

Identifying Aesthetic Judgments when Participants Anticipate an Experimenter's Evaluation

Wanhong HUANG
huangwanhong@serendip.ngo

Abstract

A participant can judge an artwork unfavorably and nevertheless report a favorable response after considering how an experimenter will interpret it. The same cue can also change attention, task meaning or judgment itself. This paper examines what aesthetic-response data identify under those possibilities. Experience, judgment, response decision, execution and recording are distinguished without assuming a universal sequence of independent mental stages. A categorical response channel yields conditional population identification under independent calibration and full column rank. Null-space countermodels expose failure, while a singular-value bound separates uniqueness from stability. A binary specialization derives sharp judgment-prevalence bounds under uncertain sensitivity and false-positive rates, including sampling uncertainty and cue-specific compatibility. Randomized cues identify total effects on recorded responses under ordinary design assumptions; a constructed pair of observationally equivalent models allocates the same effect differently between judgment and reporting. A three-indicator special case derives a possible noncircular calibration route under fixed judgment, conditional independence and homogeneous rates. A bounded nested-interpretation model separates the participant's beliefs about experimenter preferences, study objectives and evaluation of the participant. A prospective aesthetic-study design addresses refusal, category semantics, calibration transport, elicitation effects and alternative explanations. Orne's demand-characteristics account, an established demand-bounding method and semantic-context research supply explicit antecedents. The contribution is a limited identification framework and research design, with individual experience and psychological mechanism requiring further evidence.

Keywords: aesthetic judgment; demand characteristics; response measurement; nested interpretation; partial identification

Research Note

This paper develops a bounded argument through conceptual distinctions, stated assumptions, conditional derivations, worked examples, and objections. Its length follows the requirements of that argument. Formal statements concern the specified models; the application of those models to persons and institutions requires independent interpretation and evidence. Constructed examples are explicitly identified and carry illustrative status.

The manuscript follows the repository’s paper template for typography, numbered formal statements, section structure, citations, and navigable contents. The associated project files preserve research provenance, source verification, derivation records, unresolved objections, and validation results. Philosophical commitments about the ultimate nature of social facts remain suspended where the descriptive framework is sufficient for the argument.

Authorship and Research Assistance

Wanhong Huang directed the research questions and authorized this paper’s preparation. An AI research assistant helped organize the material, develop explicit formal constructions, check primary sources, and prepare the text. Agent-added arguments and formulations remain subject to the author’s review. Source positions, repository proposals, and model consequences are distinguished throughout the manuscript.

Use and Licence

Copyright © 2026 Wanhong Huang. This work is licensed under Creative Commons Attribution–NonCommercial 4.0 International, following the repository template. Third-party material remains subject to its own applicable rights. The author encourages careful criticism and responsible application; this ethical request adds no condition to the licence. Reuse implies no authorial endorsement.

1 Introduction

This paper studies the identification of aesthetic judgments when participants anticipate how an experimenter will evaluate their responses. It separates experience, judgment, response decision, execution and recording, then examines what a response distribution can establish about a declared judgment target. The argument develops conditional identification results, sharp bounds under channel uncertainty, countermodels for cue effects and a prospective study design.

Consider a participant viewing an artwork and pressing a button labeled beautiful. The participant may have judged the work beautiful and chosen to report that judgment. The participant may have judged it unappealing but wished to appear knowledgeable, cooperative or appreciative. The participant may have interpreted the question as asking whether the work deserved institutional recognition, a different target from immediate enjoyment. A motor or recording error adds another possibility. Identical records can therefore arise from different episodes.

The motivating insight concerns the participant’s interpretation of the experimental relation. The participant can consider what the experimenter expects, what a particular response will communicate, and what the experimenter will infer about the participant. That is a bounded Theory of Mind problem within the measurement situation. It gives the response decision an intelligible social role without requiring the participant to perform an unlimited hierarchy of recursive beliefs.

The existence of experimental demand is an established antecedent. Orne examined participants’ interpretations of the experimental setting [4]. The present paper’s contribution must therefore concern a more specific question: which judgments are identified by which records, under which assumptions about the path from judgment to response? Recognizing social influence alone supplies an inadequate basis for novelty.

The distinction between a judgment and its report is useful even when the two are closely connected. A person may form an evaluative judgment while articulating it. A request for a response may redirect attention and change the experience. The paper treats the five roles as analytic distinctions within a locally ordered model, with a more general temporal account available when feedback matters. It assumes no universal architecture of separate mental stages.

The central formal object is a response channel. Given a finite judgment vocabulary, a column-stochastic matrix describes the probability of each recorded response conditional on a judgment in the specified context. A known full-rank channel can identify population judgment proportions. Unknown channels admit several explanations of the same observations. Nearly indistinguishable columns can make inference unstable even when uniqueness holds.

These results lead to a bounded alternative when point identification is unavailable. In a binary specialization, independently justified ranges for reporting sensitivity and false-positive probability generate an exact compatibility interval for the prevalence of the declared judgment. The derivation separates uncertainty about the channel from sampling uncertainty. It also exposes which assumptions a comparison across experimenter cues would need.

Randomization contributes an important but different result. Under the design assumptions, it identifies an effect of a cue on recorded responses. The cue may change judgment, the decision to report, or both. Two constructed mechanisms can produce the same observed contrast while allocating it differently between those pathways. A claim about a particular mental mechanism therefore requires more than randomized assignment.

The paper concludes with a design that distinguishes response measurement, channel calibration, nested interpretation and later elicitation. Its formal examples are constructed, and its experiments remain proposed. The aim is a comprehensive account of what can be learned from an aesthetic-response task while preserving the participant’s active interpretive role and the uncertainty of individual experience.

2 Experimental Demand and Aesthetic Context

This section identifies the verified antecedents that constrain the paper’s contribution. It considers demand characteristics, an established demand-bounding method and a study of semantic context in aesthetic evaluation. The comparison distinguishes these sources from the proposed judgment-channel analysis.

Orne treats the participant as an active interpreter of the experimental situation, attending to cues and the meaning of participation [4]. His discussion makes it untenable to claim that the social meaning of an experiment has generally been overlooked. The present question concerns the inference from a particular response record to a particular aesthetic judgment, including the participant’s anticipated evaluation by the experimenter.

De Quidt, Haushofer and Roth develop a method for bounding experimenter demand by deliberately varying perceived experimenter expectations. Their account states monotonicity and bounding assumptions and distinguishes them from stronger assumptions needed for comparative conclusions [2]. The current paper borrows the discipline of explicit assumptions, while its matrix inversion and binary misclassification bounds concern a different target: a latent judgment distribution under a stipulated reporting channel.

