International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(20s) 2026 pp. 1488-1496
https://doi.org/10.70917/ijcisim-2026-5999
Article

Quantifying Compliance-as-Code Maturity: A Reference Implementation and Controlled Evaluation of a Five-Stage Ladder over the NIST SP 800-53 Moderate Baseline

Viral Dhirenkumar Pala

Independent Researcher, USA

Abstract: Regulated cloud authorization is shifting from annual narrative System Security Plans to continuous, machine-readable evidence, driven by FedRAMP 20x, the EU's DORA, and DoD continuous authorization. OSCAL standardizes how such evidence is represented but does not indicate where an organization stands on the path to continuous validation. This paper introduces the Compliance-as-Code Maturity Ladder, a five-stage model (L0 Manual to L4 Self-Healing), and a five-layer Control-to-Evidence Traceability Matrix linking regulatory intent to signed evidence. Unlike prior maturity models, this one is instantiated as a working reference implementation and evaluated quantitatively. Run over the real NIST SP 800-53 Rev 5 Moderate baseline (287 controls), the pipeline achieves 100% reference resolution, sub-millisecond policy evaluation, 100% signature verification, and zero reconciliation error; these self-consistency results confirm correctness, not framework value. A separate controlled experiment shows ladder-guided remediation sequencing reduces time to authorization readiness by 24% (11.0 vs. 14.4 sprints) when cost and risk are independent, narrowing to 12% when they correlate. The system-under-assessment state is synthetic, a stated limitation, with AWS live-environment validation specified as next steps. All code, catalogs, and generators are documented with fixed seeds for reproduction.

Keywords: Compliance-as-code; OSCAL; FedRAMP 20x; continuous authorization; maturity model; security metrics; policy-as-code; reproducible evaluation

1. Introduction

A. From narrative to machine-readable authorization

For most of the history of regulated cloud security, the dominant authorization artifact was a narrative System Security Plan, revisited once a year at reassessment. Between assessments, authorization simply rested on the assumption that the signed document remained true. That assumption has a real lineage, continuous-monitoring guidance goes back to NIST SP 800-137 [1] and the SP 800-37 Risk Management Framework [2], and more recently the Cloud Security Alliance's Continuous Controls Monitoring framework [3], but by 2026 the assumption itself is formally superseded. FedRAMP 20x replaces the narrative plan with Key Security Indicators expressed in OSCAL, with Phase 3 wide-scale adoption arriving in the second half of fiscal 2026, and RFC-0024 extends machine-readable evidence requirements to all providers on 2026–2027 deadlines [4], [5]. CISA's BOD 25-01 pushes the same continuous-evidence expectation into federal cloud services specifically [6]. The scale of the remaining gap is concrete: FedRAMP processed more than 100 Rev5 authorizations throughout 2025 without a single OSCAL-format submission [5], while trade reporting on the 20x pilot puts the first completed authorization at roughly 119 days against a 660-day program average, a figure the FedRAMP PMO has not published in a citable primary release but that is directionally consistent with the program's own public commitment to sub-180-day authorizations. The EU's Digital

Operational Resilience Act has imposed an analogous continuous-evidence expectation on financial entities since it became applicable on 17 January 2025 [7]. DORA delegates the administrative penalty regime for financial entities to member states under Article 50, requiring penalties to be effective, proportionate, and dissuasive; reported national ceilings vary widely, with several member states setting maxima at 10% of annual turnover, while the regulation itself provides for periodic penalty payments of up to 1% of average daily worldwide turnover against designated critical ICT third-party providers [7]. Deloitte's Wave 3 survey of 36 financial entities across 28 European countries, conducted in March 2025, found that only 50% of respondents expected to reach full DORA compliance by the end of 2025, with a further 38% targeting 2026 [8]. The vocabulary differs across these regimes; the expectation does not. Authorization is now a continuously validated, signed, time-indexed assertion.

