

What Development Evidence Can and Cannot Show

AAG, EMALS and model-dependent engineering tests

Zawwar Sami

Independent Researcher, Systems Engineer, and Cybersecurity Specialist

ORCID: 0009-0004-5819-3017

W Series | W2 | Version 1.2 | 17 October 2026 | Independent preprint

DOI: [10.5281/zenodo.23435836](https://doi.org/10.5281/zenodo.23435836)

Abstract

A successful demonstration does not establish dependable service. A failed test does not establish a limit of physics. I examine this distinction through documented AAG aircraft-recovery and EMALS launch-system milestones. Four worked analyses show why operation counts alone cannot establish peak loads, reliability confidence, failure mechanisms or performance after configuration changes. The most instructive example concerns test design: a contrast that separates two explanations can leave others indistinguishable when interactions are allowed. Exact calculations also show how assumed count variability changes the repetitions needed for a specified decision. These are case-based analytical results and synthetic computations, not carrier measurements. The practical contribution is a traceable assessment sequence for choosing the next useful record, measurement or test.

Keywords: aircraft recovery; reliability identifiability; experimental design; Poisson model; overdispersion; engineering evidence

Key takeaways

- Preserve the tested subsystem, aircraft and configuration when interpreting a milestone.
- Successful energy absorption does not establish a bound on interface peak load.
- A contrast that identifies one model can remain ambiguous under a credible interaction model.
- At the synthetic means studied, overdispersion changes both classification error and replication requirements.
- Complete code, CSV outputs and the captured execution log are embedded in this PDF; Appendices A-D expose the calculations.

Plain-language summary

A machine can complete many operations while important questions remain open about how it behaves, what causes interruptions, or whether a change improved it. This paper follows actual development milestones and shows, with explicit examples, why those questions need different evidence. Its purpose is to help a reader choose the next observation that would distinguish the explanations under consideration. The numerical experiments use declared synthetic models; they do not substitute for measurements of the real equipment.

1. What has actually been established?

A machine can work before its service reliability has been demonstrated. A machine can fail a test without exhausting the possibilities of its underlying principle. Engineering judgment depends on holding both facts together.

The word possible needs an object. Possible to produce an effect once? Possible to remain within an interface limit? Possible to repeat the event under the intended conditions? Possible to recover from an interruption with the available people and equipment? Each proposition needs evidence. None becomes true merely because another has been established.

Aircraft recovery makes this distinction concrete. Arresting motion, controlling the load transmitted through an interface, qualifying different aircraft and sustaining repeated operation belong to the same system, but they are different assessment questions. Collapsing them into a single verdict loses the relationships that explain the engineering.

My question is therefore practical: when evidence is insufficient for a claim, what exactly is missing, and which next observation would change the conclusion? More repetitions can reduce sampling uncertainty. They cannot reconstruct a missing event definition, measure an unrecorded force transient, or distinguish mechanisms that predict the same observations under every condition tested.

W1 examined how components, infrastructure and human organisation combine into operational capability [1]. W2 follows the evidence through development: which parts of a capability have been demonstrated, which remain conditional, and which cannot be inferred from the chosen measurements. The analysis concerns the structure of the inference as much as the volume of data.

2. Case selection and source-to-claim audit

AAG is the central case because released records connect component redesign to subsequent testing. EMALS provides a comparison within the same carrier programme. Their different energy paths prevent treating their counts as interchangeable trials.

The 2016 Inspector General investigation documents redesign of the water twister, cable shock absorber and power conditioning equipment, alongside software and test difficulties [2, pp. 8-10]. GAO's 2013 investigation had identified risks from production before maturity and from overlapping integration with operational evaluation [8]. These are documented development problems. They do not establish a limit of physics or a present readiness ranking.

I transcribed four milestone claims from three dated NAVAIR accounts, retaining the source date, event date or cutoff, count qualifier and test object in case_claims.csv. The records were selected for their distinct evidential boundaries, not sampled to estimate the prevalence of inadequate reporting. Each row below is an empirical documentary claim; no row is an independently observed hardware trial.

Table 1. Documentary claim records and the inference boundary used in this paper.

Record	Source-reported achievement	Direct evidential object	Additional inference examined
C1: EMALS, September 2008 [5]	10,000 simulated cycles	Electrical/thermal equipment and energy-storage cycling	Integrated aircraft operation; mechanism separation
C2: AAG, December 2016 cutoff [4]	More than 1,400 dead-load arrestments	Tested load arrangement	Peak load and actual-aircraft compatibility
C3: AAG, same cutoff [4]	351 Super Hornet test arrestments	Tested aircraft and configuration	Air-wing qualification and evidence transfer
C4: shipboard, 25 June 2022 [6]	10,000 launches and recoveries reported	Extensive realised operation	Eligible failure exposure and confidence

The 2017 account also reports continuation of AAG after review of a Mk-7 alternative, and further qualification work [4]. The 2008 account identifies a subsequent full-power-train phase [5]. These source boundaries matter: the same numerical count can describe different engineering achievements. I do not infer that the underlying test programmes lacked the additional records merely because a public milestone account does not reproduce them.

3. Distinct achievements and measurement boundaries

NAVAIR describes AAG as an integrated arrangement of energy absorbers, power conditioning and digital controls, succeeding Mk-7 [3]. This gives a functional architecture for assessment without requiring inaccessible drawings or control laws.