That difference has practical significance. A behavior that would occur without a motive to please an experimenter is a counterfactual action target. A judgment experienced or endorsed in the actual experimental context is a time- and context-indexed evaluative target. A measure can be adequate for one while leaving the other unresolved. Importing a demand-free action estimate directly as a measurement of aesthetic experience would require an additional bridge.

Kirk and colleagues report that gallery and computer-source labels applied to a common artwork database were associated with different aesthetic ratings and neural responses [3]. This primary study supplies a concrete reason to investigate context-sensitive evaluation. Its reported neural associations provide no independently calibrated label for the judgment of each participant. The present paper therefore uses the study as a bounded contextual antecedent, with full reanalysis of its methods outside the current evidence base.

Together the sources challenge two simplifying assumptions. First, a participant’s response can be shaped by an interpretation of the research situation. Second, changing the context can affect the evaluative episode itself. These possibilities mean that a procedure intended to reduce reporting pressure may also change the judgment being investigated. A private response is a different context whose measurement properties require evidence.

The repository’s own formulation begins with a distinction between judgment and response decision. It adds the possibility that the participant predicts how an experimenter will interpret that decision. The source comparisons preserve that formulation while narrowing its novelty claim. A defensible contribution needs an explicit inferential result, a counterexample or a design consequence beyond the general recognition of contextual influence.

The present analysis supplies those elements through conditional population identification, stability analysis, sharp compatibility bounds and a mechanism countermodel. Their mathematical ingredients are elementary linear algebra and probability. The proposed contribution lies in their coordinated use for a declared aesthetic target, together with the philosophical distinction between the experience, the judgment and the act of reporting it.

The literature pass is deliberately bounded by what was verified. It establishes relevant primary positions at the scope used here. It leaves a systematic review of aesthetic measurement, experimental demand and latent-variable identification open. That remaining obligation matters for publication-level priority claims, while the current derivations and their limitations can already be assessed on their stated terms.

3 Experience, Judgment and the Response Decision

This section defines the target distinctions before introducing a statistical channel. It identifies the role of the stimulus, personal history, experimental setting and anticipated evaluation. It also explains why a finite response vocabulary requires semantic justification.

Let E denote the aesthetic episode, including aspects of experience relevant to the task. Let J denote the participant's judgment under a declared vocabulary at a declared time. Let D denote the decision about what to communicate, A the executed action and Y the recorded response. These symbols identify analytic roles. A person may revise a judgment while preparing a response, and a response may become part of the subsequent experience.

For one local ordering, propose

$$\begin{aligned} E &= f_E(X, H, C, \epsilon_E), & J &= f_J(E, H, C, \epsilon_J), \\ D &= f_D(J, B, R, N, C, \epsilon_D), & A &= f_A(D, C, \epsilon_A), \\ Y &= f_Y(A, C, \epsilon_Y). \end{aligned} \tag{1}$$

Here X is the stimulus, H the participant's relevant history, C the setting, B an interpretation of the experimenter, R perceived consequences and N relevant norms. The disturbances summarize unmodeled variation. Distinct disturbance symbols do not assert independence.

Equation (1) is a proposed decomposition. It is useful when its ordering matches the episode being investigated. If the response question forms the judgment, the target time must follow that formation or the model must include a transition. If action changes experience before recording, the local ordering needs corresponding refinement. Writing a sequence of functions is insufficient evidence for a psychological mechanism.

Definition 3.1 (Protocol-relative aesthetic judgment). A protocol-relative aesthetic judgment is an evaluative stance toward the specified object under the task's interpreted vocabulary, context and target time. Its operational representation may be finite even when the associated experience is richer.

For example, $J \in \{+, -, u\}$ can distinguish affirmative judgment, negative judgment and declared suspension. The third category requires particular care. Uncertainty, ambivalence, indifference, unfamiliarity and refusal are different possibilities. A participant may reject the task's evaluative vocabulary while having a clear response to the artwork. Combining those possibilities into one category is an analytic choice whose adequacy depends on the question.

The recorded vocabulary can differ from the judgment vocabulary. A protocol with two buttons may force a participant with an unresolved judgment to select one. A protocol with a skip button may record refusal, delay or accidental omission under one code. The channel model can represent such mappings only after their meanings have been specified. More buttons increase possible distinctions in the record but provide no guarantee that participants use those distinctions consistently.

The response decision can itself be authentic practical conduct. A participant who chooses to be polite may be accurately expressing a wish to maintain a cooperative interaction. The measurement problem arises when the researcher interprets that action as direct evidence of aesthetic endorsement. Recognizing the practical reason for a response avoids treating every divergence from the researcher's target as meaningless noise or personal dishonesty.

Table 1 gives the role-specific questions used later.

The formal analysis primarily targets the distribution of J . It reaches E only through an additional account relating judgment to experience. Even a perfectly recovered judgment distribution can leave the qualitative character of each participant's experience open. Keeping that limit explicit makes population identification useful without assigning it a philosophical scope it cannot sustain.

Table 1: Analytic roles in an aesthetic-response episode

Role	Target	Main interpretive obligation
E	Aesthetic episode	Specify which experienced aspect is being discussed
J	Judgment under the task vocabulary	Establish meaning, context and time
D	Intended communication	Consider anticipated interpretation and consequences
A	Executed response	Separate intention from motor action
Y	Stored record	Verify coding and recording conditions

4 Categorical Response Channels

This section derives the response-channel equation from total probability and states its assumptions. It separates a description of observed associations from the stronger requirements of calibration and transport. These distinctions determine what the later inversion can mean.

Fix a context c , a judgment vocabulary with k categories and a response vocabulary with m categories. Write

$$p_j = \Pr(J = j \mid c), \quad q_y = \Pr(Y = y \mid c), \quad L_{yj} = \Pr(Y = y \mid J = j, c).$$

The vectors $p \in \Delta_{k-1}$ and $q \in \Delta_{m-1}$ lie in probability simplexes. The matrix $L \in \mathbb{R}^{m \times k}$ has nonnegative entries and each column sums to one. For a judgment category with zero population probability, the conditional column requires a specified extension if it will be used in counterfactual calibration.

By partitioning the event $Y = y$ over the possible judgments,

$$\Pr(Y = y \mid c) = \sum_{j=1}^k \Pr(Y = y \mid J = j, c) \Pr(J = j \mid c).$$

Collecting the equations gives

$$q = Lp. \tag{2}$$

No independence between judgment and reporting is assumed. Their association is precisely what the conditional columns represent.

Assumption 4.1 (Calibrated applicable channel). The matrix L is independently known or identified, its judgment and response categories have the stated meanings, and it applies to the population, context and target time for which p is inferred.

