

The Prouse Residual Liability Model, Version 2.1: Specification, Simulation Evaluation and Validation Programme

Richard G. Prouse, Principal Analyst, Confidential Risk Consulting (New Zealand); author of the Prouse Residual Liability (L_r) Model (SSRN, 2026) · richard@confidentialriskconsulting.com

Status: working paper, not peer reviewed. Refines the formulation in Prouse (2026), SSRN working paper, DOI 10.2139/ssrn.6175099.

Licence: Creative Commons Attribution 4.0 International (CC BY 4.0). **Competing interest:** the author operates a consultancy that uses the model commercially (section 12). Simulation code and results:

confidentialriskconsulting.com/assets/Lr-v2.1-simulation-code.zip

Abstract. Safety, security and compliance controls are judged after an incident by what an organisation can demonstrate, not by what it did. The Prouse Residual Liability Model expresses this by treating documentation as a multiplier on the credit a control receives. The original formulation, $L_r = (T \times V) - (C_e \times C_d)$, can yield negative values, lacks a consequence term and subtracts quantities on different scales.

This paper specifies Version 2.1: $L_r = E(1 - C_e C_d) + A(1 - C_d)$, where $E = T \cdot V \cdot I$ is the inherent exposure and A is an optional term for stand-alone record-keeping offences. The model is bounded and monotone, reduces to standard residual risk when documentation is complete, and decomposes exposure exactly into unmitigated risk, the documentation void and demonstrable mitigation. The Documentation Coefficient is defined operationally as a design-quality score multiplied by an operating-evidence rate from a random sample of records.

The form was selected by a pre-specified simulation study following the ADEMP framework: 14 candidate models, seven assumed legal environments, 2,000 organisations of 20 controls each. Version 2.1 captured a mean 94.5% of the best achievable liability reduction (worst environment 91.7%), against 89.5% for expected loss, 81.2% to 85.2% for 5×5 risk matrices and 53.3% for Version 1. It estimated total liability to within about 19% on average, against a factor of about 2.5 for a standard residual-risk calculation. These results hold in simulation only. The paper maps the model onto ISO 31000, IEC 31010 and related standards and sets out a pre-registered programme to test it against real outcomes.

1. Purpose and scope

The model is a **defensibility index**. It estimates the part of a risk exposure that an organisation could not demonstrate it had mitigated if an investigator, regulator, insurer or court examined the record. It is intended for three uses: prioritising documentation effort across controls, reporting to boards and officers, and planning audits.

It is not a probability of prosecution or conviction, and it does not replace legal advice or quantitative risk analysis. Whether the index is associated with legal outcomes is an empirical question. This paper states it as a hypothesis (H3) and sets out how to test it (section 10).

2. The Version 1 formulation and its limitations

Version 1 defined residual liability as

$$L_r = (T \times V) - (C_e \times C_d) \quad (v1)$$

Here T is threat, V vulnerability, C_e control effectiveness and C_d the Documentation Coefficient, each on a 0–1 scale. Its central insight is that the credit a control earns in a later examination is the product $C_e \times C_d$. Version 2 retains that insight. Version 1 has five limitations.

	Limitation	Consequence
L1	Unbounded below	With $T = 0.9$, $V = 0.8$, $C_e = 0.95$ and $C_d = 1$, $L_r = 0.72 - 0.95 = -0.23$. A negative liability has no interpretation.
L2	Absolute, not proportional, reduction	The same control removes the same amount from a trivial and a catastrophic exposure. Credit should be proportional to what is at stake.
L3	No consequence term	Risk is conventionally characterised by scenario, likelihood and consequence (Kaplan and Garrick, 1981; ISO 31000:2018). $T \times V$ describes likelihood only.
L4	No operational definition of C_d	Two assessors can reach different values for the same control, and there is no evidence of reliability or validity.
L5	One control, no uncertainty	Real exposures are covered by layered controls, and inputs are estimates with error.

3. Version 2.1 specification

3.1 Variables

Symbol	Definition	Scale	Typical evidence
T	Probability that the threat or hazard event occurs in the assessment period (normally one year)	0–1	Incident data, sector statistics, threat assessment, calibrated expert judgement
V	Probability that, given the event, harm or loss results in the absence of the control being assessed	0–1	Exposure analysis, vulnerability assessment, bow-tie analysis
I	Consequence if harm results: expected loss including damages, penalties, remediation and defence costs (currency), or a 1–5 ordinal index	≥ 0	Loss history, legal penalty ranges, insurance data
E	Inherent exposure, $T \cdot V \cdot I$	as I	Derived
A	Optional. Expected annual penalty for breaching a stand-alone duty to keep records, whether or not harm occurs: probability that an inspection or audit examines the records $\times$ the penalty for inadequate records. Zero where no such duty applies	≥ 0	Statute and penalty schedules (for example GDPR Art. 30 with Art. 83(4); OSHA 29 CFR Part 1904), inspection rates
C_e	Proportion of the exposure the control removes when it operates as designed	0–1	Testing, drills, inspection results, performance data, expert elicitation
C_d	Documentation Coefficient: the degree to which the control's design and operation can be demonstrated from records to an independent examiner	0–1	Measured by the protocol in section 5

3.2 Equations

$$E = T \cdot V \cdot I, \quad X = E + A \quad (1)$$

$$R = E (1 - C_e) \quad (2)$$

$$L_r = E (1 - C_e C_d) + A (1 - C_d) \quad (3)$$

$$D = L_r - R = (E C_e + A) (1 - C_d) \quad (4)$$

$$M = (E C_e + A) C_d \quad (5)$$

X is the **total exposure**. R is the **operational residual risk**: what remains after the control works as it does. L_r is the **residual liability**: the part of the exposure the organisation cannot show it has mitigated. D is the **documentation void**: mitigation that exists in practice but cannot be demonstrated, plus any records penalty that adequate records would avoid. M is the **demonstrable mitigation**. Exposure decomposes exactly:

$$X = R + D + M \quad (6)$$

Where no stand-alone record-keeping duty applies, $A = 0$ and the model is identical to Version 2.0.