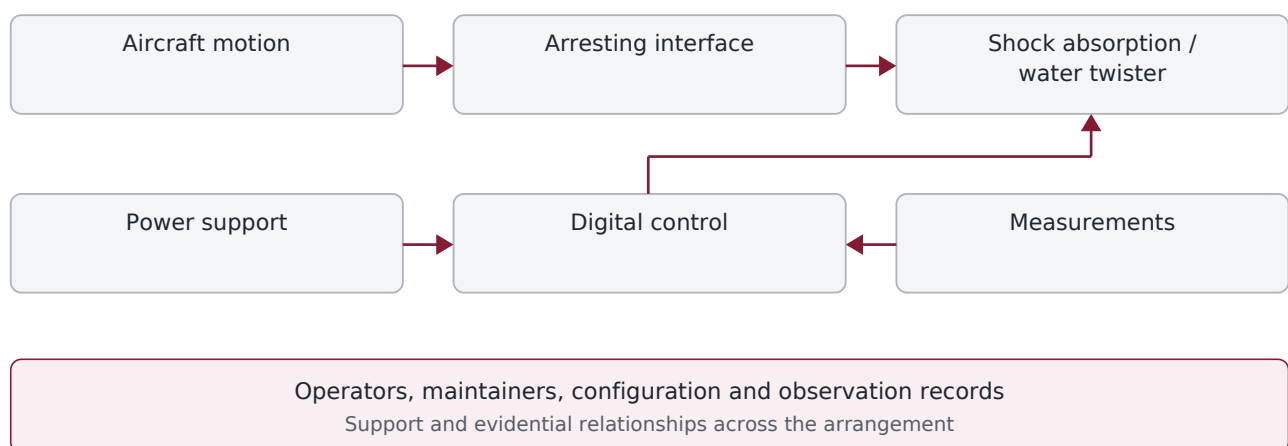


Figure 1. Functional reconstruction based on public descriptions [2,3]. Arrows indicate broad transfer and control relationships. This is an assessment diagram, not a hardware schematic or operating procedure.

The arresting interface couples motion into an absorption arrangement. Control, measurement and power support the intended behaviour. Operators and maintainers preserve the conditions under which it can be repeated. Configuration and observation records preserve what later analysis can legitimately compare.

The Inspector General identifies different roles for the fluid-based water twister, the initial-load shock absorber and the equipment supplying and conditioning electrical power [2, pp. 8-10]. One successful energy transfer does not independently qualify all three functions. The physical process and its supporting dependencies must be observed at the level relevant to the claim.

EMALS uses stored kinetic energy and solid-state electrical conversion, with computer control and monitoring [7]. Its comparison with AAG is useful because launch and recovery have distinct energy paths despite sharing shipboard support. It would be a mistake to combine their operation counts as interchangeable trials of one apparatus.

Architecture therefore changes the measurement question. Energy absorption needs energy and load observations. Repeated-use performance needs eligible exposure and events. Maintainability needs restoration records. A single favourable metric cannot stand in for all of them.

Table 2. Notation and abbreviations. Model values are illustrative unless identified as source-reported.

Symbol or term	Meaning	Units or interpretation
AAG / EMALS	Advanced Arresting Gear / Electromagnetic Aircraft Launch System	Documented engineering cases
HPP / NHPP	Homogeneous / non-homogeneous Poisson process	Constant / varying event intensity
$N; \theta$	Eligible exposure; mean interval between counted events	Cycles in the HPP illustration
$n_j; t_j$	Block cycle and elapsed exposures	Cycles; hours
$a; b$	Cycle-associated and time-associated event coefficients	Events/cycle; events/hour
$\mu_j; l$	Expected block count; Fisher information	Counts; local parameter information
$\delta; W; L$	Normalized bump width; work; displacement	Dimensionless; joules; metres
$r; \kappa; \nu$	Replicate blocks; NB shape; renewal shape	Positive counts or dimensionless parameters
$x; z; \alpha, \beta, \gamma$	Normalized exposures and interaction-model coefficients	$x=n/1,000; z=t/100$

4. Equal absorbed work can conceal unequal peak loads

Record C2 establishes repeated dead-load arrestments, while the released investigation distinguishes the initial-load shock absorber from the main absorption arrangement [2,4]. Does an arrestment count establish that every relevant peak load stayed within its limit? The following generic protective-absorber construction explains why the count alone cannot answer that question. It is not a fitted AAG model. Let W be positive absorbed work, L a positive displacement, $u = x/L$ and $f(u) = LF(x)/W$. Then

$$\int_0^1 f(u) du = 1; F_{\text{peak}} = (W/L) \max f(u). \quad (1)$$

The first equality fixes the area under the force-displacement curve. It does not fix its largest value. A constant normalized force $f = 1$ has peak 1. A continuous triangular bump of unit area over width δ , added with half weight to a constant half-unit baseline, also has total area 1. Its normalized peak is $0.5 + 1/\delta$.

For precision, define the bump $g_\delta(u) = (2/\delta)(1 - |2u/\delta - 1|)$ on $0 \leq u \leq \delta$ and zero elsewhere, with $0 < \delta \leq 1$. Set $f_\delta = 0.5 + 0.5g_\delta$. Direct integration gives unit area and peak $0.5 + 1/\delta$. The baseline keeps the force positive across the full displacement.

Table 3. Work-preserving profile sensitivity.

Normalized profile	Absorbed work / W	Peak force / (W/L)
Constant	1	1
Bump width $\delta = 1$	1	1.5
Bump width $\delta = 0.5$	1	2.5
Bump width $\delta = 0.25$	1	4.5
Bump width $\delta = 0.10$	1	10.5

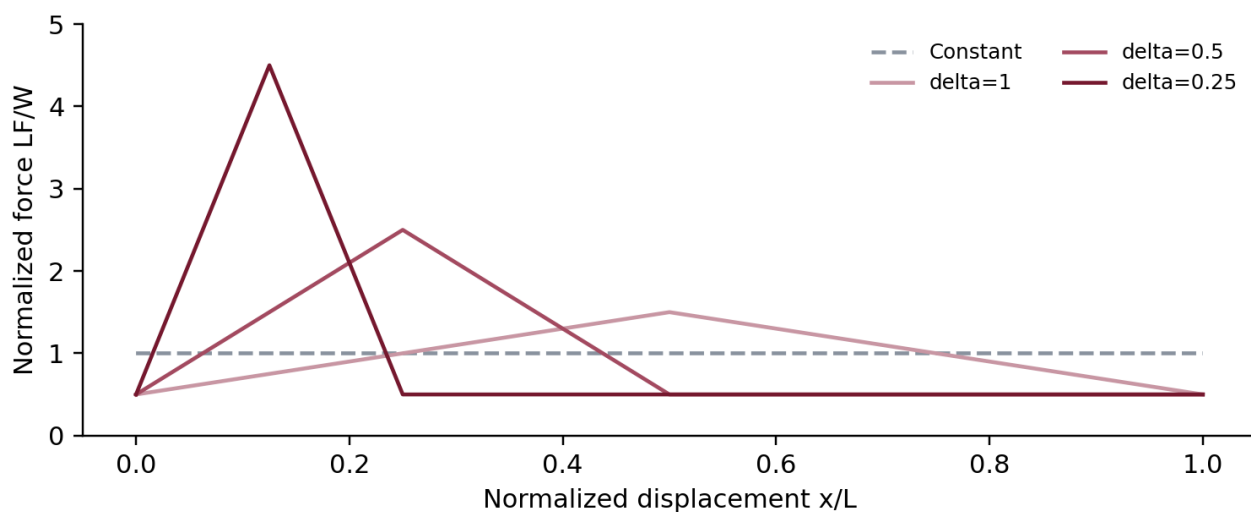


Figure 2. Dimensionless analytical counterexample. All curves absorb the same normalized work. Their peak forces differ. These are constructed profiles, not measured equipment traces or proposed control settings.

