Could the instrument have shown it? Power and saturation checks for pre-registered predictions

W. Schulz

Jul 30, 2026

Abstract

Seven pre-registered misses in one day shared a cause: the registered statistic could not have shown the effect — floors, noise, saturation, collinearity. The note adds a five-line power step to pre-registration, and argues that independent parallel attacks act as decorrelated verifiers.

2026-07-30. A third methodology doc, alongside NOVELTY-GATE.md (is the claim new?) and SOUNDNESS-AXIS.md (is the claim true?). This one asks a question neither of those can: could the instrument have shown it?

Written from a cross-track read of the 2026-07-29 runs. The program's machinery for checking *claims* is in good order — pre-registration, run-once, hostile gates, visible retraction. The failures that remain have moved down a level, into the apparatus: unpowered statistics, a single instrument mistaken for sufficient, an internal signal shipped into the one regime that defeats internal signals, and one word covering two different quantities.

That is a predictable place for a program to end up once it gets good at killing claims. It is also the cheaper of the two problems to fix.

---

1. The pre-registration protocol has no power step

The evidence

On 2026-07-29 alone, across four independent tracks, seven pre-registered predictions missed. The reason each missed is the same, and it is not substantive:

Miss	Registered	Why it could not have passed
cosmogony C3	finite-cost spread ≥ 3 bits	the 158-state floor (4.70 bits) compresses achievable spread below it
cosmogony C3r	KL-space macro-stateability	finals drawn from the *typical* set, where the effect is undetectable by construction
cosmogony D5	spatial ρ(C, credits) beats ρ(C, stock)	collinear by construction — uniform mobility makes local credits ≈ 0.5 × local stock

Miss	Registered	Why it could not have passed
cosmogony V1	CTRL at equilibrium floor by T=4000	slowest composition mode relaxes at $\tau \approx 13{,}000$ steps
language P2	ε spread across codes ≤ 0.005	Monte-Carlo $\sigma = 0.00215$, so expected max spread $\approx 3.5\sigma = 0.0075$ — the threshold was tighter than the instrument's own noise
language L1a	node(k=128) $\geq$ node(k=16) + 0.05	task reaches 0.996 by k=16; the whole window sits inside the plateau
verifier exp 13 R3/R4/R5	pool-*density* effects	at tight k, pool-*size* effects dominate the registered statistic

Sources: `cosmogony/FINDINGS.md`, `language/FINDINGS.md`, `verifier-bottleneck/FINDINGS.md`.

Every one is the registered statistic being incapable of showing the effect — floor arithmetic, sampling noise, saturation, degrees of freedom, conditioning. Not one is the hypothesis being wrong.

Why this is structural, not housekeeping

The cosmogony log notices the repetition *within its own strand* ("method lesson #1, again"; "third design lesson of the strand logged"). Nothing aggregates across tracks, so seven instances of one structural gap read as four unrelated local slips.

The consequence is a second, mechanical source of over-deflation. `SOUNDNESS-AXIS.md` corrected over-deflation at the *interpretive* layer — "not novel" was standing in as a verdict. This is the same bias arriving by a different route and it is still uncorrected: a miss caused by an unreachable threshold is scored as evidence against the hypothesis when it is no evidence at all.

Pre-registration was adopted to stop goalposts moving after the fact. It has quietly acquired the opposite failure — goalposts placed, in good faith, outside the field.

The forbidding version — and its independent confirmation, which corrected me

> Re-score every logged MISS by a single question: *under the hypothesis, could this statistic have produced a PASS?* A material fraction will resolve to void rather than negative.

This was run, independently, by a parallel session on the same day (`grounding-logic/threshold_audit.py` → `RESULTS-threshold-audit.txt`), and its verdict on that session's near-misses is "NONE were substantive." It also improves the diagnosis in two ways this document originally got wrong.

