Correction: the ε selection law is Thurstone's 1927 comparative judgment, not a discovery

W. Schulz

Jul 26, 2026

Abstract

Our ε law, $\varepsilon = \Phi(-\text{gap}/(\sigma\sqrt{2}))$, is Thurstone's Case V identity; its $R^2 = 0.997$ measured Monte-Carlo error. It is not a universality class (723× off under heavy tails), and the forgetting row is Kim et al. 2022. The consensus threshold p/(1−p), the measured constants, and the tools stand.

Issued 2026-07-26. This is the canonical correction notice for every ε claim in this repository. Evidence: `epsilon-universality/` *(three pre-registered experiment files) plus an adversarial prior-art gate. Supersedes the framing in* `EPSILON-UNIFICATION.md`*,* `FINDINGS-and-patentability.md` *§F1–F2, the two ε papers, and the various roadmap/decision logs.*

Policy, same as the ontogenic-mathematics correction: prior art is not a defect — a true claim someone else reached first is still true, and attribution is the point. False or overclaimed statements are a different matter and are corrected.

What is wrong

1. $\varepsilon(\sigma) = \Phi(-\text{gap}/(\sigma\sqrt{2}))$ is an identity, not a law — and it is 99 years old

For two independent Gaussian scores with equal variance, the difference is $N(-\text{gap}, 2\sigma^2)$, so P(wrong wins) $= \Phi(-\text{gap}/(\sigma\sqrt{2}))$ exactly, by algebra.

This is Thurstone's Law of Comparative Judgment, Case V (1927), *Psychological Review* 34:273–286. It is independently canonical in two other fields: the two-alternative-forced-choice error rate in signal detection theory (Tanner & Swets 1954; Green & Swets 1966), and the BPSK bit-error rate in additive white Gaussian noise in digital communications.

The reported $R^2 = 0.997$ was measuring the Monte-Carlo error of a Gaussian random number generator, not the quality of a fit. Verified: `exp_u` U1, deviation 3×10^{-4} across a gap×σ sweep.

2. It is not a universality class, and the three-domain agreement is the CLT

Universality requires robustness to microscopic detail — that is what the word means. Change the noise distribution and the law changes form:

Score noise	ε closed form	Name
Gaussian	$\Phi(-\text{gap}/(\sigma\sqrt{2}))$	Thurstone Case V / probit / 2AFC
Gumbel	$1/(1+\exp(\text{gap}/b))$	Bradley–Terry / multinomial logit
Cauchy	$\frac{1}{2} - \arctan(\text{gap}/2\gamma)/\pi$	robust choice model

Under Cauchy noise the Gaussian law under-predicts ε by 11× at gap/σ = 4 and 723× at gap/σ = 6 (`exp_u` U3, `exp_w` W1). That is a different fixed point, not an irrelevant operator.

The general statement subsuming all of it is ε = F_Δ(−gap) for difference-CDF F_Δ, and the distribution-free treatment already exists as the Mammen–Tsybakov margin condition. The honest name for our "class" is the Gaussian domain of attraction — the most populated and least surprising universality class in probability. The three domains agree because scores everywhere are sums of many small contributions.

3. The class-IL row is a published theorem with a proof

`acc ≈ ceiling·(1−ε)` is Kim, Xiao, Konishi, Ke & Liu, NeurIPS 2022 (arXiv:2211.02633): CIL = WP × TP, with necessity and sufficiency proved. Our $R^2 = 0.989$ was a regression onto an identity our own notes already called "exact." Inter-task confusion as the dominant class-IL failure mode is standard survey terminology (Masana et al. 2020).

4. The analog row is row 1 linearized

"ε ≈ 0.66·noise, $R^2 = 0.87$" is the probit curve Taylor-expanded near the origin; 0.66 is a Taylor coefficient. The R^2 dropping from 0.997 to 0.87 is the signature of forcing a straight line through a probit curve — two fits of one identity is a misspecification, not a generalization.