The assumption is stronger than knowing the button labels. A channel column asks how a person with a particular judgment would be recorded. If that judgment is latent, estimating the column from the same unvalidated response would be circular. An external label, another measurement model or a justified experimental structure is needed, and each brings its own assumptions.

Execution and recording can be calibrated separately in simple tasks. A participant may be asked to press a specified symbol, allowing some motor and logging errors to be observed. Such a task can help assess the final links in Equation (1). It leaves the mapping from aesthetic judgment to intended communication unresolved. It also requires a transport argument if attention, stress or device use differs in the aesthetic task.

One might factor L into several stage channels. Suppose $K_{dj} = \Pr(D = d \mid J = j, c)$, $M_{ad} = \Pr(A = a \mid D = d, c)$ and $N_{ya} = \Pr(Y = y \mid A = a, c)$. The identity $L = NMK$ requires the conditional restrictions that later stage distributions depend on the preceding stage in the stated way, with remaining relevant history adequately represented. The existence of several variables alone supplies no such factorization.

Heterogeneity creates a related transport problem. Participants with the same judgment may differ in how much they care about experimenter evaluation. The population channel averages that variation

within each judgment category. If a manipulation changes the composition of those categories or the relation between judgment and response motivation, the averaged channel may change. A calibration from one group cannot be transferred merely because both groups view the same artworks.

Equation (2) is nevertheless valuable even when calibration remains open. It identifies the unknown objects and clarifies the source of underdetermination. Observed response differences can reflect p , L or both. The next sections derive what additional rank and uncertainty assumptions permit, while retaining the obligation to justify those assumptions empirically.

5 Population Identification and Numerical Stability

This section derives uniqueness, failure and stability results for the calibrated channel. It distinguishes global identification over the simplex from identification at a boundary observation. It then explains why population recovery leaves individual judgment uncertain.

Proposition 5.1 (Full-column-rank identification). *Under Assumption 4.1, if L has full column rank, at most one $p \in \Delta_{k-1}$ satisfies $q = Lp$ for a given q .*

Proof. If $Lp = Lp'$, then $L(p - p') = 0$. Full column rank gives a trivial null space, so $p - p' = 0$ and $p = p'$. $\square$

Existence is a separate condition. An observed population vector must lie in $L\Delta_{k-1}$ for exact compatibility. A finite sample estimate may lie outside that image even when the model is adequate. Estimation should then account for sampling variation and the simplex constraints. A negative component of an unconstrained inverse indicates a need for constrained estimation or model assessment.

Proposition 5.2 (Rank deficiency and interior ambiguity). *If a column-stochastic matrix L has deficient column rank, each interior point p of the judgment simplex has a distinct nearby simplex point p' with $Lp' = Lp$.*

Proof. Choose a nonzero vector v with $Lv = 0$. Column stochasticity gives $\mathbf{1}_m^\top L = \mathbf{1}_k^\top$, hence $\mathbf{1}_k^\top v = \mathbf{1}_m^\top Lv = 0$. For sufficiently small nonzero ϵ , every component of $p + \epsilon v$ remains nonnegative because p is interior. Its components sum to one and $L(p + \epsilon v) = Lp$. Taking $p' = p + \epsilon v$ proves the claim. $\square$

With two responses and three unrestricted judgments, deficient column rank is unavoidable. Even a known channel therefore fails to identify every population distribution. A particular boundary observation can still have a unique preimage. For example,

$$L = \begin{pmatrix} 1 & 0 & 1/2 \\ 0 & 1 & 1/2 \end{pmatrix}$$

maps only $p = (1, 0, 0)^\top$ to $q = (1, 0)^\top$. This boundary case prevents overstatement of the rank result as universal ambiguity at every possible observation.

For a contrasting three-category example, take

$$L = \begin{pmatrix} 0.8 & 0.1 & 0.2 \\ 0.1 & 0.8 & 0.2 \\ 0.1 & 0.1 & 0.6 \end{pmatrix}, \quad p = \begin{pmatrix} 0.5 \\ 0.3 \\ 0.2 \end{pmatrix}.$$

Direct multiplication gives $q = (0.47, 0.33, 0.20)^\top$, and $\det L = 0.35$. The distribution is uniquely recoverable under the stipulated known channel. These numbers are constructed for calculation, with no calibration claim.

Proposition 5.3 (Stability under a fixed channel). *If L has smallest singular value $\sigma_{\min} > 0$, then for any two judgment vectors p, p' ,*

$$\|p - p'\|_2 \leq \frac{\|Lp - Lp'\|_2}{\sigma_{\min}}.$$

Proof. The defining lower bound for the smallest singular value gives $\|Lv\|_2 \geq \sigma_{\min}\|v\|_2$ for every v . Substitute $v = p - p'$ and divide by the positive constant. $\square$

A small $\sigma_{\min}$ permits large differences in inferred proportions from small response differences. Thus full rank establishes uniqueness while the singular value controls one aspect of stability. In the binary symmetric channel with diagonal 0.51 and off-diagonal 0.49, the contrast between judgment categories is attenuated by 0.02. Inverting that contrast magnifies response error by a factor of fifty.

Finally, population identification leaves individual uncertainty. In the three-category example, one affirmative response yields posterior weights proportional to $(0.8 \cdot 0.5, 0.1 \cdot 0.3, 0.2 \cdot 0.2)$. These sum to 0.47, giving probabilities 40/47, 3/47 and 4/47 under the model. A uniquely identified population can therefore contain uncertainty about the judgment behind a particular response. Direct access to individual experience requires still further interpretation and evidence.

5.1 Error in the calibrated channel

This subsection extends the stability result to uncertainty in calibration itself. The extension matters because a precisely measured response distribution can still yield a fragile judgment estimate when the reporting matrix is inaccurate.

Suppose $q = Lp$ for the reference model, while an alternative analysis uses $\tilde{q} = \tilde{L}\tilde{p}$ with $\tilde{p}$ in the probability simplex. Assume $\|q - \tilde{q}\|_2 \leq \varepsilon$ and $\|L - \tilde{L}\|_2 \leq \delta$, where the matrix norm is the induced Euclidean operator norm. The smallest singular value of L is $\sigma_{\min} > 0$.

Proposition 5.4 (Joint response and channel perturbation). *Under the stated conditions,*

$$\|p - \tilde{p}\|_2 \leq \frac{\varepsilon + \delta\|\tilde{p}\|_2}{\sigma_{\min}} \leq \frac{\varepsilon + \delta}{\sigma_{\min}}.$$

Proof. Subtract the two model equations and rearrange: $L(\tilde{p} - p) = \tilde{q} - q + (L - \tilde{L})\tilde{p}$. The triangle inequality and induced norm bound give $\|L(\tilde{p} - p)\|_2 \leq \varepsilon + \delta\|\tilde{p}\|_2$. Apply the singular-value inequality from Proposition 5.3. Finally, a nonnegative vector with unit sum has Euclidean norm at most one. $\square$

The bound can be loose, especially near the simplex boundary, but it reveals the interaction of calibration error and poor discrimination. A large response sample can reduce one source of uncertainty while leaving δ substantial. If the channel columns are close, even a small calibration discrepancy can materially change the inferred judgment distribution.