As δ approaches zero, the peak is unbounded within this mathematical family. Real hardware restricts bandwidth, compliance, force and rate of change; those restrictions would narrow the admissible family. The result is that absorbed work alone supplies no finite upper bound on peak force without such additional constraints. It does supply a lower bound: the peak cannot be smaller than the average W/L.

This is elementary mechanics used as a counterexample, not a new absorption principle. Its engineering relevance is specific: a successful stop and a correct total energy balance cannot independently establish that an interface load limit was respected. The decisive observation is a sufficiently resolved load record or a justified bound from other measurements. Another uninstrumented success cannot retrospectively supply it.

A load trace also needs a measurement argument. Sampling rate alone does not establish usable bandwidth. The assembled sensor, mounting, conditioning, filtering and acquisition chain can attenuate or distort a transient. Dynamic-force research treats the system response and dynamic sensitivity explicitly [13]. For a peak-load claim, retain calibration, transfer response, anti-alias filtering, saturation checks, timing and a declared uncertainty budget. If the retained signal lacks the required bandwidth, post-processing cannot recover an arbitrary unmeasured peak without additional assumptions. No bandwidth or sensor specification for AAG is inferred here.

5. Operation count and reliability confidence

Record C4 is the concrete trigger for this calculation. Its milestone account reports extensive operation, rather than the event log needed to fit a reliability process [6]. A count of completed operations is important evidence. To become a reliability estimate, it also needs a countable event, eligible exposure, configuration and statistical model.

For an illustrative fixed-configuration model, let events follow a homogeneous Poisson process in eligible cycle exposure N , with mean cycles between events θ . Then the expected event count is N/θ . With zero events, the probability of that observation is $\exp(-N/\theta)$. Solving for a one-sided 95 percent lower confidence bound gives [10]

$$\theta_{L,0.95} = N / [-\ln(0.05)]. \quad (2)$$

For a hypothetical 10,000 eligible event-free cycles, the bound is approximately 3,338 cycles per event. A hypothetical target of 5,000 would require at least 14,979 event-free cycles under the same assumptions. These numbers are illustrative. The NAVAIR milestone is not represented as an event-free trial, and 5,000 is not asserted to be an actual programme requirement.

The qualification is substantive. Changes in configuration, incomplete event detection, dependence between events or outcome-dependent stopping can alter the inference. A precise calculation cannot compensate for the wrong exposure definition. Nor does a mean interval between operational interruptions establish an accident probability. These are different outcomes.

The statistical lesson is constructive: when the model and records are sound, additional eligible exposure can strengthen a bound. When the observation definition is missing, the first useful intervention is to recover the record. Increasing the number printed beside an undefined denominator does not resolve its meaning.

The model choice matters even with zero recorded events. For an ordinary renewal process beginning with a newly restored unit, let independent intervals have a gamma distribution with known shape ν and mean θ . The first-interval survival probability is $Q(\nu, \nu N/\theta)$. Solving that probability equal to 0.05 gives the following bound, where G_{ν} is the unit-scale gamma CDF.

$$\theta_L = \nu N / G_{\nu}^{-1}(0.95).$$

At $N=10,000$, the lower bounds are 2,603, 3,338, 4,216 and 5,462 for $\nu=0.5, 1, 2$ and 5 . Shape one recovers the HPP. These conditional bounds are not interchangeable; unknown shape or a non-fresh observation origin requires another analysis.

For an NHPP, zero events constrain cumulative intensity $\Lambda(N)$, because their probability is $\exp[-\Lambda(N)]$. They do not identify a constant mean interval unless a stationary interpretation is separately justified. In the HPP example, hypothetical target intervals 1,000, 5,000 and 10,000 require 2,996, 14,979 and 29,958 event-free cycles respectively at the stated confidence. The complete sensitivity outputs are attached. None is a programme requirement or proposed operating schedule.

6. Why more tests can leave the explanation unresolved

Record C1 explicitly concerns repeated cycling and an energy-storage cycling rate [5]. Without block-level cycle and elapsed-time exposures, the headline does not reveal which mechanism contrasts were present. I do not assume that the real trials used a proportional design. The following example identifies the records needed to decide. Even complete records can fail to distinguish mechanisms. I use a synthetic example to establish this exactly, without attributing either mechanism to the aircraft systems.

Let independent event counts in block j obey

$$Y_j \sim \text{Poisson}(\mu_j), \mu_j = a n_j + b t_j, a, b \geq 0. \quad (3)$$

Here n_j is cycles, t_j is hours, a is events per cycle and b is events per hour. The additive model excludes interactions, thresholds, loading changes and repair-history effects. It is deliberately simple enough to make the inference transparent.

If every block has $n_j = c t_j$, then $\mu_j = (ca+b)t_j$. Only $ca+b$ is identified. Every nonnegative parameter pair with the same value of that combination gives the same full count distributions. Repetition can estimate the combination more precisely; it cannot identify its two components separately.

Take Model A with $a = 0.001$ and $b = 0.01$, and Model B with $a = 0.002$ and $b = 0$. At ten cycles per hour, both predict identical distributions.

Table 4. Additive-model alternatives; expected counts are synthetic.

Synthetic block	Cycles	Hours	Expected events: A	Expected events: B
Baseline	1,000	100	2	2
More exposure, same ratio	2,000	200	4	4
Contrasting elapsed exposure	1,000	300	4	2

The first two rows produce a rank-one exposure matrix. Pairing the baseline with the contrast produces rank two. This makes the full two-parameter family structurally identifiable. It does not guarantee that a small observed sample will distinguish it reliably.

The smallest useful change depends on the uncertainty. For this model, a contrast that changes the exposure ratio creates information that identical repetitions cannot supply. In a physical study, the proposed mechanisms must first be credible and the contrast must be safe. A change in elapsed time that also changes temperature or maintenance cannot automatically be interpreted as a pure time effect.

