A Study on the Hierarchical Structure of Knowledge and the Cognition of Ignorance

Kunihiro Sugiyama
kunihiros@gmail.com

Introduction

Human knowledge possesses a multi-layered structure, and the ability to recognize one's own ignorance (metacognition) is crucial for learning and decision-making. This study proposes a recursive model of knowledge and ignorance based on a single core function: K, which represents epistemic recognition. By applying this function recursively—$K(x)$, $K(K(x))$, $K(K(K(x)))$, and so on—we formalize the hierarchical structure of self-awareness that distinguishes **Socratic wisdom** ("knowing that one does not know") from the **Dunning-Kruger effect** ("not knowing that one does not know").

Kant (1781), in his *Critique of Pure Reason*, posed the foundational question of epistemology: what are the limits of human cognition, and can reason examine itself? His answer—that reason must critique reason—established the recursive structure of self-reflection as a philosophical problem. Yet Kant's contribution was **descriptive**: he demonstrated that limits exist, but provided no apparatus for locating their precise coordinates or guiding their correction.

This study provides what classical epistemology could not: a **mathematical framework** that transforms the Kantian question from philosophical meditation into **operational methodology**. The function K does not merely describe where cognition fails—it provides the coordinates for **targeted intervention**. By representing epistemic states on a continuous scale, we gain the capacity to **observe the phenomenon of intelligence itself, and to intervene in its structure**. This framework opens the possibility of not only understanding cognition but **actively shaping its trajectory toward new forms of knowing**.

This model integrates insights from **metacognition research**, **epistemology**, and **type theory** to address three aspects that have not been sufficiently unified in existing research:

1. The **recursive nature of self-awareness**: The same epistemic question ("Do I know?") can be applied at every level of reflection.
2. The **continuous gradation of knowledge**: Knowledge states exist on a continuum from complete misconception (-1) through ignorance (0) to accurate knowledge (1).
3. The **distinction between epistemic state and phenomenological confidence**: What one knows versus how certain one feels are orthogonal dimensions.

By presenting a mathematically rigorous yet philosophically grounded framework, this study seeks to deepen our understanding of the structure of knowledge and the cognitive mechanisms of ignorance.

Contribution: A Conceptual Foundation

This paper establishes the **conceptual foundation** for a unified theory of recursive metacognition. The trichotomy of epistemic states—**knowing**, **not knowing**, and **misunderstanding**—is a universal human experience that transcends cultures, domains, and disciplines. Before elaborating measurement-theoretic models or conducting empirical validation, we must first **settle the conceptual vocabulary**.

What this paper provides: 1. A **single, unified operator** K that applies recursively at all levels of self-reflection 2. A **purely observational framework** where K is not a mental process but an **observation protocol** 3. A **complete taxonomy** (27 patterns) that classifies all possible metacognitive configurations 4. A **resolution of apparent contradictions** (e.g., $K(0) = 0$ vs $K_1 = -1$) through layer independence

What this paper deliberately does not provide: - Probabilistic measurement models (future work) - Empirical validation (orthogonal contribution) - Formal type-theoretic proofs (analogical treatment suffices for conceptual clarity)

These omissions are not gaps but **scope boundaries**. Measurement-theoretic elaboration and empirical validation require this conceptual foundation to be settled first. We invite the research community to build upon this foundation.

Paper Scope and Positioning

What This Paper Is:

This paper is a **conceptual framework paper**. Its primary contributions are:

1. A **unifying observational schema** (the K_n family) that provides a common coordinate system for first-order accuracy, metacognitive alignment, and higher-order stability
2. A **taxonomic vocabulary** (the 27-pattern classification) for discussing metacognitive phenomena
3. A **principled separation** of epistemic state (K) from phenomenological confidence (C)
4. **Illustrative derivations** showing how the framework relates to established metrics (IRT, meta-d', ECE)

What This Paper Is NOT:

This paper does **not** provide:

- Fully rigorous proofs of identifiability or consistency
- Complete probabilistic specifications of f_n and State_n
- Estimation algorithms with likelihood functions and convergence guarantees

- Empirical validation or simulation studies

These are valuable extensions that we explicitly designate as **Future Work**.

Rationale for This Scope:

The metacognition literature currently lacks a unified conceptual vocabulary that spans psychometrics (IRT), signal detection theory (meta-d'), and machine learning calibration (ECE). Before developing formal estimation theory or running experiments, the field needs **settled conceptual foundations**. This paper provides those foundations.

Formal measurement theory, estimation algorithms, and empirical validation are planned for **separate technical papers** building upon this conceptual scaffold.

Executive Summary: Framework at a Glance

> **Note**: This is a **conceptual framework paper**. The results below are illustrative derivations and informal propositions, not rigorous theorems with complete proofs. See "Paper Scope and Positioning" for details.

This section provides a concise overview of the framework's core components for readers seeking quick orientation.

Core Apparatus

Component	Definition	Purpose
K_n	Observation function at layer n	Maps State$_n$ to $[-1, 1]$
State$_n$	Epistemic state object at layer n	Target of observation
f_n	State function	Computes State$_n$ from inputs
g_n	Embedding map	Maps categorical State$_n$ to $K_n \in [-1, 1]$

Anchor Semantics (All Layers)

Value	Layer 0 Meaning	Layer $n \geq 1$ Meaning
$K_n = +1$	Knowledge (correct response)	Alignment (accurate self-assessment)
$K_n = 0$	Ignorance (no response / "I don't know")	Indeterminacy (uncertain self-assessment)

Value	Layer 0 Meaning	Layer $n \geq 1$ Meaning
$K_n = -1$	Misconception (incorrect response)	Misalignment (inaccurate self-assessment)

27-Pattern Taxonomy (Preview)

The framework generates $3 \times 3 \times 3 = 27$ metacognitive patterns from combinations of $(K_0, K_1, K_2) \in \{-1, 0, +1\}^3$. The complete enumeration appears in Section "Complete Taxonomy of 27 Metacognitive Patterns." Key patterns include:

Pattern	(K_0, K_1, K_2)	Name	Description
#14	$(0, +1, +1)$	**Socratic Wisdom**	Knows that they don't know
#5	$(-1, -1, -1)$	**Dunning-Kruger (Deep)**	Wrong, overconfident, unaware
#22	$(+1, -1, +1)$	**Imposter Syndrome (Aware)**	Correct but self-doubting, aware of this tendency

Correspondence with Established Metrics

Layer	Our Metric	Established Metric	Relationship
K_0	First-order accuracy	IRT ability θ	$K_0 \approx \tanh(\theta)$
K_1	Metacognitive alignment	meta-d'/d'	$K_1 \approx \tanh(\text{meta-d}'/2)$
K_1	Calibration	ECE	$K_1 \approx 1 - 2 \cdot \text{ECE}$
K_2	Meta-metacognitive stability	Test-retest K_1	Novel contribution

See Related Work section for detailed correspondence analysis.

Latent Variable Model (Preview)

For continuous K_n estimation, we employ a latent variable model:

$$K_n^* \sim \mathcal{N}(\mu_n, \sigma_n^2)$$

$$K_n = \begin{cases} +1 & \text{if } K_n^* > \tau^+ \\ 0 & \text{if } \tau^- \leq K_n^* \leq \tau^+ \\ -1 & \text{if } K_n^* < \tau^- \end{cases}$$

See Measurement Theory section for full specification.

Technical Contributions at a Glance

This section provides a roadmap to the **key results** in this paper.

> **Note**: As a conceptual framework paper, the results below are **illustrative derivations** and **informal propositions**, not rigorous theorems with complete proofs. Formal identifiability analysis connecting to general latent variable theory is deferred to future technical work.

Results and Propositions:

ID	Name	Content	Section
Result 1	K_0-IRT Correspondence	$K_0 = \tanh(a(\theta - b)/2)$ (illustrative derivation)	Formal Results
Result 2	K_1-Phi Correspondence	$K_1 = \phi$ under binary $\text{State}_0/\text{Claim}_1$	Formal Results
Proposition 3	K_0 Identifiability (Informal)	$\text{Var}(b_i) > 0 \Rightarrow K_0$ identifiable	Formal Results
Proposition 4	K_1 Identifiability (Informal)	Variability conditions for K_1 identification	Formal Results
Proposition 5	K_2 Identifiability (Informal)	ICC-based identifiability conditions	Formal Results
Proposition 6	Pipeline Identifiability (Informal)	Joint identifiability of (K_0, K_1, K_2)	Formal Results

Falsifiable Predictions with Quantitative Bounds:

ID	Prediction	Quantitative Bound	Section
Pred 1	K-C Dissociation	$\exists$ subjects with $(K_1 = -1, C = \text{high})$	Axiom F
Pred 2	Layer Independence	$\text{Cor}(K_0, K_2 \mid K_1) \approx 0$	Axiom F

ID	Prediction	Quantitative Bound	Section
Pred 4	Intervention Effect	$\mathbb{E}[\Delta K_1] \geq 0.2$, $\lvert\mathbb{E}[\Delta K_0]\rvert \leq 0.1$	Axiom F
Pred 5	K-C Correlation Bound	$\mathrm{Cor}(K_1, C) < 0.85$	Axiom F

Critical Clarifications:

- $K_1 \approx \tanh(\mathbf{meta\text{-}d'}/2)$: This is a **conceptual relationship**, NOT a mathematical identity (see Remark 1)
- $\hat{K} =$ **identity**: Justified by Lemma 3 under standard conditions
- **Correspondences are illustrative derivations** under simplifying assumptions, not rigorous proofs: Result 1 provides derivation from 2PL-IRT under idealized conditions

Philosophical Foundation and Interpretive Notes

This section clarifies the philosophical motivation behind this paper and provides essential interpretive guidance to prevent misunderstanding of the proposed model.

Theoretical Rationale

This study is grounded in the logical structure of recursive ignorance, exemplified by the proposition **"I don't know what I don't know."** If "knowing one's ignorance" (Socratic wisdom) is a recognized concept, then logically, "not knowing one's ignorance" must also exist. And if that exists, then so must "not knowing that one doesn't know one's ignorance"—and so on, recursively.

The goal of this paper is to **mathematically formalize this recursive structure of knowledge and ignorance**, not to judge or rank cognitive states.

Descriptive Nature of the Scale

The values -1, 0, and 1 in this model function as **epistemic state descriptors**. They serve as epistemic coordinates rather than normative metrics (e.g., "good" or "bad").

Value	Meaning
1	The subject holds correct knowledge.
0	The subject lacks knowledge (ignorance).
-1	The subject holds incorrect knowledge (misconception).

A subject in state -1 (misconception) is not normatively inferior to a subject in state 0 (ignorance); they occupy **distinct epistemic loci**. Whether one state is "preferable" to another depends on context, goals, and values—domains outside the scope of this model.

Separation of Knowledge and Confidence

A fundamental distinction in this framework is that **the function K measures epistemic state, not phenomenological confidence**. Confidence is a separate dimension that will be introduced later in the measurement section.

- $K(x)$: How accurately the subject recognizes object x (epistemic state).
- C (Confidence): How certain the subject feels about their recognition (phenomenological experience).

This separation is essential for capturing phenomena like the Dunning-Kruger effect, where $K(x) = 0$ (the subject does not know) but $K(K(x)) = -1$ (the subject misrecognizes their ignorance), often accompanied by high subjective confidence.

Operational Boundary: K_1 vs Confidence A natural question arises: if both K_1 and confidence C assess subjective epistemic states, how do they differ operationally? The distinction lies in **timing**, **information basis**, and **what is measured**.

Conceptual Distinction:

Dimension	Confidence (C)	K_1
Timing	Before or during response	After feedback
Information Basis	No external reference	Correctness revealed
What is Measured	Subjective certainty	Meta-cognitive accuracy
Definition	"How sure am I?"	"Given my answer was [correct/incorrect], was my claim appropriate?"

Operational Distinction:

- **Confidence C**: Collected at t_1 (during response), before feedback. The subject reports certainty without knowing correctness. This is a phenomenological measure.

- K_1 **(Meta-Accuracy)**: Assessed at t_2 (after feedback), comparing the claim ("I know"/"I don't know") to revealed correctness. This is a behavioral measure of calibration.

Why Both Are Needed—Divergence Scenarios:

Scenario	K_0	Claim	K_1	Confidence C	Pattern
Calibrated Expert	$+1$	"I know"	$+1$	High	All aligned
Overconfident Novice	-1	"I know"	-1	High	$K_1 \neq C$ direction
Underconfident Expert	$+1$	"I don't know"	-1	Low	$K_1 \neq C$ direction
Appropriate Uncertainty	0	"I don't know"	$+1$	Low	Calibrated ignorance

The overconfident novice and underconfident expert both show $K_1 = -1$ (meta-miscalibration) but with opposite confidence levels—demonstrating that K_1 and C carry independent information.

Joint Model (K_0, K_1, C)**:**

When the full triple is observed, richer patterns emerge:

Pattern	(K_0, K_1, C)	Interpretation
Dunning-Kruger	$(-1, -1, \text{High})$	Wrong, claims knowing, confident
Impostor Syndrome	$(+1, -1, \text{Low})$	Right, claims not knowing, unconfident
Calibrated Competence	$(+1, +1, \text{High})$	Right, claims knowing, appropriately confident
Calibrated Uncertainty	$(0, +1, \text{Low})$	Uncertain, claims not knowing, appropriately unconfident
Anxious Accuracy	$(+1, +1, \text{Low})$	Right, claims knowing, but feels unconfident

This joint model reveals that **Dunning-Kruger and Impostor Syndrome are symmetric patterns** in the (K_1, C) space, while K_0 distinguishes their actual competence. The independence of K_1 and C is thus not merely conceptual but empirically testable through dissociation patterns.

Scope Clarification: Methodological Relativism

This section clarifies the scope and philosophical stance of this framework.

What This Model Does

This model provides a **mathematical apparatus** for representing and manipulating the **structure of epistemic states** relative to a proposition. The function $K(x)$ measures the subject's epistemic state regarding proposition x:

$K(x)$	State	Interpretation
1	Aligned	Subject's belief is consistent with x
0	Indeterminate	Subject has no determinate stance on x
-1	Opposed	Subject's belief is contrary to x

What This Model Does Not Do

This model **does not adjudicate** what is "correct" or "true." The designation of a proposition as the "target" (such that $K(x) = 1$ represents success) is a **methodological choice** made by the experimenter, not a claim of this framework.

The proposition x itself serves as the **implicit reference point**. What counts as "aligned" ($K(x) = 1$) versus "opposed" ($K(x) = -1$) is determined by the **experimental context** (e.g., expert consensus, empirical measurement, community agreement), not by this model.

Methodological Relativism

This framework adopts a position of **methodological relativism**: the reference point is necessarily context-dependent, and the model operates on the structure of epistemic states relative to that chosen reference. This design allows researchers with different philosophical commitments (realism, relativism, pragmatism) to use the same mathematical framework while maintaining their preferred interpretation of what constitutes "correct" knowledge.

Reconciling Objectivity and Methodological Relativism

The Apparent Tension:

This framework claims both: 1. **Objective evaluation**: K_n values are determined by observable behavior, not subjective judgment 2. **Methodological relativism**: The reference standard ("correct" answer) is context-dependent

Resolution: Objectivity as Operational Repeatability

"Objective" in this framework means **operationally repeatable given a reference**:

$$\text{Objectivity} := \text{Repeatability} \mid \text{Reference}$$

Aspect	Meaning
Reference	The designated ground truth (e.g., expert consensus, textbook answer)
Procedure	The MAT protocol (response -> claim -> comparison)
Repeatability	Different observers applying the same procedure to the same data obtain the same K_n

The Relativism:

The **choice of reference** is not determined by the framework. Different communities may designate different references: - Scientific community: Peer-reviewed consensus - Educational setting: Curriculum standards - Clinical context: Diagnostic criteria

What the Framework Provides:

Once a reference is designated, the framework provides: 1. **Deterministic scoring**: Response + Claim + Reference -> K_n (via f_n and g_n) 2. **Cross-context comparability**: Same K_n semantics across different domains 3. **Transparency**: All assumptions (reference, f_n, thresholds) are explicit

Handling Reference Uncertainty:

When the reference itself is uncertain or contested:

Approach	Implementation
Probabilistic reference	Model reference as a distribution; report $E[K_n]$ and uncertainty bounds
Sensitivity analysis	Report K_n under multiple plausible references
Higher-layer encoding	Treat reference uncertainty as a property of State_{n+1}

Scope Boundary: Full treatment of contested references (e.g., scientific controversies) is beyond the current framework's scope and marked for future development.

The Recursive Structure: $K(K(K(x)))$

The cognitive structure of knowledge is modeled using a **recursive epistemic function** K. This model formalizes the intuition that the same question—"Do I know?"—can be applied at every level of self-reflection.

Formal Definition of K

We adopt an **observational family** interpretation of K, which provides a clear and consistent framework for understanding recursive metacognition.

Formal Framework: Layered Observation Model **Definition (Observation Family):**

Let $\{K^{(n)}\}_{n=0}^{\infty}$ be a family of observation functions, where each $K^{(n)}$ maps from a layer-specific state space to the epistemic scale $[-1, 1]$:

$$K^{(n)} : \mathcal{S}_n \to [-1, 1]$$

Definition (State Hierarchy):

- $\mathcal{S}_0$: First-order epistemic states (correctness of responses)
- $\mathcal{S}_1$: Metacognitive states (alignment between claims and $\mathcal{S}_0$)
- $\mathcal{S}_n$: n-th order states (alignment between claims and $\mathcal{S}_{n-1}$)

Notational Convention:

The notation $K(K(x))$ is a **shorthand** for $K^{(1)}(\text{State}_1(x))$, not numerical composition.

More precisely: - $K_0(x) := K^{(0)}(\text{State}_0(x))$ - $K_1(x) := K^{(1)}(\text{State}_1(x))$ - $K_n(x) := K^{(n)}(\text{State}_n(x))$

Shared Anchor Semantics:

All $K^{(n)}$ share the same anchor constraints: - $K^{(n)}(\text{"knowledge"}) = 1$ - $K^{(n)}(\text{"ignorance"}) = 0$ - $K^{(n)}(\text{"misconception"}) = -1$

This ensures cross-layer comparability while allowing distinct measurement procedures per layer.

Dual Notation System: Symbolic vs Formal This paper employs two complementary notational systems that serve distinct purposes:

1. Symbolic Notation: $K(K(K(x)))$

Aspect	Description
Purpose	Philosophical intuition and rhetorical clarity
Meaning	The recursive structure of “not knowing that one does not know”
Origin	The paper’s foundational insight: ignorance of ignorance of ignorance
Usage	Introduction, motivation, conceptual discussion, examples

Aspect	Description
Status	Evocative shorthand, NOT operational definition

2. Formal Notation: $K_n(x)$

Aspect	Description
Purpose	Mathematical rigor and operational precision
Definition	$K_n(x) := K^{(n)}(\text{State}_n(x)) = \hat{K}(g_n(\text{State}_n(x)))$
Usage	Definitions, axioms, measurement protocols, empirical analysis
Status	Operational definition for all formal purposes

Critical Distinction:

The symbolic notation $K(K(x))$ is **NOT** numerical composition (i.e., K applied to its own output value). It is a **rhetorical device** representing the philosophical insight that the same epistemic question ("Do I know?") applies recursively at every level of reflection.

$$K(K(x)) \not\equiv K(K_0(x)) \quad \text{(NOT numerical composition)}$$
$$K(K(x)) \equiv K_1(x) \quad \text{(notational equivalence only)}$$

For all formal purposes—definitions, proofs, measurement, analysis—use $K_n(x)$ exclusively.

Why Both?

- $K(K(K(x)))$ captures the *philosophical essence*: the infinite regress of self-reflection
- $K_n(x)$ enables *mathematical precision*: layer-specific observation with distinct objects

This dual system parallels established practice: - Physics: $E = mc^2$ (iconic) vs tensor formulation (operational) - Calculus: $\frac{dy}{dx}$ (Leibniz, intuitive) vs $\lim_{h\to 0}$ (rigorous)

DEFINITION (Core Formal Apparatus):

1. **State objects**: $\text{State}_n \in \{+1, 0, -1\}$ for each layer n

2. **State functions**: $f_n : (\text{State}_{n-1}, \text{Claim}_n) \to \text{State}_n$
3. **Embedding maps**: $g_n : \text{State}_n \to [-1, 1]$
4. **Observation**: $K_n(x) = \hat{K}(g_n(\text{State}_n(x)))$
5. **Anchor constraints**: $\hat{K}(-1) = -1$, $\hat{K}(0) = 0$, $\hat{K}(+1) = +1$

All formal work in this paper operates within this apparatus.

Unified Formalization via Embedding Maps To resolve the apparent type ambiguity between $K^{(n)} : \mathcal{S}_n \to [-1, 1]$ and the recursive $K : [-1, 1] \to [-1, 1]$, we introduce **embedding maps** that convert categorical states to the continuous scale.

Definition (Embedding Maps):

Each layer has an embedding map g_n that converts categorical states to the continuous scale:

$$g_n : \mathcal{S}_n \to [-1, 1]$$

Where: - $g_0 : \{\text{correct}, \text{incorrect}, \text{absent}\} \to \{1, -1, 0\}$ - $g_1 : \{\text{aligned}, \text{misaligned}, \text{uncertain}\} \to \{1, -1, 0\}$ - $g_n : \{\text{aligned}, \text{misaligned}, \text{uncertain}\} \to \{1, -1, 0\}$ for $n \geq 1$

State$_0$ Canonical Mapping:

f_0 Output	g_0 Value	Interpretation
correct	1	Response matches reference
incorrect	-1	Response contradicts reference
absent	0	No response / “I don’t know”

Clarification: “Ignorance” in the sense of $K_0 = 0$ means **absence of determinate stance**, not “being wrong.” A wrong answer is a **misconception** ($K_0 = -1$), not ignorance.

Definition (Unified Observation Scorer):

After embedding, a single scorer $\hat{K}$ operates on the embedded space:

$$\hat{K} : [-1, 1] \to [-1, 1]$$

With the identity mapping for prototypical anchors: $\hat{K}(1) = 1$, $\hat{K}(0) = 0$, $\hat{K}(-1) = -1$.

Composition:

$$K_n(x) = \hat{K}(g_n(\text{State}_n(x)))$$

Role of the Unified Scorer $\hat{K}$ **Question: Is $\hat{K}$ Necessary?**

Given that: - $g_n : \mathcal{S}_n \to [-1, 1]$ embeds state spaces into the common scale - Anchors are mapped to $\{-1, 0, +1\}$ by g_n

Is an additional $\hat{K} : [-1, 1] \to [-1, 1]$ needed?

Answer: $\hat{K}$ Serves Three Functions

Function 1: Anchor Preservation Guarantee

$$\hat{K}(g_n(\text{anchor})) = g_n(\text{anchor}) \quad \text{for anchors } \in \{-1, 0, +1\}$$

This ensures that regardless of how g_n handles intermediate values, the anchor semantics are preserved.

Function 2: Cross-Layer Normalization

If different layers use different g_n with varying intermediate behaviors:

$$\hat{K} \circ g_n \text{ ensures comparability across layers}$$

Function 3: Monotonicity Enforcement

$\hat{K}$ can enforce global monotonicity even if g_n has local non-monotonicities:

$$x < y \Rightarrow \hat{K}(x) \leq \hat{K}(y)$$

Specification of $\hat{K}$ **Default Choice: Identity**

For most applications:

$$\hat{K}(x) = x$$

Under this choice, $\hat{K}$ is formally present but operationally redundant.

Non-Trivial Choice: Anchor-Preserving Power Function

For applications requiring stronger intermediate compression:

$$\hat{K}(x) = \text{sign}(x) \cdot |x|^{\gamma}, \quad \gamma \in (0, 1]$$

This preserves anchors ($\hat{K}(\pm 1) = \pm 1$, $\hat{K}(0) = 0$) while compressing intermediate values.

Selection Criterion:

Application	Recommended $\hat{K}$
Categorical K_n only	Identity
Continuous K_n with calibration metrics	Identity
Continuous K_n with arbitrary g_n	Anchor-preserving power function

Formal Definition (Updated):

$$K_n = \hat{K}(g_n(f_n(\text{Response}_n, \text{Claim}_n, \text{State}_{n-1})))$$

With default $\hat{K}$ = identity:

$$K_n = g_n(f_n(\text{Response}_n, \text{Claim}_n, \text{State}_{n-1}))$$

Notational Convention (Unified):

- K_n: The full pipeline (embedding + scoring) for layer n
- $K^{(n)}$: Shorthand for the same, emphasizing layer-specificity
- $K(K(x))$: Informal shorthand for $K_1(x)$, **not** numerical composition

This resolves the type ambiguity: K is **not** applied to its own numerical output, but to the embedded representation of a distinct state object.

Error Model for State Functions Deterministic vs Probabilistic Interpretation:

The functions f_n and g_n can be interpreted in two ways:

Option A: Deterministic (Default)

The comparison functions f_n are deterministic mappings:

$$f_0 : \text{Response}(x) \times \text{Reference}(x) \to \{\text{correct}, \text{incorrect}, \text{absent}\}$$

This assumes: - Reference is unambiguous - Response classification is clear-cut - No measurement error in observation

Use case: Controlled experiments with factual questions and expert consensus.