5. Stronger prior work exists on the Best-of-N side

- Gao, Schulman & Hilton (ICML 2023), arXiv:2210.10760 — *quantitatively fitted* scaling laws R_bon(d) = d(α−βd), KL_bon = log n − (n−1)/n. These predict where the peak is; ε only predicts that there is one.
- Khalaf et al. 2025, arXiv:2506.19248 — the distribution-free version via a rank/quantile transform, proving the interior optimum is unique. This is the genuinely universal route, and it is the one we did not take.
- The phenomenology is Goodhart's regressional/extremal variants (Manheim & Garrabrant 2018) and the optimizer's curse (Smith & Winkler 2006).

What is NOT wrong — do not "fix" these

ε* = p/(1−p) is CORRECT. An earlier draft of this notice said otherwise; that was retracted.

A prior-art gate objected that p/(1−p) is an odds ratio while ε is a probability, making the condition vacuous for p > ½. That objection is false, and so was the correction built on it. In `verifier-bottleneck/FINDINGS.md` F3b, p is the base rate of correct answers among sampled candidates (p =

0.10) — not a Condorcet voter competence — so $p/(1-p) = 0.111$ is an ordinary probability, and the published empirical crossover (0.11–0.15) brackets it.

Derivation: correct votes $\approx pN$, shared-exploit votes $\approx (1-p)qN$, so the exploit takes the plurality iff $q > p/(1-p)$. Verified exact for $p \in \{0.05, 0.10, 0.20, 0.30\}$, max deviation 0.009, reproducing F3b at crossover 0.115 (`exp_v`, 4/4).

The real defect is notational. ε denotes the conditional exploit rate (among wrong answers) in this formula, threshold $p/(1-p)$; but the unconditional rate elsewhere in the same documents, threshold p. Both are correct for their own reading. Define which ε you mean at each use. Do not change the formula.

`verifier-bottleneck/FINDINGS.md` already flagged this crossover as "NOT novel as a theorem — List & Goodin 2001" and advised "frame, don't claim." That was right and stands.

Also not wrong

- The measured ε values. $\varepsilon = 0.38$ for a naive class-IL head, the real-RM calibration $\varepsilon \approx 0.002$–$0.05$, the AIMC routing numbers — data is not prior art. The values are ours; the functional form is Thurstone's.
- The shipped artifacts. `verifier-budget` CLI and `adaptive_halt.py` work, and their value is calibration know-how, which is unaffected by the form being classical.
- The engineering guidance. Select by consensus rather than argmax; watch N; widen the gap or shrink σ to lower ε. All sound *within the Gaussian regime*.

What survives, and how to say it

Ship as: *"Classical selection theory — Thurstone's Law of Comparative Judgment / signal detection theory — correctly applied to three modern systems, with calibrated constants and a working budget tool."* That is defensible, useful, and will survive review.

Do not ship as: a new cross-domain law, a unification, or a universality class. It dies on the first reviewer who has taken a psychophysics or digital communications course.

One genuinely useful operational rule, which does survive

Never extrapolate a Φ fit down to small ε without checking the tail first. The law is uniformly optimistic under heavy tails and catastrophically so exactly where ε looks negligible — which is where safety arguments live. A reported $\varepsilon = 10^{-5}$ can conceal a true 8×10^{-3}. One statistic detects the regime: excess kurtosis, -0.07 (Gaussian) vs $+26.8$ (Student-t, df=3) at $n = 20{,}000$ (`exp_w` W3). Estimate ε directly; do not fit Φ and extrapolate.

Retracted from our own analysis

The claim that adversarial search *manufactures* the heavy tails is false (`exp_w` W2). Argmax selection leaves excess kurtosis at 0.03 — Gaussian — and instead produces a mean shift of 1.25, which is the optimizer's curse. Selection biases; it does not fatten tails.

The transferable diagnostic

> An $R^2 \approx 0.99+$ that holds across unrelated domains should trigger "I have derived an identity," not "I have found a law."

Real empirical laws in messy domains do not fit that well. A fit that good, that portable, almost always means the functional form was implied by a distributional assumption shared by all the simulators. This is the single most useful thing to come out of the ε campaign, and it generalizes far beyond it — see `NOVELTY-GATE.md`.

The nameable pattern across this whole repository: every "unification" claimed here turned out to be a shared functional form arising from a shared *distributional assumption*, not a shared mechanism.

Published by NeuroGenesys (AW3 Technology, Inc.). Canonical version: https://neurogenesys.si/research/epsilon-thurstone-correction