Continuity with Version 1. The defensibility value of a control, $C_e \times C_d$, is unchanged. The illustrative 95%-effective control with documentation worth 10% still earns $0.95 \times 0.10 = 9.5\%$ credit. Version 2 changes how that credit is applied to the exposure: proportionally, never below zero, and with consequence included.

What changed in Version 2.1, and why

- **Records term A .** In several jurisdictions inadequate records are an offence in their own right. Without A , the model missed that exposure; with it, decision value in that environment rose from 87.9% to 95.4% (section 8).
- **Best estimate of $S = (x+1)/(n+2)$.** Using the Wilson lower bound for ranking cost accuracy and biased results upwards. The lower bound is kept for conservative external reporting only.
- **Guidance on uncertainty ranges and sample sizes**, from the coverage and sample-size experiments.

4. Properties

For $C_e, C_d \in [0,1]$ and $E, A \geq 0$ the following hold.

P1 Boundedness. $0 \leq R \leq L_r \leq X$.

$$L_r - R = (E C_e + A)(1 - C_d) \geq 0 \text{ and } X - L_r = (E C_e + A) C_d \geq 0, \text{ while } R = E(1 - C_e) \geq 0. \quad \square$$

P2 Monotonicity. $\partial L_r / \partial C_d = -(EC_e + A) \leq 0$; $\partial L_r / \partial C_e = -EC_d \leq 0$; $\partial L_r / \partial E = 1 - C_e C_d \geq 0$; $\partial L_r / \partial A = 1 - C_d \geq 0$. Better documentation or a more effective control never increases residual liability, and a larger exposure never decreases it.

P3 Limiting cases. If $C_d = 0$ then $L_r = X$: no credit without proof. If $C_d = 1$ then $L_r = R$: complete proof earns full operational credit. If $C_e = 0$ then $L_r = E + A(1 - C_d)$: **documenting an ineffective control earns nothing against the harm**; records reduce only the separate records penalty, as the law intends. This property guards against "paper compliance".

P4 Complementarity. $\partial^2 L_r / \partial C_e \partial C_d = -E < 0$. Documentation is worth most on the most effective controls, and effectiveness is worth most when it is documented. The reduction available by fully documenting control i is exactly its void $D_i = (EC_{e,i} + A_i)(1 - C_{d,i})$, which gives a prioritisation rule: **document controls in descending order of D_i** .

P5 Proportionality. With $A = 0$, $L_r/E = 1 - C_e C_d$ depends only on the control. Controls can therefore be compared across exposures of any size, and the ratio is reported as the *residual liability ratio*.

P6 Layered controls. For n controls acting independently on the same exposure:

$$L_r = E \prod_i (1 - C_{e,i} C_{d,i}), \quad R = E \prod_i (1 - C_{e,i}) \quad (7)$$

Both are bounded, monotone in every control and independent of order. Independence is optimistic when controls share common causes of failure. Report the interval between the independent value (7) and the perfectly dependent bound $E \cdot \min_i (1 - C_{e,i} C_{d,i})$, in which only the strongest demonstrable control counts. Records terms add across controls: $\sum_i A_i (1 - C_{d,i})$.

P7 Units. If T is an annual probability and I is in currency, then E , A , X , R , D , M and L_r are expected annual amounts in that currency, and the void D can be compared directly with the cost of closing it. With an ordinal I , set $A = 0$; results are an index and must not be presented as money.

5. Measuring the Documentation Coefficient

External auditing distinguishes a **test of design**, which asks whether the control is properly specified, from a **test of operating effectiveness**, which asks whether it ran as specified throughout the period. Examples are PCAOB AS 2201 and the Type 2 reports of ISAE 3402. Version 2 defines C_d on the same two axes:

$$C_d = Q \times S \quad (8)$$

Here $Q \in [0,1]$ is design and documentation quality and $S \in [0,1]$ is the operating-evidence rate. The product means neither can compensate for the absence of the other: a perfect document with no records of operation earns nothing, and neither do records of an undocumented practice.

5.1 Design quality, Q

The assessor scores ten criteria from 0 to 3 against written anchors (Appendix A), and $Q = \Sigma \text{ scores} / 30$. The criteria put into practice the requirements for documented information in clause 7.5 of ISO management system standards and the guidance in ISO 10013:2021. They also cover the

"verify" element of officers' due-diligence duties. Weights are equal until the expert panel in Study 1 sets them.

5.2 Operating evidence, S

The assessor draws a random sample of n occurrences of the control from the assessment period (Appendix B gives indicative sample sizes) and tests each for four attributes. A record is:

- **complete**: every required step is recorded;
- **attributable**: it shows who acted;
- **contemporaneous**: it was made at the time;
- **reviewed**: it was checked where a check is required.