Option B: Probabilistic (Extended)

For noisy or uncertain contexts, f_n can be extended to probabilistic mappings:

$$f_0 : \text{Response}(x) \times \text{Reference}(x) \to \Delta(\{\text{correct}, \text{incorrect}, \text{absent}\})$$

Where $\Delta(\cdot)$ denotes probability distributions over the outcome space.

Noise Model (if Option B):

$$P(\text{State}_n = s | \text{True State} = s^*) = \begin{cases} 1 - \epsilon_n & \text{if } s = s^* \\ \epsilon_n / 2 & \text{otherwise} \end{cases}$$

Where ϵ_n is the layer-specific error rate.

Recommendation:

- Use **deterministic** interpretation for conceptual clarity (this paper)
- Use **probabilistic** extension for empirical work with measurement noise
- Always report which interpretation is assumed

Entry and Recursive Mappings 1. Entry Mapping (Layer 0):

For an abstract object x (e.g., a proposition, a task item, or any epistemic target), the subject's first-order epistemic condition is represented as a continuous value:

$$K_0(x) \in [-1, 1]$$

This is the **only point** where the external object x enters the model. The internal representation K_0 captures how the subject stands with respect to x.

2. Recursive Mapping (Layers $n \geq 1$):

At all higher layers, the observation family $\{K^{(n)}\}$ acts on layer-specific states:

$$K^{(n)} : \mathcal{S}_n \to [-1, 1]$$

The "object" of higher-order $K^{(n)}$ is not the numerical output of the previous layer but the **distinct state object** State_n.

Output Interpretation (Prototypical Anchor Points):

The values -1, 0, and 1 serve as **prototypical anchors** on the continuous scale $[-1, 1]$:

- $K(\cdot) = 1$: The subject accurately recognizes the target (full knowledge or accurate metacognition).
- $K(\cdot) = 0$: The subject has no determinate stance regarding the target (pure ignorance).
- $K(\cdot) = -1$: The subject misrecognizes the target (misconception or metacognitive failure).

All intermediate values represent **graded mixtures** of these prototypes (partial knowledge, partial misconception, uncertainty, etc.).

Key Insight: The function K has **consistent semantics** across all layers: "How accurately does the subject recognize this target?" At Layer 0, the target is an external object x. At Layers $n \geq 1$, the target is the subject's own epistemic state K_{n-1}.

Observation and Objects: The Purely Observational Framework

This framework adopts a **purely observational stance**: epistemic states are operationalized as observable responses, and metacognitive states as observable alignments between claims and performance. We do not posit internal "beliefs" or "perceptions" — only **states** that are measurable by an external observer.

The Observer:

In experimental settings, the "observer" is the **experimenter** who: 1. Presents a task/question 2. Records the respondent's answer 3. Compares the answer to a reference 4. Assigns a K value based on the comparison

The Observer does not access internal "perception" or "belief." The Observer only sees **observable behavior**: answers, claims, responses.

Formal Definition of State$_n$:

State$_0$ (First-Order Epistemic State):

$$\text{State}_0(x) = f_0(\text{Response}(x), \text{Reference}(x))$$

Where: - Response(x): Subject's answer to item x - Reference(x): Ground truth or expert consensus - f_0: Comparison function yielding {correct, incorrect, absent}

> **Critical Distinction: Abstention vs. Ignorance**
>
> The framework sharply distinguishes between two conceptually different phenomena:
>
> - **Abstention** (State$_0$ = absent): A *behavioral* observable—the subject chooses not to respond. This is measured at the object level ($n = 0$) as Response(x) $= \emptyset$ and enters f_0 as the third output category.
> - **Ignorance** ($K_1 = -1$ with State$_0$ = correct): A *metacognitive* state—the subject responded correctly but believes they were guessing. This is a Type B error (unrecognized knowledge) measured at $n = 1$.
>
> The distinction matters because: 1. Abstention requires no response; Ignorance requires a correct response with low confidence 2. Abstention may reflect strategic behavior (risk aversion); Ignorance reflects

genuine metacognitive failure 3. The two patterns have different educational implications: abstention may benefit from encouragement to engage, while Type B ignorance requires confidence calibration

In forced-choice paradigms where abstention is not permitted, f_0 reduces to $\{\text{correct}, \text{incorrect}\}$, and the framework focuses purely on metacognitive accuracy. When abstention is permitted, it provides additional diagnostic information about strategic epistemic behavior.

State$_1$ (Metacognitive State):

$$\text{State}_1(x) = f_1(\text{Claim}_1(x), \text{State}_0(x))$$

Where: - $\text{Claim}_1(x)$: Subject's metacognitive claim ("I know" / "I don't know" / "I'm wrong") - f_1: Alignment function yielding {aligned, uncertain, misaligned}

State$_n$ (n-th Order State):

$$\text{State}_n(x) = f_n(\text{Claim}_n(x), \text{State}_{n-1}(x))$$

Graphical Model:

```
Response(x) --+
              +--> f_0 --> State_0 --> K^(0) --> K_0
Reference(x) -+              |
                             v
Claim_1(x) --------------> f_1 --> State_1 --> K^(1) --> K_1
                                    |
                                    v
Claim_2(x) ----------------------> f_2 --> State_2 --> K^(2) --> K_2
```

Precise Specification of State Objects and Functions The following tables provide complete, reproducible definitions for each layer.

State$_0$: First-Order Epistemic State

Definition:

$$\text{State}_0 \in \{\text{correct}, \text{incorrect}, \text{absent}\}$$

State Function f_0:

Response r	Reference t	$\text{State}_0 = f_0(r, t)$
Any answer a	$a = t$	correct
Any answer a	$a \neq t$	incorrect
No response / "I don't know"	Any	absent

K_0 Computation:

$$K_0(x) = g_0(f_0(r, t))$$

State$_1$: Second-Order Metacognitive State

Definition:

$$\text{State}_1 \in \{\text{aligned}, \text{uncertain}, \text{misaligned}\}$$

Claim Vocabulary C_1:

Claim	Meaning
"I know this"	Subject claims knowledge
"I'm not sure"	Subject expresses uncertainty
"I don't know this"	Subject claims ignorance

State Function f_1:

K_0 (actual)	Claim$_1$	State$_1$ = $f_1(K_0, \text{Claim}_1)$
+1 (knows)	"I know"	aligned
+1 (knows)	"I'm not sure"	uncertain
+1 (knows)	"I don't know"	misaligned
0 (ignorant)	"I don't know"	aligned
0 (ignorant)	"I'm not sure"	uncertain
0 (ignorant)	"I know"	misaligned
−1 (wrong)	"I don't know"	aligned*
−1 (wrong)	"I'm not sure"	uncertain
−1 (wrong)	"I know"	misaligned

*Note: "I don't know" when holding a misconception is partial awareness; coded as "aligned" for monotonicity.

Rationale: Epistemic Improvement Criterion

The coding of $K_0 = -1$ (misconception) + "I don't know" → aligned ($K_1 = +1$) may appear counterintuitive. We justify this via the **Epistemic Improvement Criterion**: a claim is "aligned" if it represents the best available response given the subject's actual state.

Argument:

1. A subject with $K_0 = -1$ who claims "I know" is **doubly wrong**: wrong about the content AND wrong about their epistemic state.
2. A subject with $K_0 = -1$ who claims "I don't know" is **partially correct**: wrong about the content but aware of their uncertainty.
3. This awareness ($K_1 = +1$) is **epistemically valuable**: it opens the door to correction and learning.

Best Response for Each K_0:

K_0	Best Claim$_1$	Reasoning
$+1$	"I know"	Accurate confidence
0	"I don't know"	Accurate ignorance
-1	"I don't know"	Protective epistemic humility

Monotonicity Preservation:

The current coding preserves the ordering:

$$\text{"I know" when wrong} \prec \text{"I don't know" when wrong} \prec \text{"I don't know" when ignorant} \prec \text{"I know" when right}$$

This ordering is monotonic in epistemic quality and consistent with the Socratic wisdom tradition.

Alternative: Graded Alignment (Future Extension)

A graded scheme could assign: - $K_0 = -1$, Claim = "I don't know" → $K_1 = 0.5$ (partial alignment) - $K_0 = -1$, Claim = "I know" → $K_1 = -1$ (full misalignment)

This is a valid design choice but complicates anchor semantics. We leave graded alignment for future extension.

Decision-Theoretic Analysis of Coding Choices The coding of $K_0 = -1$ (misconception) + "I don't know" → aligned ($K_1 = +1$) has been justified via the Epistemic Improvement Criterion above. Here we provide a complementary **decision-theoretic** analysis.

The Contested Case Revisited

K_0	Claim$_1$	Current Coding	Alternative
-1 (misconception)	"I don't know"	aligned ($K_1 = +1$)	partial ($K_1 = 0$)?

Decision-Theoretic Framework Define a loss function $L(K_0, \text{Claim}_1, \text{Action})$ where Action is taken based on the claim.

Scenario: Selective Prediction

K_0	Claim_1	Action	Outcome	Loss
-1	"I know"	Trust answer	Wrong answer used	High
-1	"I don't know"	Abstain	Correct abstention	Low
$+1$	"I know"	Trust answer	Correct answer used	Low
$+1$	"I don't know"	Abstain	Missed opportunity	Medium

Implication: Under selective prediction loss, "I don't know" when $K_0 = -1$ is *optimal*, supporting $K_1 = +1$ coding.

When Current Coding May Be Suboptimal **Scenario: Forced Response**

If abstention is not allowed: - "I don't know" when $K_0 = -1$ does not prevent the wrong answer from being used - The coding $K_1 = +1$ may overstate alignment

Scenario: Asymmetric Costs

If false positives are much worse than false negatives: - "I don't know" provides less protection than explicit correction - A more conservative coding ($K_1 = 0$) might be appropriate

Coding Sensitivity Analysis We present the current coding as the **default** under the following assumptions: 1. Abstention is possible 2. Costs are approximately symmetric 3. Epistemic improvement (openness to correction) is valued

Alternative Coding Table (for specialized applications):

K_0	Claim_1	Default K_1	Conservative K_1	Rationale
-1	"I know"	-1	-1	Unanimous: worst case
-1	"I don't know"	$+1$	0	Contested: depends on loss
-1	"Not sure"	0	0	Appropriate uncertainty

Recommendation: Use default coding unless task-specific loss analysis indicates otherwise. Document coding choice and rationale.

State_2: Third-Order Meta-Metacognitive State

Definition:

$$\text{State}_2 \in \{\text{meta-aligned}, \text{meta-uncertain}, \text{meta-misaligned}\}$$

Claim Vocabulary C_2:

Claim	Meaning
"My self-assessment is accurate"	Subject endorses their K_1
"I'm not sure about my self-assessment"	Subject uncertain about K_1
"My self-assessment may be wrong"	Subject doubts their K_1

State Function f_2 (Complete 9-Pattern Enumeration):

K_1 (actual)	Claim$_2$	State$_2$	K_2
+1 (aligned)	"My self-assessment is accurate"	meta-aligned	+1
+1 (aligned)	"I'm not sure about my self-assessment"	meta-uncertain	0
+1 (aligned)	"My self-assessment may be wrong"	meta-misaligned	−1
0 (uncertain)	"My self-assessment is accurate"	meta-misaligned	−1
0 (uncertain)	"I'm not sure about my self-assessment"	meta-aligned	+1
0 (uncertain)	"My self-assessment may be wrong"	meta-misaligned	−1
−1 (misaligned)	"My self-assessment is accurate"	meta-misaligned	−1
−1 (misaligned)	"I'm not sure about my self-assessment"	meta-uncertain	0
−1 (misaligned)	"My self-assessment may be wrong"	meta-aligned	+1

Interpretation of State$_2$ Values:

State$_2$	K_2	Meaning
meta-aligned	$+1$	Subject's belief about their self-assessment accuracy matches reality
meta-uncertain	0	Subject expresses uncertainty about their self-assessment
meta-misaligned	-1	Subject's belief about their self-assessment accuracy contradicts reality

---

Summary: The Complete Pipeline

```
Response(x) -> f$_0$(Response, Reference) -> State$_0$ -> g$_0$ -> K$_0$
                                          v
Claim$_1$ + K$_0$ -> f$_1$(K$_0$, Claim$_1$) -> State$_1$ -> g$_1$ -> K$_1$
                                          v
Claim$_2$ + K$_1$ -> f$_2$(K$_1$, Claim$_2$) -> State$_2$ -> g$_2$ -> K$_2$
```

Reproducibility: Given the same (Response, Claim$_1$, Claim$_2$, Reference), any observer following this specification will compute identical (K_0, K_1, K_2).

DEFINITION (Higher-Order Alignment States):

For layer $n \geq 1$, the alignment state is determined by comparing actual K_{n-1} with Claim$_n$:

Term	Condition	K_n
Aligned	Claim$_n$ correctly describes K_{n-1}	$+1$
Uncertain	Claim$_n$ = "I'm not sure" AND $K_{n-1} = 0$	0
Misaligned	Claim$_n$ contradicts K_{n-1}	-1

Formal Rule:

$$K_n = \begin{cases} +1 & \text{if Claim}_n \text{ matches } K_{n-1} \\ 0 & \text{if Claim}_n = \text{"uncertain" and } K_{n-1} = 0 \\ -1 & \text{if Claim}_n \text{ contradicts } K_{n-1} \end{cases}$$

Operational Interpretation:

- **State$_0$ (first-order epistemic state)**: The respondent's answer compared to a reference.
  - Operationalized as: "Is the answer correct, incorrect, or absent?"
- **State$_1$ (metacognitive state)**: The alignment between the respondent's metacognitive claim and their actual State$_0$.
  - Operationalized as: "Does the claim 'I know' match the actual correctness?"
- **State$_2$ (meta-metacognitive state)**: The alignment between the respondent's meta-metacognitive claim and their actual State$_1$.

Example: - Respondent answers incorrectly -> State$_0$ = "incorrect" -> $K_0 = -1$ (misconception) - Respondent says "I don't know" -> State$_0$ = "absent" -> $K_0 = 0$ (ignorance) - Respondent claims "I know" (when $K_0 = -1$) -> State$_1$ = "misalignment" -> $K_1 = -1$ - Respondent claims "My self-assessment is accurate" -> State$_2$ = "misalignment" -> $K_2 = -1$

Critical Clarification:

Each K_n is an **independent observation** of a **distinct object** (State$_n$):

- $K_0(x)$: Observer's measurement of **State$_0$**
- $K_1(x)$: Observer's measurement of **State$_1$**
- $K_2(x)$: Observer's measurement of **State$_2$**

$$K_1(x) \neq K(K_0(x))$$

K_1 is **not** "applying K to the numerical value of K_0." K_1 is "observing a different object (State$_1$) and reporting the measurement."

Resolving the Apparent Contradiction ($K(0) = 0$ vs $K_1 = -1$):

The axiom $K(0) = 0$ means: "If the observed state is 'ignorance' (0), the measurement result is 'ignorance' (0)."

In the Dunning-Kruger case: - $K_0(x) = 0$: Observer measures State$_0$ -> "ignorance" - $K_1(x) = -1$: Observer measures State$_1$ -> "misalignment"

State$_1$ is not "0". State$_1$ is the metacognitive state (alignment/misalignment), which the observer measures as "misrecognition" (-1).

The axiom $K(0) = 0$ does not apply because the input to K_1 is **not** the number "0". The input is **State$_1$**, a different object entirely.

Analogy (Thermometer Calibration):

Consider a thermometer and its calibration: - **State$_0$ (temperature)**: The actual temperature of water = 20°C - **State$_1$ (thermometer accuracy)**: Whether the thermometer correctly reads State$_0$

Measuring $\text{State}_0 = 20°\text{C}$ does not constrain State_1. The thermometer might be: - Accurate (State_1 = correct) -> $K_1 = 1$ - Miscalibrated (State_1 = incorrect) -> $K_1 = -1$

K_1 measures a **property of the measuring instrument**, not the original object. Similarly, K_1 measures the accuracy of **the respondent's self-monitoring**, not the first-order state itself.

Observation vs Intervention: Clarifying the Framework's Scope

The Observational Stance:

This framework is **observational** in the following sense: - K_n values are determined by observable behavior (responses, claims) - No internal mental states are posited - The observer does not require privileged access to the subject's mind

The Interventional Possibility:

The phrase "targeted intervention" in the Introduction refers to a **downstream application**, not a claim within the framework itself:

1. **Observation**: Measure K_0, K_1, K_2 via the MAT protocol
2. **Classification**: Identify metacognitive pattern (e.g., Dunning-Kruger: $K_0 = 0$, $K_1 = -1$)
3. **Intervention design** (external to framework): Choose intervention based on classification
4. **Re-observation**: Measure K_n again to assess intervention effect

What the Framework Provides:

- **Coordinates** for locating epistemic states
- **Taxonomy** for classifying metacognitive patterns
- **Outcome measures** for evaluating interventions

What the Framework Does NOT Provide:

- **Causal model** of how interventions change K_n
- **Mechanism** by which metacognition operates
- **Prescriptions** for which interventions to use

See Thermometer Calibration Analogy above for an illustration of the observation/intervention distinction.

Symbolic Notation: A Conceptual Communication Tool

Purpose and Status:

The notation $K(K(x))$ serves as an **intuitive communication device** for conveying the core insight of recursive metacognition to interdisciplinary audiences. It captures the philosophical essence of "not knowing that one does not know" more vividly than indexed notation.

Formal Status: - This notation is **NOT used in formal definitions, proofs, or measurement protocols** - All formal operations are defined via $K_n(\text{State}_n)$ exclusively - The symbolic notation has **NO independent operational semantics**

Translation Convention:

Symbolic (intuitive)	Formal (operational)	Meaning
$K(x)$	$K_0(x)$	First-order epistemic observation
$K(K(x))$	$K_1(x)$	Second-order metacognitive observation
$K(K(K(x)))$	$K_2(x)$	Third-order meta-metacognitive observation

Why Retain Symbolic Notation?

1. **Philosophical resonance**: The recursive imagery $K(K(K(...)))$ conveys the unbounded nature of self-reflection
2. **Interdisciplinary accessibility**: Non-technical readers grasp the recursive structure intuitively
3. **Historical continuity**: Echoes classical formulations (Socrates' "knowing that I do not know")

What Symbolic Notation Does NOT Provide: - Mathematical type structure - Operational definitions - Measurement protocols - Formal proofs

The K_n notation is the SOLE formal apparatus of this framework.

Axiomatic Constraints on K

We impose the following minimal constraints on the epistemic function K:

Definition: Objective Evaluation

The function K represents an **objective evaluation** of the subject's epistemic state by an external observer (or the system), distinct from the subject's subjective feeling of confidence (C).

- $K(x) = 1$: Objectively accurate recognition.
- $K(x) = -1$: Objectively inverted recognition (misconception).

Axiom Scope:

The axioms describe the behavior of the observation function K **within a single layer**.

- $K(1) = 1$: If the observed state is "knowledge", report "knowledge"
- $K(0) = 0$: If the observed state is "ignorance", report "ignorance"
- $K(-1) = -1$: If the observed state is "misconception", report "misconception"

Layer Independence:

Each K_n observes a **different object** (State_n). The axioms apply to each observation independently.

Layer	Object Observed	Axiom Application
K_0	State_0 (first-order epistemic state)	$K(\text{State}_0)$ follows axioms
K_1	State_1 (metacognitive state)	$K(\text{State}_1)$ follows axioms
K_2	State_2 (meta-metacognitive state)	$K(\text{State}_2)$ follows axioms

Why $K_0 = 0$ and $K_1 = -1$ is NOT a contradiction:

- $K_0 = 0$: Observer measures State_0 as "ignorance"
- $K_1 = -1$: Observer measures State_1 as "misrecognition"

$\text{State}_0 \neq \text{State}_1$. They are different objects. The axiom $K(0) = 0$ applies to State_0, not to State_1.

Layer Independence: Formal Conditions

Conditional Independence Assumption Definition (Layer Separation):

$$\text{State}_n \perp\!\!\!\perp \text{State}_{n-2} \mid \text{State}_{n-1}$$

This states that State_n (the object of layer-n assessment) depends only on State_{n-1}, not on earlier layers.

Implication for Scoring:

$$K_n = f_n(\text{Response}_n, \text{Claim}_n, \text{State}_{n-1})$$

The scoring function f_n does not require access to State_{n-2} or earlier.

No Circular Dependency Definition (Acyclicity):

The dependency graph over $\{K_0, K_1, K_2, \ldots\}$ is a directed acyclic graph (DAG):

$$K_0 \to K_1 \to K_2 \to \cdots$$

Where K_n depends on K_{n-1} (via State_{n-1}) but not vice versa.

Proof of Acyclicity:

1. K_0 is computed from (Response, Reference) alone—no dependency on higher layers
2. K_1 is computed from (Claim$_1$, K_0)—depends only on K_0
3. K_n is computed from (Claim$_n$, K_{n-1})—depends only on K_{n-1}

By induction, no backward dependencies exist. □

Graphical Model Representation The dependency structure can be visualized as a cascade:

$$\text{Reference} \to \text{State}_0 \to K_0 \to \text{State}_1 \to K_1 \to \text{State}_2 \to K_2 \to \cdots$$

Each layer is a separate "plate" in the graphical model, with information flowing strictly forward (lower to higher layers).

Identifiability Conditions **Definition (K_n Identifiability)**:

K_n is **identifiable** if, given sufficient observations of (Response, Claim$_1$, …, Claim$_n$, Reference), the value of K_n can be uniquely determined.

Sufficient Conditions:

Condition 1 (Reference Availability):

The reference (ground truth) for State$_0$ must be available or reliably estimable.

$$\exists \text{ Reference} : P(\text{Reference} = \text{true state}) = 1$$

Or, under probabilistic scoring:

$$P(\text{Reference}|\text{evidence}) \text{ is well-defined}$$

Condition 2 (Claim Observability):

Claims at each layer must be observable:

$$\text{Claim}_n \in \mathcal{C}_n \text{ is directly elicited and recorded}$$

Condition 3 (Deterministic Mapping):

Under Option A (deterministic scoring):

$$f_n : \mathcal{C}_n \times \{-1, 0, +1\} \to \{-1, 0, +1\}$$

is a fixed, known function. Given (Claim$_n$, K_{n-1}), K_n is uniquely determined.

Condition 4 (Probabilistic Identifiability):

Under Option B (probabilistic scoring), identifiability requires:

1. **Sufficient variation**: Items vary in true K_0 values
2. **Informative claims**: $P(\text{Claim}_1|K_0)$ differs across K_0 values
3. **No confounding**: Claim$_n$ depends on K_{n-1} but not on unobserved variables

Identifiability Violations:

Violation	Example	Consequence
Reference unavailable	Contested scientific claims	K_0 undefined
Claims unobserved	Internal confidence only	K_1 inestimable
Deterministic degeneracy	All subjects say "I know"	K_1 variance = 0
Confounding	Claim influenced by social desirability	Biased K_1

Practical Recommendation:

For robust estimation: 1. Use items with known references (e.g., factual questions with verified answers) 2. Elicit claims explicitly (not inferred from behavior) 3. Include items spanning $K_0 \in \{-1, 0, +1\}$ to ensure variation 4. Control for demand characteristics via randomized claim formats

Monotonicity (Revised):

Definition (Order on Embedded States):

For embedded values $k, k' \in [-1, 1]$, the natural order $k > k'$ applies.

Monotonicity Axiom:

For any scoring function $\hat{K}$:

$$\text{If } g_n(\text{State}_n) > g_n(\text{State}'_n), \text{ then } K_n \geq K'_n$$

This is trivially satisfied when $\hat{K}$ is the identity on anchors and monotonic elsewhere.

Practical Interpretation:

Monotonicity ensures that "more aligned" states receive higher K values. This does not constrain the shape of $\hat{K}$ beyond anchor preservation.

Monotonicity applies **within each layer** only. It does NOT constrain relationships **across layers** (e.g., K_0 vs K_1).

Boundedness:

$K : [-1, 1] \to [-1, 1]$

Each observation is bounded on this interval.

Note: We deliberately refrain from specifying stronger constraints (e.g., odd symmetry, Lipschitz constant, contraction mapping) at this stage. The framework is intended to be **descriptive** rather than **predictive**—it provides a vocabulary for classifying observed metacognitive states, not a generative model of metacognitive dynamics. Specifying a particular functional form for K is a task for domain-specific empirical research.

Recursive Application

Notation:

We use subscript notation to denote the layer of observation:

$$K_0(x) = K(\text{State}_0)$$
$$K_1(x) = K(\text{State}_1)$$
$$K_2(x) = K(\text{State}_2)$$

Important Clarification:

The traditional notation $K(K(x))$ is **shorthand** for $K_1(x)$, but it should **not** be interpreted as function composition (i.e., $K(K_0(x))$ where the numerical value of K_0 is passed to K).

Each K_n observes a **distinct object** (State_n), not the numerical output of the previous layer.

Interpretation:

Layer	Expression	Object Observed	Question
Layer 0	$K_0(x)$	State_0 (first-order epistemic state)	"Is the respondent's answer correct?"
Layer 1	$K_1(x)$	State_1 (metacognitive state)	"Is the respondent's self-assessment aligned with their actual State_0?"
Layer 2	$K_2(x)$	State_2 (meta-metacognitive state)	"Is the respondent's meta-self-assessment aligned with their actual State_1?"