(a) It is two defects, not one. I lumped them; the audit separates them, and the distinction matters because the remedies differ:

Defect	Meaning	Instances
MALFORMED	the threshold was unreachable *by construction*	P3 (threshold placed at the null value, 0.50 — fails ~50% of the time when the null is exactly true); T1 (0.900 sits above the reference instrument's own measured ceiling of 0.863); E4 (0.160 sits below the ~0.130 null band-width of a range statistic over 38 noisy points)
UNDERPOWERED	observed-vs-threshold shortfall sits inside the noise	B3 (1.6 SE), and T1 (0.5 SE) and E4 (0.6 SE) *also* — so those two carry both defects

(b) Void is *worse* than negative, not better. My original phrasing — that the scorecards "understate their hypotheses" — was wrong in spirit, and the audit says so plainly: *"This does NOT rescue the claims — three of these experiments measured nothing, which is worse than a clean refutation, and the reformulated versions are post-hoc and must be re-registered before they count."* A void miss does not return the hypothesis to open; it burns the run and leaves the question untested. The over-deflation consequence in §1 stands — a void is not evidence against — but the correct feeling is a *wasted experiment*, not a *rescued claim*.

The audit's own rule, which supersedes my five-item list where they overlap: before registering a threshold, compute (1) the null SE of the exact statistic, (2) the ceiling of the instrument on ideal data, (3) for range/band statistics, the expected spread under a flat truth. And never register a one-sided threshold *at* a null value — register the two-sided equivalence band that was actually meant.

A third convergent arrival: `NOVELTY-GATE.md` was amended the same day with the floor/ceiling habit, derived independently from breath B1/B6 (a "drops ≥ 0.20" test on a statistic starting at 0.052) plus cosmogony C3. Three independent derivations of one gap in one day is itself the strongest evidence that the gap is structural rather than four local slips.

Falsifier (now largely discharged): the re-score finds $\leq$1–2 void misses, i.e. the seven above are coincidence rather than a pattern.

The fix already exists in this repo, unlabelled

Exp 14 registered a per-query regime detector and scored 2/4 — R1 and R2 both failed because the statistic was too noisy per query (`RESULTS-14-regime-guard.txt`). Exp 15 changed *nothing about the hypothesis*. It pooled the same statistic over T queries, obtained $\sqrt{T}$ separation, and scored 4/4 — FA 0.000, miss 0.010 at T=50 (`RESULTS-15-deployment-monitor.txt`).

That is the missing protocol step, discovered once in one track and never generalized.

Protocol amendment (proposed)

Add to the pre-registration header, before running, one line per prediction:

1. Floor/ceiling. What is the achievable range of this statistic given the event definition, state count, or discretization? (cosmogony C3, C3r)
2. Noise floor. What is the sampling σ at this N? The threshold must exceed it with margin. (language P2)
3. Saturation. Is the comparison window inside a plateau or a ceiling? (language L1a)
4. Degrees of freedom. Are the two quantities being contrasted independent, or collinear by construction? (cosmogony D5)
5. Horizon. Does the run length exceed the slowest relevant relaxation time? (cosmogony V1)

If a prediction fails any of these, it is not registerable as stated — pool, re-scale, or re-site it first, per exp 15.

---

2. Parallel-session collisions are decorrelated replication, not a process defect

The evidence

Twice, parallel sessions independently attacked the same question:

- R1e: Round 3 (two-field engine, by theory) and Round 4/D′ (two-species instrument, by measurement), each unaware of the other.
- C3: closed by two parallel runs (collision #2).

The logged prescription is: *"serialize strand work, or at minimum check `git log` before opening a round."*

What actually happened contradicts the prescription. Round 2's §B surrogate instrument failed validation — vacuum-growth regimes read negative C_str because the shuffle null flattened the density profile, destroying the nuisance statistic along with the signal. The parallel session's two-species instrument had an exact null (the 256-particle composition string is conserved in every regime), and it validated. The two rounds did not duplicate; they partitioned: dissipative structure tracks flux at +0.235, relic structure at −0.922, and the gap's *stock* turned out causal for neither. Neither session finds that partition alone.

In the first collision, having a second independently-built instrument is precisely what revealed that the first one was invalid.