If x of the n occurrences meet all four, the **best estimate** used for ranking and planning is $S = (x+1)/(n+2)$, the posterior mean under a uniform prior (Laplace's rule of succession). It never reaches 0 or 1 from a small sample. Small samples are uncertain, so the **conservative** C_d uses the lower bound of the Wilson 95% score interval (Wilson, 1927; Agresti and Coull, 1998) around $\hat{s} = x/n$:

$$S_L = [\hat{s} + z^2/2n - z \sqrt{(\hat{s}(1-\hat{s})/n + z^2/4n^2)}] / (1 + z^2/n), \quad z = 1.96 \quad (9)$$

Both values are reported. Ranking and planning use the best estimate; any external claim uses the conservative value. In simulation, using the lower bound for ranking lowered decision value from 93.4% to 92.6% and overstated liability by about 12% (section 8). Sample size barely affected ranking (95.1% to 95.8% for n from 5 to 50), but it sets the conservative floor: a perfect sample gives a lower bound of 0.57 at $n = 5$, 0.72 at 10, 0.84 at 20 and 0.93 at 50. Use $n \geq 10$ for internal ranking and $n \geq 20$ where the conservative figure will be reported externally.

5.3 Critical-item caps

Caps stop strong scores elsewhere from masking a fundamental failure. In simulation they made no material difference once a records sample was taken, but without a sample they improved accuracy where courts look for the weakest link and where paperwork outruns practice (section 8):

- if criterion 1 (current, approved written procedure) scores 0, then $C_d \leq 0.20$;
- if no records of operation exist for the period, then $S = 0$ and $C_d = 0$;
- if criterion 7 (records of each operation) scores 0 or 1, then $C_d \leq 0.40$.

5.4 Screening and audit grades

The ten-question yes/no self-test in the companion handbook gives a **screening** C_d equal to the number of yes answers divided by 10. It is a prioritisation aid and must not be reported as an audit result. In simulation it still captured 92.2% of the achievable reduction, more than any conventional method tested. An **audit-grade** C_d requires the full protocol (sections 5.1 to 5.3), an assessor competent and independent of the control's owner in the sense of ISO 19011:2018, and retention of the sampled evidence.

6. Estimating the other inputs and the uncertainty

Inputs should come from data wherever possible. C_e should come from testing, drills, inspection results or incident history. T and V should come from incident and sector data. I should come from loss history, insurance data and the penalty ranges set by sentencing guidance. Where data are thin, use structured expert judgement and record each input as three points: a low value, the most likely value and a high value, chosen so that the expert is 90% sure the truth lies between them.

Propagate uncertainty by Monte Carlo simulation with at least 10,000 draws from triangular or PERT distributions, and report the 10th, 50th and 90th percentiles of L_r and D together with a rank-correlation sensitivity analysis. This follows the uncertainty guidance in IEC 31010:2019. Single point values should not be presented without their range.

Ranges must be generous. In simulation, intervals built from correctly sized expert ranges contained the true total only 64% of the time against a stated 80%; when experts were overconfident by half, 41%. Widening ranges by half raised this to 74%. Report percentile ranges as indicative until they have been checked against real outcomes (Study 5), and use calibration training for estimators (Lichtenstein, Fischhoff and Phillips, 1982).

Spend estimation effort where it matters. A variance-based (Sobol) sensitivity analysis attributed most of the variation in L_r to impact (total-effect index 0.74), threat (0.41) and vulnerability (0.23), and less to C_d (0.08) and C_e (0.01) (Figure 3). Impact and threat estimates therefore deserve the most care. C_d matters mainly for deciding which controls to fix, not for the size of the total.

7. Worked example

This example is illustrative: it is the biometric access control from the original paper, given a consequence of NZ\$1,000,000. Take $T = 0.9$, $V = 0.8$ and $C_e = 0.95$, so $E = \text{NZ\$}720,000$ and $R = \text{NZ\$}36,000$.

	C_d	Demonstrable mitigation M	Void D	Residual liability L_r	L_r / E
Before: no enrolment criteria or review records	0.10	NZ\$68,400	NZ\$615,600	NZ\$651,600	90.5%
After: Q = 0.90, S = 0.80	0.72	NZ\$492,480	NZ\$191,520	NZ\$227,520	31.6%
Complete documentation	1.00	NZ\$684,000	NZ\$0	NZ\$36,000	5.0%

All figures satisfy $E = R + D + M$. Without any change to the control itself, documentation reduces residual liability by NZ\$424,080, which is the decrease in the void. If the same procedure were also subject to a stand-alone record-keeping duty with a 25% annual chance of inspection and a NZ\$40,000 penalty, then $A = \text{NZ\$}10,000$, and residual liability would be NZ\$660,600 before and NZ\$230,320 after.

8. Simulation evaluation

A model for real decisions should be chosen by evidence, not preference. Before validation data exist, the accepted way to compare candidate methods is a simulation study in which the truth is known by construction. This study followed the ADEMP framework (Morris, White and Crowther,

2019). The protocol, including the rule for choosing the winning model, was written before any results were generated (Appendix E). Code and results are published so that anyone can reproduce or challenge them.

8.1 Design

Worlds. Seven legal environments define how much credit a court gives for documentation: proportional credit; a threshold, below which documentation earns almost nothing; a weakest link, in which credit depends on the worst of approval, records and review; partial credit for unrecorded practice given in testimony; strong correlation between documentation and control quality; a stand-alone records offence; and paper-only gaming, in which a quarter of procedures have perfect paperwork but poor records. Each world contained 2,000 organisations with 20 controls each.

Assessors. Every model saw the same imperfect inputs. Expert estimates carried error ($\sigma = 0.3$ on the logit or log scale; 0 and 0.6 in sensitivity runs), criteria were scored with rater error, and records were sampled with $n = 20$ (5 to 50 in sensitivity runs).