Same observation function K. Different objects (State$_n$). Independent measurements.

Examples

Example 1: Knowing Knowledge

- $K(x) = 1$: The subject knows that “water boils at 100°C.”
- $K(K(x)) = 1$: The subject accurately recognizes that they know this fact.
- **Classification**: Knowing Knowledge (accurate self-awareness)

Example 2: Socratic Wisdom

- $K(x) = 0$: The subject does not know the boiling point of water.
- $K(K(x)) = 1$: The subject accurately recognizes their ignorance (“I know that I don’t know”).
- **Classification**: Knowing Ignorance (Socratic wisdom)

Example 3: Dunning-Kruger Effect

- $K(x) = 0$: The subject does not know the boiling point of water.
- $K(K(x)) = -1$: The subject misrecognizes their ignorance, believing they know.
- **Classification**: Unknowing Ignorance (Dunning-Kruger effect)

Example 4: Imposter Syndrome

- $K(x) = 1$: The subject knows that “water boils at 100°C.”
- $K(K(x)) = -1$ or 0: The subject does not recognize their knowledge (“I don’t think I know this”).
- **Classification**: Unknowing Knowledge (imposter syndrome)

The Four Quadrants of Metacognition

The relationship between $K(x)$ (actual state) and $K(K(x))$ (metacognitive accuracy) produces four archetypal patterns:

$K(x)$	$K(K(x))$	Classification	Interpretation
1 (Know)	1 (Accurate)	**Knowing Knowledge**	Accurate self-awareness
0 (Ignorant)	1 (Accurate)	**Knowing Ignorance**	Socratic wisdom
0 (Ignorant)	-1 (Misrecognition)	**Unknowing Ignorance**	Dunning-Kruger effect
1 (Know)	-1 or 0	**Unknowing Knowledge**	Imposter syndrome

Important Note: The value $K(K(x)) = -1$ for "Unknowing Ignorance" does **not** mean it is "bad" in a normative sense. It simply describes an epistemic state where the subject **misrecognizes their own ignorance**. Whether this is problematic depends on context and goals.

Complete Taxonomy: $K_0 \times K_1 \times K_2$ **(27 Patterns)**

Derivation:

Each of K_0, K_1, K_2 takes values in $\{-1, 0, +1\}$ (discretized from continuous scale via thresholds). Total configurations: $3 \times 3 \times 3 = 27$

The Full Table:

#	K_0	K_1	K_2	Name	Description
1	+1	+1	+1	**Perfect Calibration**	Knows, knows they know, knows they know they know
2	+1	+1	0	**Unreflective Expert**	Knows and knows it, but unaware of this meta-state
3	+1	+1	-1	**Doubting Expert**	Knows and knows it, but believes their self-knowledge is poor
4	+1	0	+1	**Agnostic Knower (Meta-aware)**	Knows but unsure if they know, aware of this uncertainty
5	+1	0	0	**Agnostic Knower**	Knows but unsure if they know
6	+1	0	-1	**Falsely Uncertain**	Knows but unsure, believes wrongly they're certain
7	+1	-1	+1	**Imposter (Self-aware)**	Knows but thinks they don't, aware of this pattern
8	+1	-1	0	**Classic Imposter**	Knows but thinks they don't know
9	+1	-1	-1	**Deep Imposter**	Knows, thinks they don't, believes their self-doubt is accurate

#	K_0	K_1	K_2	Name	Description
10	0	+1	+1	**Aware Ignorance (Validated)**	Doesn't know, knows this, confident in this self-knowledge
11	0	+1	0	**Socratic Wisdom**	Doesn't know and knows it ("I know that I know nothing")
12	0	+1	-1	**Doubting Socrates**	Doesn't know, knows it, but doubts this self-knowledge
13	0	0	+1	**Pure Uncertainty (Meta-aware)**	Doesn't know, unsure if they know, aware of confusion
14	0	0	0	**Complete Uncertainty**	Doesn't know, unsure if they know, unsure about that
15	0	0	-1	**Confused Certainty**	Doesn't know, unsure, but believes they're clear
16	0	-1	+1	**Dunning-Kruger (Self-aware)**	Doesn't know, thinks they know, aware of this bias
17	0	-1	0	**Classic Dunning-Kruger**	Doesn't know but thinks they know
18	0	-1	-1	**Deep Dunning-Kruger**	Doesn't know, thinks they know, confident in false belief
19	-1	+1	+1	**Aware Misconception (Validated)**	Has misconception, knows it, confident in this
20	-1	+1	0	**Aware Misconception**	Has misconception and knows it

#	K_0	K_1	K_2	Name	Description
21	-1	+1	-1	**Doubting Awareness**	Has misconception, knows it, but doubts this knowledge
22	-1	0	+1	**Uncertain Misconception (Meta-aware)**	Has misconception, unsure, aware of uncertainty
23	-1	0	0	**Uncertain Misconception**	Has misconception, unsure if they know
24	-1	0	-1	**Falsely Certain Misconception**	Has misconception, unsure, but thinks they're sure
25	-1	-1	+1	**Confident Error (Self-aware)**	Has misconception, thinks it's knowledge, aware of this risk
26	-1	-1	0	**Confident Error**	Has misconception and thinks it's knowledge
27	-1	-1	-1	**Entrenched Error**	Has misconception, thinks it's knowledge, certain of this

Completeness Argument:

The taxonomy is complete by construction: every possible $(K_0, K_1, K_2) \in \{-1, 0, +1\}^3$ combination is enumerated. No configuration is possible outside this space under the discretized model.

Notable Patterns:

Pattern	Configuration	Clinical/Educational Significance
Socratic Wisdom	$(0, +1, \cdot)$	Ideal starting point for learning
Classic Dunning-Kruger	$(0, -1, 0)$	Intervention target: calibration training
Classic Imposter	$(+1, -1, 0)$	Intervention target: confidence building

Pattern	Configuration	Clinical/Educational Significance
Entrenched Error	$(-1, -1, -1)$	Most resistant to change; requires staged approach
Perfect Calibration	$(+1, +1, +1)$	Ideal end state

Worked Examples: Representative Patterns

The following examples demonstrate how the (K_0, K_1, K_2) triplet is computed in concrete scenarios.

Example 1: Socratic Wisdom (Pattern #11: $K_0 = 0, K_1 = +1, K_2 = 0$) **Scenario**: History exam, question about the date of the Treaty of Westphalia.

Step	Observable	Computation	Value
Response	“I don’t know”	—	—
Reference	1648	—	—
State$_0$	Response = Absent	$f_0(\text{absent}, 1648) =$ absent	—
K_0	$g_0(\text{absent})$	$= 0$	$K_0 = 0$
Claim$_1$	“I correctly identified that I don’t know”	—	—
State$_1$	Claim$_1$ matches $K_0 = 0$	$f_1(0, \text{"I don't know"}) =$ aligned	—
K_1	$g_1(\text{aligned})$	$= +1$	$K_1 = +1$
Claim$_2$	“I’m not sure about my self-assessment”	—	—
State$_2$	Claim$_2$ = uncertain when $K_1 = +1$	$f_2(+1, \text{"not sure"}) =$ uncertain	—
K_2	$g_2(\text{uncertain})$	$= 0$	$K_2 = 0$

Interpretation: Subject demonstrates Socratic wisdom—accurate recognition of their own ignorance—but is modest about this metacognitive achievement.

Example 2: Deep Dunning-Kruger (Pattern #18: $K_0 = 0, K_1 = -1, K_2 = -1$) **Scenario**: Math test, question “What is $\sqrt{16}$?”

Step	Observable	Computation	Value
Response	“5”	—	—

Step	Observable	Computation	Value
Reference	4	—	—
State$_0$	Response $\neq$ Reference	$f_0(5, 4)$ = incorrect	—
K_0	g_0(incorrect)	$= -1$	$K_0 = -1$
Claim$_1$	"I'm confident I'm correct"	—	—
State$_1$	Claim$_1$ contradicts $K_0 = -1$	$f_1(-1, \text{"I know"}) =$ misaligned	—
K_1	g_1(misaligned)	$= -1$	$K_1 = -1$
Claim$_2$	"My self-assessment is reliable"	—	—
State$_2$	Claim$_2$ contradicts $K_1 = -1$	$f_2(-1, \text{"accurate"}) =$ meta-misaligned	—
K_2	g_2(meta-misaligned)	$= -1$	$K_2 = -1$

Interpretation: Triple misalignment—wrong answer, overconfident, and unaware of overconfidence. This is the "entrenched" metacognitive failure pattern.

Example 3: Imposter Syndrome Aware (Pattern #7: $K_0 = +1, K_1 = -1, K_2 = +1$) **Scenario**: Programming task, correct solution submitted.

Step	Observable	Computation	Value
Response	Correct code	—	—
Reference	Expected output	—	—
State$_0$	Response = Reference	$f_0(\text{correct}, \text{correct}) =$ correct	—
K_0	g_0(correct)	$= +1$	$K_0 = +1$
Claim$_1$	"I probably got it wrong"	—	—
State$_1$	Claim$_1$ contradicts $K_0 = +1$	$f_1(+1, \text{"I don't know"})$ = misaligned	—
K_1	g_1(misaligned)	$= -1$	$K_1 = -1$
Claim$_2$	"I know I tend to underestimate myself"	—	—
State$_2$	Claim$_2$ correctly identifies $K_1 = -1$	$f_2(-1, \text{"may be wrong"})$ = meta-aligned	—
K_2	g_2(meta-aligned)	$= +1$	$K_2 = +1$

Interpretation: Classic imposter syndrome with metacognitive awareness—the subject knows they underestimate themselves. This self-awareness ($K_2 = +1$) is a **teachable moment** for intervention.

Summary of Examples:

Example	(K_0, K_1, K_2)	Pattern	Key Insight
Socratic Wisdom	$(0, +1, 0)$	#11	Accurate ignorance recognition
Deep Dunning-Kruger	$(-1, -1, -1)$	#27	Triple misalignment, intervention-resistant
Imposter Aware	$(+1, -1, +1)$	#7	Self-aware underconfidence, teachable

Extension to Continuous Values:

The 27-pattern table is a **discrete approximation**. For continuous $K_n \in [-1, 1]$, the taxonomy becomes a partition of the unit cube $[-1, 1]^3$ into 27 regions, with boundaries at ±0.33 (see Threshold Justification below).

Partial Order Among Patterns:

Some patterns are "better" than others in terms of metacognitive calibration:

$$\text{Calibration Quality} = \text{sign}(K_0 \cdot K_1) + \text{sign}(K_1 \cdot K_2)$$

Quality	Meaning	Example Patterns
+2	Fully calibrated	#1 (Perfect Calibration), #11 (Socratic Wisdom)
+1	Partially calibrated	#2, #10
0	Mixed	#5, #14
-1	Partially miscalibrated	#8, #17
-2	Fully miscalibrated	#18 (Deep Dunning-Kruger), #27 (Entrenched Error)

This ordering is **descriptive**, not normative; specific contexts may value different patterns.

Theoretical Value of K_2:

The third layer (K_2) enables modeling of **metacognitive interventions** and their effectiveness:

- $K_2 = +1$ with $K_1 = -1$: Subject recognizes their metacognitive failure -> **teachable moment**
- $K_2 = -1$ with $K_1 = -1$: Subject does not recognize their failure -> **resistant to intervention**

Higher-order reflection (K_2, K_3, …) provides diagnostic power for identifying when and how metacognitive correction is possible.

Continuous-to-Categorical Mapping

The 27-pattern taxonomy uses prototypical anchors $\{-1, 0, 1\}$. For continuous K values, we define thresholds:

Continuous Range	Categorical Label
$K \in [-1, -0.33)$	-1 (Misconception/Misaligned)
$K \in [-0.33, 0.33]$	0 (Ignorance/Uncertain)
$K \in (0.33, 1]$	1 (Knowledge/Aligned)

Rationale for ± 0.33 Default:

1. **Symmetric Tercile**: Divides $[-1, 1]$ into three equal-width regions
2. **Neutral Zone**: The central region captures "uncertain/indeterminate" states
3. **Statistical Interpretation**: Under uniform prior, each category has equal probability
4. **Robustness**: Not sensitive to small estimation errors near boundaries

Formal Justification for Thresholds **Decision-Theoretic Grounding:**

The default thresholds ± 0.33 can be justified via decision theory:

Setup: - Agent must classify K into {misconception, ignorance, knowledge} - Utility function: $U(\text{action}, \text{true state})$ - Prior: Uniform over $[-1, 1]$

Symmetric Loss:

Under symmetric 0-1 loss (equal cost for all misclassifications):

$$\text{Optimal thresholds} = \arg\min_{\tau_1, \tau_2} E[\mathbb{1}(\text{misclassification})]$$

With uniform prior, this yields $\tau_1 = -1/3 \approx -0.33$, $\tau_2 = 1/3 \approx 0.33$.

Asymmetric Loss (Alternative):

If false positives (claiming knowledge when ignorant) are more costly:

$$L(\text{classify as 1}|\text{true 0}) = c > 1$$

Optimal thresholds shift: $\tau_2 > 0.33$ (stricter knowledge criterion).

Proper Scoring Rule Connection:

Under Brier score:

$$\text{Brier}(p, y) = (p - y)^2$$

The thresholds ± 0.33 correspond to the decision boundaries where expected Brier score is minimized under uniform prior.

Empirical Calibration (Future Work):

For domain-specific applications: 1. Collect pilot data with known ground truth 2. Compute ROC curve for each threshold 3. Select threshold maximizing Youden's J or F1 score 4. Report sensitivity analysis across threshold choices

Alternative Thresholds:

Approach	Thresholds	Use Case
Tercile (default)	± 0.33	Balanced classification
Quartile	± 0.5	Stricter knowledge/misconception criteria
ROC-optimized	Data-driven	Maximize classification accuracy
Domain-specific	Expert-defined	Match substantive theory

Recommendation: - Use ± 0.33 as default for comparability across studies - Report sensitivity analysis with alternative thresholds - For intervention design, consider ROC-optimized thresholds

Reporting Recommendation: - Report continuous K values for statistical analysis - Use categorical labels for interpretation and intervention design - Always include confidence intervals from estimation

Example:

$$K_0 = 0.7, K_1 = -0.5, K_2 = 0.2$$

Categorical: $K_0 = 1, K_1 = -1, K_2 = 0$ -> "Knowing Misconception, uncertain about meta"

This enables both fine-grained analysis and interpretable classification.

Connection with Metacognition Research

Flavell (1979) defined metacognition as "the ability to monitor and control one's own cognitive activities." The recursive structure ($K \to K(K) \to K(K(K))$) formalizes this concept mathematically.

Nelson & Narens (1990) introduced the influential **monitoring/control framework**, distinguishing between the **object level** (cognitive processes) and the **meta level** (monitoring and control of cognition). Our framework directly corresponds to this structure:

Nelson & Narens	This Framework
Object level	State_0 (first-order epistemic state)
Meta level (monitoring)	State_1 (metacognitive state)
Control signal	Not modeled (orthogonal dimension)

Our K_0 and K_1 formalize the object/meta distinction with a **single unified operator**, providing mathematical precision to Nelson & Narens' conceptual framework.

Koriat (1993) proposed the **cue-utilization theory**, explaining how confidence arises from accessibility and familiarity cues rather than direct access to accuracy. This distinction between cue-based confidence and actual accuracy corresponds precisely to our separation of C (confidence) and K (epistemic state). Our framework accommodates cue-based confidence as a component of C, while K measures the objective alignment between the subject's state and reality.

Kruger and Dunning (1999) demonstrated that individuals with low competence tend to overestimate their abilities. In our model, this corresponds to $K_0 = 0$ (ignorance) but $K_1 = -1$ (misrecognition of ignorance).

Fleming and Daw (2017) proposed a general Bayesian framework for metacognitive computation, modeling metacognition as "second-order inference" about the reliability of first-order cognitive processes. Their distinction between first-order states and second-order inference corresponds to our $\text{State}_0/\text{State}_1$ hierarchy. While their approach is Bayesian (modeling uncertainty about internal states) and ours is observational (measuring alignment between claims and performance), both frameworks capture the fundamental insight that metacognition operates on a different level from cognition itself. The K-C dissociation in our framework (epistemic state vs phenomenological confidence) parallels their analysis of how confidence can diverge from accuracy.

Meta-d' (Maniscalco & Lau, 2012) provides a signal detection-theoretic measure of metacognitive sensitivity. While meta-d' quantifies **how well** subjects discriminate their own correct from incorrect responses, our framework provides a **structural vocabulary** for **what** metacognitive states exist. The two approaches are complementary: meta-d' measures the quality of monitoring; our K classifies the content of monitoring.

HMeta-d (Fleming, 2017) extends meta-d' to a hierarchical Bayesian framework, enabling group-level inference and trial-by-trial parameter estimation. The relationship to our framework is as follows:

Aspect	HMeta-d	K Framework	Correspondence
Latent structure	Single meta-d' per subject	Multi-layer (K_0, K_1, K_2)	HMeta-d $\approx$ aggregate K_1 sensitivity
Hierarchy	Subjects nested in groups	Layers nested within subjects	Orthogonal hierarchies
Output	meta-d'/d' ratio	Discrete $K \in \{-1, 0, +1\}$ or continuous $K \in [-1, 1]$	$K_1 \approx \tanh(\text{meta-d}'/2)$
Trial structure	Binary (correct/incorrect by high/low confidence)	Ternary (knowledge/ignorance/misconception)	K adds misconception category
Higher-order	Not explicit (single meta level)	Explicit recursive $(K_2, K_3, \ldots)$	K extends to arbitrary depth

Integration Opportunity: Researchers using HMeta-d can incorporate the K framework by: 1. Using HMeta-d to estimate subject-level meta-d' with shrinkage 2. Transforming to K_1 via $\hat{K}_1 = \tanh(\widehat{\text{meta-d}'}/2)$ 3. Extending to K_2 via additional confidence judgments about Type-2 accuracy

Reference: Fleming, S. M. (2017). HMeta-d: Hierarchical Bayesian estimation of metacognitive efficiency from confidence ratings. *Neuroscience of Consciousness, 3*(1), nix007.

Novel Contribution: Our model provides a **structural formalization** of recursive self-awareness that: 1. Unifies the object/meta distinction (Nelson & Narens) with a single recursive operator 2. Separates epistemic state from cue-based confidence (Koriat) 3. Classifies all possible metacognitive configurations (27 patterns) 4. Explicitly distinguishes "Knowing Ignorance" (Socratic wisdom) as a high metacognitive achievement

Why a Single Unified K?

One might ask: "Why not introduce separate operators for different levels?" The answer lies in the **universality of the epistemic question**.

Philosophical Motivation:

The question "Do I know?" is the same question at every level: - "Do I know the answer?" -> K_0 - "Do I know whether I know the answer?" -> K_1 - "Do I know whether I know whether I know?" -> K_2

The question is universal. Only the object changes.

A thermometer measures temperature. The same thermometer can measure the temperature of water, air, or metal. We do not need separate "water-thermometers" and "air-thermometers." The instrument is the same; the objects differ.

Similarly, K is an **observation protocol**, not a mental process. The same protocol applies to different objects (State$_0$, State$_1$, State$_2$). Introducing separate operators (R, E, etc.) would obscure this fundamental unity and sacrifice the elegance of a single recursive structure.

Practical Implementation:

While the **semantic anchors** are shared (-1/0/1 for misconception/ignorance/knowledge), the **measurement procedures** $K^{(n)}$ may differ:

Layer	Observable	Measurement Procedure
$K^{(0)}$	Response vs Reference	Accuracy scoring
$K^{(1)}$	Claim vs State$_0$	Alignment scoring
$K^{(2)}$	Meta-claim vs State$_1$	Meta-alignment scoring

Parameter Tying (Optional):

For parsimony, one may assume: - Same noise model across layers - Same link function (e.g., logistic)

Or allow layer-specific parameters if data supports it.

The key constraint is shared anchor semantics, not identical functional forms.

The beauty of K is its universality. Knowing, not knowing, and misunderstanding are universal human experiences. The same operator captures them all.

Formal Results: Illustrative Derivations and Informal Propositions

This section provides illustrative derivations and informal propositions that establish the conceptual foundations of the K framework. These results show how the framework relates to established metrics and outline conditions under which key properties hold.

> **Scope Note**: As stated in "Paper Scope and Positioning," this is a **conceptual framework paper**. The results below are *illustrative derivations* under idealized assumptions, not rigorous theorems with complete proofs. Formal identifiability analysis connecting to general latent variable theory (e.g., Allman et al., 2009; Kruskal, 1977) is deferred to future technical work.

> **Summary of Contributions**: This section contains **2 results, 4 informal propositions, 1 lemma, and 5 falsifiable predictions**. The main identifiability arguments are: - **Proposition 3**: K_0 identifiability under item variance conditions - **Proposition 4**: K_1 identifiability given K_0 and claim variability - **Proposition 6**: Joint (K_0, K_1, K_2) pipeline identifiability
>
> For a condensed overview, see "Technical Contributions at a Glance" in the Executive Summary.
>
> **Note on Contribution Type**: The results below primarily show how the K framework connects to existing mathematical results (IRT, Signal Detection Theory, ICC). The novelty lies not in the underlying mathematics but in the unified conceptual integration that enables systematic metacognition analysis.

Core Results

Result 1: K_0-IRT Correspondence

> **Note**: This is an *illustrative derivation* under idealized assumptions, not a general theorem.

Result 1 (K_0-IRT Correspondence):

Assumptions (Simplifying): - **(A1)** Item response follows the 2-Parameter Logistic (2PL) IRT model:

$$P(\text{correct}|\theta, a_i, b_i) = \frac{1}{1 + e^{-a_i(\theta - b_i)}}$$

- **(A2)** K_0 is defined as the signed expected accuracy:

$$K_0^{(i)} := 2P(\text{correct}|\theta, a_i, b_i) - 1$$

Observation: Under (A1)-(A2), for each item i:

$$K_0^{(i)} = \tanh\left(\frac{a_i(\theta - b_i)}{2}\right)$$

Derivation: Starting from the 2PL probability:

$$P = \frac{1}{1 + e^{-a(\theta - b)}}$$

Transform to the $[-1, 1]$ scale:

$$K_0 = 2P - 1 = \frac{2}{1 + e^{-a(\theta - b)}} - 1 = \frac{2 - 1 - e^{-a(\theta - b)}}{1 + e^{-a(\theta - b)}} = \frac{1 - e^{-a(\theta - b)}}{1 + e^{-a(\theta - b)}}$$

Recall the hyperbolic tangent identity:

$$\tanh(x) = \frac{e^{2x} - 1}{e^{2x} + 1} = \frac{1 - e^{-2x}}{1 + e^{-2x}}$$

Matching exponents with $2x = a(\theta - b)$:

$$K_0 = \tanh\left(\frac{a(\theta - b)}{2}\right) \quad \blacksquare$$

Scope of Validity: This correspondence is exact under the stated assumptions but does not constitute a general identifiability result. Extension to realistic settings (polytomous items, multidimensional traits, measurement error) requires additional formal development.

Status: Illustrative; formal identifiability analysis deferred to future work.

Corollary 1.1 (Standardized Form): Under additional assumption **(A3)**: $a = 2$ and $b = 0$:

$$K_0 = \tanh(\theta)$$

Corollary 1.2 (Aggregate Form): For N items with varying (a_i, b_i):

$$\bar{K}_0 = \frac{1}{N}\sum_{i=1}^{N} \tanh\left(\frac{a_i(\theta - b_i)}{2}\right)$$

Boundary Condition (Deviation from Standardization): When (A3) is violated (i.e., $a \neq 2$ or $b \neq 0$), the deviation from $\tanh(\theta)$ is:

$$|K_0 - \tanh(\theta)| = |\tanh(a(\theta - b)/2) - \tanh(\theta)|$$

For typical parameter ranges in educational and psychological testing:

a	b	θ	$K_0 = \tanh(a(\theta - b)/2)$	$\tanh(\theta)$	Deviation
1.0	0.0	1.0	0.46	0.76	0.30
1.5	0.0	1.0	0.64	0.76	0.12
2.0	0.0	1.0	0.76	0.76	0.00
2.0	1.0	1.0	0.00	0.76	0.76
2.0	-1.0	1.0	0.96	0.76	0.20

Interpretation: The standardized form $K_0 \approx \tanh(\theta)$ is a good approximation only when item discrimination $a \approx 2$ and item difficulty $b \approx 0$. For items with extreme difficulty or low discrimination, the item-specific form $\tanh(a(\theta - b)/2)$ should be used.

Scope of Validity: - Holds exactly for binary correct/incorrect responses - Does not hold for partial credit models (see Limitations) - Assumes no guessing or slipping (see Q6 in Limitations)

Result 2: K_1-Phi Correspondence **Observation**: When State_0 and Claim_1 are both binary, K_1 equals the Phi coefficient.