Why the prescription is backwards

Set that against the program's own deepest result:

Finding	Source	Says
same-context critic 12/12 miss vs fresh 0/12	`correlated-critique/FINDINGS.md`	a critic sharing the generator's context shares its standard of correctness
correlated prefilter (0.051) < no prefilter (0.077)	verifier exp 13	a checker correlated with the checked has negative value, not merely zero
displacement	`ontogenic-mathematics/THEOREM-displacement.md`	the checking instrument cannot be absorbed into the thing checked

Two sessions with no shared context, building different instruments against the same question, are maximally decorrelated verifiers — the exact thing the program has proven is the only reliable epistemic move, and the thing it otherwise cannot buy.

Serializing enforces a *single* instrument and a *same-context* check. The cost the log actually complains about is duplicated compute — which this program's own results price as cheap relative to a claim resting on one unvalidated instrument.

> Amendment: for the highest-stakes question in a strand, *deliberately* commission two independent attacks with different instruments and no shared context. Treat agreement as replication, disagreement as an instrument-validation signal, and partition as the best case. Serialize the *bookkeeping* (one canonical findings file), not the *attack*.

Falsifier: a collision where the two attacks agree on a wrong answer for a shared reason — i.e. the sessions were not actually decorrelated (shared priors from `CLAUDE.md`/memory would be the route). This has not yet been observed but is the failure mode to watch, and it is the reason the two attacks must not share a pre-registration.

Bound: this is a STATIC-regime prescription only

`grounding-logic/REVELATIONS-two-displacements.md` (2026-07-30, exp 16, 7/8) splits displacement into two mechanisms that merely resemble each other, and the split bounds everything above:

	Correlation-displacement	Diagonal-displacement
Regime	static — exploits are *sampled*	adaptive — exploits are constructed against the checker's definition
Fix	decorrelation	decorrelation fails; only coverage/surplus survives

Measured (exp 16, B2): the same fully-decorrelated panel misses 0.032 static vs 0.559 adaptive.

So "commission two decorrelated attacks" buys what it claims *only where the failure being hunted is sampled rather than constructed*. Against an adversary optimizing against your checking criterion, two decorrelated instruments are not two chances — they are two blueprints. §1's power problem is squarely static (a threshold does not adapt to your registration), so the §1 amendment is unaffected. §3's danger zone is arguably adaptive, which strengthens rather than weakens it.

This is also collision #3, and it is the mechanism argument for §2 running live: a parallel session with a different instrument both *confirmed* the decorrelation prescription and *drew its boundary*, which no single session did.

---

3. **`adaptive_halt.py`** is a same-context critic, in the regime that defeats them

The tension

The shipped tool halts on an internal consistency residual. Two measurements bear on it:

- `language/FINDINGS.md` L1d: AUC(residual → error) = 0.589, and 82% of wrong instances are *fully* self-consistent (residual exactly 0).
- Exp 12 (`12_halting_verifier_eps.py`): under verifier blind spots, degradation is silent *in the strongest sense* — the compute profile improves while accuracy erodes.

The standing defense is correct as far as it goes and is stated in-file: the residual is a halt signal, not a correctness signal. Exp 50's result stands on its own terms (gated beats best-fixed depth on 10/10 seeds).

Where the defense has a hole

In the danger zone exp 12 identifies — imperfect solver × imperfect verifier — halting and correctness stop being separable. The tool halts *because* the checker is blind; every observable it exposes says it is doing better. And per exp 12, both imperfections are label-free-invisible: you cannot tell from deployment telemetry that you are in the zone.

So the tool is an internal check, applied by a system to itself, in the one regime this program has proven internal checks fail. The scoping defense holds *outside* the zone and is unfalsifiable *inside* it.

The patch, already written

Exp 15's monitor detects the regime from a decorrelated reference channel and reaches FA 0.000 / miss 0.010 at T=50. That is the missing ingredient — an external, non-same-context signal for "am I in the danger zone." Nothing currently wires it to the halting path.