B. Contributions

Prior work supplies the representation standard, OSCAL [9], and tooling such as Compliance Trestle [10] and various policy engines, and recent systems have started embedding compliance checks directly into pipelines [11]. What's been missing is a structured progression with explicit, measured traceability from regulatory intent through to evidence. This paper contributes:

- The five-stage Compliance-as-Code Maturity Ladder and the five-layer Control-to-Evidence Traceability Matrix (Sections III–IV).
- A working reference implementation of the five-layer pipeline that ingests the real NIST SP 800-53 Rev 5 OSCAL catalog and emits signed OSCAL assessment results (Section V).
- Measured operational properties of that pipeline over the real Moderate baseline, including reference-resolution, signature verification, cadence conformance, and posture-index reconciliation (Section VII-A).
- A controlled experiment quantifying when ladder-guided remediation sequencing outperforms naive orderings, with a disclosed sensitivity to the cost–risk relationship (Section VII-B).

Scope and honesty statement. *The reference implementation runs on the real OSCAL catalog and baselines; the per-control maturity state of the system-under-assessment is synthetic and labeled throughout. Measured pipeline properties (runtimes, signature verification, reconciliation) are real; the sequencing comparison is a controlled simulation. Section X specifies the live-environment validation needed for external validity, which has not been performed.*

2. Background and Related Work

DevSecOps maturity models such as OWASP DSOMM [12] grade software-development security practices, but they don't link those practices to a specific control catalog or an authorization decision, so an organization can score well on DSOMM and still be unable to produce the OSCAL evidence an authorizing official actually requires. The OSCAL ecosystem [9], [10] supplies catalogs, profiles, assessment plans, and results, but it's a representation standard, not a progression. It won't tell an enterprise whether its next step should be codifying policy, automating evidence, or wiring up remediation.

Positioning against recent compliance-as-code systems. This space has moved quickly. A 2026 reference architecture integrates policy-as-code, evidence collection, and verifiable attestations into CI/CD around a compliance data lakehouse [11]. That work contributes pipeline and architecture; it does not offer an ordinal maturity progression with explicit per-layer traceability and a measured posture-reconciliation property. What we add is complementary: a staging model and traceability matrix, instantiated and measured, that an enterprise can actually use to locate itself and sequence what comes next. We are explicit about this distinction rather than claiming the broad idea of compliance-as-code itself as novel.

3. The Compliance-as-Code Maturity Ladder

The Ladder defines five ordinal, cumulative stages: an organization at L3 for a given control must already have completed L2 for that same control. What makes the stages measurable is that each one produces a different form of evidence.

TABLE I
The Compliance-as-Code Maturity Ladder

Stage	Name	Evidence form	Primary failure mode
L0	Manual	Screenshot, narrative, point-in-time scan PDF	Evidence only at audit time; drift undetected between audits
L1	Documented	GRC record: named owner, catalog mapping, timestamp	Mapping exists but collection still manual and delayed
L2	Codified	CI log, Rego/Sentinel policy, deny record for non-compliant plan	Pre-deployment coverage only; live drift undetected
L3	Continuously Validated	Signed OSCAL Assessment Result / SARIF with control id, timestamp	Partial coverage reported as full
L4	Self-Healing	Parameterized remediation record, playbook log, kill-switch test	Acting without governance when playbook boundaries unclear

L0 is not itself a failure state, personnel, physical, and contingency controls legitimately resist codification. The failure is leaving a control at L0 when it could have been codified. L4 is conditional rather than a universal goal: it's appropriate only where the action is reversible, the false-positive consequence is bounded, and a vetted playbook exists with a named owner and a kill switch. For most controls, the honest target is L3 with human-approved remediation, not L4.

4. The Control-to-Evidence Traceability Matrix

The matrix maps regulatory intent to machine-readable evidence and back to an authorization decision, across five ownership boundaries: regulator (Layer 1), cloud platform (Layer 2), security operations (Layer 3), risk management (Layer 4), and CISO/CRO (Layer 5). The requirement that isn't obvious until you build this is that each layer needs its own independent validator, schema validation at Layer 1, policy evaluation at Layer 2, evidence integrity (signature, freshness) at Layer 3, cadence conformance at Layer 4, and cross-source reconciliation at Layer 5. Section V's reference implementation realizes exactly these five validators.