6.1. A contrast can fail when the mechanism model changes

A more demanding test allows an interaction. Normalize $x=n/1,000$ and $z=t/100$ and write $\mu = \alpha x + \beta z + \gamma xz$, with nonnegative coefficients. The product is a local phenomenological term: cycle-associated damage may depend on accumulated exposure or thermal state. It is not a claim that calendar time itself causes the actual faults, nor a validated damage law.

The additive model is the $\gamma=0$ subfamily. Now compare $C=(1,1,0)$ with $D=(1,0,1)$. Whenever $x=1$, both give $\mu=1+z$. Consequently, the baseline and dwell contrast that separate A from B still give identical full Poisson distributions under C and D. No number of additional dwell-only repetitions at that fixed cycle exposure can separate the time term from the interaction.

Table 5. An interaction alternative survives the dwell contrast. Both models are synthetic.

x	z	Expected count under C	Expected count under D
1	1	2	2
1	3	4	4
2	1	3	4
2	3	5	8

The predictor matrix has columns x , z and xz . Along a fixed ratio $z=x$, its rank is two for distinct positive exposures: the interaction can vary, but the two main effects remain confounded. At fixed $x=1$, the z and xz columns coincide, again giving rank two. The two-by-two exposure combination in Table 5 has rank three. It removes this specific structural ambiguity, while still requiring replication for useful precision.

This is the main refinement to the additive result: a successful contrast is relative to a declared model family. Interactions and factorial design are established experimental-design topics [14]. The present counterexample makes the consequence explicit for interpreting cycle/time evidence. It does not establish that a two-by-two design identifies every credible mechanism or is feasible on the aircraft systems.

7. Structural information and finite observations

The separation condition has an explicit information expression. For independent Poisson counts with positive means, the expected Fisher information for (a,b) is the sum of the outer products of the exposure vectors, weighted by $1/\mu_j$. Its diagonal entries sum the squared cycle and time exposures divided by their corresponding means. Its off-diagonal entry sums cycle-time products with the same weights. When the determinant is expanded, equal-index terms cancel. Grouping the remaining terms by unordered pairs gives

$$\det I = \sum_{i < j} \frac{(n_i t_j - n_j t_i)^2}{\mu_i \mu_j} \quad (4)$$

Each term is nonnegative. The determinant is positive exactly when at least one pair of exposure vectors is nonproportional, assuming finite positive means. This is a standard weighted-Gram identity applied to Eq. (3). It shows why repeating a proportional design increases some information without recovering the missing parameter direction. Nearly proportional exposures can also be poorly conditioned even when the determinant is positive.

With event counts treated as dimensionless, the determinant has units of cycle-squared hour-squared; its magnitude depends on parameter units and is not a universal cross-system score. The zero-versus-positive distinction is the relevant structural statement here. Boundary cases with zero means require separate treatment and are outside this regular-information calculation.

For the two specified alternatives, finite-sample performance can be calculated exactly. Across r independent contrasting blocks, the total S has Poisson mean $4r$ under A and $2r$ under B. With equal prior probabilities and equal error costs, the likelihood ratio is $\exp(-2r)$ multiplied by 2^S . Choose A when $S \ln(2) > 2r$.

Table 6. Exact Poisson classification errors under two specified alternatives and equal error costs.

Contrasting blocks	Choose A when total S is at least	P(choose A under B)	P(choose B under A)
1	3	32.33%	23.81%
5	15	8.35%	10.49%
10	29	3.43%	2.94%
20	58	0.44%	0.43%

These are exact probabilities under synthetic fixed alternatives. They are not measured trial results or a proposed carrier test schedule. One contrasting block separates the expected

values but still permits substantial classification error. Replication becomes useful after the contrast has removed structural ambiguity.

At the unchanged ratio the distributions are identical, so the minimum average classification error with equal priors remains 50 percent regardless of repetitions. Under the contrast it falls as the independent observations accumulate. Structural separation and statistical confidence therefore require different arguments. Unknown parameters, nuisance effects, unequal costs or non-Poisson events would require a different analysis.

Burclova and Pazman study identifiability and stable estimation through information divergence and experimental design in generalized regression, including Poisson models [12]. The present calculation applies established ideas; it does not claim to originate experimental identifiability or optimal design.

7.1. Executed robustness checks: overdispersion and replication

A Poisson count has variance equal to its mean. To challenge that assumption, I used independent negative-binomial blocks with mean μ and variance $\mu + \mu^2/\kappa$. This is a gamma-Poisson mixture with an independent gamma multiplier for each block. It represents one form of heterogeneity, not a fit to carrier data. A common random multiplier across multiple blocks would introduce dependence and invalidate the sum calculation used here [15,16].

For r independent blocks with common κ and mean μ , their total is negative binomial with shape $r*\kappa$ and success parameter $\kappa/(\kappa+\mu)$. I derived the equal-prior likelihood-ratio threshold for the two means 4 and 2, calculated exact tails, and searched integer r from 1 to 500 for the first block count with both classification errors at most 5 percent. The search is a fixed-alternative planning calculation, not general power for an unknown composite model.

Table 7. Error-model sensitivity. At $r=10$ the model-matched integer cutoff happens to be $S \geq 29$ in all four rows.

Block-count model	Exact P(choose A under B), $r=10$	Exact P(choose B under A), $r=10$	First r with both errors $\leq 5\%$
Poisson	3.43%	2.94%	9
NB, $\kappa=5$	6.22%	7.91%	13
NB, $\kappa=2$	9.67%	14.23%	22
NB, $\kappa=1$	13.69%	21.55%	34

A plan adequate under the Poisson assumptions can therefore be inadequate under a specified overdispersed alternative. For $\kappa=2$, the larger error at ten blocks rises from 3.43 to 14.23 percent. A model-matched rule first brings both errors below 5 percent at 22 blocks; retaining the Poisson threshold sequence instead requires 24. This difference is about assumed observation variability and decision rules, not a change in the equipment.

