

Coordinated Revision and Review Record

Package v1.0 | 10 October 2026 | DRAFT / NOT FINAL / NOT LOCKED

This package combines Crosswalk v0.6.2, SEC-CDP-001 Pilot v0.2 and SEC-CDP-002 Infrastructure Requirements v0.1.1. It applies the latest review findings while retaining the distinction between an evidence map, a proposed human study and a proposed engineering extension. The original v0.6.1 and SEC-CDP-002 v0.1 packages are preserved byte-for-byte.

1 Corrections and their limits

Finding	Disposition	Location	Action
MA-1	ACCEPT WITH QUALIFICATION	Crosswalk 14.1-14.2; Pilot 8-9	Both resampling and mixed models face item-selection limits. Main reporting targets the fixed bank; model superpopulation summaries are hypothetical and bank-referenced, not representative by fit.
MA-2	ACCEPT	Crosswalk 13.2 and 14.7; Pilot 2, 6, 8-9	Freeze 50% challenge prevalence; prespecify position bands 1-8, 9-24 and 25-48, balanced denominators, and descriptive contrasts. Add condition-by-position terms for exploratory feasibility only.
Minor 1	ACCEPT REVIEWER CORRECTION	Crosswalk S31 table	Separate helpfulness-driven underperformance from rare, less-nudged scheming. The earlier conflation is corrected.
Minor 2	ACCEPT WITH QUALIFICATION	Crosswalk 13.3 and 14.5; Pilot 7-9; scoring code	Retain known events for bounds and timing. Mask every interrupted record in the complete-item binary-model field. Exclusion does not cure informative loss; full-data point inference needs a validated missingness method.
Minor 3	ACCEPT	Crosswalk 13.3; Pilot 6-7; event schema	Retraction is an explicit logged action; any previously submitted valid target flag remains detection, and any benign flag remains a false alarm.
Minor 4	ACCEPT WITH SPECIFICATION	Crosswalk 14.2; Pilot 9	Terminal miss is a competing event. Report detection cumulative incidence and restricted mean time without detection. No hazard model is silently selected.

Finding	Disposition	Location	Action
Minor 5	ACCEPT WITH QUALIFICATION	Pilot 10	Report exposure error flags, benign false flags and technical delivery. Low flagging is weak observed engagement, not proof that a participant did not notice an error.
Minor 6	ACCEPT	Pilot 2	Unattributed fictional materials do not test reactions specific to AI attribution.
Minor 7	ACCEPT	Pilot 4	Fix identical compensation terms before enrollment, independent of speed, flags and accuracy; acknowledge residual incentive to finish early.
Minor 8	ACCEPT	Crosswalk 14.1; Pilot 8 and 10	Fix delta_star and justification before pilot unmasking; otherwise later confirmatory planners receive only arm-pooled nuisance summaries until the threshold is fixed.
Minor 9	ACCEPT	Pilot 10	Use point estimates for provisional feasibility decisions; show uncertainty separately and predefine missing or undefined metrics as not passing.
Minor 10	ACCEPT WITH BOUNDARY	Pilot 6-7; reference scorer and fixtures	Normalize deadline closure within a declared 1 ms logging tolerance; retain raw time, reject larger discrepancies, and do not extend the flag acceptance deadline.
S27 prior n7	WITHDRAWAL RECORDED	Crosswalk S27 and Appendix C	Keep 133-140 and note index disagreement and reviewer withdrawal; no unresolved substantive defect inferred from pagination alone.
Prior provenance 3 and 4	OPEN AUTHOR DISPOSITION	Review record	Search scope documented. Detailed v0.3 history and threshold origin not reconstructed or declared globally irrecoverable.
Google operational claims	CORRECTED	Crosswalk 12.4 and 13.6-13.8; SEC-CDP-002	Reject guaranteed unprofitability, compulsory compliance and invulnerable infrastructure. Address all three risk areas with requirements and residual risks.
Dynamic generator and DCE	INTEGRATED AS PROPOSED	SEC-CDP-002 v0.1.1	Define generation, independent ground truth, separate response streams, audit controls and acceptance evidence; no runtime implementation claimed.

2 Decisions requiring careful interpretation

Population inference: probability sampling from a defined item pool supports a design-based item-population target. Purposeful selection does not acquire that status from a stable fit. Model-based extrapolation is not prohibited, but its hypothetical target and transport assumptions must be explicit and separately labeled. Pilot primary reporting remains finite-bank.

Interrupted outcomes: raw observed detections and benign flags remain known facts. The binary-model analysis view masks all interrupted records symmetrically, while bounds and timing preserve known events. This removes one outcome-selection rule; it does not establish missing at random. A naive complete-item model is a sensitivity fit, not a validated full-data estimator. No numerical loss-bias estimate from Claude is adopted or claimed replicated.