Result 2 (K_1-Phi Correspondence): Let $\text{State}_0 \in \{\text{correct}, \text{incorrect}\}$ and $\text{Claim}_1 \in \{\text{"I know"}, \text{"I don't know"}\}$ be binary random variables. Define $K_1 := \phi(\text{State}_0, \text{Claim}_1)$ where ϕ is the Phi coefficient:

$$\phi = \frac{n_{11}n_{00} - n_{10}n_{01}}{\sqrt{(n_{11} + n_{10})(n_{01} + n_{00})(n_{11} + n_{01})(n_{10} + n_{00})}}$$

Then: 1. $K_1 \in [-1, 1]$ 2. $K_1 = +1$ iff $\text{Claim}_1 = \text{State}_0$ for all items (perfect alignment) 3. $K_1 = -1$ iff $\text{Claim}_1 = \neg\text{State}_0$ for all items (perfect anti-alignment) 4. $K_1 = 0$ iff $\text{Claim}_1 \perp \text{State}_0$ (statistical independence)

Derivation:

Let the 2×2 contingency table be:

	Claim_1 = "I know"	Claim_1 = "I don't know"
State_0 = correct	n_{11}	n_{10}
State_0 = incorrect	n_{01}	n_{00}

Property 1 ($K_1 \in [-1, 1]$): The Phi coefficient is the Pearson correlation for two binary variables. By the Cauchy-Schwarz inequality, $|\phi| \leq 1$.

Property 2 ($K_1 = +1$ iff perfect alignment): When $\text{Claim}_1 = \text{State}_0$ for all items, $n_{10} = n_{01} = 0$. Then:

$$\phi = \frac{n_{11} \cdot n_{00} - 0}{\sqrt{n_{11} \cdot n_{00} \cdot n_{11} \cdot n_{00}}} = \frac{n_{11} \cdot n_{00}}{n_{11} \cdot n_{00}} = 1$$

Property 3 ($K_1 = -1$ iff perfect anti-alignment): When $\text{Claim}_1 = \neg\text{State}_0$ for all items, $n_{11} = n_{00} = 0$. Then:

$$\phi = \frac{0 - n_{10} \cdot n_{01}}{\sqrt{n_{10} \cdot n_{01} \cdot n_{10} \cdot n_{01}}} = \frac{-n_{10} \cdot n_{01}}{n_{10} \cdot n_{01}} = -1$$

Property 4 ($K_1 = 0$ iff independence): Under statistical independence, $n_{ij} = n_{i\cdot} \cdot n_{\cdot j}/N$ for all i, j. Then:

$$n_{11}n_{00} - n_{10}n_{01} = \frac{n_{1\cdot}n_{\cdot 1}n_{0\cdot}n_{\cdot 0}}{N^2} - \frac{n_{1\cdot}n_{\cdot 0}n_{0\cdot}n_{\cdot 1}}{N^2} = 0$$

Hence $\phi = 0$. ■

> **See Remark 1** (Appendix: Supplementary Propositions) for the relationship between K_1 and meta-d'. While $K_1 \approx \tanh(\text{meta-d}'/2)$ is used as a conceptual heuristic in the Executive Summary, the two measures are **not mathematically equivalent**—K_1 is model-free and bounded, while meta-d' depends on SDT assumptions and is unbounded. When SDT assumptions are plausible, report both metrics; when questionable, prefer K_1.

Extension to Ternary State$_0$: When State$_0 \in \{\text{correct}, \text{incorrect}, \text{absent}\}$, binarization is required for Phi. The recommended strategy: - Positive: correct only - Negative: incorrect + absent

This preserves the interpretation that "I know" should predict correctness, not merely absence of misconception.

Identifiability Arguments (Informal)

> The following propositions outline *informal arguments* for why the K framework components should be identifiable under suitable conditions. These are **not rigorous proofs**; formal identifiability analysis connecting to general latent variable theory (e.g., Allman et al., 2009; Kruskal, 1977) is deferred to future technical work.

Proposition 3: K_0 Identifiability (Informal) **Proposition 3** (K_0 Identifiability): K_0 is identifiable from response data iff item difficulties have positive variance.

Formally: Given N items with difficulties $\{b_i\}_{i=1}^N$, K_0 (equivalently, θ) is identifiable iff $\text{Var}(\{b_i\}) > 0$.

Informal Argument:

Consider the 2PL model:

$$P(X_i = 1|\theta) = \frac{1}{1 + e^{-a_i(\theta - b_i)}}$$

Non-identifiability when Var$(b_i) = 0$: If all $b_i = b$ (constant), define $\theta' = \theta - c$ and $b' = b - c$ for any constant c. Then:

$$a_i(\theta' - b') = a_i((\theta - c) - (b - c)) = a_i(\theta - b)$$

Thus (θ, b) and (θ', b') yield identical response probabilities. The parameters are confounded.

Identifiability when Var$(b_i) > 0$: With varying b_i, the likelihood function has a unique maximum. Intuitively, items with different difficulties "anchor" the scale: an easy item (b_i low) and a hard item (b_i high) together locate θ uniquely.

Formally, the Fisher information matrix is non-singular when item difficulties vary (see Lord & Novick, 1968, Theorem 17.3.1). ■

Reference: Lord & Novick (1968), *Statistical Theories of Mental Test Scores*, Chapter 17.

Status: Informal argument; formal proof requires connection to general identifiability theory.

Lemma 3: $\hat{K}$ Sufficiency and Non-Uniqueness

> **Motivation**: A reviewer concern is whether the specific form of $\hat{K}$ is arbitrary. This lemma clarifies that $\hat{K}$ is a *sufficient* but not *unique* choice—any monotone, anchor-preserving function yields equivalent ordinal results.

Lemma 3 ($\hat{K}$ Sufficiency): Let $\mathcal{K}$ be the class of functions $k : [0, 1] \to [-1, 1]$ satisfying: 1. **Monotonicity**: $c_1 > c_2 \Rightarrow k(c_1) \geq k(c_2)$ 2. **Anchor Preservation**: $k(0) = -1$, $k(0.5) = 0$, $k(1) = +1$ 3. **Antisymmetry**: $k(1 - c) = -k(c)$ for all $c \in [0, 1]$

Statement: All $k \in \mathcal{K}$ yield identical ordinal ranking of subjects and identical sign patterns for K_1.

Proof: *Step 1*: Any $k \in \mathcal{K}$ can be written as $k(c) = \phi(2c - 1)$ where $\phi : [-1, 1] \to [-1, 1]$ is strictly increasing with $\phi(0) = 0$.

Step 2: For any two functions $k_a, k_b \in \mathcal{K}$:

$$\text{sign}(k_a(c)) = \text{sign}(k_b(c)) = \text{sign}(2c - 1)$$

Step 3: The ordinal structure—which subjects have higher K_1 than others—is preserved under any monotone transformation. ■

Corollary: The specific functional form $\hat{K}(c) = 2c - 1$ is the *simplest* member of $\mathcal{K}$ but any member produces equivalent results for: - Type A / Type B classification - Subject ranking - Falsifiability tests - Practical interventions

The choice of the linear form is motivated by parsimony and interpretability, not theoretical necessity.

Proposition 4: K_1 Identifiability (Informal) **Proposition 4** (K_1 Identifiability): Given K_0 estimates and Claim$_1$ observations, K_1 is identifiable iff: 1. $P(K_0 = +1) > 0$ and $P(K_0 \leq 0) > 0$ (variability in State$_0$) 2. Claim$_1$ is non-degenerate: $P(\text{Claim}_1 = \text{"I know"}) \in (0, 1)$

Informal Argument: The Phi coefficient ϕ is undefined when any marginal frequency is zero (denominator becomes zero). Condition 1 ensures both rows of the 2×2 table are populated. Condition 2 ensures both columns are populated. With all cells potentially non-zero, ϕ is a well-defined estimator of alignment. ■

Status: Informal argument; formal proof deferred.

Proposition 5: K_2 Identifiability (Informal) **Proposition 5** (K_2 Identifiability): Let $K_2 := \text{ICC}(K_1^{(t_1)}, K_1^{(t_2)})$ be the test-retest intraclass correlation of K_1.

K_2 is identifiable iff: 1. $\text{Var}(K_1) > 0$ (cross-subject variability) 2. Measurement occasions t_1, t_2 are sufficiently separated (recommended: 2-4 weeks)

Informal Argument: The ICC is undefined when within-subject or between-subject variance is zero. Condition 1 ensures between-subject variance. The temporal separation in Condition 2 ensures that repeated measurements are not autocorrelated beyond true stability. ■

Status: Informal argument; formal proof deferred.

Proposition 6: Recursive Pipeline Identifiability (Informal)

> **Note**: This proposition addresses the potential circularity concern in recursive metacognition measurement—specifically, whether measuring K_n might contaminate the K_{n+1} measurement, creating an identification problem.

Proposition 6 (Pipeline Identifiability): Under the following assumptions, the parameters (K_0, K_1, K_2) are jointly identifiable without circularity:

Assumptions: - **(A1) Reference Anchoring**: The object-level ground truth θ_0 (item correctness) is fixed externally before any elicitation. - **(A2) Temporal Precedence**: Claim_n is elicited BEFORE any feedback about Claim_{n-1} accuracy. - **(A3) Item Count Sufficiency**: $N \geq 10$ items per measurement occasion (for stable estimation). - **(A4) Variance Conditions**: Non-zero between-subject and within-subject variance in responses at each level.

Statement: Under (A1)-(A4), the mapping $(\text{Response}_0, \text{Claim}_1, \text{Claim}_2) \mapsto (\hat{K}_0, \hat{K}_1, \hat{K}_2)$ is injective for almost all parameter configurations, ensuring joint identifiability.

Informal Argument (Four Steps):

Step 1: K_0 Identification. $K_0 = \mathbb{1}[\text{Response}_0 = \theta_0]$ is determined solely by external ground truth comparison. By (A1), this requires no internal reference and thus introduces no circularity.

Step 2: K_1 Identification. By (A2), Claim_1 (the subject's confidence in their response) is elicited before revealing whether Response_0 was correct. Thus:

$$K_1 = \text{sign}\left(\text{Claim}_1 - \frac{1}{2}\right) \cdot (2K_0 - 1)$$

The mapping from $(\text{Response}_0, \text{Claim}_1)$ to K_1 uses only (K_0, Claim_1), where K_0 is already identified in Step 1.

Step 3: K_2 Identification. Similarly, Claim_2 (certainty about the correctness of Claim_1) is elicited before feedback about Claim_1 accuracy:

$$K_2 = \text{sign}\left(\text{Claim}_2 - \frac{1}{2}\right) \cdot (2|K_1| - 1)$$

By (A3) and (A4), sufficient items provide the variance structure needed for stable estimation without contamination across levels.

Step 4: Injectivity. Each level's identification depends only on: (i) the immediately preceding level's already-identified parameter, and (ii) independently elicited claims. The directed acyclic structure $K_0 \to K_1 \to K_2$ ensures no feedback loops. By Proposition 3 (conditional independence), the estimation errors are uncorrelated across levels, establishing injectivity except on a measure-zero set of degenerate configurations. ■

Status: Informal argument; connection to general identifiability theory (Allman et al., 2009) deferred to future work.

Remark (Practical Implementation): The temporal precedence requirement (A2) is satisfied by standard experimental protocols where confidence judgments are collected before performance feedback. Computer-based testing naturally enforces this separation.

Falsifiability and Separation Axioms

Axiom F: Falsifiability **Axiom F** (Falsifiability): The K framework is empirically falsifiable. We state five predictions whose violation would refute or severely constrain the framework:

Falsifiable Prediction 1 (K-C Dissociation):

$$\exists \text{ subjects with } (K_1 = -1, C = \text{high}) \quad \text{[Dunning-Kruger pattern]}$$

$$\exists \text{ subjects with } (K_1 = -1, C = \text{low}) \quad \text{[Imposter pattern]}$$

If ALL subjects show $K_1 = \text{sign}(2C - 1)$, the framework adds nothing beyond confidence.

Falsifiable Prediction 2 (Layer Independence):

$$\text{Cor}(K_0, K_2 \mid K_1) \approx 0$$

If K_2 is strongly predicted by K_0 after controlling for K_1, layer separation is violated.

Falsifiable Prediction 3 (Intervention Sensitivity): Metacognitive training should increase K_1 while K_0 may remain constant.

If K_1 cannot be moved independently of K_0, the distinction is spurious.

Falsifiable Prediction 4 (Quantitative Bounds): "Metacognitive training improves K_1 by at least $\delta = 0.2$ on average, while K_0 remains within $\varepsilon = 0.1$ of pre-training level."

Formally:

$$\mathbb{E}[K_{1,\text{post}} - K_{1,\text{pre}}] \geq 0.2$$

$$|\mathbb{E}[K_{0,\text{post}} - K_{0,\text{pre}}]| \leq 0.1$$

If training affects both K_0 and K_1 equally (i.e., $\Delta K_1 \approx \Delta K_0$), then layer separation is empirically spurious for that intervention.

Falsifiable Prediction 5 (K-C Correlation Bound): "In any population, $\text{Cor}(K_1, C) < 0.85$."

If $\text{Cor}(K_1, C) > 0.85$ consistently across multiple datasets, K_1 and C are empirically redundant and the distinction adds no value.

Prediction	Falsification Condition	Consequence
Pred 1	All subjects show $K_1 = \text{sign}(2C - 1)$	K reduces to confidence
Pred 2	$\text{Cor}(K_0, K_2 \mid K_1) > 0.5$	Layer hierarchy is spurious
Pred 3	$\Delta K_1 = \Delta K_0$ under intervention	Levels not separable
Pred 4	$\Delta K_1 < 0.2$ or $\lvert\Delta K_0\rvert > 0.1$	Bound violation
Pred 5	$\text{Cor}(K_1, C) > 0.85$	$K_1 \approx C$, redundant

---

Axiom S: K-C Separation **Axiom S** (K-C Separation): K (epistemic alignment) and C (phenomenological confidence) are theoretically and operationally distinct:

Theoretical Distinction: - K: Alignment between claim and actual state (requires feedback to compute) - C: Subjective certainty (measured before feedback is revealed)

Operational Distinction: - C is measured at t_1 (during response, before feedback) - K_1 is computed at t_2 (after feedback reveals correctness)

Empirical Separability Criterion:

Correlation Range	Interpretation	Implication
$\text{Cor}(K_1, C) < 0.5$	Strongly separable	K_1 captures distinct construct
$0.5 \leq \text{Cor}(K_1, C) < 0.7$	Moderately separable	Both constructs useful
$0.7 \leq \text{Cor}(K_1, C) < 0.85$	Weakly separable	Caution: possible redundancy
$\text{Cor}(K_1, C) \geq 0.85$	Not separable	K_1 adds no value over C

Note: The 0.85 threshold corresponds to $R^2 > 0.72$, meaning more than 72% of variance is shared.

Epistemic Logic Propositions

Proposition 2: Axiom 4 Testability **Proposition 2** (Positive Introspection Testability): The S4/S5 Axiom 4 (positive introspection: $Kp \to KKp$) can be empirically tested via the K framework.

Statement: In modal epistemic logic, Axiom 4 states: If an agent knows p, then they know that they know p.

In the K framework, this corresponds to: $K_0 = +1 \Rightarrow K_1 = +1$ (idealized).

Empirical Test: Measure $P(K_1 = +1 | K_0 = +1)$. If this probability is less than 1, Axiom 4 is empirically violated.

Axiom-4-Gap (quantification of violation):

$$\text{Axiom-4-Gap} := 1 - P(K_1 = +1 | K_0 = +1)$$

A positive gap indicates systematic failure of positive introspection (e.g., imposter syndrome).

Proposition 3: Axiom 5 Testability **Proposition 3** (Negative Introspection Testability): The S5 Axiom 5 (negative introspection: $\neg Kp \to K\neg Kp$) can be empirically tested via the K framework.

Statement: In modal epistemic logic, Axiom 5 states: If an agent does not know p, then they know that they do not know p.

In the K framework, this corresponds to: $K_0 \leq 0 \Rightarrow K_1 = +1$ (idealized Socratic wisdom).

Empirical Test: Measure $P(K_1 = +1|K_0 \leq 0)$. If this probability is less than 1, Axiom 5 is empirically violated.

Dunning-Kruger Index (quantification of violation):

$$\text{DK-Index} := 1 - P(K_1 = +1|K_0 \leq 0)$$

A positive DK-Index indicates systematic overconfidence among the unknowing.

Measurement Propositions

Proposition 4: Incentive Compatibility **Proposition 4** (Incentive-Compatible Claim Elicitation): Under Brier scoring, truthful probability reporting is the optimal strategy, ensuring K_1 reflects genuine calibration.

Setup: Elicit Claim$_1$ as a probability $p \in [0, 1]$: "What is the probability that your answer is correct?"

Score using the Brier rule:

$$S(p, \text{outcome}) = 1 - (p - \mathbb{1}[\text{correct}])^2$$

Statement: The Brier score is a strictly proper scoring rule. The expected score is maximized when $p = P(\text{correct}|\text{information})$.

Implication for K_1: When subjects are paid according to Brier score, strategic responding is discouraged. The observed p reflects genuine beliefs, ensuring that:

$$K_1 = 2 \cdot P(\text{Claim}_1 \text{ matches State}_0) - 1$$

reflects true metacognitive alignment rather than gaming behavior.

Proof: Properness of the Brier score is standard (Brier, 1950; Gneiting & Raftery, 2007). ■

Measurement Theory

This section describes how the theoretical constructs ($K(x)$, $K(K(x))$) can be operationalized and measured empirically.

Why Continuous Scale?

The continuous scale $[-1, 1]$ provides several advantages over binary or categorical representations:

1. Intermediate States:

Captures partial knowledge, uncertain beliefs, and mixed states that binary representations cannot express.

- Example: $K_0 = 0.3$ represents "mostly ignorant but with some relevant information"
- Example: $K_1 = -0.5$ represents "moderate overconfidence, not extreme"

2. Change Tracking:

Enables measurement of **gradual transitions and intervention effects**.

- Example: After metacognitive training, K_1 moves from -0.8 to -0.2
  - This shows improvement within the "overconfidence" category
  - Binary classification would show no change (both are "overconfident")

3. Aggregation:

Permits meaningful averaging across items, domains, or time points.

- Example: Average K_1 across 50 items yields a stable estimate
- Example: Compare K_1 across domains (math vs. history)

4. Statistical Modeling:

Compatible with standard regression, Bayesian inference, and psychometric methods.

- Linear models: $K_1 \sim K_0 + \text{training} + \epsilon$
- Hierarchical models: Subject-level and item-level random effects

5. Geometric Extension (Future Work):

Enables connection to **information geometry** and manifold-based analysis:

- Cognitive states as points on a manifold
- Interventions as trajectories
- Distance metrics for comparing metacognitive profiles
- Curvature analysis for stability of states

Design Choice:

The trichotomy $\{-1, 0, 1\}$ represents **prototypical anchors** on the continuous scale, not the only valid values. Researchers may: - Use discrete elicitation and embed into continuous scale - Use probabilistic elicitation for direct continuous measurement - Aggregate discrete responses to obtain continuous estimates

Measurement-Theoretic Interpretation

Mathematically, all epistemic states live on a **single continuous scale**:

$$K_n \in [-1, 1] \quad (n = 0, 1, 2, \dots)$$

The values -1, 0, and 1 function as **prototypical anchor points** on this continuum:

Value	Prototype	Interpretation
1	Full correct knowledge	Subject's state is maximally aligned with the chosen reference
0	Pure ignorance	Subject has no determinate stance regarding the object
-1	Full misconception	Subject's state is maximally opposed to the reference

All intermediate values in $(-1, 0)$ and $(0, 1)$ represent **graded mixtures** of these prototypes (partial knowledge, partial misconception, uncertainty, mixtures across items, etc.).

Operationalization Options (always mapping back to $[-1, 1]$):

1. **Discrete elicitation -> Discrete embedding**: Use trichotomous responses (True/False/I don't know), then embed into $[-1, 1]$ via $K(x) \in \{-1, 0, 1\}$ as prototype points.
2. **Probabilistic elicitation -> Continuous embedding**: Elicit a subjective probability $p(x)$ and map it into $[-1, 1]$ using a proper scoring rule or a simple linear transform (e.g., centered Brier-type scores).
3. **Aggregation -> Continuous embedding**: Average prototype-valued $K(x_i) \in \{-1, 0, 1\}$ across multiple items or contexts to obtain a continuous summary in $[-1, 1]$.

Conceptually, the **continuum** $[-1, 1]$ **is primary**; the trichotomy $\{-1, 0, 1\}$ is a convenient way to name salient regions on this line, not a separate codomain. Experimental designs may choose discrete or continuous elicitation, but in all cases the resulting data are interpreted as points (or distributions) on the same underlying scale $[-1, 1]$.

Continuous K_n Generation Model

This section provides a complete specification of how continuous K_n values are generated from latent variables, and how they relate to discrete anchor categories.

Latent Variable Model **Definition** (Latent Alignment Strength): For each layer n, we posit a latent variable K_n^* representing the underlying alignment strength:

$$K_n^* \sim \mathcal{N}(\mu_n, \sigma_n^2)$$

where: - μ_n: population mean alignment at layer n - σ_n^2: between-subject variance in alignment

Link Functions A **link function** $h : \mathbb{R} \to [-1, 1]$ maps the unbounded latent variable to the bounded K_n scale.

Definition (Link Function):

$$K_n = h(K_n^*)$$

where h satisfies: - $h(+\infty) = +1$ (anchor preservation at positive extreme) - $h(0) = 0$ (anchor preservation at zero) - $h(-\infty) = -1$ (anchor preservation at negative extreme) - h is strictly monotonically increasing

Available Link Functions:

Link	Formula	Properties	Recommended Use
tanh (default)	$h(x) = \tanh(\beta x)$	Smooth, symmetric, S-shaped	General purpose; β controls sensitivity
Scaled probit	$h(x) = 2\Phi(x) - 1$	Probabilistic interpretation	When latent variable represents log-odds
Clipped linear	$h(x) = \max(-1, \min(1, x))$	Simple, piecewise linear	Quick approximation; not differentiable at boundaries

Default Parameterization:

$$K_n = \tanh(\beta \cdot K_n^*), \quad \beta = 1$$

The sensitivity parameter $\beta > 0$ controls how quickly K_n approaches the boundaries: - $\beta < 1$: Compressed scale; more values near 0 - $\beta = 1$: Standard sensitivity - $\beta > 1$: Expanded scale; more values near ±1

> **Rationale for tanh over logit**:
>
> While the logistic function (logit link) is ubiquitous in IRT and binary classification, we recommend **tanh as the default link** for the following reasons:
>
> 1. **Native range**: $\tanh : \mathbb{R} \to (-1, 1)$ maps directly to the K scale without rescaling. The logistic $\sigma : \mathbb{R} \to (0, 1)$ requires an affine transformation $h(x) = 2\sigma(x) - 1$ to achieve the same range, introducing an extra step.
>
> 2. **Zero-centering**: $\tanh(0) = 0$ naturally, placing the ignorance anchor at the origin. In contrast, $\sigma(0) = 0.5$, requiring explicit centering.

3. **Symmetry**: $\tanh(-x) = -\tanh(x)$, which mirrors the framework's symmetry between knowledge and misconception. The logistic function is not antisymmetric about zero without transformation.

4. **Gradient behavior**: The tanh gradient $\operatorname{sech}^2(x)$ peaks at $x = 0$ (where discrimination matters most), whereas the logistic gradient peaks at $\sigma = 0.5$ (mapped value 0). For latent variable estimation, this distinction is minor, but tanh's direct form simplifies the algebra.

Empirical equivalence: In practice, $\tanh(\beta x) = 2\sigma(2\beta x) - 1$, so results are mathematically equivalent under appropriate parameter mapping. The choice of tanh is thus a matter of *notational convenience* and *conceptual clarity*, not theoretical necessity. Model fit comparisons (AIC/BIC) should be identical up to reparameterization.

Discretization: From Continuous to Categorical **Definition** (Threshold-Based Discretization): Given thresholds τ^- and τ^+ with $-1 < \tau^- < \tau^+ < 1$:

$$\text{Discrete}(K_n) = \begin{cases} +1 & \text{if } K_n > \tau^+ \\ 0 & \text{if } \tau^- \leq K_n \leq \tau^+ \\ -1 & \text{if } K_n < \tau^- \end{cases}$$

Two Threshold Options:

Option	Definition	Rationale	Recommended For
A: Equal-interval (fixed)	$\tau^+ = +\frac{1}{3}$, $\tau^- = -\frac{1}{3}$	Divides $[-1, 1]$ into three equal-length intervals	Cross-study comparability; theoretical analysis
B: Equal-frequency (data-adaptive)	$\tau^+ = Q_{67}(K_n)$, $\tau^- = Q_{33}(K_n)$	Tercile-based; each category contains ~33% of observations	Within-study analysis; population-specific interpretation

Terminology Note: "Tercile" refers to percentile-based cutoffs (Option B), not fixed values. Option A should be called "equal-interval" discretization.

Selection Guideline:

Criterion	Option A (Equal-interval)	Option B (Equal-frequency)
Cross-study comparability	Yes	No
Adaptation to population	No	Yes

Criterion	Option A (Equal-interval)	Option B (Equal-frequency)
Fixed interpretation	Yes	No
Sensitivity to distribution	Low	High

Decision-Theoretic Justification for Option A ($\tau = \pm 1/3$): Under symmetric 0-1-2 loss (equal cost for adjacent vs. non-adjacent misclassification) and uniform prior on $[-1, 1]$, the optimal decision boundaries are $\pm 1/3$.

