

ER 13 Corrective Review Record

Draft v0.6.1 | 10 October 2026 | NOT FINAL / NOT LOCKED

This record explains the changes from v0.6, the disposition of Claude’s and Google’s reviews, and the limits of the companion pilot. It supports checking the revision without treating AI review or integrity checks as scientific validation.

1 What changed

The two major gaps were accepted: Control now receives corrected exposure outputs, and early final submission receives a terminal binary outcome. The timing interpretation is tied to that task policy. The revision also makes target/benign item variances and exposure-by-family effects explicit, adds a routine sensitivity estimate without overclaiming population inference, clarifies S23’s denominator, adds S22’s DOI and the less-nudged S31 case, and strengthens the verifier.

The main manuscript retains all eight canonical questions and eight author-approved positions. v0.5 and the complete v0.6 package are preserved byte-for-byte. Structural changes are recorded by original JSON element, including before/after content. Appendix A is historical; Appendix B is the current disposition.

2 Claude finding dispositions

Finding	Disposition	Location	Action
N1	ACCEPT	13.2 and 14.7	Corrected-output Control selected; task and no-feedback policy identical; conspicuousness cannot be isolated.
N2	ACCEPT WITH CLARIFICATION	13.3 and 14.2	Terminal early misses are zero; no detection through tau under a terminal policy. Benign endpoints and known events are treated consistently.
n1a	ACCEPT	14.2	Separate target/benign item covariance matrices and marginalization. This is an explicit modeling choice, not proof of validity.
n1b	ACCEPT	14.2	Exposure-by-stratum terms included within item types; reductions require preregistered justification.
n2	PARTLY ACCEPT	14.2	Routine sensitivity estimate added. Population inference still needs sampling assumptions and validation; bootstrap is not exact randomization inference.
n3	ACCEPT	13.1 S31	Rare less-nudged helpfulness case added, retaining constructed-evaluation boundary.

Finding	Disposition	Location	Action
n4	ACCEPT AS CLARIFICATION	5.4 and 5.6 S23	Explicit 38/560 denominator and 19/280 per condition. Prior wording omitted rather than explicitly assigned a denominator.
n5	ACCEPT	10 S22	RSS DOI added to the register.
n6	ACCEPT	Verification companion	Baseline PDF is the source for frozen text; arithmetic illustrations labeled; pymupdf import; critical manuscript assertions added.
n7	DISPUTED METADATA	10 S27	Retain 133-140 supported by PubMed. Claude reports Crossref 133-139; direct Crossref access failed this round. Resolve against publisher scan before finalization.
Prior Minor 3	OPEN PROVENANCE	1.1	v0.3 detailed history remains unrecovered; no invented reconstruction.
Prior Minor 4	OPEN PROVENANCE	13.4	Original threshold/script proposer remains unrecovered. No declaration of global irrecoverability.
Prior Minor 10	RESOLVED	5.2	Requested S19 chance formula supplied. Dataset-specific reanalysis was not part of that correction.
S01 access	ABSTRACT CONFIRMED	Review record	PubMed abstract confirms 239 workers and reported associations; full methods not audited.
Google verdict	REJECT OVERCLAIMS	18 and review record	No bulletproof or operational certification; numerical rating has no supplied validated rubric.
Google S19 pilot demand	NOT REQUIRED HERE	Pilot section 2	Current human detection task does not measure model wrong-answer agreement. Add a chance baseline only if such an endpoint is introduced.

3 Google review disposition

Accepted observations: v0.6 added relevant machine evidence, separated H-S and H-C, selected a marginal risk difference, required false-alarm reporting and improved timing analysis. Its explicit participant-unblinding limitation is retained. These improvements do not erase the two omissions identified by Claude.

Rejected conclusions: bulletproof, unassailable, no remaining scientific vulnerability, maximum operational maturity and peer-review-grade compliance are not established. The 9.75/10 rating has no supplied validated

rubric. The supplied Google text uses references such as S72 and S96 that do not map to the crosswalk’s S01-S34 register; their underlying source map was not supplied, so those labels are not imported as verified citations.

Publication precision: S30 has an ICLR proceedings record; the cited S31 is arXiv v2 and is retained as a preprint. Evaluation awareness in S32 is performance on transcript classification, not mathematical certainty about an actual execution environment. S33 concerns intrinsic correction of tested reasoning tasks, and S34 concerns measured self-preference; neither proves a universal external-human-check requirement.