The result assumes exact compatibility of each fitted pair with its own response vector. An estimation residual would add another term through the same triangle inequality. A reported interval should therefore distinguish response sampling error, calibration uncertainty and model misfit. Combining them under one standard error without stating the estimation procedure would conceal their different origins.

The practical priority follows the bottleneck. If the reporting channel is weakly discriminating, additional identical responses may contribute less than better evidence about category meaning or response behavior. If calibration is strong and sampling error dominates, a larger sample can help. The inequality supplies a way to articulate that comparison within a model; choosing a research investment still requires costs, feasible methods and substantive judgment.

6 Sharp Bounds under Channel Uncertainty

This section derives a binary compatibility interval when the reporting channel is only bounded. It states the independence of the permitted parameter ranges, proves attainability and extends the result to uncertainty in the response probability. The result supplies a limited inference target when exact calibration is unavailable.

Let $\theta = \Pr(J = 1)$ and $s = \Pr(Y = 1)$. Define $\alpha = \Pr(Y = 1 \mid J = 1)$ and $\beta = \Pr(Y = 1 \mid J = 0)$. Total probability gives

$$s = \alpha\theta + \beta(1 - \theta) = \beta + (\alpha - \beta)\theta. \quad (3)$$

When α and β are known and distinct, inversion yields $\theta = (s - \beta)/(\alpha - \beta)$, subject to membership in $[0, 1]$. If they coincide, the response probability is constant across all judgment proportions.

Assumption 6.1 (Separated rectangular channel ranges). The admissible channel class is the entire rectangle

$$\alpha \in [a_-, a_+], \quad \beta \in [b_-, b_+], \quad 0 \leq b_- \leq b_+ < a_- \leq a_+ \leq 1.$$

Every pair in the rectangle is allowed. The target judgment and response semantics remain fixed.

The strict separation gives a positively discriminating channel. It is an empirical and interpretive assumption requiring evidence beyond the button label. The rectangle also matters. If calibration evidence couples α and β , the set of admissible pairs can be smaller and the interval below may cease to be sharp.

Proposition 6.2 (Sharp binary compatibility interval). *Under Assumption 6.1, if $s \in [b_-, a_+]$, the set of θ compatible with Equation (3) is exactly*

$$\Theta(s) = \left[\max \left\{ 0, \frac{s - b_+}{a_+ - b_+} \right\}, \min \left\{ 1, \frac{s - b_-}{a_- - b_-} \right\} \right]. \quad (4)$$

For $s \notin [b_-, a_+]$, the compatibility set is empty.

Proof. For fixed $\theta \in [0, 1]$, the attainable response probabilities range from $b_- + (a_- - b_-)\theta$ to $b_+ + (a_+ - b_+)\theta$. This follows because Equation (3) is linear in (α, β) with nonnegative coefficients, and a continuous image of the rectangle fills the interval between its minimum and maximum. Compatibility therefore requires and suffices for

$$b_- + (a_- - b_-)\theta \leq s \leq b_+ + (a_+ - b_+)\theta.$$

The positive denominators permit rearrangement into the two bounds in Equation (4), followed by intersection with $[0, 1]$. Every point in that interval is attained by some allowed parameter pair. Across all θ the attainable response range is $[b_-, a_+]$, proving the final statement. $\square$

For $s = 0.60$, $\alpha \in [0.8, 0.9]$ and $\beta \in [0.1, 0.2]$, the interval is

$$\Theta(0.60) = [4/7, 5/7].$$

The lower endpoint is attained at $\alpha = 0.9, \beta = 0.2$; the upper at $\alpha = 0.8, \beta = 0.1$. The result expresses uncertainty about a population proportion within the model. It gives no classification of an individual participant.

Suppose a response-probability uncertainty interval is $S = [s_-, s_+]$. The compatible judgment set is $\bigcup_{s \in S} \Theta(s)$. First intersect S with $[b_-, a_+]$; an empty intersection indicates incompatibility. If the remaining endpoints are $\tilde{s}_-, \tilde{s}_+$, monotonicity of the two bounds gives

$$\Theta(S) = \left[\max \left\{ 0, \frac{\tilde{s}_- - b_+}{a_+ - b_+} \right\}, \min \left\{ 1, \frac{\tilde{s}_+ - b_-}{a_- - b_-} \right\} \right]. \quad (5)$$

For $S = [0.56, 0.64]$ in the example, the interval becomes $[18/35, 27/35]$. If S has a stated confidence coverage for s and the fixed channel bounds contain the true channel, the induced set inherits at least that coverage for θ . Estimated channel bounds require a joint coverage argument.

The derivation makes the evidential burden visible. Narrow bounds can come from strong calibration evidence or from unjustifiably restrictive assumptions. The numerical width alone cannot distinguish those origins. A useful report should therefore present the source and transport conditions of the channel ranges alongside the resulting judgment interval.

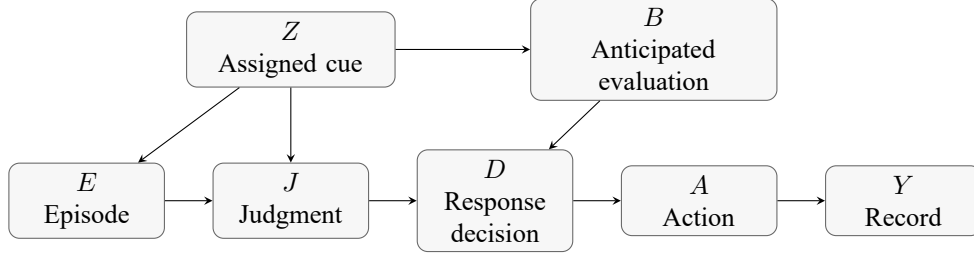


Figure 1: Selected candidate pathways for an experimenter cue in the local model. The arrows display rival explanatory possibilities, with fixed timing and additional dependencies requiring explicit specification.

7 Randomized Cues and Competing Pathways

This section separates a randomized response contrast from an interpretation of its mechanism. It derives an algebraic decomposition and constructs two observationally equivalent accounts. It then shows how cross-cue compatibility can test a bundle of assumptions without identifying which assumption failed.