Candidates. Fourteen models: inherent and residual 5×5 risk matrices, as used widely in practice; inherent expected loss; residual expected loss crediting control effectiveness only, which is the logic of FAIR-style quantification (Freund and Jones, 2014); Version 1; Version 2.0 with each way of measuring C_d ; a version with a guessed threshold curve; and Version 2.1.

Measures. The primary measure is **decision value**: an organisation fixes the documentation of the five controls a model ranks highest, and the measure is the share of the best achievable liability reduction it obtains. Secondary measures are the rank correlation with true liability, the share of the true top five identified and calibration, the absolute log ratio of estimated to true total liability (0.10 is about 10% error). The largest Monte Carlo standard error of any reported mean is 0.008.

8.2 Results

Model	Decision value (mean)	Worst world	Rank correlation	Top-5 precision	Calibration error
Lr v2.1 (selected)	94.5%	91.7%	0.92	82%	0.17
Lr v2.0, $C_d = Q \times S$	93.4%	87.9%	0.91	81%	0.19
Lr v2.0, screening C_d	92.2%	87.9%	0.90	80%	0.24
Expected loss, inherent ($T \times V \times I$)	89.5%	83.8%	0.87	77%	0.37
Risk matrix 5×5, inherent	85.2%	80.7%	0.82	72%	–
Expected loss, residual (FAIR-style)	84.3%	80.0%	0.86	75%	0.90
Risk matrix 5×5, residual	81.2%	77.5%	0.81	71%	–
Lr v1 ($T \times V - C_e \times C_d$)	53.3%	48.2%	0.33	40%	–

Means across the seven worlds at $\sigma = 0.3$. Calibration is shown for models that produce a currency amount. Full results by world are in the published results file.

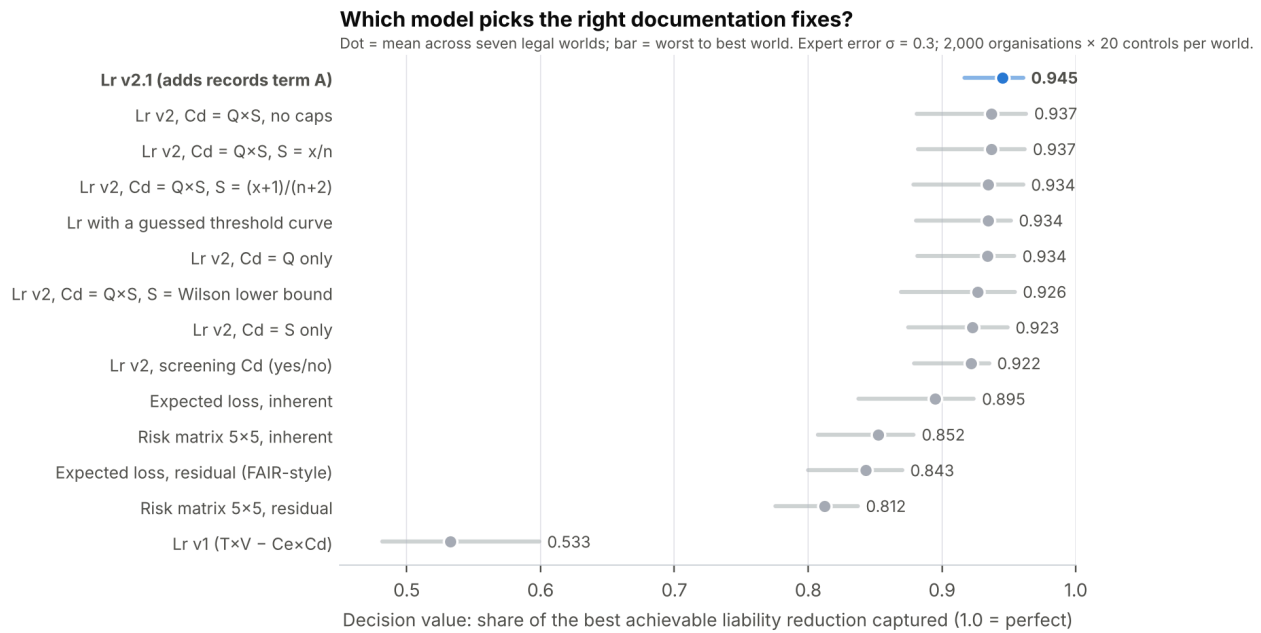


Figure 1. Decision value by model. Version 2.1 is highest on average and in its worst world.

Version 1 fails. Its ranking correlated only 0.33 with true liability. Multiplying it by impact made matters worse (correlation -0.11 to -0.20), so the subtractive form, not just the missing impact term, is the fault.

Documentation-aware models beat conventional methods. Every Version 2 variant beat every conventional method on all measures. The best conventional method, inherent expected loss, captured 89.5% of the achievable reduction; Version 2.1 captured 94.5%. Residual expected loss underestimated total liability by a factor of about 2.5 (calibration error 0.90), because it credits controls that could not be demonstrated. Residual 5×5 matrices were least accurate of the conventional methods, consistent with Cox (2008).

Both parts of C_d earn their place. Design quality alone and records alone each did worse than their product, and the records sample mattered most where paperwork outran practice (95.0% without it, 96.1% with it, in the gaming world).

The threshold question is open. A threshold curve helped when courts really behaved that way (94.7% against 92.0%) and hurt when they did not, with the same average. The linear form is therefore kept as the safer choice until Study 3 estimates the real curve from sentencing data.