Recommended: gate `adaptive_halt.py` on the exp-15 monitor, or at minimum ship the zone warning in its interface docs the way the pool-relativity caveat was shipped into `verifier_budget.py`.

Falsifier: the two-factor surface exp 12 proposes but never ran (train k checkpointed models to varying convergence, sweep ε per checkpoint) shows trap rate is negligible at all achievable model qualities — in which case the zone is theoretical and the caveat is enough.

---

4. "Routing quality" is two quantities, and ε is provably blind to one

The equivocation, caught late

The tempting synthesis — *L3 proves ε is blind to compositionality; XD6c says ~80% of achieved ε is placement slack; therefore our instrument is blind to our main lever* — is wrong, and worth recording as wrong because the setup for making it is live.

There are two distinct placement questions:

	Question	Does ε see it?
Geometry	where the key vectors sit in space	Yes — this is what XD6c measures; headroom above the achievable ETF floor *is* an ε quantity (`xd6c_gauss_floor.py`, 5/5, d*=16, headroom 0.80 at deployment dims)

	Question	Does ε see it?
Assignment	which meaning attaches to which point, at a fixed point-set	No — provably. All codes induce the same codeword set, verified at $\max\|\Delta\| = 0.000e{+}00$; a nearest-codeword decoder cannot see the difference (`language/FINDINGS.md` L3)

The live risk

`XD6c` concluded that routing quality, not capacity, is the lever — correctly. The repo now uses "routing quality" and "placement" as single terms covering both columns. They are not one quantity, and the instrument that certified the conclusion is provably blind to the second.

> Any routing improvement that works by re-assignment rather than re-placement is invisible to every ε measurement in this program. It reads as a null.

Given the program has just concluded routing is *the* lever, this is a concrete way to discard a real gain as nothing. Fix: when a routing intervention is proposed, classify it as geometry or assignment *first*, and if assignment, measure it with an assignment-sensitive statistic (topographic similarity, transmission-to-a-second-learner — L3's own instruments) rather than ε.

Scope and attribution — what is NOT claimed

- §1's individual misses are all self-reported, in-file, by the runs that made them. What is added here is the aggregation across tracks, the diagnosis as one structural gap, the over-deflation consequence, and the protocol amendment. The exp-14→15 fix is likewise the repo's own; only its generalization is new. The MALFORMED/UNDERPOWERED decomposition and the "void is worse than negative" correction are **`threshold_audit.py`**'s, not this document's — this file originally had both wrong and is corrected in place rather than silently amended.
- §2 inverts a written prescription, and rests on three instances. Three is not a distribution. The mechanism argument (parallel sessions are decorrelated verifiers) is stronger than the sample, but this is a *serious argument*, not a settled result, in `SOUNDNESS-AXIS.md`'s sense — and it is now explicitly bounded to the static regime by exp 16.
- §3 asserts no error in exp 50. Its result stands as run. The claim is narrower: a shipped artifact carries an unstated risk in a named, label-invisible regime, and the mitigation exists unconnected.
- §4 is a definitional correction, not a finding. Its whole value is preventing a null from being believed.
- Novelty axis: not applicable. None of this is filable or publishable; §1 is close to standard power analysis, which is exactly the point — the gap is that a program this careful about claim-level rigor never imported it. Per Check 2, statistical power is far older than the literature it would be published in.
- Soundness axis: §1 and §4 are checkable against the existing record today. §2 and §3 are arguments with named falsifiers.

The through-line

Claim-level rigor is solved here. Instrument-level rigor is not, and the four failures above are one failure wearing four costumes: the apparatus was never asked whether it could answer the question.

`NOVELTY-GATE.md` asks whether an idea is worth investigating. `SOUNDNESS-AXIS.md` asks what a verdict means once you have one. This gate goes between them, and it is the cheapest of the three: five lines of arithmetic before a run, one deliberate duplicate attack per strand, and a habit of asking which quantity a word denotes.

Published by NeuroGenesys (AW3 Technology, Inc.). Canonical version: https://neurogenesys.si/research/instrument-gate