SCIENTIFIC MANUSCRIPT

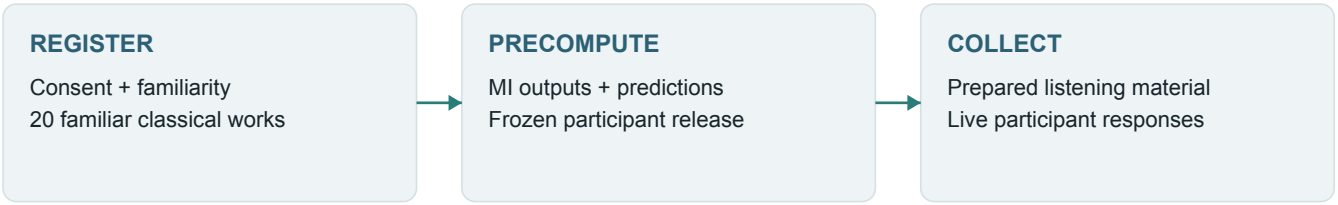
Musical Pleasure Across Interpretations of Familiar Classical Works

A prospective within-work prediction protocol for Musical Intelligence

Amaç Erdem | Independent researcher

10 October 2026 | Scientific protocol - Revision 4 / classical design

Abstract. This protocol tests whether Musical Intelligence (MI) improves person-specific prediction of pleasure differences between interpretations of the same familiar classical work. The recruitment target is 1,000 registrations, with screening, session readiness and completion reported separately. Each participant identifies exactly 20 distinct familiar composition families from a curated catalogue, confirms prior familiarity after a brief reference preview, and rates one reference performance per selected work during onboarding. A general listening-history export is not collected. The primary experiment compares two additional, previously unpresented performances of the same passage and instrumentation for each selected work. Shared MI representations and all participant-specific primary predictions are computed and committed before the session. During the experiment, the platform only presents prepared stimuli and collects responses. The primary endpoint is the improvement over a matched acoustic baseline in predicting within-work pleasure-rating differences. Arrangement comparisons and offline feedback replay are secondary extensions. The manuscript specifies stimulus identity, exposure accounting, participant flow, prediction cutoffs, missingness, analysis and launch gates. It reports no participant outcomes or validated MI effect.



Accepted scope. Classical repertoire; 20 familiar works rather than a general music history; different interpretations of the same passage; 1,000 total registrations; preparation before the session and response collection during it.

This revision supersedes the earlier history-driven experimental design for the proposed study [1]. The companion O-Lab systems protocol specifies implementation and operations. Previous PDFs remain historical records. This document is a design draft, not a completed preregistration or evidence of ethics approval, recruitment, deployed infrastructure, or adequate statistical power.

1 / SCIENTIFIC SCOPE

Question, estimand and interpretation

1.1 The primary question

Can a frozen personal MI predictor anticipate which of two interpretations of a familiar work a listener will enjoy more, and the magnitude of that difference, beyond an equally informed acoustic predictor? Work identity and nominal passage are matched within each comparison. The object of prediction is a subjective response to a particular performance, not an intrinsic pleasure value assigned to a composition.

Let y(i,w,B) and y(i,w,C) be participant i's post-excerpt pleasure ratings on a 0-100 scale for two test performances of work w. The observed contrast is d(i,w) = y(i,w,B) - y(i,w,C); the predicted contrast is the corresponding difference between predictions committed before session start. B and C are arbitrary recording IDs, independent of the randomized presentation order. They do not denote the earlier protocol's model conditions.

1.2 One primary comparison

The primary quantity is the participant-weighted reduction in mean absolute contrast error of the personal MI model relative to a personal acoustic baseline. A positive reduction favours MI. All eligible completed work pairs contribute; absolute pleasure predictions, ordinal pair preferences, uncertainty calibration and continuous responses are secondary. The primary comparison, estimator and exclusions must be preregistered before confirmatory data are accessed.

1.3 What natural recordings can establish

Different pianists also differ in instrument, room, microphone, mastering and sometimes score edition. Matching a composition does not isolate rubato or any neural mechanism. The primary result is predictive evidence conditional on this catalogue and recruitment sample. A causal claim about timing, dynamics or timbre requires a separately specified manipulation with other dimensions controlled.

Classical works need not have an authoritative original recording. We use the terms reference performance, alternative performance and arrangement. Familiarity with a melody or passage does not imply familiarity with the entire work or with the particular performer.

2 / RECRUITMENT AND SCREENING

Participants and twenty familiar works

Recruitment targets 1,000 unique, email-verified adult study registrations. Current language support, recruitment jurisdictions, compensation and the responsible institution are launch decisions. Consent is required before collecting musical background or screening responses. Record training years, current music-making, classical-listening frequency, age band, country, instruction language and listening setup, with optional fields explicitly marked.