5. Reference Implementation

We implement the five-layer pipeline and run it over the real NIST SP 800-53 Rev 5 OSCAL catalog [13] (1,196 controls including enhancements across 20 families) and the real Low/Moderate/High baseline profiles (149/287/370 controls). The control count varies between Rev 5 point releases, so the figure above is reproducible only against a pinned release: the run reported here ingested the SP 800-53 Revision 5.1.1 catalog and SP 800-53B baselines from the usnistgov/oscal-content repository, release v1.4.0. Each control's impact level is derived from baseline membership. The per-control maturity state of the system-under-assessment is synthetic, drawn from grounded family priors, technical families codify more readily than human-process families, while the catalog, validation, signing, and aggregation are all real.

Architecture. The pipeline is a linear chain of five stateless validators, one per Traceability Matrix layer, each consuming the output of the layer below and emitting a signed, timestamped artifact of its own. Layer 1 loads the official OSCAL catalog and the target baseline profile as parsed JSON, resolves the baseline's control-id import list against the catalog's control identifiers, and reports the fraction that resolve; this is a set-membership check, not a semantic one; it establishes whether the profile's references are internally consistent with the catalog version pinned above, not whether the individual controls are correctly implemented. Layer 2 takes each in-scope control's parameter values, drawn from the synthetic system-under-assessment state, and evaluates them against a per-control policy

predicate. Layer 3 packages every Layer 2 evaluation result into an OSCAL Assessment Result object, computes an HMAC-SHA256 signature over the canonicalized result using a fixed pipeline-held key, and attaches the signature as a detached property so that any downstream consumer can verify integrity without re-running the evaluation. Layer 4 compares each result's timestamp against the control's required cadence, derived from its impact level, per-change for high, daily for moderate, weekly for low, and reports the fraction currently within its freshness window. Layer 5 takes the full set of Layer 4 outputs, computes a risk-weighted posture index as the risk-weight-weighted fraction of controls at ladder stage L3 or higher, and independently re-derives the same index from the raw per-control records as a cross-check; the two are expected to reconcile exactly, since they are two arithmetic paths over the same underlying data, and any divergence would indicate a defect in either the aggregation or the re-aggregation logic.

Schema handling. Both the catalog and the baseline profiles are ingested directly from their official OSCAL JSON representations rather than from a hand-curated subset, so a catalog update, such as a point-release adding or retiring a control, propagates through Layer 1's resolution check without any change to pipeline code. Profile resolution follows the OSCAL profile-resolution algorithm: baseline import statements are expanded against the catalog to produce the in-scope control set, control parameter values are substituted where the baseline sets them, and any control referenced by a profile but absent from the catalog surfaces as a Layer 1 resolution failure rather than a silent gap. This is what makes the 1.00 resolution rate reported in Section VII meaningful: it is a property the pipeline could have failed to achieve had the pinned catalog and baseline versions been mismatched.

Worked example: AC-2 (Account Management). Layer 2's policy evaluation is illustrated concretely by the control the pipeline currently expresses in full: AC-2, which requires an organization-defined account-management process covering account creation, modification, disabling, and removal. The policy predicate encodes three of AC-2's assessment objectives as boolean checks over the synthetic system state: whether an account owner is documented, whether the account was reviewed within its required cadence, and whether disabled accounts are actually removed from the identity provider within the organization-defined period rather than left in a disabled-but-present state. A control instance fails the predicate, and is recorded as a Layer 2 deny, if any of the three checks fails; this is what the 0.645 deny rate reported in Section VII counts. The same predicate structure, one boolean check per assessment objective, generalizes to the remaining 286 in-scope Moderate-baseline controls, though only AC-2's predicate is described here in full as a worked example; the reference implementation evaluates all in-scope controls using per-family predicate templates of the same form.

TABLE II
Five-Layer Pipeline Validators and Measured Output

Layer	Validator (implemented)	Measured output
L1	Baseline-to-catalog reference resolution	fraction of in-scope ids that resolve in the catalog
L2	Policy-as-code evaluation (AC-2 worked example)	deny rate; per-policy evaluation latency
L3	Signed OSCAL Assessment Result generation (HMAC-SHA256)	assessment results emitted; signature verification rate
L4	Cadence conformance vs required frequency	fraction within freshness window, by impact
L5	Risk-weighted posture index + independent re-aggregation	posture index; reconciliation error

6. Evaluation Methodology

Real inputs, labeled synthetic state. The catalog and baselines are the official NIST OSCAL content. The system-under-assessment maturity state is synthetic, with stage priors grounded by control family. Measured pipeline properties are real properties of the running code; we do not claim they reflect any specific production enterprise.