Figure 1 displays selected pathways that the response contrast leaves unresolved. The cue can affect the evaluative episode, the judgment or the anticipated interpersonal consequences of reporting. The diagram is an explanatory proposal whose arrows require substantive evidence; its layout supplies no empirical proof of mediation.

Let $Z \in \{0, 1\}$ be a randomized cue concerning the experimenter. Under consistency, suitable treatment definition, assignment independence and the relevant interference assumptions, the contrast

$$\Delta_y = \Pr(Y = y \mid Z = 1) - \Pr(Y = y \mid Z = 0)$$

identifies the average effect of the assigned cue on the recorded category. It is a total response effect. The causal interpretation depends on the design assumptions, including how participants encounter the cue and whether one participant's condition affects another's response.

For each cue, write $q_z = L_z p_z$. Adding and subtracting $L_0 p_1$ gives

$$q_1 - q_0 = L_0(p_1 - p_0) + (L_1 - L_0)p_1. \quad (6)$$

The first term represents a distributional judgment change under a fixed reference channel; the second represents a channel change evaluated at the later distribution. Adding and subtracting $L_1 p_0$ gives another valid decomposition. These algebraic identities are not by themselves identified causal mediation effects.

Example 7.1 (Equivalent observed cue effects). Suppose affirmative response probabilities are $s_0 = 0.50$ and $s_1 = 0.70$. In Model A, reporting is perfect and judgment prevalence changes from $\theta_0 = 0.50$ to $\theta_1 = 0.70$. In Model B, judgment prevalence remains 0.50, reporting is perfect at $Z = 0$, and at $Z = 1$ the channel has $\alpha_1 = 1$, $\beta_1 = 0.4$. Equation (3) gives the same response probabilities in both models.

Both models can be implemented with random assignment. In the first, the cue changes attention or the interpreted evaluative task. In the second, it changes the response decision among participants with negative judgments. The example establishes underidentification from those response probabilities alone. It leaves the actual prevalence of either mechanism open.

A claim that a cue affects reporting while preserving judgment requires an exclusion or invariance assumption concerning the target. Privacy, anonymity or reassurance might make that assumption more plausible in some settings. They can also alter comfort, attention and the meaning of the encounter. The design should therefore justify the restricted pathway for the particular target and examine counterconditions.

If a common judgment prevalence θ is assumed across cue conditions and each condition has its own admissible channel class, the compatible set is

$$\Theta_{\text{common}} = \bigcap_z \Theta_z(s_z), \quad (7)$$

provided channel parameters can vary independently across conditions within their specified classes. A shared channel parameter would impose additional coupling and can make this intersection too permissive.

For a constructed test, suppose both cue conditions use the ranges in Section 6. At $s_0 = 0.60$, $\Theta_0 = [4/7, 5/7]$. At $s_1 = 0.80$, $\Theta_1 = [6/7, 1]$. The sets are disjoint. The data and assumptions therefore fail to admit one stable judgment prevalence. At least one element of the bundle needs revision: stable judgment, the channel bounds, the category semantics or, with estimated probabilities, the treatment of sampling error.

The empty intersection supplies no unique mechanism diagnosis. It cannot establish that Theory of Mind changed the response, because a different judgment pathway or inaccurate calibration could explain the failure. A nonempty intersection is also limited: it shows compatibility with the bundle, not proof of the bundle. These distinctions identify the discriminating evidence needed to assess a favored psychological explanation.

The practical implication is to report the total cue effect first, then state which further assumptions permit a judgment or reporting interpretation. A response contrast can be informative even when its decomposition remains unresolved. Preserving that result avoids both overclaiming mechanism and discarding an experiment merely because it leaves a richer mental target unidentified.

8 Anticipated Evaluation and Nested Interpretation

This section specifies the participant’s model of the experimenter. It separates beliefs about taste, research objectives and personal evaluation, then relates them to response decisions. The construction gives the Theory of Mind insight a bounded explanatory role.

Let i be the participant and e the experimenter. Distinguish three candidate beliefs held by i : what e personally likes, what pattern of responses e wants the study to produce, and what e will infer about i from a particular response. These beliefs can differ. A participant may think the experimenter dislikes an artwork personally but hopes to find favorable ratings, or may believe that an unfavorable rating will be interpreted as independence of judgment.

Represent the third belief by a function $\hat{V}_i(d, c)$ describing the evaluation the participant anticipates after response decision d in context c . Its range can be a finite set of attributed evaluations, a probability distribution over them, or a practical ranking. Choosing a numerical scale requires further justification. The function models the participant’s interpretation, which can be inaccurate about the experimenter.

A schematic response decision is

$$D_i \in \operatorname{argmax}_{d \in \mathcal{D}_i} \{U_i(d; J_i, c) + \lambda_i V_i(d; c)\}. \quad (8)$$

Here U_i represents other response considerations and V_i the participant’s evaluation of anticipated interpersonal consequences. The coefficient λ_i is a modeling parameter. Equation (8) is a proposed local choice representation, not an established utility law. Its existence requires a defensible action set, comparability and a model of practical motivation.

The equation helps generate distinct countermodels. A participant can endorse the work and report endorsement with little concern for evaluation. A participant can judge it negatively yet select endorsement because an anticipated consequence outweighs the value of accurate reporting in the constructed representation. A participant can deliberately report disagreement because the anticipated evaluation rewards independence. Opposite responses can therefore arise from concern with the same evaluator.

The model also distinguishes concern from accuracy. A participant can respond strongly to an inaccurate belief about the experimenter. Measuring the experimenter’s actual preference would leave that mechanism unresolved. Conversely, correctly predicting the experimenter’s taste need not affect the response if the participant treats it as irrelevant. A useful study therefore separates the content of the participant’s belief, confidence in it and its practical significance.

Nested interpretation need not continue indefinitely. The relevant question is whether a limited attributed evaluation helps explain a contrast that simpler accounts leave unresolved. An added level such as what the experimenter thinks the participant expects requires a concrete discriminating role. Otherwise, recursion can produce an increasingly elaborate description without improved identification.

Eliciting the belief creates its own intervention. Asking participants how they think the experimenter will judge them may make evaluation salient, supply a new hypothesis or encourage a socially legible explanation. A later verbal account can be informative about the participant's present interpretation while leaving the earlier response decision uncertain. The time-indexed inquiry problem parallels the repository's dialogue analysis.

The philosophical significance is that the measured person participates in constructing the measurement situation. The participant interprets the artwork, the task and the interpersonal relation. A response can therefore express a practical stance toward the experiment as well as an aesthetic stance toward the object. The inferential task is to identify which stance the available evidence supports, while remaining open to interactions among them.