The target is not 1,000 completers. Report registrations, consented participants, screening attempts, people reaching 20 familiar works, calibration completion, readiness, starts, completions and analysis eligibility. Withdrawals and technical failures remain separate accounting categories. A convenience sample is not made representative by its size.

2.1 Count compositions, not audio files

Each selected item points to a canonical composition_family_id, a specified movement and a recognizable passage. Different recordings, arrangements, excerpts or movements of one composition count once toward the required 20 families. A respondent may recognize the selected passage without recognizing the title. The interface may provide common titles and composer filters, but a name alone cannot establish familiarity.

2.2 Verification and incomplete screening

Play a short, licensed preview from reference performance A, then ask whether the participant knew this music before today. Responses are yes, unsure or no, followed by familiarity on a labelled 1-5 scale. Only explicit yes responses count. Familiarity intensity is retained as a covariate rather than silently imposing an unapproved cut-off. The exact screening passage, preview duration and permitted replay count are frozen after the pilot. Viewing and playback exposures are logged.

Screening ends when exactly 20 distinct qualifying works are selected. If fewer are found, offer additional catalogue candidates and saved progress; allow an honest pause or exit. The person remains registered but is not session-ready. Do not teach unfamiliar pieces and relabel them as pre-existing familiarity, force a yes response, invent entries, or reduce the threshold silently.

> The expectation that every recruit can identify 20 familiar classical works is a feasibility hypothesis. Measure its success rate, time burden and background dependence in the pilot. Report failures and withdrawals; do not treat universal coverage as an established fact.

3 / STIMULUS IDENTITY AND RIGHTS

Build a catalogue of performance families

The primary catalogue contains Western classical works with at least three usable performances of the same passage and instrumentation. A work can be offered in screening only after reference A and alternatives B/C pass rights, identity, audio-quality and alignment review. Begin with solo-piano comparisons to limit instrumentation differences. Expanding the classical tradition or ensemble type creates explicit catalogue strata rather than an unrecorded change in the target population.

Entity	Required distinction
Composition family	Composer and catalogue number; grouping for all movements, arrangements and derivative excerpts.
Movement / passage	Score edition, movement, measure range, repeat path and phrase-boundary annotation.
Performance / recording	Performer, instrument, performance date if known, recording session and original source.
Audio asset / excerpt	Waveform hash, source sample rate/channels, boundaries, gain, derivation chain and permissions.
Participant role	Reference A for screening/calibration; alternative B/C for confirmation; separate arrangement flag.

IMSLP lists candidate first-movement recordings of Beethoven's Op. 27 No. 2 by Paul Pitman (public-domain dedication), German Kitkin (CC BY 4.0) and Philip Addison Jones (CC BY 3.0), plus a guitar arrangement by Ramón León Egea (CC BY-SA 4.0) [2]. These are catalogue candidates, not audio-quality-verified experimental stimuli. One matching family is not evidence that a sufficiently broad multi-performance pool already exists.

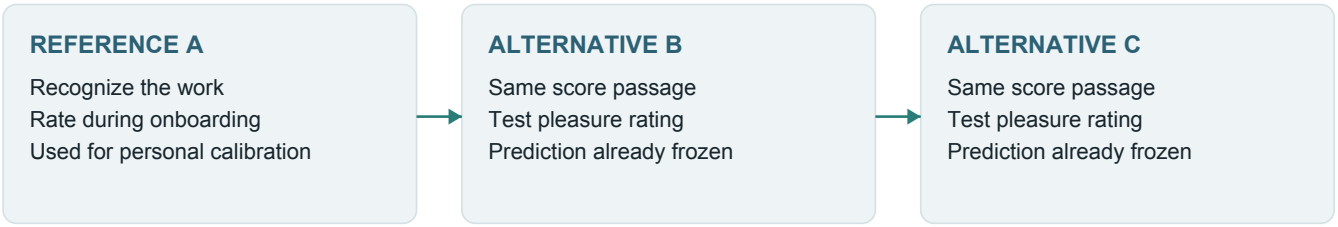
Musopen and IMSLP are discovery sources, not blanket permission for every file. Verify composition, arrangement, performance and recording rights, relevant jurisdictions, redistribution, excerpting, analysis, attribution and any expiry. Archive the specific permission evidence [2,3]. Exclude unclear assets. A public-domain composition does not make a modern commercial recording public domain.