Robustness. The ordering of methods held at every level of expert error (Figure 2). All methods degrade as estimates worsen, which is why section 6 asks for data over opinion wherever possible.

Accuracy as expert estimates get worse

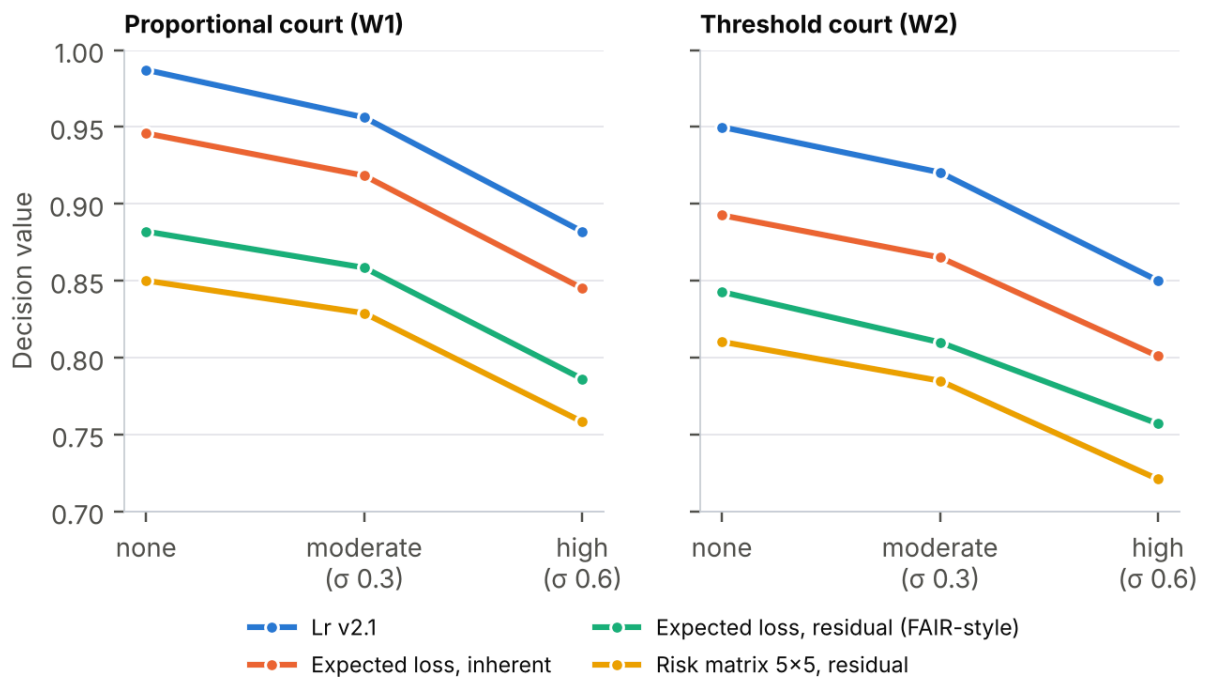


Figure 2. Decision value as expert error rises, in the proportional and threshold worlds.

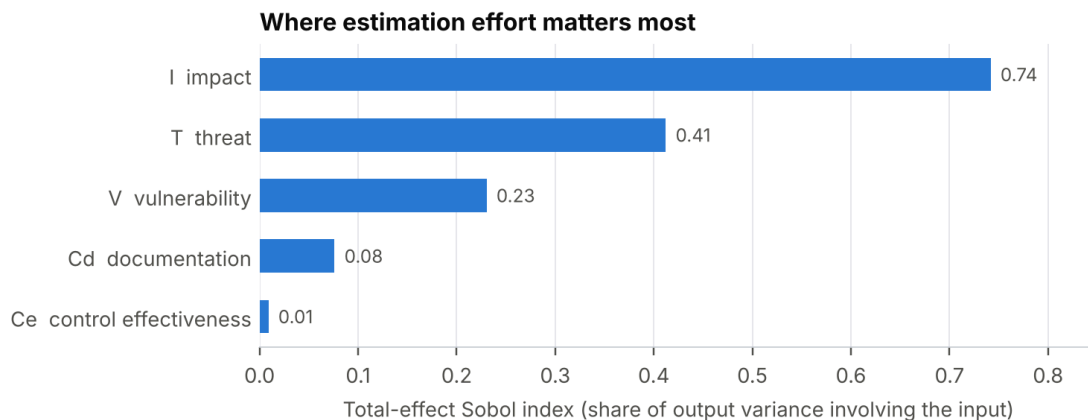


Figure 3. Total-effect Sobol indices for L_r (Saltelli and colleagues, 2010; Jansen, 1999), 200,000 base samples. Indices sum to more than 1 because the inputs interact multiplicatively. Shares depend on the assumed input ranges.

8.3 Selection

Under the pre-specified rule, Version 2.1 had the highest mean decision value (94.5%). Its margin over the next model (93.7%) was within the 0.01 tie band, so the tie-break applied: the best worst-case world. Version 2.1 scored 91.7% there, against 88.2%, and was selected. The records term costs nothing when no records duty applies, because A is then zero.

What this does and does not show. The simulation shows that, if courts and regulators behave in any of these seven ways, Version 2.1 ranks documentation priorities and estimates exposure better than the methods in common use. It does not show that real courts behave in these ways, or that the model predicts real penalties. That is the purpose of Study 3.