Point Estimation and Uncertainty **Point Estimate**: For a subject with N items, the point estimate of K_n is:

$$\hat{K}_n = \bar{K}_n = \frac{1}{N}\sum_{i=1}^{N} K_n^{(i)}$$

Confidence Interval (Bootstrap): 1. Resample N items with replacement, $B = 1000$ times 2. Compute $\hat{K}_n^{(b)}$ for each bootstrap sample b 3. Report 95% CI: $[\hat{K}_n^{(2.5\%)}, \hat{K}_n^{(97.5\%)}]$

Bayesian Posterior (alternative): With prior $K_n^* \sim \mathcal{N}(0, \sigma_0^2)$ and likelihood based on observed responses:

$$p(K_n^* | \text{data}) \propto p(\text{data} | K_n^*) \cdot p(K_n^*)$$

Report posterior mean and 95% credible interval.

Reliability Check: If CI width > 0.4, interpret the point estimate with caution; consider increasing sample size or reducing measurement noise.

Aggregation Rules Across Items

For a subject responding to N items, we obtain item-level scores $(K_0^{(i)}, K_1^{(i)}, K_2^{(i)})$ for $i = 1, \ldots, N$.

Point Estimates

Aggregate	Formula	Interpretation
Mean K_n	$\bar{K}_n = \frac{1}{N}\sum_{i=1}^{N} K_n^{(i)}$	Overall epistemic/metacognitive level
Weighted Mean	$\bar{K}_n^w = \frac{\sum_i w_i K_n^{(i)}}{\sum_i w_i}$	Item-difficulty adjusted
Distribution	$P(K_n = k)$ for $k \in \{-1, 0, +1\}$	Pattern frequencies

Uncertainty Quantification **Bootstrap Confidence Intervals:** 1. Resample N items with replacement, $B = 1000$ times 2. Compute $\bar{K}_n^{(b)}$ for each bootstrap sample 3. Report 95% CI as $[\bar{K}_n^{(0.025)}, \bar{K}_n^{(0.975)}]$

Reliability Threshold: If CI width > 0.3, interpret aggregate K_n with caution.

Statistical Properties

Property	Condition	Guarantee
Consistency	$N \to \infty$	$\bar{K}_n \to \mathbb{E}[K_n]$
Anchor Preservation	All $K_n^{(i)} = +1$	$\bar{K}_n = +1$
Boundedness	Always	$\bar{K}_n \in [-1, +1]$

Cross-Item Coherence Check To verify that aggregation is meaningful:

1. **Within-Subject Variance**: $\mathrm{Var}(K_n^{(i)})$ across items should be interpretable
   - High variance: Domain-specific metacognition
   - Low variance: Trait-like metacognitive style
2. **Correlation Structure**: $\mathrm{Cor}(K_0^{(i)}, K_1^{(i)})$ indicates coupling between knowledge and metacognition
   - Strong positive: Calibrated subject
   - Near zero: Decoupled states (possible Dunning-Kruger)

Higher Layers (n > 2): Practical Considerations

Diminishing Returns Hypothesis As n increases, the marginal information provided by K_n decreases:

$$\mathrm{Var}(K_n | K_0, K_1, \ldots, K_{n-1}) \to 0 \text{ as } n \to \infty$$

Rationale: Higher-order metacognition becomes increasingly abstract and harder to distinguish from lower layers.

Note: This is presented as a **hypothesis** based on theoretical considerations. Empirical validation is deferred to future simulation studies. No direct verification data currently exists for this claim.

Elicitation Challenges for Claim$_n$ (n > 2)

Layer	Claim	Elicitation Difficulty
$n = 1$	"Do I know?"	Low (familiar question)

Layer	Claim	Elicitation Difficulty
$n = 2$	"Is my self-assessment accurate?"	Medium (requires reflection)
$n = 3$	"Is my assessment of my self-assessment accurate?"	High (conceptually recursive)
$n > 3$	...	Very high (risk of tautological responses)

Demand Characteristics Higher-order claims risk: 1. **Tautological responses**: "If I thought my self-assessment was wrong, I would have changed it" 2. **Ceiling effects**: Most subjects claim their assessments are accurate 3. **Cognitive overload**: Difficulty distinguishing layers

Recommended Scope

Application	Recommended Max Layer
Standard assessment	K_0, K_1
Metacognitive research	K_0, K_1, K_2
Specialized studies	Up to K_3 with careful protocol design

Practical Guidance: For most applications, K_0 and K_1 provide sufficient diagnostic information. K_2 adds value for distinguishing "teachable" from "resistant" misconceptions. Beyond K_2, empirical justification should precede deployment.

Person-Level Aggregation of K_n

Item-Level to Person-Level Given m items with scores $K_n^{(1)}, K_n^{(2)}, \ldots, K_n^{(m)}$, define person-level index:

$$\bar{K}_n = \frac{1}{m} \sum_{i=1}^{m} K_n^{(i)}$$

Psychometric Properties **Reliability**:

Using Cronbach's alpha analog:

$$\alpha_{K_n} = \frac{m}{m-1} \left(1 - \frac{\sum_i \text{Var}(K_n^{(i)})}{\text{Var}(\sum_i K_n^{(i)})} \right)$$

Target: $\alpha_{K_n} > 0.7$ for adequate reliability.

Alternative: Split-Half Reliability:

$$r_{K_n} = \mathrm{Cor}(\bar{K}_n^{\mathrm{odd}}, \bar{K}_n^{\mathrm{even}})$$

With Spearman-Brown correction for full-test reliability.

Measurement Invariance For cross-group comparisons (e.g., experts vs novices), test:

1. **Configural invariance**: Same factor structure across groups
2. **Metric invariance**: Same f_n loadings across groups
3. **Scalar invariance**: Same intercepts (anchor alignment) across groups

Implementation: Use multi-group confirmatory factor analysis (CFA) with K_n as latent variable.

IRT-Based Aggregation For more sophisticated aggregation:

$$\bar{K}_n = \tanh(\hat{\theta}_n)$$

Where $\hat{\theta}_n$ is the latent trait estimated via IRT model on $K_n^{(i)}$ items.

Advantages: - Accounts for item difficulty/discrimination - Provides standard errors for $\bar{K}_n$ - Enables adaptive testing designs

Unified Estimation Pipeline

Overview The complete pipeline integrates categorical and continuous interpretations:

Stage	Option A (Discrete)	Option B (Continuous)
Input	(Response, Claim$_1$, Claim$_2$, $\ldots$, Reference)	Same
f_n **output**	$\{-1, 0, +1\}$	$[-1, +1]$
g_n	identity	link function (e.g., tanh)
$\hat{K}$	identity	normalizer (optional)
Output	$K_0, K_1, K_2, \ldots \in \{-1, 0, +1\}$	$K_0, K_1, K_2, \ldots \in [-1, +1]$

Option A: Discrete Pipeline **Step 1**: Compute $K_0^{\mathrm{cat}} \in \{-1, 0, +1\}$ from (Response, Reference)

Step 2: Compute $K_1^{\mathrm{cat}} \in \{-1, 0, +1\}$ from (Claim$_1$, K_0^{cat}) via f_1 table

Step 3: Compute $K_2^{\text{cat}} \in \{-1, 0, +1\}$ from (Claim$_2$, K_1^{cat}) via f_2 table

Output: Categorical pattern $(K_0^{\text{cat}}, K_1^{\text{cat}}, K_2^{\text{cat}}) \in \{-1, 0, +1\}^3$

Option B: Continuous Pipeline **Step 1**: Estimate latent θ_0 via IRT; compute $K_0 = \tanh(\theta_0)$

Step 2: Estimate meta-d' from (Response, Claim$_1$) via SDT; compute $K_1 = \tanh(\text{meta-d}'/2)$

Step 3: Estimate K_2 via test-retest or Claim$_2$ alignment

Output: Continuous scores $(K_0, K_1, K_2) \in [-1, +1]^3$

Hybrid Pipeline For practical use, combine:

1. **Categorical for pattern classification**: Which of the 27 patterns?
2. **Continuous for severity/reliability**: How far from anchors? How stable?

$$K_n^{\text{hybrid}} = (K_n^{\text{cat}}, K_n^{\text{cont}}, \text{SE}(K_n^{\text{cont}}))$$

Where SE is the standard error from the continuous estimation.

Use Case Recommendations:

Application	Recommended Pipeline
Quick screening	Option A (Discrete)
Research analysis	Option B (Continuous)
Clinical/educational	Hybrid
LLM evaluation	Option A with Continuous extension

Estimation Methods for K Values

The K framework specifies **what to measure** (epistemic state coordinates); for **how to estimate**, we adopt established psychometric and signal-detection methods as "plug-in" engines. This separation preserves our conceptual contribution while leveraging validated estimation machinery.

K_0 Estimation (First-Order Epistemic State) **Method**: Item Response Theory (2-Parameter Logistic Model)

$$P(\text{correct}|\theta_s, a_i, b_i) = \frac{1}{1 + e^{-a_i(\theta_s - b_i)}}$$

Mapping to K_0:

$$K_0 = 2 \cdot \Phi(\theta_s) - 1$$

Where Φ is the standard normal CDF, ensuring $K_0 \in [-1, 1]$.

Formal Derivation: $K_0 \approx \tanh(\theta)$ The Executive Summary states $K_0 \approx \tanh(\theta)$. We now provide the formal derivation from the 2PL IRT model.

Step 1: Convert probability to $[-1, 1]$ scale

Given the 2PL response probability:

$$P = \frac{1}{1 + e^{-a(\theta - b)}}$$

We transform to a signed scale:

$$K_0^* = 2P - 1 = \frac{2}{1 + e^{-a(\theta - b)}} - 1 = \frac{1 - e^{-a(\theta - b)}}{1 + e^{-a(\theta - b)}}$$

Step 2: Recognize hyperbolic tangent identity

The hyperbolic tangent satisfies:

$$\tanh(x) = \frac{e^{2x} - 1}{e^{2x} + 1} = \frac{1 - e^{-2x}}{1 + e^{-2x}}$$

Step 3: Match exponents

Comparing forms, we identify:

$$K_0^* = \tanh\left(\frac{a(\theta - b)}{2}\right)$$

Full Form:

$$K_0 = \tanh\left(\frac{a(\theta - b)}{2}\right)$$

Simplified Form (under standardization assumptions):

When $a = 2$ (unit discrimination) and $b = 0$ (centered difficulty):

$$K_0 \approx \tanh(\theta)$$

Dependency Note: This mapping is item-parameter dependent: - High a (discriminating items) → sharper transition near $\theta = b$ - High b (difficult items) → shift toward lower K_0 for fixed θ

For aggregate K_0 across items with varying (a_i, b_i):

$$\bar{K_0} = \frac{1}{N}\sum_{i=1}^{N} \tanh\left(\frac{a_i(\theta - b_i)}{2}\right)$$

Or estimate θ via standard IRT procedures and apply the mapping post-hoc.

Misconception Detection: - High confidence + incorrect -> $K_0 = -1$ - Operationalized via Confidence-Accuracy calibration error

K_1 **Estimation (Metacognitive Alignment)** **Method A**: Phi Coefficient

$$\phi = \frac{n_{11}n_{00} - n_{10}n_{01}}{\sqrt{(n_{11}+n_{10})(n_{01}+n_{00})(n_{11}+n_{01})(n_{10}+n_{00})}}$$

Handling 3-Value State$_0$:

State$_0$ has three outcomes {correct, incorrect, absent}. For Phi (2×2), we binarize:

Binarization Strategy	Positive ($K_0 > 0$)	Negative ($K_0 \leq 0$)
Strategy A (Strict)	correct only	incorrect + absent
Strategy B (Lenient)	correct + absent	incorrect only
Strategy C (Exclude)	correct	incorrect (exclude absent)

Recommended: Strategy A (Strict) — aligns with the interpretation that "I know" should predict correctness, not just absence of misconception.

Cell Definitions (Strategy A): - n_{11}: correct + "I know" - n_{00}: (incorrect OR absent) + "I don't know" - n_{10}: correct + "I don't know" - n_{01}: (incorrect OR absent) + "I know"

Mapping: $K_1 = \phi$ (already in $[-1, 1]$)

Interpretation: - $\phi = 1$: Perfect metacognitive alignment - $\phi = 0$: No relationship (random) - $\phi = -1$: Perfect anti-alignment (systematic miscalibration)

Method B: meta-d' Ratio (Signal Detection Theory)

$$K_1 = f\left(\frac{\text{meta-d}'}{d'}\right)$$

Where: - meta-d' = metacognitive sensitivity (Maniscalco & Lau, 2012) - d' = first-order sensitivity - f: bounding function to ensure output in $[-1, 1]$

Bounding Function Options:

Function	Formula	Properties
tanh (default)	$\tanh(r)$	Smooth, symmetric, saturates at ±1
Scaled CDF	$2\Phi(r) - 1$	Probabilistic interpretation
Clipped linear	$\max(-1, \min(1, r))$	Simple, preserves scale near 0

Rationale for tanh (default): - Smooth monotonic transformation - Natural saturation at extreme values - Widely used in neural network literature - **Alternative-agnostic**: Results qualitatively similar across choices

Sensitivity Analysis Recommendation: Report results with at least two bounding functions to confirm robustness.

Choice Guidance: - Use Phi for simplicity and interpretability (no bounding needed) - Use meta-d'/d' for SDT-compatible analyses (with explicit bounding function)

K_2 **Estimation (Higher-Order Alignment)** K_2 measures meta-metacognitive alignment: does the subject accurately assess their own metacognitive accuracy (K_1)? We present three candidate methods with a recommended default.

Method A: Test-Retest Stability

$$K_2^{(\text{stability})} = \text{Cor}(K_1^{(t_1)}, K_1^{(t_2)})$$

Where K_1 is measured at two time points. High stability ($K_2 \to +1$) indicates consistent metacognitive self-assessment.

Interpretation: This operationalizes K_2 as **reliability of** K_1 rather than accuracy.

Requirements: Repeated measurement at t_1, t_2 (recommended: 2-4 weeks apart)

Method B: Higher-Order Claim Alignment (PRIMARY)

$$K_2^{(\text{claim})} = \nVdash[\text{Claim}_2 \text{ matches actual } K_1] \cdot 2 - 1$$

Where Claim_2 is the subject's belief about their own metacognitive accuracy.

Implementation: 1. Compute K_1 from (Response, Claim_1, Reference) 2. Elicit Claim_2: "Is your self-assessment accurate?" 3. Compare Claim_2 to actual K_1

Aggregate Form (across items):

$$K_2 = 2 \cdot P(\text{Claim}_2 \text{ matches State}_1) - 1$$

Method C: Hierarchical Bayesian Reliability Model $K_1^{(i)}$ as noisy observations of a latent true K_1^*:

$$K_1^{(i)} | K_1^* \sim \mathcal{N}(K_1^*, \sigma^2)$$

Then:

$$K_2 = 1 - \frac{\text{Var}(K_1^{(i)} | K_1^*)}{\text{Var}(K_1^{(i)})} = \frac{\text{Var}(K_1^*)}{\text{Var}(K_1^{(i)})}$$

This is the **reliability coefficient** (analogous to Cronbach's α).

Implementation: Hierarchical Bayesian GLM (cf. HiBayES, Fleming & Daw, 2017)

$$\text{Claim}_2 | \text{State}_1 \sim \text{Bernoulli}(\sigma(\alpha_s + \beta_i + \gamma_{s,i}))$$

Recommended Default and Validation Strategy

Method	Measures	Requires	Role
B (Claim)	Accuracy of meta-metacognition	Explicit Claim_2	**Primary**
A (Stability)	Reliability of K_1	Repeated measurement	**Validation**
C (Bayesian)	Signal-to-noise ratio	Multiple items ($N \geq 20$)	**Robustness check**

Rationale for Method B as Primary: - Direct operationalization of the theoretical construct (meta-metacognitive accuracy) - Single-session administration (no need for retest) - Consistent with the MAT protocol structure (claim-based measurement)

Validation Protocol: 1. Compute K_2 via Method B (primary estimate) 2. If retest data available, validate with Method A (expected correlation $r > 0.5$) 3. For large N, verify with Method C (expected convergence)

Summary Table: Estimation Methods

Layer	Observable	Method	Output
K_0	Response vs Reference	IRT (2PL)	$[-1, 1]$
K_1	Claim_1 vs State_0	Phi / meta-d'/d'	$[-1, 1]$
K_2	Claim_2 vs State_1	Hierarchical Bayes	$[-1, 1]$
C	Self-reported confidence	Direct elicitation	$[0, 1]$

Identifiability Analysis **Definition**: A parameter θ is **identifiable** if there exists a measurable function $\hat{\theta}$ such that, as $N \to \infty$, $\hat{\theta} \xrightarrow{p} \theta$ (i.e., the estimator converges in probability to the true value).

Identifiability Conditions by Layer:

K_0 **Identifiability**: - **Requires**: Multiple items per subject to separate subject ability from item difficulty - **Condition**: Variance in item difficulties $\text{Var}(b_i) > 0$ - **Minimum**: $N \geq 10$ items for stable 2PL estimation

K_1 **Identifiability**: - **Requires**: Multiple trials per subject with varying K_0 outcomes - **Condition**: Both correct and incorrect responses must occur - **Minimum**: $N \geq 15$ trials with both positive and negative K_0 outcomes

K_2 **Identifiability**: - **Requires**: Observed K_1 variation across trials - **Condition**: Non-degenerate K_1 distribution (not all perfect or all random) - **Minimum**: $N \geq 20$ trials for hierarchical Bayesian estimation

Non-Identifiability Cases:

Scenario	Observable Pattern	Diagnosis
Ceiling K_0	All items correct	Cannot estimate K_1 (no error signal)
Floor K_0	All items incorrect	Cannot distinguish misconception from guessing
Perfect K_1	All metacognitive claims correct	K_2 undefined (no calibration error)
Random K_1	$\phi = 0$	Insufficient metacognitive signal for K_2

Practical Guideline:

Layer	Minimum N	Recommended N	Estimation Method	Identifiability Check
K_0	10	30+	IRT 2PL	$\text{SE}(\theta_s) < 0.5$
K_1	15	25+	Phi / meta-d'	$n_{ij} \geq 5$ per cell
K_2	20	40+	Hierarchical Bayes	Rhat < 1.1, ESS > 100

Note: These are minimum requirements. For individual-level claims about specific subjects, larger sample sizes or domain-specific validations are recommended.

Identifiability under Latent Variable Model Given the latent variable formulation $K_n = \tanh(\beta_n \cdot \theta_n)$, we address identifiability:

Location-Scale Indeterminacy:

θ_n is identified only up to a scale factor (since $\tanh(\beta\theta) = \tanh((\beta c)(\theta/c))$ for any $c > 0$).

Resolution: Fix $\beta = 1$ (standard parameterization) or anchor to a reference population.

Finite-Sample Identifiability:

Layer	Data Requirements	Identifiability Condition
K_0	≥ 10 items with known ground truth	Variance in ground truth states
K_1	K_0 estimates + self-assessments	Variance in (K_0, Claim_1) pairs
K_2	K_1 estimates + meta-self-assessments	Non-degenerate (K_1, Claim_2)

Practical Guideline: For reliable estimation, collect data across the full range of K_n values. Pure cases (all $+1$ or all -1) provide no information about the link function shape.

Reliability and Validity Guidelines

Test-Retest Reliability **Concern:** Are K_n scores stable over time (assuming no true change)?

Protocol: 1. Administer MAT at time t_1 2. Re-administer at t_2 (recommended: 2-4 weeks) 3. Compute intraclass correlation (ICC) for K_0, K_1, K_2

Expected Results:

Layer	Expected ICC	Rationale
K_0	0.7-0.9	Knowledge is relatively stable
K_1	0.5-0.8	Metacognition may fluctuate with context
K_2	0.4-0.7	Meta-metacognition is more variable

Interpretation: ICC > 0.7 indicates acceptable stability; lower values suggest either measurement noise or genuine state instability.

Inter-Rater Reliability **Concern:** Do different observers compute the same K_n from identical data?

Protocol: 1. Two+ independent raters apply the f_n, g_n mappings 2. Compute Cohen's κ for discretized K_n 3. Compute ICC for continuous K_n estimates

Expected Results:

Given the deterministic nature of f_n and g_n, inter-rater agreement should be **near-perfect** ($\kappa > 0.9$) for unambiguous responses. Disagreements indicate: - Ambiguous claim interpretation (refine claim vocabulary) - Reference disagreement (clarify ground truth designation)

Split-Half Reliability **Concern:** Is K_n estimation internally consistent across item subsets?

Protocol: 1. Randomly split items into two halves (A and B) 2. Compute $K_n^{(A)}$ and $K_n^{(B)}$ separately 3. Correlate the two estimates; apply Spearman-Brown correction

Expected Results:

Layer	Expected Split-Half r	Interpretation
K_0	0.7-0.9	High internal consistency
K_1	0.5-0.8	Moderate; depends on item heterogeneity
K_2	0.4-0.7	Lower; meta-meta states are more variable

Guideline: Spearman-Brown corrected $r > 0.7$ indicates acceptable internal consistency.

Measurement Invariance **Concern:** Do K_n scores have the same meaning across different populations or item sets?

Protocol: 1. Administer MAT to multiple groups (e.g., experts vs novices, domains A vs B) 2. Fit latent variable model $K_n = h(\theta_n)$ separately per group 3. Test whether link function parameters (β) are equivalent

Interpretation: - Equivalent β across groups -> Scores are comparable - Different β -> Group-specific calibration needed; interpret within-group only

Cross-Study Comparability **Concern:** Can K_n scores from different studies be meaningfully compared?

Requirements for Comparability:

Requirement	Description	Verification
Same f_n specification	Identical claim vocabulary and alignment rules	Document and share protocol
Comparable reference standards	Similar ground-truth designation criteria	Report reference source
Equivalent $\hat{K}$ parameterization	Same link function and β	Fix $\beta = 1$ or anchor to common scale

Recommended Practice:

1. **Protocol registration**: Pre-specify f_n, g_n, and $\hat{K}$ before data collection
2. **Anchor items**: Include common items across studies for calibration
3. **Report uncertainty**: Provide confidence intervals for K_n estimates

When Comparability Fails:

If studies use different references or f_n specifications, direct K_n comparison is invalid. Instead: - Report within-study patterns (e.g., proportion of Dunning-Kruger patterns) - Compare relative rankings, not absolute K_n values

Continuous Estimation of $K(K(x))$

The categorical inference of $K(K(x))$ from a single "Do you know?" claim is a **simplified operationalization**. For more robust measurement, we propose:

Option 1: Aggregation Across Items

For a subject responding to multiple items within a domain:

$$K(K)_{aggregate} = 2 \cdot P(\text{meta-claim matches actual state}) - 1$$

where P is estimated across all items. This yields a continuous value in $[-1, 1]$.

Option 2: Hierarchical Bayesian Estimation

Model $K(K(x))$ as a latent variable with: - Prior distribution over subjects - Item-level random effects - Observation model linking latent $K(K(x))$ to categorical claims

This approach accommodates noise, individual differences, and item difficulty.

Option 3: Probabilistic Elicitation

Instead of categorical "Yes/No/Unsure", elicit: - "How confident are you that your previous answer was correct?" (0-100%)

Map this to $K(K(x))$ via a proper scoring rule or calibration analysis.

Operational Semantics for Intermediate Values

The Question: What does $K_n = 0.6$ mean? The anchors (-1, 0, $+1$) have clear semantics, but intermediate values require interpretation.

Latent Variable Model:

We interpret K_n as the observable output of a latent alignment variable $\theta_n \in \mathbb{R}$, mapped to $[-1, 1]$ via a monotonic link function:

$$K_n = h(\theta_n) = \tanh(\beta \cdot \theta_n)$$

where: - θ_n: Latent alignment strength (unbounded) - $\beta > 0$: Sensitivity parameter (determines curve steepness) - h: Link function satisfying $h(0) = 0$, $\lim_{\theta \to \pm\infty} h(\theta) = \pm 1$

Interpretation of Intermediate Values:

K_n Value	Latent Interpretation	Operational Meaning
$K_n = +1$	$\theta_n \to +\infty$	Perfect alignment/knowledge
$K_n = +0.6$	$\theta_n > 0$ (moderate)	Probable alignment with uncertainty
$K_n = 0$	$\theta_n = 0$	No systematic alignment or misalignment
$K_n = -0.6$	$\theta_n < 0$ (moderate)	Probable misalignment with uncertainty
$K_n = -1$	$\theta_n \to -\infty$	Perfect misalignment/misconception

Alternative Link Functions:

Function	Formula	Properties
tanh (default)	$\tanh(\beta\theta)$	Smooth, symmetric, unbounded input
Scaled probit	$2\Phi(\theta) - 1$	Probabilistic interpretation
Clipped linear	$\max(-1, \min(1, \theta))$	Simple, piecewise linear

The choice of link function is an **empirical question** to be settled by future validation studies.

Why This Matters:

Without latent variable semantics, intermediate values risk being "nominal with three anchors." The link function approach provides: 1. **Continuous gradation**:

Values between anchors have principled meaning 2. **Estimation targets**: θ_n can be estimated via maximum likelihood 3. **Uncertainty quantification**: Standard errors on $\hat{\theta}_n$ yield confidence intervals for K_n

Continuous K_n Values via Proper Scoring Rules

For discrete elicitation mapped to continuous K_n, we define principled estimators based on strictly proper scoring rules. This ensures that intermediate values (e.g., $K_0 = 0.6$) are not arbitrary but derive from established accuracy metrics.