9 A Prospective Aesthetic-Response Study

This section develops a concrete study architecture from the formal distinctions. It separates an immediately estimable response contrast from more demanding judgment and mechanism targets. The design is a proposal, with no participant data or completed calibration claimed.

The basic task presents a specified set of artworks under controlled viewing conditions. Participants receive an explicit question whose vocabulary has undergone preliminary comprehension work. For illustration, the question concerns a present aesthetic endorsement of the displayed work. The design distinguishes endorsement, rejection, declared suspension and refusal to answer. Whether those categories are adequate is an empirical and participatory question for the intended population.

The primary manipulation concerns a declared audience for the response. One condition describes an experimenter who will see the individual answer; another describes a procedure in which the answer contributes to an aggregate without ordinary individual evaluation by that experimenter. Any statements about privacy must accurately describe the procedure. The manipulation targets anticipated evaluation, while its total effects can include changes in comfort, attention and judgment.

A separate expectation cue could describe the researcher's predicted pattern of responses. This factor should remain distinct from audience visibility. A participant can know that a researcher expects high ratings while believing that individual responses remain private. Another participant can anticipate personal evaluation without knowing a favored aggregate pattern. Treating those situations as one demand manipulation would obscure the nested-interpretation question.

The primary estimand is the assigned cue's effect on recorded responses, with the randomization unit and possible interference specified. Judgment prevalence is a secondary conditional estimand requiring channel assumptions. An estimate of the effect mediated by anticipated evaluation is a further target requiring a causal design and additional restrictions. Stating this hierarchy in advance prevents a response effect from acquiring a stronger interpretation after results are seen.

Table 2 links design components to their contribution and remaining obligations.

A pilot could examine how participants understand the task and category labels. Its aim would be to identify ambiguities, including whether beautiful is taken to mean enjoyable, skillful, culturally important or personally moving. If the vocabulary produces systematically different tasks, the main study should either narrow it or explicitly model that variation. A common word alone provides insufficient evidence of a common latent category.

The numerical example in Section 6 illustrates the reporting format once channel ranges have an independent basis. If the affirmative response probability is estimated near 0.60 and the stated calibration ranges apply, the judgment compatibility interval can be reported alongside the observed response estimate. The source of each channel range should be visible. If calibration remains inadequate, the response contrast retains its primary interpretation and the judgment interval remains a sensitivity analysis.

Table 2: Proposed design components and their inferential roles

Component	Intended contribution	Remaining obligation
Audience visibility	Total effect of anticipated individual evaluation	Cue can also change the evaluative episode
Expectation cue	Contrast in perceived research objectives	Personal evaluation and task meaning may also change
Explicit response categories	Better-defined recording vocabulary	Refusal and suspension require semantic checks
Motor and logging task	Evidence about execution and recording	Transport to the aesthetic task
Belief elicitation	Evidence about interpreted experimenter aims	Elicitation can alter the relevant belief
Follow-up account	Participant's explanation and correction	Historical target differs from present account

Order requires attention. Repeated exposure to an artwork can change familiarity, interpretation and response strategy. Within-participant comparisons may reduce some between-person variation while introducing carryover. A between-participant design may reduce carryover while requiring more evidence about population comparability. The choice should follow the target and anticipated threats to interpretation, with order counterbalancing used only where its assumptions are defensible.

The design's value lies in aligning claims with evidence. It can establish a response effect, reveal a mismatch between category meanings, challenge a channel assumption or motivate a richer account of anticipated evaluation. A study that leaves the individual aesthetic episode uncertain can still contribute useful knowledge, provided that uncertainty remains part of the reported result.

9.1 Estimands and interpretation of possible results

This subsection specifies how the proposed study should report different patterns of evidence. The purpose is to assess results against the declared target and preserve competing explanations of an observed contrast.

The first result is a response-distribution contrast with uncertainty from the sampling design. A small or imprecisely estimated contrast can coexist with individual changes in opposite directions. It therefore leaves open the possibility of heterogeneous response effects. A large contrast supplies evidence of a change in the recorded distribution while leaving the judgment/reporting division unresolved.

The second result is a judgment compatibility set conditional on a documented channel class. The report should display both the response data and the assumed calibration ranges, including sensitivity to wider ranges. If the set is empty, the report should identify the inconsistent assumption bundle. If it is wide, the report should state which target remains weakly constrained. Both outcomes can guide the next study.

The third result concerns participant interpretation. Accounts of experimenter expectations can be compared with the delivered cues, their timing and the participant's subsequent actions. Agreement between the cue and the reported belief supports one part of a mechanism account. The causal role of that belief remains a further question, particularly when the elicitation itself was an intervention.

The fourth result concerns measurement adequacy. If participants use endorsement to refer to several different evaluations, the study may have identified a response convention whose relation to the intended judgment is unstable. Revising the vocabulary can then be a substantive research outcome. Preserving the distinction avoids treating better coding consistency as sufficient proof of shared psychological meaning.

This reporting structure also clarifies an unsuccessful manipulation. If participants fail to interpret the cue in the intended way, the assigned-cue response effect remains a well-defined design target, while a claim about the intended belief change loses support. A mechanism analysis should retain that failure and label any revised interpretation of the manipulation as a subsequent hypothesis.

10 Calibration, Missing Responses and Temporal Validation

This section examines the evidence needed to apply the channel results. It considers circular calibration, transport, missing responses and later elicitation. The aim is to distinguish a promising measurement strategy from a justified identification claim.

Calibration is difficult because judgment is the latent target. One cannot define an affirmative judgment by an affirmative button press, estimate a perfect channel from that definition, and then claim to have independently validated the button as a measure of judgment. The resulting identity would concern the coding rule. It would leave the intended psychological target unresolved.

Several sources of evidence can help without resolving the problem automatically. Participants can explain what a response meant, compare alternative descriptions, correct the researcher's interpretation or distinguish a judgment from a response strategy. Multiple indicators can provide additional constraints when their dependencies are modeled. A separate task can check whether the response device records intended actions. Each source has a target and a limitation that should be recorded.

Calibration transport requires particular scrutiny. Suppose a channel is estimated in a low-pressure setting and then applied when participants anticipate expert evaluation. The manipulation may change the very response mapping that was calibrated. A study cannot assume invariance of L while claiming that the cue changes reporting behavior through that channel. It can instead specify cue-specific channels or make the invariance assumption part of a testable bundle.

Population aggregation introduces a related issue. If response motivation differs across participants and correlates with judgment, a channel averaged over one population can fail in another. Conditioning on relevant covariates can help when the resulting categories have adequate data and interpretation. It can also create small samples and unstable estimates. A comprehensive report should state the population to which the calibration is intended to apply.