9. Alignment with standards

Standard or duty	How Version 2.1 aligns
ISO 31000:2018 Risk management – Guidelines	The model sits in risk analysis and evaluation, and supports treatment and monitoring. L_r and D are evaluation outputs; documentation is a treatment option.
IEC 31010:2019 Risk assessment techniques	Version 2 is specified as a technique, with inputs, outputs, uncertainty treatment and stated limitations, as the standard's descriptions require.
ISO 31073:2022 Risk management – Vocabulary	Terms such as risk, consequence, likelihood and control are used in their standard senses.
ISO 9001, 14001, 45001 and ISO/IEC 27001 (clause 7.5)	Q criteria 1, 2, 5 and 10 put the requirements for creating, updating and controlling documented information into practice.
ISO 10013:2021 Guidance for documented information	The source of good practice for document structure, approval and control that informs the Q anchors.
ISO 19011:2018 Auditing management systems	Governs assessor competence and independence, audit evidence and sampling for S.
ISO 37301:2021 Compliance management systems	The model extends to compliance controls: the same evaluation of design and operating evidence applies.
COSO Internal Control – Integrated Framework (2013)	C_d corresponds to evidence of control activities and monitoring activities.
PCAOB AS 2201; ISAE 3402 (Type 2)	The split of C_d into Q and S mirrors the distinction between test of design and test of operating effectiveness.
Officers' due diligence: NZ HSWA 2015 s 44(4); model WHS Act (Australia) s 27(5)	Criteria 7 to 10 and the S sample evidence the duty to verify that resources and processes are provided and used.

Clause references to management system standards relate to the common harmonized structure for ISO management system standards. This crosswalk is the author's mapping; it implies no endorsement by any standards body.

10. Validation programme

The programme follows the validity framework of the *Standards for Educational and Psychological Testing* (AERA, APA and NCME, 2014). Evidence is gathered on content, response processes, reliability, relations with other variables, and the consequences of use. Every study is pre-registered on the Open Science Framework before any data are collected. Protocols, coding frames, anonymised data and analysis code are published under CC BY 4.0 with a DOI, and all pre-registered outcomes are reported.

Study	Hypothesis and success criterion	Design and sample	Analysis and standards
1. Content validity	H1: at least 80% of the ten criteria reach an item content validity index (I-CVI) of 0.78 or more, and the scale average (S-CVI/Ave) is 0.90 or more	Modified Delphi study, 2 to 3 anonymous rounds, with 10 to 15 experts: health and safety lawyers, lead auditors, former regulators and insurers. Experts rate each criterion's relevance on a 4-point scale and set weights by swing weighting	Lynn (1986); Polit and Beck (2006). Criteria failing the threshold are revised or removed and re-rated

Study	Hypothesis and success criterion	Design and sample	Analysis and standards
2. Inter-rater reliability	H2: ICC(2,1) for absolute agreement on C_d is 0.75 or more, with a lower 95% bound above 0.50; Fleiss' kappa for each criterion is 0.60 or more	30 procedures from consenting organisations, or realistic anonymised composites, each scored independently by 3 trained assessors who cannot see each other's scores	$n = 29$ is required for $k = 3$, $\rho_0 = 0.50$, $\rho_1 = 0.75$, $\alpha = 0.05$ and power 0.80 (Walter, Eliasziw and Donner, 1998). Analysis per Shrout and Fleiss (1979) and Koo and Li (2016); reporting per GRRAS (Kottner and colleagues, 2011); Bland–Altman plots
3. Retrospective outcomes	H3: in published sentencing decisions, a higher coded documentation deficiency is associated with a higher culpability band, after adjusting for harm, organisation size and sector	Published New Zealand District Court sentencing decisions in WorkSafe prosecutions, and cases on the UK HSE public register of convictions, over a defined window; target $n \geq 150$. Two blinded coders apply the coding frame (Appendix C). NZ outcome: Stumpmaster culpability band; UK outcome: culpability category under the Sentencing Council guideline	Proportional-odds ordinal logistic regression with at least 10 cases per parameter; inter-coder kappa of 0.70 or more required before analysis. Reported per STROBE (von Elm and colleagues, 2007). Interpreted as association only (section 11). Also compares a linear with a threshold (logistic) documentation-credit curve by AIC, to settle the open question in section 8
4. Known groups	H4: organisations with a certified ISO 45001, 9001 or 27001 system score higher C_d than comparable uncertified ones (one-sided test; expected effect size $d \geq 0.5$)	Paired comparison matched by sector and size; about 50 procedures per group for 80% power	Welch t-test or Mann–Whitney U test; effect size with a 95% confidence interval
5. Prospective pilot	H5: C_d rises after the 30-day plan, and nonconformities in external audits fall within 12 months, compared with a matched comparison group	10 to 20 volunteer organisations; before-and-after measurement with a comparison group where feasible	Mixed-effects models; pre-registered primary outcome; all outcomes reported, including null results

Response processes

Before Study 2, run think-aloud interviews with 5 to 8 intended users to confirm that each criterion and anchor is understood as intended, and revise the wording where it is not.

Ethics, privacy and independence

Studies 1, 2, 4 and 5 involve people and organisations. They require approval from an accredited human research ethics committee (for example through a university partner), informed consent, confidentiality agreements and data minimisation consistent with the Privacy Act 2020 (New Zealand) and, where relevant, the GDPR. Study 3 uses published judgments only.

To offset the author's commercial interest:

- an independent statistician conducts the analyses;
- coders and assessors are not employed by the author's consultancy;
- the competing interest is declared in every output.