Model precision: Order is a position variable rather than a direct fatigue measurement. A planned crossed mixed model has not been implemented or validated. Paired hit/false-alarm reporting constrains interpretation but cannot prevent every misleading claim. Wolfe and Berbaum provide adjacent evidence, not direct proof of the proposed AI-output carryover effect. Google’s S19 chance-baseline request is conditional on adding a model-agreement endpoint, which this pilot does not do.

Next-step disposition: pilot drafting accepted after the definition fixes. A dynamic challenge generator is deferred because the proposed first pilot uses a fixed bank. Ordinary file hashes will document the final material version; no unbuilt seed-commitment logic is claimed as delivered.

4 Source checks completed for this correction

Source	Check and result	Limit
S01	PubMed PMID 17617239 abstract: 239 workers; safety-production compatibility and safe behavior relationships confirmed.	Abstract-level confirmation; full methods and causal identification not audited.
S14	Issue number 3 added to existing DOI record.	Core source reversal was corrected in v0.6; no new universal inference.
S22	Official RSS proceedings lists DOI 10.15607/RSS.2026.XXII.181.	Verification feedback remains foundation-model dependent.
S23	Full text section 4.2 and Table 2: 38/560; 19/280 in each time condition.	Denominator clarification; domain-specific outcome.
S27	PubMed reports 25(2):133-140. Claude reports Crossref 133-139.	Crossref endpoint unavailable; retain current range and log discrepancy.
S31	arXiv v2 abstract reports rare less-nudged scheming and helpfulness-related strategic underperformance.	Constructed setting, not prevalence in natural deployment.
S32	arXiv v3 abstract reports AUC 0.83 versus simple human baseline 0.92.	Classification benchmark, not guaranteed awareness.
S33	ICLR proceedings describes difficulty with intrinsic reasoning correction.	No universal impossibility result.
Resampling	Owen’s pigeonhole bootstrap supports crossed resampling under assumptions.	Does not validate this proposed pilot or give representativeness to a purposive bank.

Primary access routes checked 10 October 2026: <https://pubmed.ncbi.nlm.nih.gov/17617239/> ; <https://www.roboticsproceedings.org/rss22/p181.html> ; <https://arxiv.org/html/2603.11821v2> ; <https://pubmed.ncbi.nlm.nih.gov/2312249/> ; <https://arxiv.org/abs/2412.04984> ; <https://arxiv.org/abs/2505.23836> ; https://proceedings.iclr.cc/paper_files/paper/2024/hash/8b4add8b0aa8749d80a34ca5d941c355-Abstract-Conference.html ; <https://arxiv.org/abs/0712.1111> .

This is a targeted audit, not a new full review of all 34 sources. Previous source-check records remain in the preserved v0.6 package. No source metadata disagreement is resolved merely by agreeing with another AI reviewer.

5 Verification meaning

The new verifier extracts canonical-question and author-position spans directly from the preserved v0.5 PDF, compares them with the v0.6.1 JSON and DOCX/PDF text, checks baseline hashes and manuscript policy clauses, recalculates the independent-observation planning table, tests the scoring reference on explicit software fixtures, and checks the file manifest. It imports pymupdf directly. Arithmetic illustrations are labeled as such and are not evidence about human behavior.

The verifier cannot establish source truth, ethical approval, adequate sample size, correct causal attribution, participant behavior, mixed-model identifiability or interval coverage. It does not execute a power simulation. Its actual results and the visual QA record are separate files, written only after the checks and inspection finish.

6 Remaining gates

The original threshold/script attribution and exact v0.3 change history remain unrecovered. The supplied records do not justify declaring them globally irrecoverable. S27's page range remains a metadata discrepancy. Finalization should resolve these items or record an explicit author disposition with the search scope and residual uncertainty.

The pilot has proposed settings and blank data schemas, but no validated material bank, live interface, allocation sequence, participant data, fitted model, power curve, named research staff, ethics determination or independent statistical approval. Review the protocol before recruiting. The proposed 60-person allocation is for feasibility and may be changed only in a new pre-enrollment version. Independent non-AI review remains desirable before final promotion; no such review is claimed.

Research circulation: ready as a labeled corrective draft. Pilot protocol drafting: completed at Draft v0.1 for review. Recruitment: not ready. Final Archive promotion: not performed and not claimed ready. No Archive, Mirror Engine, Master Hash Manifest, GitHub, Arweave, website or preservation-network modification is part of this package.