Carryover: the overall estimand is unchanged. Bands 1-8, 9-24 and 25-48 are prespecified descriptive reports, not three new primary tests. Family and expertise weights remain common across arms. Six order lists must satisfy explicit item-type and family quotas and rotate item identity across bands; their actual construction and validation remain execution prerequisites.

Manipulation checks: the draft distinguishes whether material was displayed from whether an error was flagged. Low flagging is not proof of invisibility. Compensation is fixed independently of speed and accuracy, while the residual incentive to finish early is acknowledged. Feasibility decisions use registered point rules, with missing outcome bounds and unavailable metrics handled explicitly.

Timing: terminal submission is a competing event for detection under the task policy. Cumulative incidence and restricted mean time without detection are the summaries. No cause-specific or subdistribution hazard regression is selected in the pilot. The scorer accepts exact-deadline flags only under the assumed authoritative ordering and normalizes closure logs within 1 ms without extending the flag deadline.

3 Infrastructure integration

The infrastructure addition is now section 12.4 of the crosswalk, and sections 13.6-13.8 identify SEC-CDP-002 by its actual title and purpose. A hash commitment is no longer described as an interchangeable seed input. The proposed dynamic profile freezes realized blocks before exposure. The fixed-bank pilot does not silently adopt a dynamic generator or Differential Consensus Engine.

SEC-CDP-002 v0.1.1 changes the companion references and marks the crosswalk addition as integrated. Its 20 requirements remain NOT IMPLEMENTED / NOT TESTED. It addresses upstream compromise, known model-class violations, and institutional shortcuts through proposed controls, evidence and incident rules. It does not establish mathematical unprofitability of deception, compulsory compliance or invulnerable infrastructure. No firewall, HSM, network, production workflow or Archive service was changed.

4 Primary source checks and access limits

S31: the arXiv v2 abstract was reopened and supports separating helpfulness-driven underperformance from rare scheming without strong nudging. <https://arxiv.org/abs/2412.04984v2> . **S27:** the latest Claude review withdraws pagination as a defect because indexes disagree; 133-140 is retained with a note. No new publisher scan was obtained.

S01: the prior v0.6.1 record reports a successful PubMed abstract check. In this round direct PubMed retrieval yielded no abstract and the publisher returned 403. Primary-source indexed search extracts again supported the title, 239-worker sample and abstract context, but this is not a full-paper or independent human validation. That source-access distinction is preserved rather than relabeled as a new full audit.

<https://pubmed.ncbi.nlm.nih.gov/17617239/> ;

<https://www.sciencedirect.com/science/article/pii/S0022437507000588> .

M01: Austin, Lee and Fine (2016), Circulation 133(6):601-609, DOI 10.1161/CIRCULATIONAHA.115.017719. Primary-paper abstract/search extracts support the distinction among cumulative incidence, cause-specific hazards and subdistribution hazards. Full PMC retrieval hit an access challenge. This is an auxiliary method reference, not a new empirical source S35 or validation of the pilot.

Infrastructure sources R1-R10 and their access limits are carried in SEC-CDP-002 and SOURCE_RECORD.json. No exhaustive re-audit of S01-S34 was performed in this round. Google labels such as S99 and S117 were not accompanied by a source mapping, so they remain unverified labels and are not imported into the reference register.

5 Provenance and independent review

The available lineage checked for this revision comprises the preserved v0.6.1 package, its embedded v0.6 package and v0.5 PDF, their change records, the supplied Claude review files, and the current SEC-CDP-002 package. No new broad search of private chats, external Archive mirrors or unavailable v0.3 files is claimed. This scope does not establish the original threshold proposer or reconstruct the detailed v0.3 change list. Author disposition of these two unresolved provenance items remains open before final promotion; no author approval or global irrecoverability is invented.

Claude reports independently reproducing 91/91 checks in the previous package. That is a reviewer-reported result, preserved in the original review. The current verification report records only checks actually executed in this revision. Independent human methodological and engineering review remains pending. AI review and file integrity do not establish a substantive safety benefit.

6 What verification establishes

The verifier extracts all eight questions and eight author-approved positions directly from the preserved v0.5 PDF, compares manuscript and companion text across formats, replays structural edits against the previous JSON sources, checks baseline and package hashes, and exercises scoring software on explicitly synthetic fixtures. It checks proposed arithmetic and settings. Policy-presence checks are labeled as such. No participant behavior, model behavior, power curve, fitted inference pipeline or engineering attack test is supplied.

7 Remaining execution and promotion conditions

The eight human-pilot execution gates remain open, including investigator and ethics, recruitment and governance, validated bank, interface and order lists, allocation, calibrated analysis, independent statistical review and registration/approval. Infrastructure execution separately requires actual implementations, selected products/configurations, independent review and its required adverse tests. Final Archive promotion additionally requires unresolved provenance dispositions and an appropriate independent review record. Research circulation as labeled drafts does not authorize recruitment or deployment.