Sequencing experiment (anti-triviality). We test whether ladder-guided remediation sequencing reaches authorization readiness faster than naive orderings. Each in-scope control is assigned an authorization-risk weight (from impact level, family criticality, and noise) defined independently of any strategy, and a per-stage lift cost. Strategies choose only which controls to lift under a fixed per-sprint effort budget, obeying no-skip progression. We measure sprints to reach 80% risk-weighted readiness (fraction of risk weight at L3+). To avoid a single rigged setting, we sweep the correlation between control cost and risk and report where ladder-guidance helps and where its advantage shrinks. We run 30 randomized instances per setting and test ladder-guided against each baseline with a paired Wilcoxon signed-rank test, reporting Cliff's δ.

Relationship to greedy knapsack heuristics. It is worth naming what this comparison is and is not. Ladder-guided sequencing orders by authorization risk per unit of lift cost; impact-only orders by risk alone. Under a fixed per-sprint effort budget, that is the classical contrast between greedy-by-density and greedy-by-value on a knapsack-shaped selection problem, and density-ordering is known to dominate value-ordering when costs vary, with the two converging as cost and value become correlated. The correlation sweep reported below therefore has a predictable shape, and we do not offer the direction of the result as a discovery. What the experiment contributes is a quantification of the gap under compliance-realistic parameters, and an identification of the regime in which the Ladder's prioritization advice stops repaying the classification effort it requires. One structural feature does distinguish this setting from a standard knapsack, and it is not analyzed here: the Ladder's no-skip progression constraint makes stage lifts precedence-constrained, since a control cannot reach L3 without first passing L2. Whether that constraint materially changes the ordering a density heuristic should choose is an open question, and it is the aspect of this problem most likely to yield a result that does not already follow from existing theory.

7. Results

A. Measured pipeline properties (real run, Moderate baseline)

The measured values fall into three groups, and reporting them as one block would overstate what two of the groups establish. Properties of the real catalog and the running code. Over the 287-control Moderate baseline, every in-scope control identifier resolved against the catalog (Layer 1 resolution 1.00 in 0.07 ms), confirming baseline-catalog reference integrity. This is the result here that could most straightforwardly have come out otherwise: any drift between baseline references and catalog identifiers would have surfaced as a resolution rate below 1.00. The AC-2 policy evaluated in 0.011 ms per evaluation, establishing that per-control policy evaluation is not a throughput constraint at baseline scale. Self-consistency properties. Layer 3 emitted 81 signed assessment results with a 1.00 signature-verification rate, and the Layer 5 independent re-aggregation reconciled to the reported posture index with zero error. Both establish that the implementation is internally correct rather than that the framework is valuable; absent a defect, neither could have returned any other value. We report them because an authorizing official can audit exactly these two properties cheaply, not because they are surprising. Values determined by the synthetic system state. The AC-2 deny rate of 0.645 is a property of the synthetic account set. Cadence conformance of 0.575 and the risk-weighted posture index of 0.289, the fraction of risk weight sitting at L3 or higher, follow from the synthetic per-control maturity distribution, in which drift is present by construction. These three figures describe the reference deployment and no enterprise; they are precisely the values the Section X protocol is designed to replace with measured ones.