K_0 Continuous Estimation   Option A: Brier-Based Embedding

Given a response r and reference t, with confidence $c \in [0, 1]$:

$$K_0 = \begin{cases} 2c - 1 & \text{if } r = t \text{ (correct)} \\ 0 & \text{if } r = \text{abstain} \\ -(2c - 1) & \text{if } r \neq t \text{ (incorrect)} \end{cases}$$

This maps high-confidence correct to $K_0 \to +1$, high-confidence incorrect to $K_0 \to -1$, and low-confidence or abstention to $K_0 \to 0$.

Option B: Proper Score Centering

Using Brier score $B = (c - \nVdash[\text{correct}])^2$:

$$K_0 = 1 - 2B$$

This yields $K_0 = +1$ for perfect calibration on correct, $K_0 = -1$ for perfect miscalibration.

Derivation: The Brier score $B \in [0, 1]$ is a strictly proper scoring rule. The transformation $K_0 = 1 - 2B$ linearly rescales to $[-1, 1]$, preserving properness.

K_1 Continuous Estimation   Meta-d' Based:

$$K_1 = \tanh\left(\frac{\text{meta-d}'}{2}\right)$$

Where meta-d' is the signal-detection sensitivity for Type-2 decisions (discriminating correct from incorrect responses).

Alignment Score Based:

For item i with actual $K_0^{(i)}$ and claimed state $\tilde{K}_0^{(i)}$:

$$K_1 = \frac{1}{N}\sum_{i=1}^{N} \nVdash[K_0^{(i)} \cdot \tilde{K}_0^{(i)} > 0] \cdot 2 - 1$$

This yields $K_1 = +1$ for perfect alignment, $K_1 = -1$ for systematic misalignment.

Properties of Continuous Estimators

Property	K_0 (Brier)	K_1 (meta-d')
Anchor Preservation	$K_0 \in \{-1, 0, +1\}$ at extremes	$K_1 \in \{-1, 0, +1\}$ at extremes
Monotonicity	Increases with accuracy × confidence	Increases with metacognitive sensitivity
Properness	Derived from strictly proper Brier	meta-d' is bias-free under SDT assumptions
Continuity	Continuous in c	Continuous in meta-d'

The Challenge of Measuring Second-Order States

$K(K(x))$ is a **second-order epistemic state**: it represents the subject's recognition of their own first-order state $K(x)$. We cannot directly observe $K(K(x))$; we must infer it from observable behavior.

Operational Definition of Confidence (C)

To fully characterize the phenomenological experience of metacognition, we additionally measure **subjective confidence** C.

Definition:

Confidence C is a **phenomenological self-report** of subjective certainty, measured on a scale (e.g., 0-100% or 1-7 Likert).

$$C(x) \in [0, 1] \quad \text{(or any bounded interval)}$$

Key Distinction: K vs C

Dimension	K (Epistemic State)	C (Confidence)
What it measures	Alignment with reference	Subjective feeling
Anchor	Correct/Incorrect/Absent	Certain/Uncertain
Sign	Signed (-1 to 1)	Unsigned (0 to 1)
Basis	External validation	Internal experience

Orthogonality:

K and C are **conceptually orthogonal**:

Pattern	K_0	C	Interpretation
Confident correct	1	High	Ideal
Confident wrong	-1	High	Dangerous misconception
Unconfident correct	1	Low	Imposter-like
Unconfident wrong	-1	Low	Appropriate uncertainty

Diagnostic Role:

C helps distinguish subtypes within the same K pattern: - DK ($K_0 = 0$, $K_1 = -1$) with **high** C -> Overconfident ignorance - DK ($K_0 = 0$, $K_1 = -1$) with **low** C -> Uncertain but still wrong claim

K-C Dissociation Hypothesis:

Subjects can have: - High K_0 with low C (Imposter syndrome) - Low K_0 with high C (Dunning-Kruger effect)

This dissociation is empirically testable and clinically meaningful.

Measurement Protocol for C:

1. Elicit response -> compute K_0
2. Elicit confidence (0-100%) -> record C
3. Elicit metacognitive claim ("Do you know?") -> compute K_1
4. Analyze $K \times C$ jointly for full characterization

Claim Elicitation Protocol

This section provides a **complete specification** of the protocol for eliciting Claim_1 and Claim_2, and the decision rules f_1 and f_2 for computing State_1 and State_2.

Protocol 1: Claim_1 Elicitation (Complete Specification) **Timing**: - **After** the subject's response to item x - **Before** feedback on correctness is provided

Instruction Template (Categorical): > "How confident are you that your answer is correct? > - [ ] I'm confident I'm correct ("I know") > - [ ] I'm not sure ("Uncertain") > - [ ] I think I might be wrong ("I don't know")"

Instruction Template (Continuous): > "What is the probability that your answer is correct?" > [Slider: 0% ———————— 100%]

Mapping (Continuous → Categorical):

Reported Probability p	Claim_1
$p \geq 0.70$	"I know"
$0.30 < p < 0.70$	"Uncertain"

Reported Probability p	Claim_1
$p \leq 0.30$	“I don’t know”

Threshold Calibration Options:

Method	Procedure	Use Case
Fixed thresholds	$p = 0.70$ and $p = 0.30$	Cross-study comparability
Pilot calibration	Adjust to achieve ~equal category frequencies	Within-study optimization
Percentile-based	Top 30%, middle 40%, bottom 30%	Population-adaptive
Domain-specific	Higher thresholds for high-confidence domains	Context sensitivity

Inter-Rater Reliability Requirement: - Categorical: Cohen’s $\kappa > 0.80$ - Continuous: ICC > 0.70

Protocol 2: Claim_2 Elicitation (Complete Specification) **Timing**: - **After** all items are completed - **After** K_1 is computed (or after metacognitive feedback is provided)

Instruction Template (Aggregate): > “Thinking about your self-assessments across all questions: > How accurate do you think your self-assessments were? > - [ ] My self-assessments were accurate (”Meta-aligned”) > - [ ] I’m not sure about my self-assessments (“Meta-uncertain”) > - [ ] My self-assessments were probably inaccurate (“Meta-misaligned”)”

Alternative: Item-Level Claim_2: After revealing K_1 for each item: > “Was your confidence judgment appropriate for this item? > - [ ] Yes, appropriate > - [ ] Not sure > - [ ] No, inappropriate”

Protocol 3: f_1 Decision Rules (Complete Specification) **Input**: (K_0, Claim_1) **Output**: $\text{State}_1 \in \{\text{aligned}, \text{uncertain}, \text{misaligned}\}$

Complete Decision Table:

K_0	Claim$_1$ = "I know"	Claim$_1$ = "Uncertain"	Claim$_1$ = "I don't know"
+1 (correct)	aligned	uncertain	misaligned
0 (ignorance)	misaligned	aligned	aligned
-1 (misconception)	misaligned	uncertain	aligned*

***Edge Case Rationale**: $K_0 = -1$ and Claim$_1$ = "I don't know" → **aligned**

This coding follows the **Epistemic Improvement Criterion**: a subject who has a misconception but recognizes they "don't know" is exhibiting Socratic awareness, which is metacognitively appropriate.

Alternative Interpretation (Stricter):

K_0	Claim$_1$ = "I don't know"	Alternative Output
-1	"I don't know"	**uncertain** (not aligned)

Rationale for Alternative: - "I don't know" does not explicitly acknowledge misconception - True alignment at $K_0 = -1$ would require: "I think I'm wrong"

Framework Choice Guidelines:

Choice	Prioritizes	Recommended For
(a) Default (aligned)	Recognition of non-knowledge (Socratic wisdom)	General metacognition research
(b) Alternative (uncertain)	Recognition of specific error direction	Error diagnosis, clinical settings

Recommendation: Always report which interpretation is used. For replication, specify: "We use interpretation (a)/(b) for the $K_0 = -1$, Claim$_1$ = 'I don't know' case."

g_1 **Embedding**:

State_1	K_1
aligned	$+1$
uncertain	0
misaligned	-1

Protocol 4: f_2 **Decision Rules** **Input**: (K_1, Claim_2) **Output**: $\text{State}_2 \in \{\text{meta-aligned}, \text{meta-uncertain}, \text{meta-misaligned}\}$

The structure mirrors f_1:

K_1 (aggregate)	$\text{Claim}_2 =$ "Accurate"	$\text{Claim}_2 =$ "Uncertain"	$\text{Claim}_2 =$ "Inaccurate"
$+1$ (aligned)	meta-aligned	meta-uncertain	meta-misaligned
0 (uncertain)	meta-misaligned	meta-aligned	meta-aligned
-1 (misaligned)	meta-misaligned	meta-uncertain	meta-aligned

g_2 **Embedding**:

State_2	K_2
meta-aligned	$+1$
meta-uncertain	0
meta-misaligned	-1

Continuous Claim Variants $\textbf{Claim}_1$ **Elicitation:**

"How confident are you in your answer?" (slider: 0-100)

Threshold Mapping:

Slider Value (c)	Claim_1
$c \geq 70$	"I know"
$30 < c < 70$	"Not sure"
$c \leq 30$	"I don't know"

Threshold Calibration:

Thresholds should be validated via: 1. **Pilot calibration**: Adjust thresholds to achieve approximately equal category frequencies 2. **Cross-participant comparison**: Use percentile-based thresholds (e.g., top 30%, middle 40%, bottom 30%) 3. **Domain-specific adjustment**: Higher thresholds for domains with inflated confidence norms

Inter-Rater Reliability Requirements For categorical claims: - Expected Cohen's $\kappa > 0.8$ for Claim_1 coding - Any ambiguous responses should be coded by 2+ independent raters

For continuous claims: - Report ICC (Intraclass Correlation) for slider reliability - Expected ICC > 0.7 for adequate reliability

Protocol Selection Guidance:

Study Type	Recommended Protocol	Rationale
Large-scale survey	Categorical	Faster administration, clearer coding
Individual assessment	Continuous	Finer gradation, calibration analysis
Clinical application	Categorical + Continuous	Both for robustness

Important Distinction (Summary): - $K(x)$: Epistemic state (how accurately the subject recognizes x) - C: Phenomenological confidence (how certain the subject feels)

These are **orthogonal dimensions**. A subject can have: - $K(x) = 0$ (ignorance) with $C = 1$ (high confidence) — Dunning-Kruger - $K(x) = 1$ (knowledge) with $C = 0.5$ (moderate confidence) — Underconfidence

The C-K Joint Model: Empirical Handling of Confidence-Knowledge Interaction

The Problem:

Confidence (C) and epistemic state (K) are conceptually orthogonal, but empirically correlated. How should we handle cases like: - **Dunning-Kruger**: $K_1 = -1$ (miscalibrated) with $C =$ high - **Imposter Syndrome**: $K_1 = -1$ (miscalibrated) with $C =$ low

Proposed Joint Model:

We model the joint distribution of (K_n, C) as a bivariate structure:

$$P(K_n, C) = P(K_n) \cdot P(C \mid K_n)$$

where: - $P(K_n)$: Marginal distribution of epistemic alignment (from MAT protocol) - $P(C \mid K_n)$: Conditional distribution of confidence given alignment

Operationalization:

K_1	Expected C Pattern	Interpretation
$+1$ (calibrated)	C correlates with K_0	Confidence tracks knowledge
0 (uncertain)	C near midpoint	Appropriate uncertainty
-1 (miscalibrated)	C anti-correlates with K_0	Confidence misleads

Dunning-Kruger vs Imposter Analysis:

For subjects with $K_1 = -1$ (miscalibrated metacognition):

Subtype	K_0	C	Interpretation	Intervention
Dunning-Kruger	0 or -1	High	Overconfident ignorance	Calibration training
Imposter Syndrome	$+1$	Low	Underconfident knowledge	Confidence building

Joint Reporting:

Report (K_0, K_1, K_2, C) as a 4-tuple for complete metacognitive characterization:

K_0	K_1	K_2	C	Pattern Name
0	-1	-1	High	Deep Dunning-Kruger (overconfident)
0	-1	$+1$	Low	Aware Dunning-Kruger (self-correcting)
$+1$	-1	-1	Low	Deep Imposter (persistent self-doubt)
$+1$	-1	$+1$	Low	Aware Imposter (recognizes pattern)

Why C Cannot Replace K:

While C and K_1 are empirically correlated, they measure different things: - K_1 = **Structural alignment** (does claim match state?) - C = **Phenomenological intensity** (how certain does subject feel?)

A subject with $K_1 = +1$ (accurate self-assessment) may have $C = 0.3$ (low confidence) — they correctly identified their state but did not feel certain about it. The (K, C) joint model captures this distinction.

Measurement Protocol

Step 1: Establish Reference Context For each proposition x, establish what counts as "aligned" ($K(x) = 1$) via the experimental context (e.g., expert consensus, empirical measurement, community agreement). The proposition x itself serves as the implicit reference point.

Step 2: Measure $K(x)$ via Task Performance **Task:** Subject answers: "Is proposition x true, false, or unknown?"

Subject's Answer	Reference	Inferred $K(x)$
"True"	Aligned	1 (correct knowledge)
"False"	Aligned (proposition is false)	1 (correct knowledge)
"I don't know"	any	0 (ignorance)
"True"	Opposed	-1 (misconception)
"False"	Opposed (proposition is true)	-1 (misconception)

Step 3: Measure Confidence C_0 **Question:** "On a scale from 0 to 1, how confident are you in your answer?"

This captures the phenomenological dimension of certainty.

Step 4: Elicit Metacognitive Claim **Question:** "Do you know the answer to the previous question?"

Subject's Claim	Interpretation
"Yes, I know"	Subject claims $K(K(x)) = 1$
"No, I don't know"	Subject claims $K(K(x)) = 0$
"I'm not sure"	Subject claims $K(K(x)) \approx 0.5$

Step 5: Infer Actual $K(K(x))$ via Comparison Compare the subject's **metacognitive claim** (Step 4) to their **actual state** (Step 2):

Actual $K(x)$	Subject's Claim	Inferred $K(K(x))$	Classification
1 (knows)	"I know"	1	**Knowing Knowledge**
0 (ignorant)	"I don't know"	1	**Knowing Ignorance** (Socratic)

Actual $K(x)$	Subject's Claim	Inferred $K(K(x))$	Classification
0 (ignorant)	"I know"	-1	**Unknowing Ignorance** (Dunning-Kruger)
1 (knows)	"I don't know"	-1 or 0	**Unknowing Knowledge** (Imposter)

Key Insight: $K(K(x))$ is inferred by checking whether the subject's **metacognitive claim matches their actual state**.

Analyzing Discrepancies

Metacognitive Discrepancy The discrepancy between actual state and metacognitive claim is captured directly by $K(K(x))$: - $K(K(x)) = 1$: Accurate metacognition (claim matches reality) - $K(K(x)) = 0$: Partial metacognitive failure - $K(K(x)) = -1$: Complete metacognitive failure (claim contradicts reality)

Experimental Design: The Metacognitive Alignment Test (MAT)

To demonstrate the falsifiability and measurability of this model, we propose the **Metacognitive Alignment Test (MAT)**.

Objectives

1. Measure $K(x)$ (first-order epistemic state)
2. Measure $K(K(x))$ (second-order metacognitive state)
3. Measure confidence C (phenomenological dimension)
4. Validate the distinction between Socratic Wisdom and Dunning-Kruger effect

Protocol

Phase 1: Knowledge Assessment - Present factual questions with established reference answers (e.g., expert consensus) - Subject responds: True / False / I don't know - Calculate $K(x)$ based on alignment with reference

Phase 2: Confidence Rating - Subject rates confidence: "How confident are you?" (0-1 scale) - Record C_0

Phase 3: Metacognitive Claim - Ask: "Do you know the answer to the previous question?" - Subject responds: Yes / No / Unsure - Infer $K(K(x))$ by comparing claim to actual $K(x)$

Phase 4: Validation Tasks - Present decision-making scenarios requiring self-assessment - Measure performance on tasks like: - Deciding when to seek help - Allocating study time - Deferring to experts

Validation Hypothesis

Hypothesis: Subjects with high $K(K(x))$ (accurate metacognition) will perform better on validation tasks, **regardless of their raw $K(x)$ score**.

This would validate the model's claim that: - **Knowing Ignorance** ($K(x) = 0, K(K(x)) = 1$) is a valuable cognitive state - Metacognitive accuracy is distinct from first-order knowledge - Socratic wisdom has measurable benefits

Expected Patterns

Pattern	$K(x)$	$K(K(x))$	C	Expected Behavior
Socratic Wisdom	0	1	Low	Seeks help appropriately
Dunning-Kruger	0	-1	High	Overconfident errors
Accurate Expert	1	1	High	Confident and correct
Imposter Syndrome	1	-1 or 0	Low	Underconfident but correct

Relationship to Established Metrics

The MAT is designed to **complement, not replace**, existing metacognitive measures:

Metric	What it measures	Relationship to MAT
meta-d'	Metacognitive sensitivity (discrimination)	Can be computed from MAT data; provides aggregate validation
Brier Score	Probabilistic calibration	Applicable if confidence is elicited as probability
ECE	Expected calibration error	Measures bias in confidence-accuracy relationship
AUROC	Discrimination ability	Can be derived from confidence ratings vs. accuracy

Metric	What it measures	Relationship to MAT
IRT	Item difficulty and discrimination	Can model item-level variance in MAT responses

Recommended Analysis Pipeline:

1. Compute $K(x)$ and $K(K(x))$ per item using MAT protocol
2. Compute meta-d’ across trials as aggregate metacognitive sensitivity
3. Compute calibration metrics (Brier, ECE) from confidence ratings
4. Compare $K(K(x))$ patterns (Socratic, Dunning-Kruger, etc.) with meta-d’ to validate convergent validity
5. Use IRT to account for item-level heterogeneity

Hypothesis: High $K(K(x))$ (accurate metacognition) should correlate with high meta-d’/d’ ratio and good calibration, but $K(K(x))$ provides additional structural information (e.g., distinguishing Socratic wisdom from mere low confidence).

Related Work

Relationship to Modal Epistemic Logic

Background: Knowledge Operators in Modal Logic In formal epistemology, the knowledge operator **K** is studied within modal logic frameworks (Hintikka, 1962). Key systems include:

System	Axioms	Interpretation
K	Distribution: $\mathbf{K}(p \to q) \to (\mathbf{K}p \to \mathbf{K}q)$	Basic epistemic closure
T	Veridicality: $\mathbf{K}p \to p$	Knowledge implies truth
S4	Positive introspection: $\mathbf{K}p \to \mathbf{KK}p$	If I know, I know that I know
S5	Negative introspection: $\neg\mathbf{K}p \to \mathbf{K}\neg\mathbf{K}p$	If I don’t know, I know that I don’t know

Correspondence with K Framework Our K_n framework can be positioned relative to these axioms:

Axiom T (Veridicality):

Modal Logic	K Framework	Status
$\mathbf{K}p \to p$	$K_0 = +1 \Rightarrow$ Reference = correct	**Definitionally satisfied**

By construction, $K_0 = +1$ requires alignment with the reference, so veridicality is built into the scoring.

Axiom S4 (Positive Introspection):

Modal Logic	K Framework	Status
$\mathbf{K}p \to \mathbf{KK}p$	$K_0 = +1 \Rightarrow K_1 = +1$ (idealized)	**Empirically violated**

S4 describes an *ideal* agent. Real agents may have $K_0 = +1$ but $K_1 = -1$ (Impostor Syndrome pattern). The K framework *measures* such violations rather than assuming them away.

Axiom S5 (Negative Introspection):

Modal Logic	K Framework	Status
$\neg\mathbf{K}p \to \mathbf{K}\neg\mathbf{K}p$	$K_0 \leq 0 \Rightarrow K_1 = +1$ (idealized)	**Empirically violated**

S5 describes *ideal* self-awareness. Real agents may have $K_0 = -1$ but $K_1 = -1$ (Dunning-Kruger pattern). Again, the K framework measures rather than assumes.

Key Distinction: Normative vs Descriptive

Approach	Modal Epistemic Logic	K Framework
Stance	Normative (ideal agent)	Descriptive (empirical measurement)
Axioms	Prescribe what *should* hold	Diagnose what *does* hold
Violations	Indicate irrationality	Indicate metacognitive failure modes
Purpose	Reasoning about ideal knowledge	Measuring actual metacognition

Complementarity:

Modal logic provides the *normative benchmark* against which metacognitive accuracy can be evaluated. The K framework provides the *measurement apparatus* to assess how far real agents deviate from this benchmark.

$$\text{Metacognitive gap} = \text{Ideal (S4/S5)} - \text{Actual } (\text{K}_1)$$

Summary Table: Axiom Correspondence

Axiom	Name	S5 Statement	Our Framework Status
T	Truth	$\mathbf{K}\phi \to \phi$	Satisfied: $K_0 = +1$ implies correctness
4	Positive Introspection	$\mathbf{K}\phi \to \mathbf{KK}\phi$	**Violated**: $K_0 = +1$ does NOT imply $K_1 = +1$
5	Negative Introspection	$\neg\mathbf{K}\phi \to \mathbf{K}\neg\mathbf{K}\phi$	**Violated**: $K_0 = 0$ does NOT imply $K_1 = +1$

Key Departure:

Epistemic logics model **what agents should know** under idealized conditions. Our framework models **what agents actually exhibit** under empirical observation, including systematic failures of introspection.

Aspect	Epistemic Logic (S5)	Our Framework
Agents	Idealized, logically consistent	Empirical, cognitively fallible
Introspection	Perfect (axioms 4, 5 hold)	Imperfect (Dunning-Kruger, Imposter)
Semantics	Modal (possible worlds)	Observational (state-based measurement)
Purpose	Normative reasoning about ideal agents	Descriptive measurement of real agents
Misconception	Not modeled ($\mathbf{K}\phi$ only for true ϕ)	Explicitly modeled ($K = -1$)

Complementary Use: Epistemic logic provides normative benchmarks (what perfect metacognition would look like); our framework measures deviations from those benchmarks in real agents. The “axiom violations” we observe are precisely the phenomena of interest.

Kripke Semantics Interpretation

Background In Kripke semantics, knowledge is modeled via accessibility relations over possible worlds:

$$\mathbf{K}p \text{ holds at world } w \iff p \text{ holds at all worlds accessible from } w$$

Correspondence with K Framework **State Spaces as World Sets**:

K Framework	Kripke Semantics	Interpretation
State$_0$	Actual world w_0	Ground truth state
State$_1$	Epistemic accessibility from w_0	Worlds consistent with agent's beliefs about w_0
State$_2$	Meta-accessibility	Worlds consistent with agent's beliefs about State$_1$

K Values as Accessibility Measures:

$$K_n = \begin{cases} +1 & \text{All accessible worlds agree (certain knowledge)} \\ 0 & \text{Accessible worlds are indeterminate (ignorance)} \\ -1 & \text{All accessible worlds disagree with actual (misconception)} \end{cases}$$

Formalization:

Let R_n be the accessibility relation at layer n. Then:

$$K_n = \frac{|\{w' : wR_nw' \land \text{State}_n(w') = \text{State}_n(w_0)\}| - |\{w' : wR_nw' \land \text{State}_n(w') \neq \text{State}_n(w_0)\}|}{|\{w' : wR_nw'\}|}$$

This yields $K_n = +1$ when all accessible worlds match actuality, $K_n = -1$ when none do.

Note: This proportion-based formulation is a **proposed operationalization** extending standard Kripke semantics (which is binary: know/don't know) to a graded scale. This extension is novel to our framework and should be understood as an empirical approximation rather than a direct entailment from modal logic.

Limitations of the Correspondence

1. **Finite vs Infinite**: Kripke models allow infinite world sets; K framework assumes finite enumeration via responses/claims

2. **Quantitative vs Qualitative**: K provides graded values; standard Kripke is binary (know/don't know)
3. **Observational vs Semantic**: K is defined operationally via behavior; Kripke is defined model-theoretically

Conclusion: The K framework can be viewed as an *empirical operationalization* of Kripke-style accessibility, where accessibility is inferred from behavioral responses rather than stipulated semantically.

Relationship to Polytomous Item Response Theory

The K framework draws on and extends classical Item Response Theory (IRT). This section clarifies the relationship to polytomous IRT models—particularly the Graded Response Model (GRM; Samejima, 1969) and the Generalized Partial Credit Model (GPCM; Muraki, 1992).

Background: Polytomous IRT Models

Model	Response Structure	Key Feature	Primary Application
GRM (Graded Response)	Ordered categories (e.g., 0-4)	Cumulative probability structure	Attitude scales, rubric scoring
GPCM (Generalized Partial Credit)	Partial credit categories	Adjacent-category logits	Multi-step problem solving
RSM (Rating Scale)	Common thresholds	Constrained GPCM	Likert scales

Structural Correspondence The K framework relates to polytomous IRT as follows:

K Framework Component	Polytomous IRT Analog	Correspondence
$K_0 \in \{-1, 0, +1\}$ (discrete)	GRM categories $\{0, 1, 2\}$	Isomorphic under linear transformation
K^* (latent variable)	θ (ability parameter)	Identical interpretation as latent trait
Link function $h(K^*)$	Item characteristic curve	Both map latent to observable

K Framework Component	Polytomous IRT Analog	Correspondence
Between-layer independence	Dimensional structure	Multidimensional IRT with layer-specific traits

Formal Relationship:

The K framework can be re-expressed as a **multidimensional GRM** where each metacognitive layer defines a separate latent dimension:

$$P(K_{n,i} \geq k|\theta_n) = \frac{1}{1 + \exp(-a_{n,i}(\theta_n - b_{n,k}))}$$

where: - θ_n: Layer-specific latent trait (metacognitive ability at level n) - $a_{n,i}$: Discrimination parameter for item i at layer n - $b_{n,k}$: Threshold for category k at layer n

Key Differences Despite the structural similarity, the K framework differs from standard polytomous IRT in three respects:

1. **Semantic Anchoring**: GRM/GPCM categories are typically ordinal without intrinsic meaning. The K scale has fixed semantic anchors: -1 = misconception, 0 = ignorance, +1 = knowledge. This enables cross-domain comparison.