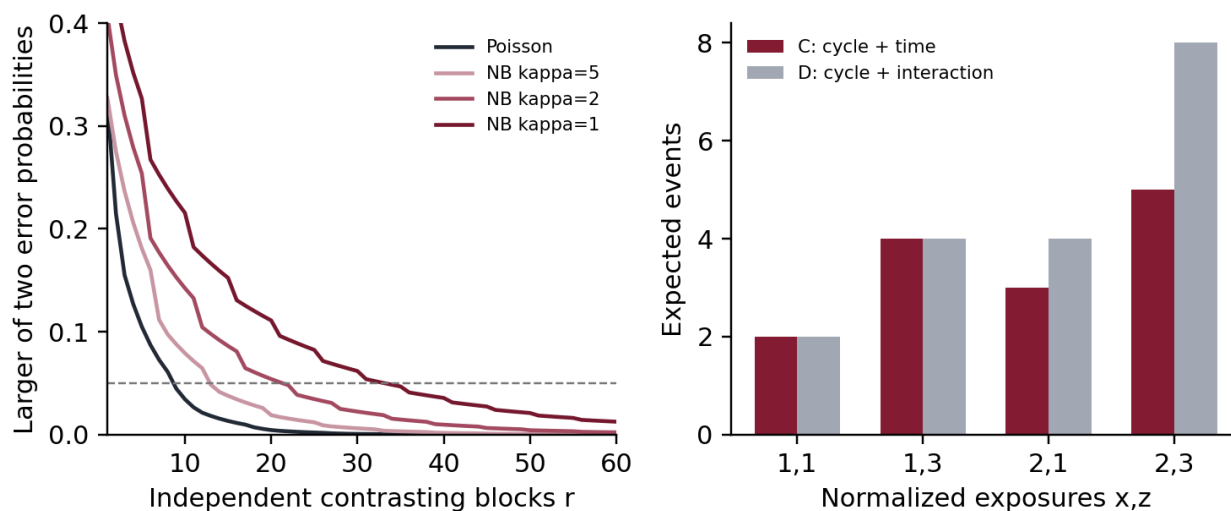


Figure 3. Left: exact model-matched error curves, using the larger of the two errors. Right: C and D remain identical at the dwell-only settings but diverge at changed cycle exposure. These are synthetic calculations.

A seeded simulation and probability-mass recurrences check the implementation; Appendix D reports the execution details. These checks do not validate the count models for AAG or EMALS. The prior literature already covers identifiability and gamma-Poisson design [12,16]. The contribution is the connected case application and reproducible sensitivity study.

8. A pooled history can hide change in either direction

The final construction concerns configuration. Suppose a record contains 10,000 eligible cycles and ten events, with 1,000 cycles on the final configuration. The pooled ratio is 1,000 cycles per event.

Now allocate all ten events before the configuration change. Alternatively, allocate all ten afterwards. Both records have the same pooled summary. On the final configuration, however, the first has zero events and the second has ten. Its estimate in the second record is 100 cycles per event. Under Eq. (2), the first record's one-sided lower bound is only approximately 334, despite its zero observed events.

The pooled summary does not determine the final-configuration result. Earlier failures can conceal improvement; earlier successful exposure can conceal deterioration. This symmetry matters. An assessment that interprets every missing detail pessimistically is no more justified than one that interprets it optimistically.

No claim is made that the real programme combined its records this way. The construction identifies what a pooled number cannot establish without a transfer argument. National Research Council guidance already addresses the importance of configuration, environment and appropriate use of developmental reliability evidence [11]. The exact example makes that issue inspectable.

Separating every minor revision also has a cost: the resulting strata may be too small to support useful inference. An evaluator should preserve the configuration record and then justify which evidence can be combined. The answer is not automatic pooling or automatic rejection of earlier work.

The documentary link is C3: hardware changes and software revision make transfer between test configurations a real assessment question [2,4]. The pooling construction shows why an undifferentiated history cannot answer it; it does not assert that the programme actually pooled data incorrectly.

For a transfer argument, specify the target configuration and event outcome, identify modified hardware and software functions, compare load/environment and observation rules, and justify which failure mechanisms remain exchangeable. Preserve a sensitivity analysis with and without transferred exposure. If event detection changes, apparent reliability growth may be an observation change. These are proposed review criteria, rather than a statistical certificate of equivalence.

Record an interruption, a counted failure, repair, abort and no-fault-found outcome separately, then declare which enter the estimator. Administrative end-of-test censoring differs from removal triggered by an emerging fault. Preserve at-risk exposure, interruption reason and event ascertainment; do not silently count missing records as successes. Informative censoring or missingness needs an explicit sensitivity model.

Repairable equipment permits recurrent events and restoration. First-failure lifetime data for non-repairable units require a different observation structure. A software version is also a configuration: fixes can change the event process, and repeated execution of similar states does not cover unvisited conditions. Human maintenance, spares and diagnosis affect restoration time and observation completeness. NASA's availability definitions are useful here [9], but a time-availability claim still cannot replace a cycle-event record.

9. Choose the observation that addresses the uncertainty

The four constructions lead to different next actions. Their value lies in keeping those actions aligned with the claim.

Table 8. Assessment actions aligned with the obstacle to inference.

Obstacle to the claim	What repeated operation alone cannot recover	Decisive next evidence
Aggregate physical measurement	Unobserved transient extremes	Resolved load record or justified physical bound
Undefined or unmatched record	Eligible denominator, event and configuration	Matched definitions and traceable observation record
Sampling uncertainty under a sound model	A tighter confidence bound without more eligible data	Additional observations under the justified plan
Proportional exposure design	Separate cycle and time contributions	Controlled contrast with distinct predictions
Pooled configurations	Final-configuration behaviour	Configuration-linked events and exposure; transfer argument

This is an assessment sequence, not a substitute for reliability engineering or a new certification standard. First state the claim and the tested system boundary. Then identify whether the obstacle is measurement, record interpretation, sampling, mechanism

separation or transfer between configurations. Select the observation that changes the disputed inference. Finally check whether its result survives the alternative explanation it was intended to distinguish.

Applied to the cases, a subsystem endurance milestone should retain its subsystem meaning. Aircraft compatibility evidence should retain its aircraft and configuration scope. A cumulative shipboard milestone should remain evidence of extensive operation unless the corresponding event records support a stronger statistical statement. This reading preserves the achievements while preventing unsupported transfer between them.