MAESTRO supplies paired acoustic piano and performance MIDI recordings; ASAP adds score-performance alignment resources [4,5]. Pin versions and inspect the actual audio subset; MIDI-only entries are not recordings. Do not assume dataset annotations, source audio and rendered derivatives have identical permission conditions.

4 / EXPOSURE AND PERSONAL INPUTS

Separate recognition, calibration and testing

After selection, the participant hears the planned reference passage A for each of the 20 works and gives a 0-100 pleasure rating, work-familiarity rating and a separate yes/unsure/no report about having heard this exact performance before. Cannot-rate is a distinct missing value. These are baseline observations, collected before the confirmatory session. A reference rating is direct feedback; it must not be described as a passive listening history.



One composition family; three recordings; B/C order counterbalanced.

The same A performance is used for recognition and calibration. Both exposures and all replays are logged. Primary alternatives B and C are withheld from onboarding. Allocate eligible recordings to A/B/C through a seeded, versioned counterbalanced schedule; reference status is not attached permanently to the most famous or most preferred recording. Allocation must not use the participant's later test ratings.

All 20 familiarity confirmations are mandatory for selection completion. Missing calibration pleasure ratings do not erase familiarity or get filled with predicted responses. The model uses an explicit missingness policy and a shrinkage prior; report the number of observed calibration ratings. Readiness requires completion of all questions, with cannot-rate allowed, and a valid predeclared prediction path even when fewer numerical ratings are available.

4.1 Generalization claim

The primary task is transfer to new performances of personally familiar and calibrated works. It is not a test on entirely unseen compositions: A has already supplied a rating for each selected work. An additional unseen-composition analysis would require withholding whole composition families from personal calibration and is outside the primary claim. Development/test separation for population parameters remains mandatory.

Selections may reflect exposure, musical training and the catalogue offered. Record candidate presentation order and selection opportunities. Do not infer that a missing selection means dislike or that successful screening covers the listener's full musical experience.

5 / AUDIO AND PRECOMPUTATION

Prepare matched passages before listening

Choose the same score-defined musical passage in A, B and C. A provisional planning range is approximately 45-75 seconds per rendition, but musical boundaries take priority. Repeats, ornaments, omissions and edition differences require a documented alignment decision. Use beat/score alignment to identify correspondence and human review to confirm it. Do not select each recording's most dramatic segment independently.

Preserve performance timing. Never time-stretch recordings solely to equalize excerpt duration. Apply only a frozen constant-gain presentation rule to address gross level differences, with peak headroom and no dynamic compression. Store pre/post-gain audio properties, decoder settings and hashes. Analyze the exact presented waveform. Record technical noise, clipping, channel balance and ambience; reject unusable recordings using prespecified quality criteria.

5.1 Two preparation stages

First, compute shared MI outputs for every released catalogue recording and exact excerpt, keyed by waveform, preprocessing, engine commit and configuration. Second, after a participant submits the frozen 20-work profile and calibration responses, fit the allowed personal parameters and generate predictions for every scheduled B/C stimulus. Shared audio need not be recomputed for each person.

The participant release binds catalogue, consent, protocol, baseline snapshot, stimulus order, all asset hashes, output schema, model version, normalization, prediction values, uncertainty and input cutoff. The release timestamp must precede the first confirmatory presentation. A new baseline or a changed waveform invalidates the old readiness token and requires a new release.

> The live session never calls MI, an LLM or a GPU for scientific inference. Ordinary CPU playback, timing, authorization and data persistence remain necessary. All primary predictions remain frozen through the session.

Population parameters and feature choices are frozen using independent development data. New confirmatory responses cannot enter another participant's primary predictions during recruitment. The data-management contract must make this restriction auditable, not rely solely on researcher memory.

6 / PRESENTATION AND RESPONSE COLLECTION

The online session

Begin with current consent confirmation, headphone/device setup, comfortable volume and an unrelated practice trial. Use a quiet interface without model predictions, performer names or decorative motion. Show accessible instructions and allow breaks and withdrawal. Do not collect diagnostic health free text through the listening task.

The proposed primary session has 20 work trials, each presenting alternatives B and C for one familiar work. Counterbalance B-first/C-first within participant (10/10 where all trials are completed), randomize work order with a stored seed, and distribute trials into four blocks of five. Use the same neutral separation interval, provisionally 3 seconds, between alternatives. Exact timing and instructions are frozen after the pilot.

After each excerpt, collect a 0-100 pleasure rating with endpoints not pleasurable and extremely pleasurable; the control starts unanswered. After the pair, collect B preferred, C preferred or no preference, with cannot judge recorded separately. Ask exact-recording familiarity after the pleasure judgement to limit cueing. Prior exposure is recorded, not automatically treated as noncompliance. A sensitivity analysis can restrict to pairs whose two recordings were not previously known.