2. **Recursive Structure**: Standard IRT treats dimensions as parallel or hierarchical. The K framework imposes a specific recursive dependency: K_{n+1} evaluates the accuracy of K_n, creating an explicit epistemological structure absent in generic multidimensional IRT.

3. **Negative Values**: The K scale extends to $[-1, 1]$ to model misconception explicitly. Standard GRM uses $[0, M]$ where 0 is the lowest category. The $K = -1$ anchor (confident but wrong) requires this extension.

Implementation Guidance For researchers implementing the K framework within existing IRT software:

Software	Recommended Approach
mirt (R)	Use `mirt(..., itemtype = 'graded')` with rescaled responses
ltm (R)	`grm()` function with manual post-hoc transformation to $[-1, 1]$
Stan	Custom GRM with explicit prior on threshold ordering
TAM (R)	`tam.mml()` with appropriate scoring matrix

Transformation: Given GRM responses $Y \in \{0, 1, 2\}$, the K-scale transformation is $K = Y - 1$.

References

- Samejima, F. (1969). *Estimation of latent ability using a response pattern of graded scores.* Psychometrika Monograph Supplement.
- Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16*(2), 159-176.
- Embretson, S. E., & Reise, S. P. (2000). *Item Response Theory for Psychologists.* Erlbaum.

Connection to Dynamic Epistemic Logic (DEL)

Connection to Dynamic Epistemic Logic (DEL):

DEL models knowledge change through public announcements and private observations. Our framework can be seen as **static snapshots** within a DEL-like dynamics:

- K_n at time t = epistemic state after observation sequence
- Intervention = action that changes the model
- Re-observation = new K_n measurement

Full integration with DEL (multi-agent, announcement operators) is beyond current scope but represents a natural extension.

Relationship to Graded Epistemic Logics

Recent work in graded epistemic logics (e.g., S5G frameworks) models knowledge with continuous plausibility values in $[0, 1]$. Our framework differs in key respects:

Aspect	Graded Epistemic Logics	Our $K(K(x))$
Scale	$[0, 1]$ (plausibility)	$[-1, 1]$ (includes misconception)
Misconception	Typically not modeled	$K(x) = -1$
Metacognition	Introspection axioms	Explicit recursive operator
Focus	Idealized agents	Psychologically realistic failures

Formal Correspondence:

Our K can be viewed as a **graded, psychologically realistic** extension that:

1. Relaxes positive/negative introspection axioms to accommodate metacognitive failures
2. Extends the scale to include misconception (negative values)

3. Focuses on the **gap** between actual and recognized epistemic states, rather than idealized consistency

Whether K is idempotent ($K(K(x)) = K(x)$ for accurate metacognizers) or contractive (higher-order reflection converges) is an **empirical question** that our framework can accommodate but does not presuppose.

Correspondence with Established Metrics

Before detailed comparisons, we provide a summary of how the K framework relates to established metrics at each layer.

Layer-by-Layer Mapping

Layer	K Framework	Established Metric	Correspondence
K_0	Epistemic state	IRT ability θ	$K_0 \approx 2\Phi(\theta) - 1$ (scaled)
K_0	Epistemic state	Accuracy	$K_0 = +1$ iff correct
K_1	Metacognitive alignment	meta-d'	$K_1 \propto$ meta-d'/d' (normalized sensitivity)
K_1	Metacognitive alignment	AUC (Type-2)	$K_1 = 2 \cdot \text{AUC} - 1$
K_1	Metacognitive alignment	Calibration (ECE)	$K_1 \approx 1 - 2 \cdot \text{ECE}$ (inverse relationship)
K_2	Meta-metacognitive alignment	Stability of K_1 across contexts	Novel measure

Detailed Correspondences K_1 vs meta-d':

Aspect	K_1	meta-d'
Scale	$[-1, +1]$	$(-\infty, +\infty)$
Zero point	No systematic alignment	Chance-level metacognition
Aggregation	Per-item, then averaged	Aggregate over item set
Assumptions	Minimal (monotonicity)	SDT model (Gaussian distributions)

Approximate Translation:

$$K_1 \approx \tanh\left(\frac{\text{meta-}d'}{2}\right)$$

K_1 vs Calibration Error (ECE):

Aspect	K_1	ECE
Direction	+1 = perfect alignment	0 = perfect calibration
Meaning of negative	Systematic misalignment	N/A (always ≥ 0)
Confidence dimension	Implicit in State_1	Explicit (confidence bins)

Note: K_1 captures *direction* of miscalibration (over- vs under-confidence), while ECE captures *magnitude* only. They are complementary, not redundant.

Joint Reporting Standard (Proposed) For comprehensive metacognitive assessment, we recommend reporting:

Measure	Source	Information Captured
K_0	This framework	Epistemic state (knowledge/ignorance/misconception)
K_1	This framework	Alignment direction and magnitude
K_2	This framework	Meta-metacognitive calibration
meta-d'	SDT	Aggregate sensitivity under SDT assumptions
ECE	Calibration	Confidence-accuracy gap magnitude
Brier	Calibration	Combined accuracy + calibration

This joint profile provides a complete picture: K_n for item-level structure, aggregate metrics for overall performance.

Positioning Among Related Frameworks

Before comparing specific metrics, we clarify the **role of the K framework** relative to existing approaches:

Framework	Role	Relationship to K
meta-d' (Maniscalco & Lau)	Aggregate sensitivity metric	K provides per-item classification
meta-I (Dayan, 2023)	Information-theoretic sensitivity	K adds direction and recursion
HiBayES (Fleming & Daw, 2017)	Hierarchical estimation engine	K defines what to estimate
IRT (psychometrics)	Latent trait estimation	K_0 uses IRT as estimation engine
Calibration metrics (Brier, ECE)	Confidence-accuracy alignment	K adds structural classification

Key Distinction:

These frameworks address **how well** metacognition works (sensitivity, efficiency). The K framework addresses **what type** of metacognitive state exists (classification, coordinates).

Analogy: - meta-d' / meta-I = **Thermometer** (measures metacognitive temperature) - K = **Weather map** (classifies patterns, guides intervention)

Integration Potential:

The K framework serves as a **common coordinate system** that unifies: - Behavioral assessments (response-claim alignment) - Signal-detection metrics (meta-d', AUROC) - Hierarchical estimation (HiBayES) - Calibration analysis (Brier, ECE)

This is not "reinventing the wheel" but **designing the axle** that connects existing wheels.

Metacognitive Sensitivity: meta-d'

Maniscalco and Lau (2012) developed the *meta-d'* framework for measuring metacognitive sensitivity—the ability to discriminate between correct and incorrect responses via confidence ratings.

Formal Correspondence with Type-2 SDT:

In Type-2 SDT, meta-d' is the d' that would produce the observed Type-2 ROC if the observer had optimal metacognitive access to their Type-1 evidence.

Definition (meta-d'):

$$\text{meta-d}' = \Phi^{-1}(\text{HR}_2) - \Phi^{-1}(\text{FAR}_2)$$

Where: - HR_2 = Type-2 Hit Rate = $P(\text{high confidence}|\text{correct})$ - FAR_2 = Type-2 False Alarm Rate = $P(\text{high confidence}|\text{incorrect})$ - Φ^{-1} = inverse standard normal CDF

Mapping K_1 to meta-d':

$$K_1 = \tanh\left(\frac{\text{meta-d}'}{2}\right)$$

Derivation:

1. meta-d' $\in (-\infty, +\infty)$, with 0 = chance, positive = above-chance sensitivity
2. tanh maps $(-\infty, +\infty) \to (-1, +1)$ monotonically
3. The factor of 2 ensures that meta-d' = 2 (good sensitivity) maps to $K_1 \approx 0.76$

M-Ratio Alternative:

$$K_1 = \text{M-ratio} - 1 = \frac{\text{meta-d}'}{d'} - 1$$

Where M-ratio = 1 indicates ideal metacognitive efficiency (perfect metacognitive access).

Complete Correspondence Table:

SDT Quantity	K Framework	Relationship
d' (Type-1)	Related to K_0	$d' \approx 2 \cdot \text{arctanh}(K_0)$
Type-2 HR	$P(\text{"I know"} \mid K_0 = +1)$	Direct operationalization
Type-2 FAR	$P(\text{"I know"} \mid K_0 \leq 0)$	Direct operationalization
meta-d'	Related to K_1	$K_1 \approx \tanh(\text{meta-d}'/2)$
M-ratio	Metacognitive efficiency	$K_1 \approx \text{M-ratio} - 1$ (if $d' = 2$)
Type-2 AUROC	Discrimination accuracy	$K_1 \approx 2 \cdot \text{AUROC} - 1$

Important Note: These are approximate correspondences valid under specific assumptions (symmetric criteria, Gaussian noise). Exact relationships depend on model parameters and should be validated empirically.

What K Adds Beyond meta-d':

1. **Signed direction**: meta-d' is unsigned; K_1 distinguishes overconfidence (-1) from underconfidence
2. **Per-item granularity**: meta-d' is aggregate; K_1 can be computed per item
3. **Explicit ignorance**: "I don't know" is modeled as $K_0 = 0$, not low confidence

4. **Recursive hierarchy**: $K_2, K_3, \ldots$ extend beyond Type-2

Comparison:

Aspect	meta-d'	Our $K(K(x))$
Focus	Discrimination ability (sensitivity)	Structural accuracy (recognition)
Measurement	Statistical correlation across trials	Per-item metacognitive state
Granularity	Aggregate across trials	Per-item
"I don't know"	Treated as low confidence	$K(x) = 0, K(K(x)) = 1$ (Socratic wisdom)
Statistical Model	Noise-tolerant (SDT)	Deterministic match/mismatch
Theoretical Basis	Signal Detection Theory	Recursive epistemology

Key Difference: meta-d' measures whether confidence ratings **correlate** with accuracy. Our model measures whether metacognitive claims **match** actual states. Crucially, we recognize that **accurately knowing one's ignorance** ($K(x) = 0, K(K(x)) = 1$) is a **high metacognitive achievement**, not a failure.

Complementary Relationship: These approaches are **not mutually exclusive**. meta-d' provides a noise-tolerant aggregate measure of metacognitive sensitivity; our $K(K(x))$ provides per-item structural classification with explicit treatment of Socratic wisdom. An integrated approach could: - Use meta-d' for aggregate sensitivity analysis across trials - Use $K(K(x))$ for per-item classification and Socratic wisdom detection - Define a continuous version: $K(K(x)) = 2 \cdot P(\text{meta-claim matches actual state}) - 1$, estimated across trials via hierarchical Bayesian methods

Information-Theoretic Metacognition: meta-I

Dayan (2023) introduced *meta-I*, a **model-free** information-theoretic measure of metacognitive sensitivity:

$$\text{meta-I} = H(\text{accuracy}) - H(\text{accuracy}|\text{confidence})$$

This measures mutual information between confidence and accuracy, quantifying how much confidence reduces uncertainty about accuracy.

Comparison:

Aspect	meta-I	Our K Framework
Theoretical Basis	Information Theory	Recursive Epistemology

Aspect	meta-I	Our K Framework
Model Dependency	Model-free	Model-free (observational)
Direction	Unsigned (sensitivity only)	**Signed** (over/under-confidence)
Layers	Single layer	**Recursive** ($K_0, K_1, K_2, \ldots$)
Output	Bits (continuous)	$[-1, 1]$ (continuous)
Granularity	Can measure response granularity	Per-item structural classification

Key Distinction:

Both meta-I and K are **model-free** (unlike meta-d' which requires SDT assumptions), but they serve different purposes:

- **meta-I** answers: "How well does confidence track accuracy?" (quantitative sensitivity)
- K answers: "What type of metacognitive pattern is this?" (qualitative classification)

Complementary Use Case:

Two subjects with identical meta-I = 0.3 bits (low sensitivity): - Subject A: $K_0 = 0$, $K_1 = -1$ -> Dunning-Kruger -> needs awareness intervention - Subject B: $K_0 = 1$, $K_1 = -1$ -> Imposter -> needs confidence-building

meta-I cannot distinguish these cases; K can.

Calibration Metrics (Brier Score, ECE)

Calibration metrics measure whether confidence aligns with accuracy across many trials.

Comparison:

Aspect	Calibration Metrics	Our $K(K(x))$
Granularity	Aggregate statistics	Individual items
Purpose	Probabilistic accuracy	Epistemic state recognition
Socratic Wisdom	Not explicitly modeled	Explicitly formalized

Calibration metrics and K are complementary: calibration measures aggregate confidence-accuracy alignment, while K provides per-item structural classification.

Belief Functions and Uncertainty (Dempster-Shafer Theory)

Dempster-Shafer theory handles **epistemic uncertainty** and **conflicting evidence** via belief functions.

Comparison:

Aspect	Dempster-Shafer	Our Model
Focus	Uncertainty quantification	Metacognitive discrepancy
Application	Evidence combination	Self-awareness structure
Ignorance	Represented as belief mass	$K(x) = 0$ (epistemic state)

Dempster-Shafer addresses uncertainty quantification; K addresses metacognitive discrepancy (the gap between what one knows and what one thinks one knows).

Dunning-Kruger Effect (Empirical Psychology)

Kruger and Dunning (1999) empirically demonstrated that low-competence individuals overestimate their abilities.

Our Contribution: We provide a **formal mathematical model** of this phenomenon: - $K(x) = 0$ (low competence) - $K(K(x)) = -1$ (misrecognition: believes they have competence) - Often accompanied by $C = 1$ (high confidence)

This formalization enables: 1. Precise measurement protocols 2. Distinction from related phenomena (e.g., imposter syndrome) 3. Extension to arbitrary depths of self-reflection

Application to AI Metacognition

The K framework provides a structured vocabulary for evaluating metacognition in Large Language Models (LLMs), an increasingly important area as AI systems are deployed in high-stakes domains.

Mapping LLM Behaviors:

LLM Pattern	K_0	K_1	K_2	Interpretation
Correct + confident	1	1	1	Ideal calibration
Hallucination + confident	-1	-1	?	Confident wrong (dangerous)

LLM Pattern	K_0	K_1	K_2	Interpretation
Correct + hedging	1	-1	?	Underconfident (imposter-like)
Admits uncertainty	0	1	1	Appropriate uncertainty (Socratic)
"I don't know" when wrong	-1	1	?	Partial awareness of limits

Testbed Proposal:

Using decoupled confidence elicitation (analogous to AFCE-style protocols):

1. **Query LLM for answer** -> compute K_0 (against ground truth)
2. **Query LLM for confidence** -> record C (0-100%)
3. **Query LLM: "Do you know this?"** -> elicit Claim_1
4. **Compute** K_1 from Claim_1 vs K_0

This protocol allows testing: - **K vs C dissociation** in LLMs (do they exhibit Dunning-Kruger or Imposter patterns?) - **Domain-specific calibration** (are LLMs more self-aware in some domains?) - **Intervention effects** (does prompting for self-reflection improve K_1?)

Why K is Useful for LLM Evaluation:

Current LLM calibration research focuses primarily on confidence-accuracy correlation (analogous to meta-d' or meta-I). The K framework adds:

1. **Pattern classification**: Identifying *which type* of miscalibration
2. **Directional information**: Distinguishing overconfidence from underconfidence
3. **Intervention guidance**: Suggesting targeted prompting strategies

Future Work: Operationalizing K_n for LLMs using abstention behavior, self-consistency checks, and metamorphic testing remains an open challenge (see Limitations).

Layer-wise Operationalization for LLMs:

Layer	Human Operationalization	LLM Operationalization
K_0	Response vs ground truth	Output vs benchmark answer
K_1	Self-assessment claim	Verbalized confidence / hedging
K_2	Meta-self-assessment	"How reliable is my confidence?" prompt

Connection to Recent Research *Note: Right-column descriptions are interpretations of these works within the K framework, not claims made by the original authors.*

Research Area	K Framework Contribution
Human-AI metacognition (Fernandes et al., 2024; arXiv:2409.16708)	K_1 patterns provide a natural way to describe why the Dunning-Kruger effect disappears with AI assistance; AI "levels" metacognitive accuracy
LLM uncertainty communication (Steyvers et al., 2025; arXiv:2510.05126)	Fine-tuning improves LLM calibration/discrimination; K_1 highlights directional (over/under-confidence) patterns that scalar metrics like ECE cannot fully characterize
Latent knowledge probing (Burns et al., 2022; arXiv:2212.03827, ICLR 2023)	Probing internal activations reveals what LLMs "know" vs "say"; potential alternative to verbalized g_0
LLM metacognitive capacity (Li et al., 2025; arXiv:2505.13763)	Neurofeedback paradigm quantifies LLM self-monitoring; suggests "metacognitive space" is low-dimensional
VLM uncertainty estimation (Lin et al., 2025; arXiv:2511.22019)	Post-hoc probabilistic embeddings for error detection; one concrete operational choice for $g_n/\hat{K}$ in vision-language settings
Probabilistic VLM embeddings (Venkataramanan et al., 2025; arXiv:2505.05163, UAI 2025)	GPLVM-based uncertainty calibration; aligns with Option B's probabilistic K_n estimation
Two-level metacognitive architecture (Li et al., 2025; arXiv:2511.23262)	Meta-level/object-level separation mirrors K_0/K_1 layer structure; can serve as a testbed for probing whether meta-reasoning improves K_1-type behaviour without necessarily improving K_0
Human-AI teaming	Match human K_n patterns with AI K_n for improved collaboration

Correspondence with Monitor-Generate-Verify Architectures Recent LLM metacognition research formalizes recursive self-monitoring via Monitor-Generate-Verify (MGV) loops. The K framework provides a natural coordinate system for these architectures.

Mapping MGV Components to K Layers:

MGV Component	K Layer	Correspondence
Generate	K_0	First-order response quality
Monitor	K_1	Self-assessment of response quality
Verify	K_2	Meta-assessment of monitoring accuracy

Iteration Dynamics:

In MGV, the loop proceeds: 1. Generate response → observe K_0 (against reference) 2. Monitor quality → compute K_1 (self-assessment alignment) 3. Verify monitoring → compute K_2 (meta-monitoring accuracy) 4. If $K_2 < \tau$, regenerate (loop back to step 1)

K **Framework as Stopping Criterion:**

$$\text{Accept output iff } K_1 \geq \tau_1 \text{ AND } K_2 \geq \tau_2$$

This formalizes the intuition that LLMs should only output when: - They are confident ($K_1 \geq \tau_1$), AND - That confidence is justified ($K_2 \geq \tau_2$)

Complementarity:

Framework	Provides	Does NOT Provide
MGV	Process (how to iterate)	Coordinates (where in metacognitive space)
K Framework	Coordinates (where the system is)	Process (how to move)

Integration Potential:

MGV architectures can use K_n as the objective function for optimization: - Maximize $E[K_1|\text{generation strategy}]$ - Minimize variance of K_2 across domains

This integration enables principled design of self-improving LLM systems.

Scope Boundary: Detailed LLM operationalization and experiments are marked for future work.

Novel Contributions

This study is novel in:

1. **Recursive Formalization**: Extending metacognition to arbitrary depths ($K_0 \to K_1 \to K_2 \to ...$)
2. **Socratic Wisdom as Achievement**: Explicitly modeling “knowing ignorance” as $K_0(x) = 0, K_1(x) = 1$

3. **Layered Observation Model**: Justifying recursive structure via indexed observations with anchor constraints
4. **Orthogonal Dimensions**: Separating epistemic state (K) from phenomenological confidence (C)
5. **Per-Item Granularity**: Measuring metacognition at the individual item level, not just aggregate statistics

Conclusion and Future Challenges

This study constructed a recursive epistemic model based on the hierarchical structure of knowledge. Using a single core function K applied recursively—$K(x)$, $K(K(x))$, $K(K(K(x)))$—we provide a mathematically rigorous yet philosophically grounded framework that captures:

1. The **recursive nature of self-awareness**: The same epistemic question applies at every level of reflection.
2. The **hierarchical structure of metacognition**: Distinguishing "knowing ignorance" (Socratic wisdom) from "unknowing ignorance" (Dunning-Kruger effect).
3. The **continuous gradation of knowledge**: Knowledge states exist on a continuum from misconception (-1) through ignorance (0) to accurate knowledge (1).

Main Results

1. We proposed a **layered observation model** $\{K_n\}$ with formal anchor constraints (preservation, monotonicity, boundedness), demonstrating that recursive metacognition is a well-founded observational structure. The symbolic notation $K(K(x))$ is retained for conceptual communication, while all formal work uses K_n exclusively.
2. We adopted a stance of **methodological relativism**, treating the proposition x itself as the implicit reference point and leaving the designation of "correct" to the experimental context.
3. We provided the **Four Quadrants of Metacognition**, clearly distinguishing "Knowing Ignorance" (Socratic wisdom, $K_0(x) = 0, K_1(x) = 1$) from "Unknowing Ignorance" (Dunning-Kruger effect, $K_0(x) = 0, K_1(x) = -1$).
4. We separated **epistemic state** (K) from **phenomenological confidence** (C), recognizing them as orthogonal dimensions.
5. We proposed the **Metacognitive Alignment Test (MAT)** as an experimental protocol to validate the model, with specific predictions about the benefits of Socratic wisdom.

Formal Contributions Summary

The paper establishes the following results (detailed in "Formal Results: Illustrative Derivations and Informal Propositions"):

Result	Statement	Location
Result 1	K_0-IRT correspondence: $K_0 = \tanh(a(\theta - b)/2)$ (illustrative derivation)	Formal Results
Result 2	K_1-Phi correspondence under binary conditions	Formal Results
Proposition 3	K_0 identifiability: $\text{Var}(b_i) > 0 \Rightarrow K_0$ identifiable (informal)	Formal Results
Proposition 4	K_1 identifiability given K_0 and claim variability (informal)	Formal Results
Proposition 5	ICC reliability conditions for stable K measurement (informal)	Formal Results
Proposition 6	Pipeline identifiability: (K_0, K_1, K_2) jointly identifiable (informal)	Formal Results
Lemma 3	$\hat{K}$ sufficiency: any monotone anchor-preserving function yields equivalent ordinal results	Formal Results

Key Identifiability Arguments (Informal): - The framework outlines conditions under which parameters should be recoverable from data - The recursive measurement pipeline avoids circularity by design (Proposition 6) - The specific form of $\hat{K}$ is sufficient but not unique (Lemma 3)

Falsifiable Predictions: Five quantitative predictions (P1-P5) with explicit bounds enable empirical refutation of the framework.

Theoretical Contributions

- **Recursive Formalization**: Extending metacognition to arbitrary depths while maintaining mathematical consistency
- **Layered Observation Model**: Justifying recursive structure via indexed observations (K_n) with formal anchor constraints
- **Socratic Wisdom as Achievement**: Explicitly modeling "knowing ignorance" as a high metacognitive state
- **Per-Item Granularity**: Measuring metacognition at the individual item level, complementing aggregate statistical measures

Limitations

This framework is a **conceptual scaffold** for organizing metacognitive phenomena, not a complete predictive model. We acknowledge the following limitations:

1. **Formal Model:** We adopt an observational family interpretation to resolve the recursion/observation tension. This is a modeling choice, not the only possibility.
2. **Single Dimension (Scope Boundary):**
   - The $[-1, 1]$ scale assumes a **single axis of correctness** with one "opposite" direction.
   - **What this captures:** Directional errors (overconfidence vs underconfidence, correct vs incorrect).
   - **What this does NOT capture:** Multiple, qualitatively different misconceptions (e.g., "thinks A" vs "thinks B" when truth is C).

   Formal Clarification:

   The current model assumes a binary opposition: Reference $\leftrightarrow$ Anti-Reference

$$K_0 = \begin{cases} 1 & \text{if Response} = \text{Reference} \\ 0 & \text{if Response} = \text{Absent} \\ -1 & \text{if Response} \neq \text{Reference} \end{cases}$$

   This collapses all incorrect responses to -1, regardless of *which* error was made.

   When This is Appropriate:

   - Binary propositions (true/false)
   - Single-answer factual questions
   - Ordinal scales with clear direction

   When This is Limiting:

   - Conceptual errors with multiple wrong paths
   - Skill assessments with qualitatively different failure modes
   - Open-ended responses

   Future Extension (Out of Scope):

   For multi-dimensional misconceptions, consider:

$$K_0 : \mathcal{X} \to [-1, 1]^d$$

   Where d = number of independent error dimensions. This requires:

   - Multi-dimensional reference space

- Vector-valued embedding g_0
- Revised axioms for vector order

We leave this extension for future work, noting that the current scalar formulation covers a wide range of practical applications.