Indicative timeline

Months	Activity
0–2	Pre-registration; ethics application; recruit expert panel and partner organisations
2–5	Study 1 (Delphi) and response-process interviews; finalise criteria and weights
4–8	Study 3 (retrospective coding); Study 2 (reliability)
8–10	Study 4; analysis; manuscript submission to a peer-reviewed journal such as <i>Safety Science</i> , <i>Journal of Risk Research</i> or <i>Risk Analysis</i>
10–22	Study 5 (12-month follow-up)

11. What may be claimed, and when

Evidence in hand	Claims that are supportable
This specification only (now)	"A mathematically consistent framework proposed in a working paper". Results are illustrative.
The simulation evaluation (now)	"In a pre-specified simulation study across seven assumed legal environments, the model prioritised documentation work better than risk matrices and expected-loss methods". Always state that this is a simulation result.
Studies 1 and 2 meet their criteria	"A content-validated instrument with demonstrated inter-rater reliability".
Study 3 meets its criterion	"Documentation deficiencies coded with the instrument are associated with more severe sentencing outcomes in published cases". This is association, not proof of cause.
Studies 4 and 5 meet their criteria and the work is peer reviewed	"Evidence that the method distinguishes stronger systems and that its use improves documentation".

12. Limitations and competing interests

- **Simulation results are conditional.** They depend on the assumed worlds, input distributions and error levels, which are plausible but not measured. They do not replace empirical validation.
- **It is not a loss prediction.** Version 2.1 is a defensibility index. Its link to legal outcomes is untested until Study 3 is complete.
- **Study 3 has built-in biases.** It sees only prosecuted cases (selection bias), and courts' narratives may emphasise documentation after the fact. Its results can show association, not causation.
- **Layered controls rely on an independence assumption** in equation 7. The dependent bound is reported alongside it for that reason.
- **Expert-elicited inputs carry bias.** Calibration training and three-point estimates reduce but do not remove it.
- **The law differs between jurisdictions.** The model does not encode any particular legal test.
- **Competing interest.** The author is the model's originator and operates Confidential Risk Consulting, which offers assessments using it. The independence measures in section 10 are intended to mitigate this.

References

1. Agresti, A. and Coull, B. A. (1998). Approximate is better than “exact” for interval estimation of binomial proportions. *The American Statistician*, 52(2), 119–126.
2. AERA, APA and NCME (2014). *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association.
3. Cox, L. A. (2008). What’s wrong with risk matrices? *Risk Analysis*, 28(2), 497–512.
4. COSO (2013). *Internal Control – Integrated Framework*. Committee of Sponsoring Organizations of the Treadway Commission.
5. European Union (2016). Regulation (EU) 2016/679 (General Data Protection Regulation), Articles 30 and 83(4).
6. Freund, J. and Jones, J. (2014). *Measuring and Managing Information Risk: A FAIR Approach*. Oxford: Butterworth-Heinemann.
7. IEC 31010:2019. *Risk management – Risk assessment techniques*.
8. IAASB. ISAE 3402, *Assurance Reports on Controls at a Service Organization*.
9. ISO 10013:2021. *Quality management systems – Guidance for documented information*. [iso.org/standard/75736.html](https://www.iso.org/standard/75736.html)
10. ISO 19011:2018. *Guidelines for auditing management systems*.
11. ISO 31000:2018. *Risk management – Guidelines*.
12. ISO 31073:2022. *Risk management – Vocabulary*.
13. ISO 37301:2021. *Compliance management systems – Requirements with guidance for use*.
14. Jansen, M. J. W. (1999). Analysis of variance designs for model output. *Computer Physics Communications*, 117(1–2), 35–43.
15. Kaplan, S. and Garrick, B. J. (1981). On the quantitative definition of risk. *Risk Analysis*, 1(1), 11–27.
16. Koo, T. K. and Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, 15(2), 155–163.
17. Kottner, J. and colleagues (2011). Guidelines for Reporting Reliability and Agreement Studies (GRRAS) were proposed. *Journal of Clinical Epidemiology*, 64(1), 96–106.
18. Lichtenstein, S., Fischhoff, B. and Phillips, L. D. (1982). Calibration of probabilities: the state of the art to 1980. In D. Kahneman, P. Slovic and A. Tversky (eds), *Judgment under Uncertainty: Heuristics and Biases*, 306–334. Cambridge University Press.
19. Lynn, M. R. (1986). Determination and quantification of content validity. *Nursing Research*, 35(6), 382–385.
20. Morris, T. P., White, I. R. and Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in Medicine*, 38(11), 2074–2102.
21. Occupational Safety and Health Administration. 29 CFR Part 1904, Recording and Reporting Occupational Injuries and Illnesses.
22. PCAOB. Auditing Standard AS 2201, *An Audit of Internal Control Over Financial Reporting That Is Integrated with An Audit of Financial Statements*.
23. Polit, D. F. and Beck, C. T. (2006). The content validity index: are you sure you know what's being reported? *Research in Nursing & Health*, 29(5), 489–497.
24. Prouse, R. G. (2026). *Bridging the Documentation Void: A Quantitative Approach to Residual Liability (Lr) and Fiduciary Defensibility*. SSRN working paper. DOI 10.2139/ssrn.6175099.
25. Saltelli, A., Annoni, P., Azzini, I., Campolongo, F., Ratto, M. and Tarantola, S. (2010). Variance based sensitivity analysis of model output: design and estimator for the total sensitivity index. *Computer Physics Communications*, 181(2), 259–270.
26. Shrout, P. E. and Fleiss, J. L. (1979). Intraclass correlations: uses in assessing rater reliability. *Psychological Bulletin*, 86(2), 420–428.
27. Stumpmaster v WorkSafe New Zealand [2018] NZHC 2020 (sentencing guideline bands for the Health and Safety at Work Act 2015).
28. von Elm, E. and colleagues (2007). The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement. *The Lancet*, 370(9596), 1453–1457.