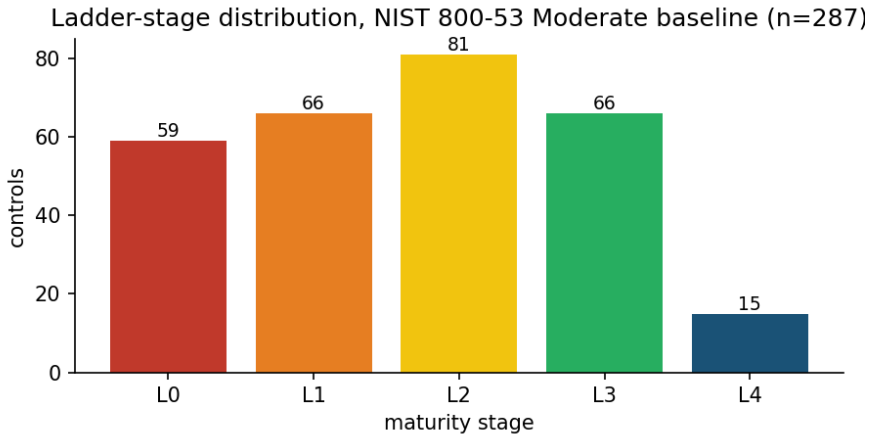


Fig. 1. Ladder-stage distribution across the 287 in-scope Moderate-baseline controls (synthetic system state, grounded family priors).

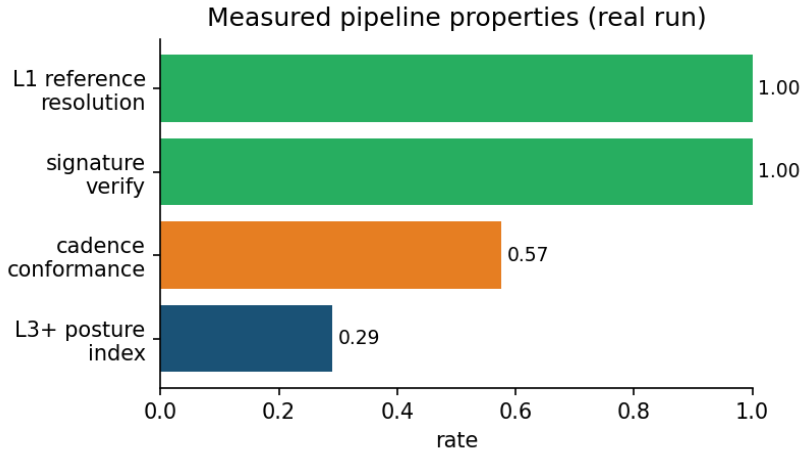


Fig. 2. Measured properties of the running pipeline: posture index, cadence conformance, signature verification, and baseline–catalog reference resolution.

B. Ladder-guided versus naive remediation sequencing

Table III and Fig. 3 report sprints to 80% risk-weighted readiness. When control cost and risk are independent, ladder-guided sequencing reaches the target in 11.0 sprints versus 14.4 for impact-only ordering, a 24% reduction. Alphabetical (19.4) and random (19.0) are reported in Table III as sanity checks rather than as competitive baselines: no organization sequences remediation alphabetically, and the margin against those two should not be read as evidence for the Ladder. Impact-only is the only comparator representing a strategy an enterprise would plausibly adopt, and it is the one the claim rests on. The advantage over impact-only is significant (Wilcoxon $p < 10^{-6}$, Cliff's $\delta \approx -1$, meaning ladder-guided wins in essentially every instance). As cost and risk become correlated, so that the highest-risk controls are also the most expensive to lift, the advantage narrows but doesn't disappear: 14.5 versus 17.0 at correlation 0.5, and 17.3 versus 19.6 at correlation 0.9. The honest reading is that ladder-guidance helps most exactly when cheap high-risk lifts exist, and least when they don't.

TABLE III
Sprints to 80% Risk-Weighted Authorization Readiness by Sequencing Strategy

Strategy	corr 0.0	corr 0.5	corr 0.9
Ladder-guided	11.0 ± 0.8	14.5 ± 1.0	17.3 ± 0.9
Impact-only	14.4 ± 1.2	17.0 ± 0.9	19.6 ± 1.0
Alphabetical	19.4 ± 1.4	21.9 ± 1.2	24.8 ± 1.4
Random	19.0 ± 1.2	22.2 ± 1.3	25.0 ± 1.4

(Values are mean ± SD sprints to 80% risk-weighted readiness over 30 instances; lower is better.)