10. First feasible experiment and practical evaluation

The first feasible physical step is a low-energy, instrumented benchtop measurement check, before any reliability-growth trial. Use a commercially rated force sensor and a compliant protective test piece within their rated range. Apply repeatable low-energy loading, retain the highest usable calibrated trace, and compare peak estimates obtained through declared filtering and downsampling. Record mounting, calibration, acquisition settings and uncertainty. This tests how retained measurement bandwidth changes the inferred peak; it does not reproduce aircraft arrestment.

For the mechanism example, first run a two-by-two set of cycle and elapsed-exposure blocks on non-weapon laboratory equipment with existing safe operating limits. Randomize block order, measure temperature and actual exposure, and preserve configuration and aborted-block records. Use initial blocks to assess whether countable events occur often enough to fit any event model. A short pilot with no informative events supports a measurement-feasibility finding, not a reliability improvement claim.

A subsequent study should compare cycle-only, additive and interaction models on held-out settings. Replicate using dispersion estimated from pilot data, then recalculate the error criterion; Table 7 is not a universal sample-size prescription. If changing elapsed time changes thermal state, record both and test the competing explanation. If the interaction improves neither prediction nor calibration, reject it for that apparatus.

The assessment sequence itself could be evaluated by blinded reviewers using the same historical claims. Outcomes would include correct boundary identification, false failure-count reconstruction, causal overclaiming, agreement and review time. No such reviewer study or institutional saving is claimed here. Appendix E provides an immediately usable one-page assessment sheet.

11. Contribution and strongest counterarguments

The contribution has three inspectable parts: a source-to-claim record that keeps the real milestones tied to their test objects; a worked interaction example showing why a dwell contrast must be assessed against the declared model family; and exact calculations showing how dispersion changes the replication needed for specified decisions. The simulation is an implementation check, not additional empirical support.

None of the component mathematics is claimed as new theory. Rank deficiency, interaction aliasing, Poisson bounds, negative-binomial tails and configuration-aware inference are established. The literature already studies design under these models [10-16]. I have not established worldwide priority for this particular combination, nor shown superiority to a competent existing test-review process.

A strong objection is that the public announcements were never intended as complete qualification datasets. That is correct. The analysis constrains a reader's inference from those announcements; it does not accuse their authors of making an unsupported reliability claim. Another objection is that the synthetic models are deliberately simple. They are constructed to prove specific inference limits and then challenged by an interaction extension and overdispersion. They remain examples, not recovered carrier mechanisms.

A third objection is that demanding every observation could make evaluation impractical. The purpose is the opposite: identify which missing record or contrast is decisive for the claim at hand. A peak-load claim does not require inventing a cycle-aging model, and a configuration-transfer claim cannot be settled by measuring an unrelated transient more precisely.

Table 9. Scope and evidence status of the conclusions.

Claim	Evidence type	Status
Milestones C1-C4 occurred as reported	Dated official public accounts	Source-reported documentary evidence
Work alone does not bound peak force in the constructed family	Algebra and numerical quadrature	Demonstrated within stated assumptions
Dwell-only contrast leaves C/D indistinguishable	Equal means and rank calculation	Exact synthetic counterexample
Dispersion changes the stated replication criterion	Exact tails and seeded Monte Carlo	Model-conditional computational result
Better carrier reliability or institutional savings	No corresponding measurements	Not demonstrated

12. Limitations

The empirical material consists of selected public documentary records, not independently collected carrier measurements. The mathematical and Monte Carlo results are conditional on specified synthetic models. Hardware failure logs, load histories, actual dispersion, exposure interactions and configuration-specific field performance are unavailable to this study. A source summary's omissions are not evidence that the programme failed to collect the corresponding data.

Independent negative-binomial blocks do not cover common-cause dependence across blocks. The gamma-renewal sensitivity assumes known shape and a fresh observation origin. The interaction model excludes further state variables and thresholds; a factorial rank check does not establish their absence. The peak construction establishes a measurement limit without predicting a physically achievable load profile for either system. No accident probability, safety certification, field reliability improvement or financial saving is estimated.

13. Conclusion

Successful operation, dependable repeated service and an explanation of failure are different achievements. AAG and EMALS development records make those boundaries visible. The analytical results explain why the next useful observation depends on the unresolved claim.

The most instructive result is model dependence: a contrast that is sufficient for one mechanism family can preserve exact ambiguity in another. Sampling uncertainty adds a separate requirement; in the synthetic example, allowing block heterogeneity changes nine qualifying replications to 22. Neither problem is solved by an impressive cumulative count without its observation model.

An evaluator can act on this now: retain the tested boundary, state the alternatives, check which remain observationally equivalent, and select the smallest safe measurement or contrast that separates them. Start the physical follow-up with a calibrated benchtop trace and a randomized exposure pilot. Let the result determine whether a larger experiment is justified.

Research materials, authorship and status

Zawwar Sami, Independent Researcher. ORCID: 0009-0004-5819-3017. Research correspondence: <https://zawwarsami.com>. Version 1.2, 17 October 2026. Independent preprint; not peer reviewed. License: CC BY 4.0 (<https://creativecommons.org/licenses/by/4.0/>), excluding cited third-party material.

Paper DOI: <https://doi.org/10.5281/zenodo.23435836>. Reproducibility files: W2-v1.2-reproducibility.zip, supplied separately and embedded in this PDF. The archive contains executable code, source-to-claim records, exact and simulation outputs, figures, requirements, the execution log and a SHA-256 manifest. Appendix D prints an executable verifier for viewers that hide attachments.

The computation uses synthetic counts and publicly released documents; no human participants or private operational data were collected. No firsthand carrier experiment or institutional affiliation is asserted.

Appendix A. Fisher-information identity

For independent Poisson counts with positive means $\mu_j = a n_j + b t_j$, differentiating the log likelihood and taking expectations gives

$$I_{aa} = \sum_j n_j^2 / \mu_j; I_{bb} = \sum_j t_j^2 / \mu_j; I_{ab} = \sum_j n_j t_j / \mu_j.$$

In the determinant, the product of the diagonal entries is reduced by the square of the off-diagonal entry. Equal-index terms cancel. Pairing each remaining term with its reversed-index counterpart gives the squared cross-product term in Eq. (4). Each term is nonnegative. The determinant is positive exactly when at least one pair of exposure vectors is nonproportional. Nearly proportional designs may remain poorly conditioned. Counts are dimensionless; the determinant's magnitude depends on parameter units and is not a universal ranking score. Zero-mean boundary cases need separate treatment.