29. Walter, S. D., Eliasziw, M. and Donner, A. (1998). Sample size and optimal designs for reliability studies. *Statistics in Medicine*, 17(1), 101–110.
30. Wilson, E. B. (1927). Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158), 209–212.
31. Data sources for Study 3: District Courts of New Zealand, published judgments (districtcourts.govt.nz/all-judgments); Health and Safety Executive, public register of convictions (resources.hse.gov.uk/convictions).

Appendix A. Design-quality criteria and anchors (Q)

Criterion	0	1	2	3
1. Written and approved	None exists	Draft, informal or unapproved	Approved, but approver or date incomplete	Current, approved by a named authority, dated, version-controlled
2. Reviewed within 12 months	Never reviewed, or unknown	Last review more than 24 months ago	12–24 months, or review not recorded on the document	Reviewed within 12 months, recorded on the document
3. Names the threats	No purpose stated	Generic purpose only	Names hazards, not linked to the risk register	Names specific threats, linked to the risk register
4. Responsibilities by role	None assigned	Named individuals only, or partial	Roles assigned for most steps	Every step has a role owner and a deputy
5. Accessible within 2 minutes	Cannot be located	Takes more than 2 minutes, or not at the point of use	Accessible during working hours only	Current version only, at the point of use, at all times
6. Training recorded	No records	Some records, version not stated	Records for most users, version not stated	All users, dated, specific to the version
7. Records of each operation	None	Sporadic	Most occurrences, with gaps	Every occurrence: complete, attributable, contemporaneous
8. Independent review of records	None	Ad hoc, unrecorded	Scheduled but irregular	Scheduled, by a second person, recorded, actions tracked
9. Tested or exercised	Never	More than 24 months ago	Within 24 months, or results unrecorded	Within 12 months, results and actions recorded
10. Version history	None	Versions exist, no change record	Change record incomplete	Complete history: reasons, approvals, superseded versions kept

Q = sum of the ten scores ÷ 30. Weights are equal until Study 1 sets them. Record the evidence behind every score.

Appendix B. Operating-evidence sample (S)

How often the control operates	Indicative sample size n
Annual	1
Quarterly	2
Monthly	2–5
Weekly	5–15

How often the control operates	Indicative sample size n
Daily	20–40
Several times a day	25–60

Indicative values reflecting common audit practice. The assessor sets and justifies n in the audit plan, draws the sample at random from the whole period, and tests each item for completeness, attribution, contemporaneity and review. Simulation supports $n \geq 10$ for ranking and $n \geq 20$ where the conservative value is reported externally (section 5.2).

Appendix C. Coding frame for Study 3

Field	Coding rule
Case identifiers	Court, citation, date, jurisdiction, offence section, sector, organisation size (employees or turnover band)
Harm	Fatality / serious injury / risk of harm only; number of people affected
Documentation findings (criteria 1–10)	For each criterion, coded as deficiency found, adequacy found, or not mentioned, using the Appendix A definitions. A deficiency counts only where the judgment refers to it explicitly
Documentation deficiency score	Deficiencies found ÷ criteria mentioned (primary); count of deficiencies (secondary)
Outcome	NZ: Stumpmaster culpability band (low up to \$250,000; medium \$250,000–\$600,000; high \$600,000–\$1,000,000; very high \$1,000,000 or more) and starting-point fine. UK: Sentencing Council culpability category and fine
Covariates	Guilty plea, previous convictions, reparation, remedial action, other mitigating and aggravating factors as stated
Coder notes	Quotation supporting each code (paragraph number); coder confidence

Appendix D. Pre-registration checklist (OSF)

- Hypotheses H1–H5 and their success criteria, exactly as worded above.
- Sampling plan, inclusion and exclusion rules, target sample sizes and stopping rules.
- The instrument version used, with frozen criteria, anchors and weights.
- The analysis plan: models, covariates, handling of missing data, sensitivity analyses and multiplicity.
- Assessor and coder training, blinding procedures and agreement thresholds.
- Ethics approval reference, consent materials and the data management plan.
- Competing interests, and the independent roles that address them.

Appendix E. Simulation protocol summary (ADEMP)

Element	Specification
Aims	Choose the model form, the measurement of C_d , the sample size and the uncertainty guidance that best support documentation decisions

Element	Specification
Data-generating mechanisms	Seven legal worlds (section 8.1). $T \sim \text{Beta}(1.5, 8)$; $V \sim \text{Beta}(2, 3)$; $I \sim \text{lognormal}(\ln 150,000, 1.0)$; $C_e = \text{logistic}(1 + 0.8z)$; documentation correlated with C_e ($\rho = 0.3$, or 0.7 in one world); ten criteria scored 0–3; records compliance rate p. 2,000 organisations $\times$ 20 controls per world; fixed random seeds
Estimands	Each control's true residual liability; each organisation's total; the reduction from fully documenting each control
Methods	14 candidate models (section 8.1), all fed the same noisy inputs
Performance measures	Decision value (primary); Spearman rank correlation; top-5 precision; calibration $ \ln(\text{estimated} / \text{true total}) $; Monte Carlo standard errors; interval coverage; Sobol indices
Selection rule (pre-specified)	Highest mean decision value; differences under 0.01 are ties, broken by the best worst-case world, then rank correlation, then fewer inputs

Full protocol, code (Python, NumPy and SciPy) and results: confidentialriskconsulting.com/assets/Lr-v2.1-simulation-code.zip, CC BY 4.0.