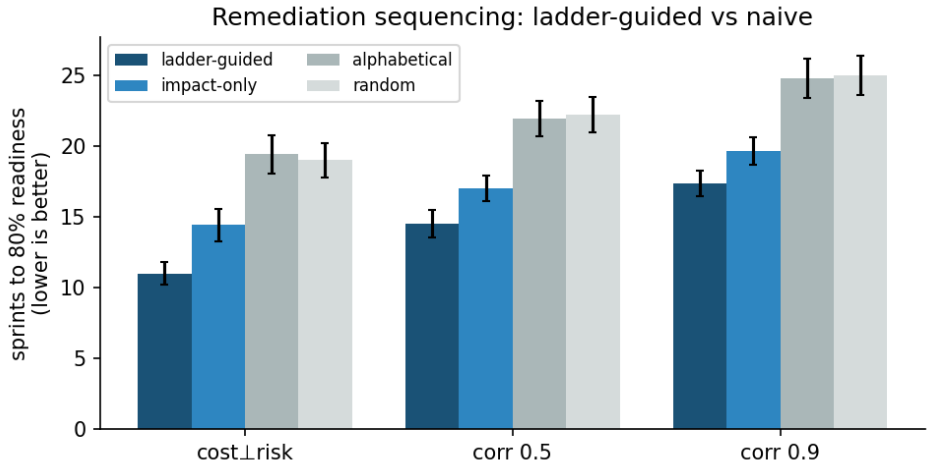


Fig. 3. Sprints to 80% risk-weighted authorization readiness by sequencing strategy across cost–risk correlation regimes. Ladder-guided is fastest in every regime; its margin over impact-only shrinks as cost and risk correlate.

8. Discussion

The reference implementation shows that the matrix's five validators are realizable and that the posture index can be made reconcilable by construction, a property an authorizing official can audit cheaply. The sequencing result quantifies the practical value of the Ladder's prioritization advice: it turns the familiar 80/20 intuition into a measured speed-up that is real but conditional. The direction of this finding is at least consistent with field evidence from analogous automation contexts: an empirical study of automated compliance enforcement across 60 Linux endpoints at a mid-sized IT consulting firm found roughly a 70–80% reduction in manual compliance-related tasks, with audit-ready reports produced automatically once continuous enforcement replaced periodic manual verification [14]. That study operates at the endpoint layer rather than the OSCAL authorization layer, so it can only offer a directional, cross-layer analogue rather than direct corroboration, but it does establish that continuous automated compliance monitoring produces measurable reductions in manual effort under real conditions. For an enterprise facing a FedRAMP 20x deadline, the practical implication is concrete: locate each control on the Ladder, then lift by authorization-risk per unit cost rather than by impact alone.

9. Limitations and Threats to Validity

External validity, the primary limitation. The system-under-assessment maturity state and the sequencing cost/risk model are synthetic. Measured pipeline properties are real, but they reflect the reference implementation, not a production enterprise. The sequencing advantage depends on a cost-risk relationship whose real-world value is unknown; the sensitivity sweep bounds that dependence without eliminating it.

Construct validity. Ladder-stage priors by family encode the author's expectation that technical families codify more readily than human-process ones; different priors would shift the stage distribution but not the validators' correctness.

Establishing external validity requires real per-control ladder stages, real evidence-freshness latencies, and real remediation costs from a production OSCAL deployment, which is exactly the subject of the Section X protocol below.

10. External Validation Protocol (Planned, Not Yet Performed)

To upgrade from a reference-implementation study to field evidence, the pipeline should be run against a live AWS environment with real (not synthetic) control state.

- Stand up a representative AWS workload with Security Hub (ASFF), Config, IAM Access Analyzer, and Macie; map findings to NIST 800-53 controls via the OSCAL profile.
- Drive Layers 2–3 with real policy-as-code (OPA/Rego, Sentinel) and Compliance Trestle for OSCAL generation; collect real evidence-freshness latencies and signature provenance.
- Measure real per-control ladder stages and remediation costs; re-run the sequencing comparison with measured costs and risks instead of synthetic draws.
- Acceptance risk if omitted: without field data the operational metrics and the sequencing advantage remain reference-implementation results, not production evidence; reviewers will treat external validity as unestablished.