The primary endpoint uses post-excerpt ratings. A separate optional continuous-rating block may be developed later; it is not silently added to the primary trial, where motor and attention load would change the task. The current protocol therefore replaces the earlier fixed one-hour continuous-slider schedule. Estimate duration from the released excerpts and pilot timing before inviting participants.

A 60-second-per-excerpt planning example gives 40 minutes of test audio, plus questions, setup and breaks; onboarding reference ratings occur beforehand. This is an arithmetic scenario, not a promised duration. A participant may complete fewer trials and must never be forced to remain for a fixed hour.

Store actual audio onset/end, visibility, interruption, device events, response times, server receipts and immutable trial IDs. Pause at safe boundaries when delivery fails. Resumed or replayed material is marked as exposed; it cannot be presented analytically as an untouched first trial.

7 / ANALYSIS PLAN

Prediction models and locked comparisons

Model	Allowed information and role
M0: reference-only	Same baseline information; no performance-specific audio contrast. Its predicted B-C difference is zero. Sanity comparator.
M1: acoustic personal baseline	Frozen acoustic descriptors, familiarity, reference A ratings and the same declared background fields. Primary comparator.
M2: personal MI	M1 information plus a small prespecified set of MI outputs. Primary candidate; matched shrinkage and tuning budget.
M3: offline relational replay	Earlier completed trial responses only, evaluated chronologically after acquisition. Secondary; never labelled prospective online adaptation.

For M1 and M2, a proposed implementation is a hierarchical regularized rating model with a population mapping and a low-dimensional personal deviation. Twenty reference ratings cannot support unrestricted fitting of hundreds of personal coefficients. All hyperparameters, normalization and parameter counts must be frozen on independent development data. Predictive distributions are saved alongside point predictions.

Select a small theory-justified MI output set and document the exact output names, code paths and temporal aggregation. The full architecture is context, not an automatic feature set or an 89-mechanism test battery. Candidate mechanisms are not validated merely because their names refer to pleasure or reward. The final executable model and candidate set are preregistration gates, not decisions to optimize on the confirmatory responses.

7.1 Primary error comparison

For each complete pair, calculate the absolute error in predicted d(i,w) for M1 and M2. Average their paired error reduction within each participant, then average across participants so people with more complete trials do not dominate. Report the reduction in 0-100 scale points with a 95% interval using a prespecified crossed participant/composition-family resampling procedure. Preserve each B/C pair and the pairing of model errors during resampling.

Report ordinal preference accuracy with ties and cannot-judge handled explicitly, absolute rating errors, prediction calibration and sensitivity to order and recording familiarity as secondary endpoints. A null or negative primary effect remains a valid outcome. Multiple exploratory feature or subgroup comparisons require explicit multiplicity handling and independent replication.

8 / VALIDITY AND FEASIBILITY

Missingness, leakage and sample-size planning

A complete work pair requires valid playback of both scheduled excerpts and two numerical pleasure ratings. Buffering that interrupts the passage, hidden-tab playback under the frozen quality rule, missing responses or unapproved replays create explicit flags. Primary analysis uses pairs that meet the preregistered playback criteria. Participants with at least one valid pair can contribute to the pair-available analysis; report coverage and a stricter completion sensitivity analysis. Do not impute outcomes from MI predictions.

Define thresholds for valid playback coverage, interruption handling and acceptable technical noise before launch using pilot evidence. Do not remove a participant because their ratings are flat, disagree with MI, or fail to show a desired effect. Missingness by condition, work, device and background is part of the result. Attrition may limit interpretation even when the model comparison is correctly computed.

8.1 Prevent leakage at the appropriate level

Keep all excerpts of one physical recording together. For any claim about population-level transfer to unseen works, all movements, versions and arrangements within a composition family must stay in one development/test partition. Confirmatory participants cannot supply global training labels. Primary within-person transfer to alternative performances explicitly allows reference A from the same work; that intended calibration is distinguished from accidental test leakage.

8.2 A registration target is not a power calculation

The 1,000-registration cap is fixed by the recruitment plan. The number of eligible listeners depends on consent, the ability to identify 20 works, calibration and attrition. Repeated trials within listeners and shared compositions are dependent observations. Do not count 20,000 work selections as 20,000 independent people.

Run an approved feasibility pilot separately from confirmation. Measure recognition yield, catalogue coverage, onboarding and session duration, missingness, between-listener variation and work-family dependence. Then simulate precision/power for the frozen primary effect under plausible effect sizes and attrition. Freeze pilot exclusion from confirmation, final recruitment feasibility and the analysis implementation before confirmatory launch. If the registration cap is insufficient, report that limitation or amend the plan before seeing confirmatory outcomes.