Missing responses need explicit treatment. Excluding everyone who refuses the task changes the target population to respondents. If refusal depends on aesthetic judgment or anticipated evaluation, the observed judgment distribution among respondents can differ from the original population. A missing-response category preserves one distinction at the recording level. Recovering the judgment distribution of nonrespondents still requires evidence or bounds.

For a simple extension, let $R = 1$ indicate that a participant supplies a classifiable response, and suppose $\Pr(R = 1) = r$. If a respondent judgment prevalence θ_R is identified while the nonrespondent prevalence remains unrestricted, the full prevalence satisfies

$$\theta = r\theta_R + (1 - r)\theta_{\bar{R}}, \quad \theta \in [r\theta_R, r\theta_R + 1 - r]. \quad (9)$$

The interval follows by allowing $\theta_{\bar{R}} \in [0, 1]$. It is sharp within this binary construction. If $\theta_R \in [\ell, u]$, the interval becomes $[r\ell, ru + 1 - r]$. A numerical example with $r = 0.8$ and $\theta_R \in [4/7, 5/7]$ gives $[16/35, 27/35]$.

The extension assumes that the binary judgment target is meaningful for nonrespondents. A person who refuses because the evaluative vocabulary is inapplicable may require a different state space. The formula cannot resolve that semantic problem by treating refusal as an unobserved yes or no. Its scope is the case where a common binary target is defensible and observation is incomplete.

Later elicitation introduces temporal uncertainty. A participant's explanation after seeing the study's purpose may be thoughtful and informative about their present interpretation. It can also incorporate new concepts or reconstruct the earlier episode. Validation should therefore specify whether it targets the earlier judgment, the earlier response decision or a current account of either. Agreement across time is evidence under a model of continuity; it is not a substitute for stating that model.

These obligations give calibration the status of a substantive research program. The formal results are ready for application only when their categories, channel ranges and transport conditions receive adequate support. Until then, they remain useful for sensitivity analysis and for identifying which additional evidence would most improve the interpretation of response data.

10.1 A conditional reconstruction from three indicators

This subsection supplies a constructive route to identification when the reporting rates are unknown. It also makes the price of that route explicit. Identification in latent structures with conditionally independent observations is an established statistical subject [1]; the following moment calculation is a transparent special case developed for the present audit.

Suppose a fixed binary judgment J generates three binary indicators Y_1, Y_2, Y_3 . Conditional on J , assume the indicators are independent and identically distributed, with success probability α when $J = 1$ and β when $J = 0$. Assume $0 < \theta = \Pr(J = 1) < 1$ and $0 \leq \beta < \alpha \leq 1$. The ordering labels the affirmative class; the semantics of that label still need an independent warrant.

Define the observable cross-moments

$$m_1 = \mathbb{E}[Y_1], \quad m_2 = \mathbb{E}[Y_1 Y_2], \quad m_3 = \mathbb{E}[Y_1 Y_2 Y_3].$$

Conditional independence and equal rates give $m_k = \theta\alpha^k + (1-\theta)\beta^k$ for $k = 1, 2, 3$. Let $d = \alpha - \beta > 0$ and define

$$v = m_2 - m_1^2, \quad t = m_3 - 3m_1 m_2 + 2m_1^3.$$

These are moments of the latent success probability, not the ordinary variance and third moment of one binary response.

Proposition 10.1 (Homogeneous three-indicator reconstruction). *Within the stated interior model,*

$$\begin{aligned} v &= \theta(1-\theta)d^2 > 0, \\ t &= \theta(1-\theta)(1-2\theta)d^3, \\ R &= t/v^{3/2}, \quad \theta = \frac{1}{2} \left(1 - \frac{R}{\sqrt{R^2 + 4}} \right), \\ d &= \sqrt{\frac{v}{\theta(1-\theta)}}, \quad \beta = m_1 - \theta d, \quad \alpha = \beta + d. \end{aligned} \tag{10}$$

Thus the three cross-moments identify the ordered parameter triple in this model.

Proof. Let $P = \beta + dJ$ be the conditional success probability. The cross-moment identities give $m_k = \mathbb{E}[P^k]$. Since $\mathbb{E}[P] = \beta + d\theta$, centering yields $P - \mathbb{E}[P] = d(J - \theta)$. The variance and third central moment of a Bernoulli variable are $\theta(1-\theta)$ and $\theta(1-\theta)(1-2\theta)$, obtained by summing its two possible values. This proves the expressions for v and t . Their ratio gives $R = (1-2\theta)/\sqrt{\theta(1-\theta)}$. Put $x = 1 - 2\theta$, so $\theta(1-\theta) = (1-x^2)/4$. Then $R = 2x/\sqrt{1-x^2}$, with the sign of x fixed by that of R . Solving gives $x = R/\sqrt{R^2 + 4}$ and the stated formula for θ . The positive value of d and the two means then recover α, β . $\square$

For a constructed check, take $\theta = 0.6$, $\alpha = 0.9$ and $\beta = 0.2$. The cross-moments are $m_1 = 0.62$, $m_2 = 0.502$ and $m_3 = 0.4406$. They yield $v = 0.1176$ and $t = -0.016464$, recovering the original parameters through Equation (10). These are exact model calculations, not observations.

The substantive assumptions are demanding. Repeated responses to one artwork may share a response strategy even conditional on judgment, violating independence. Viewing the work again may change J . Different response formats may have different rates, violating homogeneity. Identical formats may encourage participants to repeat the first answer. The construction therefore establishes a possible identification route, with empirical applicability still requiring evidence.

Degenerate cases are informative. If $\alpha = \beta$ or θ is zero or one, then $v = 0$ and the inversion above is unavailable. Near those cases, moment estimation can be unstable. Sampling estimates can also imply α or β outside $[0, 1]$, which calls for constrained estimation and model assessment. More indicators can provide checks on the imposed structure, but a good fit alone leaves semantic labeling and transport unresolved.

The result refines the earlier calibration objection. Independent validation need not always mean a perfect external judgment label. A jointly identified measurement model can sometimes supply a route, provided its restrictions have independent justification. The relevant distinction is between a noncircular evidential structure and defining the latent category directly from the response one hopes to validate.

11 Philosophical and Methodological Objections

This section examines objections about the nature of judgment, the interpretation of strategic responses, contextual change and the value of formalization. The responses retain the paper's limited targets while identifying cases in which those targets should be revised.

11.1 The status of an unreported judgment

The first objection questions whether a judgment exists separately from its expression. Some episodes may involve an evaluative stance before a response, while others may develop the stance through articulation. A single latent-state model cannot settle that philosophical issue for every aesthetic encounter.