Appendix B. Classification errors and dispersion

For Poisson totals with means $4r$ and $2r$, the likelihood ratio is $\exp(-2r)$ multiplied by 2^S . With equal priors and error costs, choose A if $S \ln(2) > 2r$. The integer threshold is $h = \text{floor}[2r/\ln(2)] + 1$. The error probabilities are $P_B(S \geq h)$ and $P_A(S < h)$, evaluated by exact distribution tails. The same-ratio design gives identical distributions for A and B, hence minimum equal-prior average error one half for every r .

For independent negative-binomial blocks sharing known κ , the total shapes are both $r \kappa$. Put $p_A = \kappa/(\kappa+4)$ and $p_B = \kappa/(\kappa+2)$. The combinatorial factors cancel, giving

$$\log LR = r\kappa \ln(p_A/p_B) + S \ln[(1-p_A)/(1-p_B)].$$

The integer threshold is one plus the floor of minus the constant term divided by the positive coefficient of S . Exact tails and the integer replication search produce Table 7. Priors, losses, nuisance parameters and shared block effects change this calculation.

Appendix C. Zero-event confidence calculation

Under fixed eligible exposure N and an HPP with rate $1/\theta$, $P(K=0) = \exp(-N/\theta)$. Inverting at α yields the one-sided lower confidence bound $\theta_L = N/[-\ln(\alpha)]$. A zero-event design for target θ_0 therefore needs $N \geq -\theta_0 \ln(\alpha)$, rounded upward. It is a confidence procedure under the declared model, not a posterior probability that θ exceeds the bound. The gamma-renewal calculation in Section 5 instead inverts the known-shape first-interval survival probability. Censoring, observation origin and model choice are part of both statements.

Appendix D. Implementation checks and reproducibility

The compact script below checks the principal ranks, exact errors, block-count search, renewal sensitivity and normalized-work integrals. The full script and unabridged execution log accompany the paper. Requires Python, NumPy and SciPy.

To check implementation, I executed eight seeded batches of 200,000 independent total-count draws, one for each truth/model combination in Table 7. The 1.6 million draws sample the analytically derived totals directly; they are not 1.6 million physical test blocks. Seed 20261016, NumPy 2.3.5 and SciPy 1.17.0, observed frequencies, Wilson intervals and full count histograms are retained. All eight estimates were within 1.51 binomial standard errors of the exact probabilities. Probability-mass recurrences supplied a separate arithmetic check of the tail calculations.

Table D1. Executed Monte Carlo counts for the Poisson and $\kappa=2$ rows. Each error count has denominator 200,000.

Truth and model	Observed classification errors	Estimated error	95% Wilson interval
B, Poisson	6,810	3.405%	3.326%-3.485%
A, Poisson	5,868	2.934%	2.861%-3.009%
B, NB $\kappa=2$	19,540	9.770%	9.641%-9.901%
A, NB $\kappa=2$	28,521	14.261%	14.108%-14.414%

The simulation checks implementation against declared distributions. It adds no observations of actual carrier performance.

D.1. Executable verification

W2 v1.2: synthetic numerical verification. Python, NumPy, SciPy.

```
import math
import numpy as np
from scipy.stats import poisson, nbinom, gamma
from scipy.integrate import quad

def errors(r, k=None):
    if k is None:
        h = math.floor(2*r/math.log(2)) + 1
        A, B = poisson(4*r), poisson(2*r)
    else:
        pA, pB = k/(k+4), k/(k+2)
        h = math.floor(-r*k*math.log(pA/pB) /
                      math.log((1-pA)/(1-pB))) + 1
        A, B = nbinom(r*k, pA), nbinom(r*k, pB)
    return float(B.sf(h-1)), float(A.cdf(h-1))

first = []
for k in [None, 5, 2, 1]:
    r0 = next(r for r in range(1,501) if max(errors(r,k)) <= .05)
    first.append(r0)
    print(k, errors(10,k), r0)
assert first == [9, 13, 22, 34]

def features(X):
    return np.column_stack((X, X[:,0]*X[:,1]))
fixed = np.array([[1,1],[1,3],[1,5]], float)
factor = np.array([[1,1],[1,3],[2,1],[2,3]], float)
C, D = np.array([1,1,0]), np.array([1,0,1])
assert np.allclose(features(fixed)@C, features(fixed)@D)
assert np.linalg.matrix_rank(features(fixed)) == 2
assert np.linalg.matrix_rank(features(factor)) == 3
for v in [0.5, 1, 2, 5]:
    lo = 10000*v/gamma.ppf(.95,v)
    assert abs(gamma.sf(10000,v,scale=lo/v)-.05) < 1e-12
assert math.ceil(-5000*math.log(.05)) == 14979
for d in [1, 0.5, 0.25, 0.1]:
    def f(u):
        return .5+(1/d)*(1-abs(2*u/d-1)) if 0<=u<=d else .5
    assert abs(quad(f,0,1,points=[d/2,d])[0]-1) < 1e-12
    assert abs(f(d/2)-(.5+1/d)) < 1e-12
print('All compact checks passed.')
```

Appendix E. One-page evaluator sheet

State the claim, system boundary, tested configuration and event definition before interpreting a milestone.