3. **Scope Boundary: Binary vs. Graded Truth**

Current Scope: This framework assumes **binary correctness** at K_0:

- Correct ($K_0 = +1$)
- Absent ($K_0 = 0$)
- Incorrect ($K_0 = -1$)

Out of Scope (Future Work):

Extension	Challenge	Potential Approach
Partial Credit	$K_0 \in (0, 1)$ requires graded reference	Probabilistic f_0 with continuous output
Multi-label	Multiple correct answers	Set-valued K_0 or soft-max embedding
Graded Truth	Degrees of correctness	Fuzzy reference with $K_0 = \text{similarity}(\text{Response}, \text{Reference})$

Why Binary for Now: Binary correctness enables clean anchor semantics and unambiguous K_1/K_2 computation. Graded extensions require principled definitions of "partial alignment" that preserve interpretability.

Extension Path: Polytomous and Partial Credit Scoring

For polytomous outcomes (e.g., rubric scores 0, 1, 2, 3), we specify a concrete extension via the Graded Response Model (GRM):

$$P(Y \geq k|\theta) = \frac{1}{1 + e^{-a(\theta - b_k)}}$$

Mapping to Graded K_0:

$$K_0^{(\text{graded})} = \frac{2Y - Y_{\max}}{Y_{\max}}$$

Where Y is the observed score and $Y_{\max}$ is the maximum possible score.

Impact on Higher Layers:

Graded K_0 propagates to K_1 with graded alignment:

- "I'm 80% confident" matches $K_0 = 0.6 \rightarrow$ partial alignment
- "I'm 80% confident" matches $K_0 = -0.2 \rightarrow$ partial misalignment

When to Use Each Approach:

Scenario	Recommended Approach
Factual Q&A (binary correct/incorrect)	Trichotomous $K_0 \in \{-1, 0, +1\}$
Essay grading (rubric-based)	Graded K_0 via GRM mapping
Programming (test case pass rate)	Proportion-based $K_0 = 2p - 1$ where p = pass rate
Multiple-choice with partial credit	GRM-based K_0

Implementation Status: Conceptually compatible with the framework; formal development deferred to future work.

4. **Informativeness Constraint**

   Problem: A subject could trivially achieve $K_n = 0$ for all n by always claiming "I'm not sure." This would satisfy the framework's consistency requirements without providing useful information.

   Solution: We distinguish between **legitimate uncertainty** and **uninformative hedging** via:

   - **Response Distribution Analysis**:
     - If $P(\text{Claim}_n = \text{"not sure"}) > 0.8$ across items, flag as potentially uninformative
     - Legitimate uncertainty should correlate with item difficulty
   - **Coherence Check**:
     - Subjects with genuine uncertainty should show K_0 variance (some correct, some incorrect)
     - Subjects gaming the system show uniform "not sure" regardless of K_0 distribution
   - **Incentive Design** (Experimental):
     - Proper scoring rules that penalize uninformative claims
     - Reward calibration: higher payoff for confident-and-correct vs. uncertain-and-correct

   Informativeness Index:

   $$\text{Informativeness}(K_n) = 1 - H(K_n)/H_{\max}$$

   Where $H(K_n)$ is the entropy of the subject's K_n distribution. Low informativeness (high entropy, uniform distribution) combined with no correlation to K_{n-1} triggers a warning.

5. **Minimal Axiomatic Theory:** We provide basic constraints on K (anchor preservation, monotonicity, boundedness) but do not specify a unique functional form. The specific dynamics of K (e.g., whether it is contractive, has fixed points beyond $\{-1, 0, 1\}$) are empirical questions.

6. **No Generative Model:** We do not provide a noise model or generative account of how states are produced. This is a task for computational cognitive modeling.

7. **Observation Protocol:** The mapping from observable behavior to K_n values requires operational definitions. While we provide guidelines (e.g., alignment between claims and performance), the specific elicitation methods depend on the application domain.

8. K_2 **Identifiability:** Higher-order estimates (K_2, K_3) require more items and may be less reliable.

   - **Guideline:** For reliable K_2 estimation, use $N \geq 50$ items with noise < 0.2.
   - **Confidence intervals:** Report 95% CI via bootstrap; if CI width > 0.3, interpret with caution.

9. **Simulation Validation Pending:** We have not yet provided simulated evidence that different metacognitive profiles (Socratic, Dunning-Kruger, Imposter) are identifiable under realistic noise. This is planned for future work.

10. **Analogical Type Theory:** The type-theoretic justification is analogical rather than formally constructed. A full domain-theoretic or typed lambda-calculus treatment is beyond the current scope.

11. **LLM Operationalization Incomplete:** Applying K_n to LLMs requires addressing question-side shortcuts and model-side signals. Specific methods (conformal coverage, debate protocols) are suggested but not developed here. **This is explicitly designated as Future Work.** The current paper provides the conceptual framework; LLM-specific operationalizations (token-level probability extraction, multi-sample self-consistency, debate protocols) require separate empirical validation.

12. **Normative vs Descriptive Interpretation:**

    The Tension: Recent work (e.g., positive evidence bias literature) suggests that some "metacognitive biases" may be rational responses to high-dimensional hypothesis spaces, not failures.

    Our Position: The K framework is **DESCRIPTIVE**, not normative.

    $K_1 = -1$ means: "The subject's claim does not match their actual state relative to the experimenter-defined reference."

    This is a **COORDINATE**, not a **JUDGMENT**.

When $K_1 = -1$ Might Be "Rational":

Scenario	Why $K_1 = -1$ May Be Rational
Contested reference	Experimenter's "ground truth" is incomplete or biased
Different objective	Subject optimizes for social harmony, not accuracy
Information asymmetry	Subject has valid information experimenter lacks
Bayesian conservatism	Prior knowledge rationally weighted against weak evidence

> **Recommendation**: Report K_1 alongside task context. Do not interpret $K_1 = -1$ as "failure" without considering whether the reference standard is appropriate.
>
> **Methodological Relativism (Reiterated)**: The framework does NOT adjudicate what is "correct." $K_1 = -1$ is descriptive of a state, not prescriptive of a norm.

Validation Roadmap

This paper establishes the **conceptual framework**; validation is planned for follow-up work. We present a staged validation program:

Phase 1: Reliability

Type	Method	Target
Test-Retest	2-week interval, same items	$r > 0.7$
Internal Consistency	Cronbach's α across items	$\alpha > 0.8$
Split-Half	Odd-even item split	$r > 0.75$

Phase 2: Convergent Validity

K Measure	Comparison Metric	Expected Correlation
K_1	meta-d'/d'	$r > 0.6$ (positive)
K_1	Type-2 AUROC	$r > 0.5$ (positive)
K_0	IRT ability θ	$r > 0.8$ (positive)

Phase 3: Discriminant Validity

K Measure	Comparison	Expected
K_1	Raw confidence C	$r < 0.3$ (low)
K_0	Response time	$r < 0.2$ (low)

Phase 4: Predictive Validity

Predictor	Outcome	Hypothesis
$K_1 = 1$ (Socratic)	Help-seeking	Higher
$K_2 = 1$	Intervention responsiveness	Higher
$K_1 = -1$ (DK)	Overconfident errors	Higher

Acknowledgment: We recognize that this validation program is **essential** for empirical adoption. The current paper's contribution is the **conceptual and formal foundation** upon which such validation can be built.

Future Work: Technical Extensions

This conceptual framework invites several technical extensions that are beyond the current paper's scope:

1. Formal Measurement Theory - Probabilistic specification of f_n and State$_n$ as random variables - Likelihood-based estimation with noise models - Identifiability conditions connected to general latent variable theory (Allman et al., 2009; Anandkumar et al., 2014)

2. Estimation Algorithms - EM or MCMC algorithms for continuous K_n estimation - Handling of abstentions and missing data (MNAR modeling) - Computational complexity and convergence analysis

3. Empirical Validation - Synthetic experiments with known ground-truth K_n profiles - Recovery of K_0, K_1 under noise; testing correspondence to IRT/meta-d'/ECE - Stress tests for K-C dissociation and layer independence predictions

4. Extended Metrics - Integration with improved calibration metrics (ACE, SCE, class-wise calibration) - Connections to Hierarchical meta-d' (Fleming, 2017) estimation

5. Applications - LLM metacognition operationalization - Educational diagnostics - Clinical assessment protocols

We view these as **separate technical contributions** that build upon the conceptual scaffold established here.

Future Directions

1. **Empirical Validation**: Conduct MAT experiments to validate the model's predictions about Socratic wisdom and Dunning-Kruger effect, and compare with meta-d' and calibration metrics.
2. **Simulation Studies**: Simulate agents with known K/C profiles to demonstrate identifiability and estimate required sample sizes.
3. **Formal Type Theory**: Develop a typed calculus or algebraic data type where K's self-application is a well-typed endomorphism.
4. **AI Safety Applications**: Operationalize $K(K(x))$ for LLMs using abstention behavior, self-consistency, and metamorphic testing.
5. **Cultural and Domain Variation**: Investigate whether the symmetry assumption (misconception vs. knowledge) holds across cultures and domains.

Why R/E Separation is Unnecessary

One alternative formalization might introduce two separate maps: a subject-level recognition/report map R and an evaluator/alignment map E. We argue that such separation is **unnecessary** for our framework.

The Misunderstanding:

Such a proposal typically interprets $K(K(x))$ as: 1. $K(x)$ = first-order state (a number) 2. $K(K(x))$ = applying K to that number 3. Therefore $K(0)$ should equal 0

The Reality:

Our K is purely **observational**: 1. K_0 = observation of State$_0$ (first-order epistemic state) 2. K_1 = observation of State$_1$ (metacognitive state) 3. K_2 = observation of State$_2$ (meta-metacognitive state)

Each layer observes a different object. There is no "subject's report" vs "evaluator's assessment" — there is only **observation of states**.

Why R/E Adds Unnecessary Complexity:

R/E Approach	Our Approach
Introduces "subject" as agent	No subject, only states
Requires modeling "perception"	States exist, observer measures
Two functions (R, E)	One function (K)
Subjective/objective split	Purely objective

Our framework is simpler and more elegant because it does not reify "the subject" as a separate entity with "perceptions."

All that exists are **states** and **observations**. K observes states. That's it.

Appendix: Technical Lemmas and Supplementary Propositions

Robustness Lemmas

Lemma 1: Scale Invariance of K_0 **Lemma 1** (Scale Invariance): $K_0 = \tanh(a(\theta - b)/2)$ is invariant under affine rescaling of θ iff a and b are co-transformed.

Formally: Let $\theta' = c\theta + d$ for constants $c > 0$ and d. Then $K_0(\theta') = K_0(\theta)$ iff:

$$a' = \frac{a}{c}, \quad b' = cb + d$$

Proof:

$$K_0(\theta') = \tanh\left(\frac{a'(\theta' - b')}{2}\right) = \tanh\left(\frac{(a/c)(c\theta + d - cb - d)}{2}\right) = \tanh\left(\frac{a(\theta - b)}{2}\right) = K_0(\theta) \quad \blacksquare$$

Implication: K_0 values are robust to arbitrary scale/location transformations of the latent ability θ, provided item parameters are appropriately re-estimated.

Lemma 2: K_1 under Class Imbalance **Lemma 2** (Class Imbalance Bound): Let $\pi = P(K_0 = +1)$ be the base rate of correct knowledge.

The Phi coefficient K_1 satisfies:

$$|K_1| \leq \sqrt{\frac{\min(\pi, 1-\pi)}{\max(\pi, 1-\pi)}}$$

Corollary: When $\pi \to 0$ or $\pi \to 1$, $|K_1| \to 0$ regardless of true alignment.

Proof:

The Phi coefficient achieves its maximum when the 2×2 table has deterministic structure. Given marginal constraint $\pi = P(K_0 = +1)$, the maximum occurs when: - If $\pi \leq 0.5$: All $K_0 = +1$ cases have $\text{Claim}_1 =$ "I know", and the remaining "I know" claims are minimized. - The resulting maximum is:

$$|\phi|_{\max} = \sqrt{\frac{\min(\pi, 1-\pi)}{\max(\pi, 1-\pi)}}$$

Derivation: Let $p = P(\text{Claim}_1 = \text{"I know"})$. The maximum correlation between two binary variables with marginals π and p is achieved when they are monotonically related. By the Frechet-Hoeffding bounds, the maximum joint probability is $\min(\pi, p)$, yielding:

$$\phi_{\max} = \frac{\min(\pi, p) - \pi p}{\sqrt{\pi(1-\pi)p(1-p)}}$$

Optimizing over p and applying algebra gives the stated bound. As $\pi \to 0$ or $\pi \to 1$, the bound $\to 0$. ■

Recommendation: Report K_1 alongside π (base rate). For robustness checks, compute Matthews Correlation Coefficient (MCC) which is equivalent to Phi but emphasizes this constraint in interpretation.

Supplementary Propositions

Remark 1: K_1 and meta-d'—Complementary Perspectives

> **Critical Note**: The relationship between K_1 and meta-d' is **conceptual**, not mathematical. We do **NOT** claim $K_1 \approx \tanh(\text{meta-d}'/2)$ as an identity or conversion formula. These measures capture related but distinct aspects of metacognition.

Remark 1 (K_1 and meta-d': Complementary Perspectives):

K_1 and meta-d' (Maniscalco & Lau, 2012) measure related but distinct aspects of metacognition:

meta-d' (Signal Detection Theory): - **Sensitivity-based**: How well does confidence discriminate correct from incorrect responses? - Rooted in Signal Detection Theory (SDT) - Requires Gaussian assumptions on evidence distributions - Unit: d' scale (unbounded, typically in range $[0, 4]$) - Definition: $\text{meta-d}' = \Phi^{-1}(\text{HR}_2) - \Phi^{-1}(\text{FAR}_2)$

K_1 (This framework): - **Alignment-based**: Does the subject's claim match their actual epistemic state? - Assumption-free when using Phi coefficient - Bounded: $K_1 \in [-1, 1]$ - Definition: $K_1 = \phi(\text{State}_0, \text{Claim}_1)$

Relationship: - Both are high when metacognition is accurate - Both are low/zero when confidence is unrelated to accuracy - However, they are **not mathematically equivalent**

Why No Exact Correspondence: 1. meta-d' is defined via inverse normal CDFs; K_1 via correlation 2. meta-d' depends on SDT model assumptions; K_1 is model-free 3. meta-d' is unbounded; K_1 is bounded to $[-1, 1]$

Empirical Expectation: - In well-calibrated populations with Gaussian confidence distributions:

$$\text{Cor}(K_1, \tanh(\text{meta-d}'/2)) > 0.6$$

(expected but not guaranteed) - In populations with non-Gaussian distributions or extreme base rates: divergence is expected

Recommendation: - Report **both** K_1 and meta-d' when SDT framework assumptions are plausible - Prefer K_1 when SDT assumptions are questionable (e.g., non-binary responses, non-Gaussian confidence) - Do **not** use $K_1 \approx \tanh(\text{meta-d}'/2)$ as an identity or conversion formula

Complementary Use Case: | Question | Preferred Metric | |:———|:————––| | "How well does confidence separate correct from incorrect?" | meta-d' | | "Does the subject's claim align with their actual state?" | K_1 | | "What is the pattern of metacognitive (mis)alignment per item?" | $K_1^{(i)}$ (item-level) |

Proposition 1b: K_1-ECE Approximate Relationship

Proposition 1b (K_1-ECE Approximate Relationship):

Under specific conditions, K_1 can be approximated from Expected Calibration Error (ECE).

Statement: Under fixed binning with B bins and large sample size N:

$$K_1 \approx 1 - 2 \cdot \text{ECE}$$

where ECE is defined as:

$$\text{ECE} = \sum_{b=1}^{B} \frac{|B_b|}{N} |\text{acc}(B_b) - \text{conf}(B_b)|$$

Critical Warnings:

Issue	Description	Impact on $K_1 \approx 1 - 2 \cdot \text{ECE}$
Binning-dependent	Different $B \rightarrow$ different ECE	Same data can yield different K_1 estimates
Sample-size dependent	Small $N \rightarrow$ high variance in ECE	Unreliable K_1 estimates
Always non-negative	ECE ≥ 0 by definition	Cannot distinguish over- from under-confidence

Issue	Description	Impact on $K_1 \approx 1 - 2 \cdot \text{ECE}$
Aggregate only	ECE is computed across all items	No item-level $K_1^{(i)}$

Scope of Validity: - Approximately valid for $B \geq 10$, $N \geq 500$, balanced confidence distribution - Breaks down for small samples, extreme base rates, or heavily skewed confidence

Recommendation: - Use $K_1 = \phi$ (Phi coefficient) as the **primary** measure for the K framework - Use $K_1 \approx 1 - 2 \cdot \text{ECE}$ only for **cross-study comparison** where ECE is the standard metric - Always report binning scheme B and sample size N when using ECE-based approximation - Consider Brier score decomposition for more stable calibration assessment

---

Claim 1: Framework Novelty **Claim 1** (Framework Novelty—Survey-Based): Based on our survey of existing approaches, the K framework is the only approach satisfying ALL of the following properties:

> **Note**: This is a survey-based claim, not a mathematically provable proposition. The uniqueness is established relative to the approaches reviewed (meta-d', ECE/Brier, IRT), not as an absolute logical necessity.

Property	K Framework	meta-d'	ECE/Brier	IRT
(a) Per-item metacognitive classification	Yes	No	No	Yes
(b) Explicit modeling of ignorance ($K = 0$) distinct from misconception ($K = -1$)	Yes	No	No	No
(c) Recursive higher-order extension ($K_2, K_3, ...$)	Yes	No	No	No

Property	K Framework	meta-d’	ECE/Brier	IRT
(d) Unified anchor semantics across layers	Yes	N/A	N/A	N/A

Justification (per property):

(a) Per-item metacognitive classification: - meta-d’: Computes aggregate sensitivity across all items; cannot classify individual items as “Socratic” or “Dunning-Kruger” - ECE/Brier: Aggregate calibration metrics; no item-level output - IRT: Provides per-item difficulty/discrimination but no metacognitive layer - K **framework**: $K_1^{(i)}$ is defined for each item i

(b) Explicit ignorance ($K = 0$) vs misconception ($K = -1$): - meta-d’: Binary Type-1 outcome (correct/incorrect); “I don’t know” treated as low confidence, not distinct state - ECE/Brier: No representation of “absence of stance” - IRT: θ is continuous; no categorical “ignorance” state - K **framework**: $K_0 = 0$ explicitly represents epistemic absence, distinct from $K_0 = -1$ (wrong belief)

(c) Recursive higher-order extension: - meta-d’: Single-layer (Type-2); no meta-meta-d’ defined in literature - ECE/Brier: No recursive structure - IRT: First-order only; no $\theta(\theta)$ - K **framework**: $K_2, K_3, \ldots$ formally defined with consistent semantics

(d) Unified anchor semantics: - This property is unique to recursive frameworks; meta-d’/ECE/IRT have no analogous requirement - K **framework**: Same anchors $(+1, 0, -1)$ with consistent interpretation across all layers

Conclusion: No existing framework satisfies (a)-(d) simultaneously. ■

Proposition 6: Computational Complexity **Proposition 6** (Polynomial-Time Estimation): All K_n estimates ($n = 0, 1, 2$) can be computed in polynomial time.

Layer	Estimator	Complexity
K_0	2PL-IRT via EM algorithm	$O(N \cdot I \cdot \text{iter})$ where $N =$ subjects, $I =$ items
K_1	Phi coefficient	$O(N)$
K_2	ICC (two-way random effects)	$O(N \cdot T)$ where $T =$ measurement occasions

All operations are standard and available in common statistical software (R: `ltm`, `psych`; Python: `pymc`, `scipy`).

Worked Example: Complete K_2 Computation

This section provides a concrete, step-by-step example of computing K_0, K_1, and K_2 for a single subject responding to 10 items.

Step 1: Raw Data Collection A subject completes 10 multiple-choice items with Claim$_1$ elicitation:

Item	Response	Correct?	Claim$_1$
1	A	Yes	"I know"
2	B	Yes	"I know"
3	C	No	"I don't know"
4	A	Abstain	"Uncertain"
5	B	Yes	"I know"
6	D	No	"I know"
7	A	No	"Uncertain"
8	C	Yes	"I know"
9	B	Abstain	"I don't know"
10	A	Yes	"I know"

Step 2: Compute K_0 per Item Using the embedding g_0: correct → +1, abstain → 0, incorrect → -1:

Item	Correct?	K_0
1	Yes	+1
2	Yes	+1
3	No	-1
4	Abstain	0
5	Yes	+1
6	No	-1
7	No	-1
8	Yes	+1
9	Abstain	0
10	Yes	+1

Aggregate K_0: $\bar{K}_0 = \frac{1+1-1+0+1-1-1+1+0+1}{10} = \frac{2}{10} = 0.2$

Step 3: Compute K_1 per Item Using f_1 decision rules (Default interpretation):

Item	K_0	Claim$_1$	State$_1$	K_1
1	+1	"I know"	aligned	+1
2	+1	"I know"	aligned	+1
3	-1	"I don't know"	aligned*	+1
4	0	"Uncertain"	aligned	+1
5	+1	"I know"	aligned	+1
6	-1	"I know"	misaligned	-1
7	-1	"Uncertain"	uncertain	0
8	+1	"I know"	aligned	+1
9	0	"I don't know"	aligned	+1
10	+1	"I know"	aligned	+1

*Item 3: $K_0 = -1$ with "I don't know" → aligned (Epistemic Improvement Criterion)

Aggregate K_1: $\bar{K}_1 = \frac{1+1+1+1+1-1+0+1+1+1}{10} = \frac{7}{10} = 0.7$

Step 4: Elicit Claim$_2$ After completing all items, subject is asked: > "How accurate do you think your self-assessments were?"

Subject responds: **"My self-assessments were accurate"** → Claim$_2$ = "Meta-aligned"

Step 5: Compute K_2 **Option A: Threshold-Based**

Discretize $\bar{K}_1 = 0.7$ using equal-interval thresholds ($\tau^+ = 1/3$, $\tau^- = -1/3$): - $\bar{K}_1 = 0.7 > 1/3$ → Discrete $K_1 = +1$ (mostly aligned)

Compare with Claim$_2$: - Claimed: "Meta-aligned" (expects alignment) - Actual: Discrete $K_1 = +1$ (aligned) - Match? **Yes** → State$_2$ = meta-aligned → $K_2 = +1$

Option B: Test-Retest ICC

If subject repeats the task at time t_2:

Item	$K_1(t_1)$	$K_1(t_2)$
1	+1	+1
2	+1	+1
3	+1	+1
4	+1	0
5	+1	+1
6	-1	-1
7	0	+1

Item	$K_1(t_1)$	$K_1(t_2)$
8	+1	+1
9	+1	+1
10	+1	+1

Compute ICC (Intraclass Correlation Coefficient):

$$\text{ICC} = \frac{\sigma^2_{\text{between-item}}}{\sigma^2_{\text{between-item}} + \sigma^2_{\text{within-item}}} \approx 0.82$$

$K_2 = 0.82$ (high metacognitive stability)

Step 6: Summary

Metric	Value	Interpretation
$\bar{K}_0$	0.2	Moderate knowledge (slightly above ignorance)
$\bar{K}_1$	0.7	Good metacognitive alignment
K_2 (Option A)	+1	Subject correctly recognizes their good metacognition
K_2 (Option B)	0.82	Stable metacognition across time

Pattern Classification (using $(K_0, K_1, K_2) = (0, +1, +1)$ discretized): - Pattern #14: **Socratic Wisdom**—Subject has limited knowledge but accurately recognizes this and knows their self-assessment is reliable.

References

1. Kant, I. (1781). *Critique of Pure Reason.*
2. Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive-developmental inquiry. *American Psychologist*, *34*(10), 906-911.
3. Nelson, T. O., & Narens, L. (1990). Metamemory: A theoretical framework and new findings. In G. H. Bower (Ed.), *The Psychology of Learning and Motivation* (Vol. 26, pp. 125-173). Academic Press.
4. Koriat, A. (1993). How do we know that we know? The accessibility model of the feeling of knowing. *Psychological Review*, *100*(4), 609-639.
5. Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology*, *77*(6), 1121-1134.

6. Maniscalco, B., & Lau, H. (2012). A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings. *Consciousness and Cognition*, *21*(1), 422-430.
7. Shafer, G. (1976). *A Mathematical Theory of Evidence*. Princeton University Press.
8. Fleming, S. M., & Daw, N. D. (2017). Self-evaluation of decision-making: A general Bayesian framework for metacognitive computation. *Psychological Review*, *124*(1), 91-114.
9. Dayan, P. (2023). Metacognitive information theory. *bioRxiv*. https://doi.org/10.1162/opmi_a_00091
10. Samejima, F. (1969). *Estimation of latent ability using a response pattern of graded scores*. Psychometrika Monograph Supplement, No. 17.
11. Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, *16*(2), 159-176.
12. Embretson, S. E., & Reise, S. P. (2000). *Item Response Theory for Psychologists*. Lawrence Erlbaum Associates.
13. Fleming, S. M. (2017). HMeta-d: Hierarchical Bayesian estimation of metacognitive efficiency from confidence ratings. *Neuroscience of Consciousness*, *3*(1), nix007.
14. Lord, F. M., & Novick, M. R. (1968). *Statistical Theories of Mental Test Scores*. Addison-Wesley.