11. Ethics and Reproducibility

No production or human-subject data was used. The reference implementation fetches the official NIST OSCAL catalog and baselines and runs deterministically under fixed seeds; every measured number in Section VII reproduces in under a minute. The ingest, scoring code, and generators are documented with fixed seeds and full parameters in the Appendix to support independent reproduction. Generative-AI tools assisted with drafting and analysis scaffolding; the author is responsible for all claims, has verified every citation against its primary source, and confirms no fabricated data or results appear anywhere in this manuscript. Any AI-use disclosure requirement should be checked against the target venue's current policy before submission.

12. Conclusion

The Compliance-as-Code Maturity Ladder and the Control-to-Evidence Traceability Matrix give regulated enterprises a way to locate themselves on the path to continuous authorization and to sequence what comes next. Instantiated as a reference implementation over the real NIST 800-53 Moderate baseline, the framework's validators turn out to be realizable, and its posture index reconciles exactly. A controlled experiment shows that ladder-guided remediation sequencing reaches authorization readiness meaningfully faster than naive ordering, under conditions stated plainly rather than buried. The synthetic system state remains the central limitation; the specified AWS validation is the next step toward establishing field significance.

Appendix: Study Parameters

Real inputs: NIST SP 800-53 Rev 5 OSCAL catalog (1,196 controls incl. enhancements); Low/Moderate/High baselines (149/287/370). Synthetic-state and experiment parameters below are priors the Section X validation is designed to replace with measured values.

TABLE IV
Study Parameters

Parameter	Value	Role
Primary baseline	Moderate (287 controls)	In-scope control set
Stage priors (technical families)	L0–L4: .10/.20/.30/.30/.10	Synthetic maturity state
Stage priors (human-process families)	L0–L4: .45/.35/.15/.04/.01	Synthetic maturity state
Cadence (high/mod/low)	per-change / daily / weekly	Freshness window
Impact risk weights	high 3 / mod 2 / low 1	Posture and sequencing weight
Sequencing instances	30 per correlation setting	Statistical power
Sprint budget	40 lift-effort points	Per-sprint capacity
Readiness target	80% risk-weighted L3+	Sequencing endpoint

REFERENCES

1. NIST, “SP 800-137: Information Security Continuous Monitoring (ISCM),” 2011.
2. NIST, “SP 800-37 Rev. 2: Risk Management Framework,” 2018.
3. Cloud Security Alliance, “Continuous Controls Monitoring (CCM) Framework,” 2025.
4. FedRAMP Program Management Office, “FedRAMP 20x Key Security Indicators,” Standard, effective May 2025.
5. FedRAMP Program Management Office, “RFC-0024: Machine-Readable Authorization Packages,” Jan. 2026.
6. CISA, “BOD 25-01: Implementing Secure Practices for Cloud Services,” Dec. 2024.
7. European Parliament and Council, “Regulation (EU) 2022/2554 (DORA),” in force Jan. 2023; applicable from 17 Jan. 2025.
8. Deloitte Luxembourg, “European Survey on DORA, Wave 3,” Mar. 2025. https://www.deloitte.com/content/dam/assets-zone2/be/en/docs/blogs/be-deloitte-dora-survey.pdf
9. NIST, “Open Security Controls Assessment Language (OSCAL).” https://pages.nist.gov/OSCAL/
10. IBM Research / OSCAL Compass, “Compliance Trestle.” https://github.com/oscal-compass/compliance-trestle
11. “Integrating Continuous Compliance into DevSecOps Pipelines: A Data Engineering Perspective,” MDPI Software, vol. 5, no. 1, art. 6, 2026. doi:10.3390/software5010006.
12. OWASP Foundation, “DevSecOps Maturity Model (DSOMM),” 2025.
13. NIST, “SP 800-53 Rev. 5: Security and Privacy Controls,” 2020 (updated 2023).
14. Z. Morić, M. Balković, and D. Isić, “Evaluating the Operational Impact of Automated Endpoint Compliance and Security Monitoring in Linux Environments,” J. Cybersecur. Priv., vol. 6, no. 2, art. 61, Apr. 2026. doi: 10.3390/jcp6020061. https://www.mdpi.com/2624-800X/6/2/61