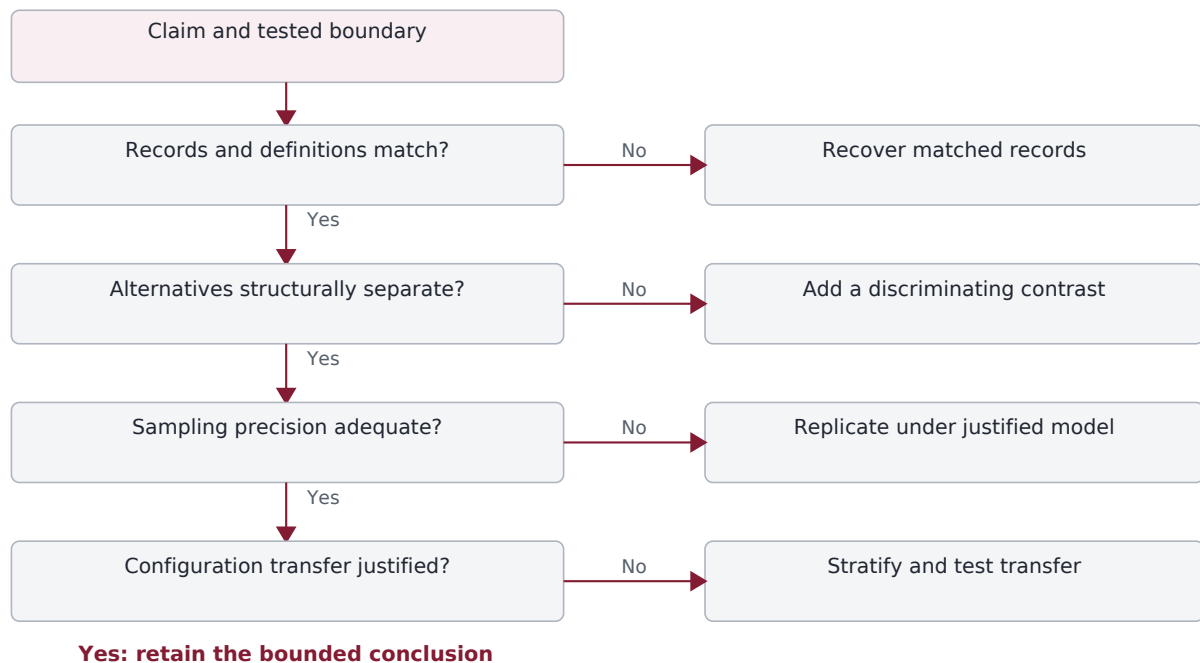


Figure 4. Choose the next evidence by the obstacle to inference. Model adequacy and measurement validity remain necessary throughout; this sheet is a review aid, not a certification procedure.

- ☐ Claim and boundary: identify the exact subsystem, aircraft or service proposition.
- ☐ Observation: retain attempts, completions, interruptions, failures, repairs, aborts and no-fault-found records separately.
- ☐ Metrology: verify calibration, usable bandwidth, filtering, saturation, timestamps and uncertainty for the claimed quantity.
- ☐ Models: name credible alternatives, including interactions or state dependence; check structural separation before calculating power.
- ☐ Sampling: justify count variability, independence, censoring and missing-data handling; recalculate replication under plausible dispersion.
- ☐ Transfer: preserve hardware/software versions, environment and observation changes; justify pooled exposure and test sensitivity without it.
- ☐ Decision: record what result would change the conclusion, and which stronger claims still remain outside the evidence.

References

- [1] Sami, Z. (2026). Weapons as Systems: Components, Infrastructure and Human Organisation. W1, v1.0, 14 October. Independent preprint.
<https://zenodo.org/records/23359788>
- [2] U.S. Department of Defense Inspector General. (2016, July 5). Advanced Arresting Gear Program Exceeded Cost and Schedule Baselines (Redacted). DODIG-2016-107. Printed pp. 4-5, 8-10.
<https://www.dodig.mil/Reports/Audits-and-Evaluations/Article/1119295/advanced-arresting-gear-program-exceed-ed-cost-and-schedule-baselines-redacted/>
Released-report mirror: https://www.defensedaily.com/wp-content/uploads/post_attachment/136651.pdf
- [3] NAVAIR. (n.d.). Advanced Arresting Gear (AAG). Product description. Accessed 16 October 2026.
<https://www.navair.navy.mil/product/advanced-arresting-gear-aag>
- [4] NAVAIR. (2017, January 23). AAG remains Navy's choice for next-gen recovery system.
<https://www.navair.navy.mil/node/24591>
- [5] NAVAIR. (2008, September 23). Carrier-based launch system reaches milestone.
<https://www.navair.navy.mil/node/10976>
- [6] NAVAIR. (2022, July 5). EMALS and AAG reach 10,000 aircraft launches and recoveries.
<https://www.navair.navy.mil/news/EMALS-and-AAG-reach-10000-aircraft-launches-and-recoveries/Tue-07052022-1019>
- [7] NAVAIR. (n.d.). Electromagnetic Aircraft Launch System (EMALS). Product description. Accessed 16 October 2026.
<https://www.navair.navy.mil/product/Electromagnetic-Aircraft-Launch-System-EMALS>
- [8] U.S. Government Accountability Office. (2013, September 5). Ford-Class Carriers: Lead Ship Testing and Reliability Shortfalls Will Limit Initial Fleet Capabilities. GAO-13-396.
<https://www.gao.gov/products/gao-13-396>
- [9] NASA Kennedy Space Center. (n.d.). Availability. Three-page technical note, revision identifier 240607. Accessed 16 October 2026.
https://extapps.ksc.nasa.gov/Reliability/Documents/Availability_What_is_it.pdf
- [10] NIST/SEMATECH. (n.d.). e-Handbook of Statistical Methods, 8.4.5.1: Constant repair rate (HPP/exponential) model. Accessed 16 October 2026.
<https://itl.nist.gov/div898/handbook/apr/section4/apr451.htm>
- [11] National Research Council. (2015). Reliability Growth: Enhancing Defense System Reliability. Chapter 7, Developmental Test and Evaluation. National Academies Press. doi:10.17226/18987.
<https://www.nationalacademies.org/read/18987/chapter/9>
- [12] Burclová, K., and Pázman, A. (2015). Optimum design via I-divergence for stable estimation in generalized regression models. arXiv:1507.07443v1. doi:10.48550/arXiv.1507.07443.
<https://arxiv.org/abs/1507.07443>
- [13] Burns, D. E., Vlajic, N., Chijioke, A., and Parker, P. A. (2022). Aerodynamic Metrologists Guide to Dynamic Force Measurement. Journal of Aircraft. doi:10.2514/1.C036676.
<https://doi.org/10.2514/1.C036676>
- [14] NIST/SEMATECH. (n.d.). e-Handbook of Statistical Methods, 5.3.3.3.1: Two-level full factorial designs. Accessed 16 October 2026.
<https://www.itl.nist.gov/div898/handbook/pri/section3/pri3331.htm>
- [15] SciPy community. (2026). scipy.stats.nbinom. SciPy v1.27.0 reference manual.
<https://docs.scipy.org/doc/scipy-1.17.0/reference/generated/scipy.stats.nbinom.html>
- [16] Schmidt, M., and Schwabe, R. (2018). Optimal Designs for Poisson Count Data with Gamma Block Effects. arXiv:1808.05412v1.
<https://arxiv.org/abs/1808.05412>