9 / SECONDARY EXTENSIONS

Arrangements and controlled follow-up studies

A separate arrangement arm can compare piano and guitar, or other explicitly matched instrumentations, for a shared work and passage. It requires an arrangement identity, composer/arranger rights and its own counterbalanced schedule. Instrument, voicing, register, texture, timing and recording conditions may all change. Label the contrast an arrangement preference, not a pure timbre intervention.

Do not pool arrangement pairs into the primary same-instrument estimate. If this arm is added before launch, preregister its sample allocation, burden, endpoint and multiplicity treatment. Otherwise it remains a subsequent study. Synthetic orchestration and human performance are distinct provenance categories and should not be mixed without an explicit design.

9.1 Controlled timing and dynamics

Performance MIDI can be rendered using the same instrument and room simulation to reduce recording-condition variation. To test a causal timing hypothesis, hold notes, velocities, pedal and renderer constant while altering a prespecified timing parameter; a dynamics manipulation requires a different controlled condition. Renderer quality and manipulation plausibility need perceptual validation. Such stimuli are new renderings, not historical original recordings [4,5].

9.2 Relational inference after acquisition

The earlier protocol's feedback and auxiliary-response proposals remain possible secondary research directions [1]. In the collection-only system, evaluate them by locked chronological replay after data capture. At each replay cutoff, fit only with earlier permitted observations; feature selection and hyperparameters cannot use future trials. An apparent replay advantage is not evidence that the deployed session adapted online.

If an adaptive protocol later changes actual stimulus selection, it requires a new scientific and operational release. It must not be simulated retrospectively and described as the randomized experience delivered to participants.

10 / GOVERNANCE

Ethics, reproducibility and launch decisions

Obtain responsible-institution review, approved consent, privacy terms and a named study contact before recruitment. Consent explains the familiar-work screening, repeated interpretations, rating tasks, expected burden, stored background information, compensation, withdrawal options and retention. The online setting and musical content do not remove the need for review.

Separate email/account identity from pseudonymous research IDs. Store consent versions, exposure logs, baseline snapshots, locked predictions and responses under controlled access. The O-Lab companion specifies Supabase, private object storage, transactional email, batch workers, backups and recovery as proposed services, not a deployed system.

Publish the protocol version, analysis code, deidentified codebook, source and rights manifest, model/configuration hashes, preprocessing recipe, composition-family partition and exclusion accounting when permissions allow. Release audio only under its own conditions. A dataset being available to researchers does not automatically authorize unrestricted redistribution of every asset.

Gate	Evidence required before confirmatory launch
Catalogue	Enough distinct familiar-work families with three licensed, reviewed, aligned performances; exact pool frozen.
Participant feasibility	Measured fraction reaching 20 works; burden and dropout reported by relevant background strata.
Scientific lock	Primary endpoint, MI output subset, baseline, fitting code, missingness, timing and sample-size assessment preregistered.
System readiness	Consent and ownership checks, frozen predictions, no live inference, durable responses, withdrawal and restore validated.
Reporting	Registration and completion denominators remain distinct; no universal familiarity or population-generalization claim.

> This revision implements the classical, 20-familiar-work design in the manuscript and data contracts. It does not claim that the full recording pool has been licensed, MI jobs executed, participants enrolled, services provisioned or the experiment validated.

SOURCE RECORD

References

Sources consulted for this revision or inherited from the checked 10 October technical draft. Dataset and provider conditions must be rechecked before use. Experimental choices and numerical scenarios are proposed, not observed results.

[1] Erdem, A. Relational Bayesian Inference for Musical Pleasure. Historical revision 3, 9 October 2026. https://musicalintelligence.app/research/learning-the-listener/Erdem-Relational-Bayesian-Visual-Manuscript-2026-10-09-r3.pdf

[2] IMSLP. Beethoven, Piano Sonata No. 14, Op. 27 No. 2. Recording and arrangement catalogue; per-file rights declarations. https://imslp.org/wiki/Piano_Sonata_No.14,_Op.27_No.2_(Beethoven,_Ludwig_van)

[3] Musopen. Frequently Asked Questions. Composition and recording rights; verification limitations. https://musopen.org/faq/

[4] Magenta. The MAESTRO Dataset. Aligned piano audio and MIDI, versions, partitions and licence. https://magenta.withgoogle.com/datasets/maestro

[5] Foscarin et al. Aligned Scores and Performances (ASAP). Scores, performance alignment, source audio and licence. https://github.com/fosfrancesco/asap-dataset