant measurement and inference contracts. The previous three compact papers and all 35 original figure sources remain in the editable archive. Supplement S1 distinguishes current requirements, pilot-set quantities, optional assays and superseded settings. The separate documentation/LLM benchmark retains its own development/evaluation tasks and primary contrast; it is neither a prerequisite nor an outcome of the human experiment. Future EEG/MEG/fMRI work requires independent measurements, synchronization and modality-specific observation models[11]; this mobile study does not measure dopamine, neural entrainment or working-memory capacity directly.

Ethics, access and declarations. Obtain appropriate institutional approval, rights for exact recordings, informed consent, withdrawal procedures and a compensation plan before recruitment. Contact records are separated from pseudonymous research data; pseudonymization is not anonymization. Release permitted materials, executable analysis, synthetic fixtures and suitably de-identified summaries under explicit access conditions[12]. Public drafts and interest registration do not establish a deployed experiment. Amaç Erdem develops MI and proposes this study; affiliation, funding and contribution statements require confirmation before submission. AI assistance supported drafting, literature organization and figure/document preparation, not empirical validation. No participant dataset, completed operational validation, achieved power or accepted preregistration is reported.

References

[1] Vuust, P., Heggli, O. A., Friston, K. J. & Kringelbach, M. L. Music in the brain. *Nature Reviews Neuroscience* **23**, 287–305 (2022). doi:10.1038/s41583-022-00578-5.

[2] Cheung, V. K. M. et al. Uncertainty and surprise jointly predict musical pleasure and amygdala, hippocampus, and auditory cortex activity. *Current Biology* **29**, 4084–4092.e4 (2019). doi:10.1016/j.cub.2019.09.067.

[3] Gold, B. P. et al. Predictability and uncertainty in the pleasure of music: a reward for learning? *Journal of Neuroscience* **39**, 9397–9409 (2019). doi:10.1523/JNEUROSCI.0428-19.2019.

[4] Mas-Herrero, E. & Marco-Pallarés, J. Predictive processes shape individual musical preferences. *Proceedings of the National Academy of Sciences USA* **122**, e2500494122 (2025). doi:10.1073/pnas.2500494122.

[5] Woods, K. J. P., Siegel, M. H., Traer, J. & McDermott, J. H. Headphone screening to facilitate web-based auditory experiments. *Attention, Perception, & Psychophysics* **79**, 2064–2072 (2017). doi:10.3758/s13414-017-1361-2.

[6] Bridges, D., Pitiot, A., MacAskill, M. R. & Peirce, J. W. The timing mega-study: comparing a range of experiment generators, both lab-based and online. *PeerJ* **8**, e9414 (2020). doi:10.7717/peerj.9414.

[7] Gold, B. P. et al. Musical reward prediction errors engage the nucleus accumbens and motivate learning. *Proceedings of the National Academy of Sciences USA* **116**, 3310–3315 (2019). doi:10.1073/pnas.1809855116.

[8] Salimpoor, V. N. et al. Anatomically distinct dopamine release during anticipation and experience of peak emotion to music. *Nature Neuroscience* **14**, 257–262 (2011). doi:10.1038/nn.2726.

[9] Erdem, A. Musical Intelligence: A Glass-Box Architecture for Computational Models of Music Listening. Author manuscript, 8 October 2026. Dated architecture record.

[10] Wilson, R. C. & Collins, A. G. E. Ten simple rules for the computational modeling of behavioral data. *eLife* **8**, e49547 (2019). doi:10.7554/eLife.49547.

[11] Kriegeskorte, N. & Douglas, P. K. Cognitive computational neuroscience. *Nature Neuroscience* **21**, 1148–1160 (2018). doi:10.1038/s41593-018-0210-5.

[12] Wilkinson, M. D. et al. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* **3**, 160018 (2016). doi:10.1038/sdata.2016.18.