The paper uses a conditional target. Where a judgment at a specified time is independently intelligible, the channel model asks how it relates to recorded responses. Where expression forms the judgment, a temporal transition model is needed. The distinction between judgment and response decision can still perform conceptual work without asserting that every judgment is a fully formed private object awaiting transmission.

This treatment also avoids reducing experience to a category. A participant may feel attraction, discomfort, recognition and uncertainty together. A binary endorsement variable can answer a narrow research question while leaving that complexity open. The adequacy of the reduction depends on whether the omitted distinctions matter to the declared purpose.

11.2 Strategic response and the meaning of authenticity

The second objection concerns the language of response error. A participant who answers politely may be acting sincerely in the interpersonal situation. Calling the response false without specifying the target can misdescribe the action.

The framework agrees with the objection. A response can be authentic communication of cooperation while providing imperfect evidence of aesthetic endorsement. The discrepancy is target relative. It becomes a measurement error only under a justified interpretation that the recorded category should represent the judgment. The participant's practical reason remains an object of understanding with its own substantive importance.

A related concern is that instructions to ignore evaluation can create a new evaluative expectation. Participants may try to appear independent, frank or resistant to influence. Such a condition needs its own channel assessment. The study should avoid defining one setting as intrinsically free of social meaning.

11.3 Context and the intended object of evaluation

The third objection asks whether contextual effects are sometimes part of the phenomenon. An artwork encountered in a gallery has a different social setting from an isolated image on a screen. A research question about situated aesthetic experience may properly include those conditions.

The target should determine the interpretation. If the study concerns judgments in a gallery interaction, removing the interaction changes the object of inquiry. If the study seeks a response under a specified counterfactual absence of evaluation pressure, that target requires a defensible intervention description. Treating every contextual effect as contamination would prematurely decide a philosophical question about the experience being studied.

11.4 Additional measures and identification

The fourth objection proposes that more measurements will solve underidentification. Additional indicators can help, but their informational contribution depends on their relation to the target and to each other. Three measures generated by the same response strategy may add little discrimination. A neural correlate or response time can vary with attention, arousal or task preparation as well as judgment.

The proper response is to state a joint measurement model and test its assumptions. A new variable can distinguish the two mechanisms in Example 7.1 only if its distribution differs across them under warranted restrictions. The number of observed variables alone provides no proof of identification.

11.5 Bounds and substantive usefulness

The fifth objection is that broad bounds may be uninformative. That is a possible result. If the available calibration permits almost every judgment prevalence, the study has weak evidence for that target. The width can identify a need for better calibration or a narrower question.

A narrow interval can also mislead when its assumptions are unsupported. The paper therefore treats the explanation of bounds as part of the result. Researchers should report how the interval changes when plausible channel ranges widen or when stability across cues is relaxed. A sensitivity analysis is informative because it displays the dependence of a conclusion on assumptions.

11.6 Novelty and the role of the formal model

The final objection concerns the modest mathematics. Matrix rank, elementary probability and interval reasoning are familiar tools. Presenting them as a new general theory would be unwarranted.

The proposed contribution is their disciplined connection to a specific interpretive problem. The framework makes category semantics, calibration, anticipated evaluation and inquiry-induced change visible in one design. Its value should be judged by whether those distinctions produce better warranted claims and useful empirical comparisons. A broader claim of originality requires a systematic literature review beyond the bounded source pass used for this manuscript.

12 Further Research and Inferential Commitments

This section consolidates the established model consequences and identifies the next research obligations. It separates formal extension, empirical calibration and philosophical development. The sequence supports continued investigation without turning a preliminary model into a general account of beauty.

Under an independently applicable channel, full column rank gives unique population recovery. Rank deficiency produces ambiguity at interior judgment distributions, while boundary observations can behave differently. The smallest singular value supplies a stability bound. These results identify what a categorical response architecture can support before any substantive psychological interpretation is added.

The binary model yields sharp compatibility bounds under a separated rectangular channel class. Adding uncertainty in the response probability enlarges the compatible judgment set in a controlled way. The nonresponse extension shows how unidentified judgments among excluded participants affect a population claim. Each result states its semantic, calibration and population assumptions, making it possible to identify which premise fails in a counterexample.

The cue analysis establishes a further limit. Random assignment can identify a total response effect while leaving its division between judgment and reporting unresolved. Cross-cue compatibility can reject a bundle of assumptions, but it cannot select the failed assumption by itself. The participant's anticipated evaluation supplies a plausible candidate pathway whose identification requires discriminating evidence.

Formal extensions could examine coupled channel uncertainty, several ordered response categories and time-varying judgments. A polyhedral channel class might encode calibration constraints more faithfully than an independent rectangle. A dynamic model could track how viewing, explanation and

response preparation change the judgment. Each extension should preserve the distinction between an algebraically admissible model and an empirically warranted psychological account.

Empirical work should begin with task interpretation and response semantics. It should then investigate execution and recording, the transport of reporting behavior across contexts, and the content and practical role of anticipated evaluation. A participant's explanation can contribute to this inquiry when its timing and formative effects are retained. Stronger claims about a mechanism should follow from a design capable of distinguishing rival accounts.

Philosophical development remains equally important. A judgment can be formed in an encounter, expressed through a social act and revised through explanation. These relations complicate a picture of measurement as the passive extraction of an already complete private item. The current model makes room for that complication by declaring the target and time. Discussion of the ultimate character of aesthetic experience remains suspended.

Research Obligation 12.1 (Calibration with interpretable targets). Develop a measurement design in which the judgment vocabulary, response mapping and participant's account can be evaluated without using the recorded response as its own independent validation label.

The paper offers a limited framework for carrying that obligation forward. It preserves the source insight that a participant interprets the experimenter while deciding how to respond. It also preserves the possibility that the experiment changes what is judged. The resulting scientific task is to explain which aspect of the episode the evidence identifies, which remains bounded or unresolved, and which further inquiry would materially improve understanding.

References

- [1] Elizabeth S. Allman, Catherine Matias, and John A. Rhodes. Identifiability of Parameters in Latent Structure Models with Many Observed Variables. *The Annals of Statistics*, 37(6A):3099–3132, 2009.
- [2] Jonathan de Quidt, Johannes Haushofer, and Christopher Roth. Measuring and Bounding Experimenter Demand. *American Economic Review*, 108(11):3266–3302, 2018.
- [3] Ulrich Kirk, Martin Skov, Oliver Hulme, Mark Schram Christensen, and Semir Zeki. Modulation of Aesthetic Value by Semantic Context: An fMRI Study. *NeuroImage*, 44(3):1125–1132, 2009.
- [4] Martin T. Orne. On the Social Psychology of the Psychological Experiment: With Particular Reference to Demand Characteristics and Their Implications. *American Psychologist*, 17:776–783, 1962.