



VEYRUM ROBOTICS STANDARD



VRS-GEN-202

Calibration and Validation Protocol

How every (prior) threshold inside a criterion rung is derived before data, fitted with data, validated out-of-sample, and — as an edition change, never a live adjustment — calibrated.

VEYRUM RESEARCH INSTITUTE

DOCUMENT
VRS-GEN-202

VERSION
1.0.0

STATUS
Draft

DATE
16 September 2026

Contents

1

Introduction

2

1 Scope

3

2 Normative references

4

3 Terms and definitions

5

4 The calibration ladder

6

5 Anchor derivation ($S_0 \rightarrow S_1$)

7

6 Structured elicitation ($S_0 \rightarrow S_1$ where no anchor exists)

8

7 Synthetic stress tests (required for every edition, pre- and post-data)

9

8 Fitting ($S_1 \rightarrow S_2$)

10

9 Out-of-sample validation ($S_2 \rightarrow S_3$)

11

10 Calibration – dropping the (prior) marker

12

11 Calibration evidence report

13

12 Independence and anti-tuning controls

14

13 Toolkit mapping (informative)

15

Bibliography

16

Change history

Foreword. In the VRS 2026 model the only numbers this Protocol touches are the **(prior)** thresholds that sit inside criterion rungs — a failure rate in a reliability rung, an availability percentage, a lead time, an adherence share (VRS-GEN-002 §7.5.8). Points, point ranges, ceilings and letter meanings are frozen by the stability contract (VRS-GEN-004 §8) and are out of scope here. A **(prior)** is the launch-edition value of such a threshold: honest about being unproven, and marked in the rung so a reader sees it. This Protocol says how a **(prior)** becomes proven: the **anchor-derivation** method that ties a threshold to a stated, citable reference; **structured elicitation** where no anchor exists; the **synthetic stress tests** every edition must pass before any data exists; the **fitting** and **out-of-sample validation** procedure once loss data exists; and the rule that a validated threshold loses its **(prior)** marker only by an **edition change** (VRS-GEN-004 §4.5), never by adjusting a live rating. It binds the independence rule of VRS-GEN-003: thresholds are fitted for predictive validity against observed loss and are never tuned to produce a wanted result for any rated party. The steps map one-to-one onto the **calibration/** toolkit that ships with each methodology edition. Requirements use “shall”.

1 Introduction

A rating scheme earns trust in exactly one way: its numbers predict outcomes. The precedent corpus behind the VRS (credit ratings, vehicle safety ratings, reliability classification) is unanimous that methodologies without outcome validation are attacked as arbitrary, and equally unanimous that a rater who tunes numbers to please the rated is finished. The VRS therefore treats calibration as a governed, versioned, published process bound to the edition cycle rather than a one-time fitting exercise. This Protocol is the method; the IR category standards (VRS-GEN-101...107) and the sector AR protocols (-201) define the criteria and hold the rungs; VRS-GEN-004 makes every threshold change an edition act; and the calibration record that ships with each edition holds every threshold this Protocol produces, with its derivation.

2 1 Scope

This Protocol applies to every numeric **(prior)** threshold inside a criterion rung that influences an IR (VRS-GEN-101...107) or a sector AR letter (VRS-GEN-006 and the sector -201 protocols). It defines:

- the **maturity ladder** a threshold passes through and the marker each stage confers (Clause 4);
- the **anchor-derivation** method for pre-data thresholds (Clause 5);
- the **structured elicitation** method for thresholds no anchor can fix (Clause 6);
- the **synthetic stress tests** an edition shall pass before release (Clause 7);
- the **fitting** procedure once labelled loss data exists (Clause 8);
- **out-of-sample validation** and its acceptance criteria (Clause 9);
- **calibration** — removal of the **(prior)** marker at an edition boundary (Clause 10);
- the **calibration evidence report** (Clause 11);
- **independence and anti-tuning** controls (Clause 12).

It excludes the criterion layout, points, point ranges and ceilings (VRS-GEN-002 §6.5, VRS-GEN-005), the aggregation and tier rules (VRS-GEN-005 §6), the edition instruments themselves (VRS-GEN-004), and the dataset collection rules (VRS-GEN-012). This Protocol changes no rung's points and adds no criterion; it changes only where inside a rung a numeric threshold sits, and only by edition.

3 2 Normative references

- **VRS-GEN-002**, Designation, Drafting and Style Rules (undated) — criterion layout; (**prior**) marking (§7.5.8).
- **VRS-GEN-003**, Governance Charter and Independence Statement (undated).
- **VRS-GEN-004**, Change, Edition and Review Policy (undated) — priors change only by edition (§4.5); the stability contract (§8).
- **VRS-GEN-005**, Intrinsic Rating — Methodology (undated) — point-sum aggregation, the three tiers, cold-start parity (§6).
- **VRS-GEN-006**, Application Rating — Scheme (undated).
- **VRS-GEN-012**, Evidence and Data Requirements (undated) — evidence grades and dataset vocabulary.
- **VRS-GEN-101 ... VRS-GEN-107**, IR Category Standards (undated) — the criterion rungs whose thresholds this Protocol calibrates.
- External anchors (dated at first use in the calibration record): US BLS *Survey of Occupational Injuries and Illnesses* incidence-rate definitions (recordable cases per 200,000 hours); OSHA 29 CFR 1904 recordability criteria; ISO 14224:2016 reliability data collection and exchange (failure-rate and MTBF conventions); IEC 61703:2016 availability and reliability measures; IEC 62443-3-3:2013 security levels SL 1–4; ISO/IEC Guide 98-3:2008 (GUM) for uncertainty statements; Cooke (1991) classical model for structured expert judgement; the Brier score of probabilistic forecast verification.

4 3 Terms and definitions

Terms per VRS-GEN-001 (including **tier**, **edition stamp**, **rung**, **evidence bar** and **prior**). Locally:

- **prior** — a numeric threshold inside a criterion rung, marked (**prior**) per VRS-GEN-002 §7.5.8, whose value has not been calibrated against assessment data; the launch-edition value.
- **anchor** — a published, citable reference quantity or convention from which a threshold can be derived by a stated rule (e.g. a sector benchmark failure rate).
- **anchor statement** — the rule, in words and in the calibration record, that maps an anchor to a named rung of a named criterion. Anchor statements are the auditable unit of derivation.
- **elicitation** — the structured collection of expert judgements to set a threshold no anchor can fix, per Clause 6.
- **labelled loss dataset** — the set of rated models with observed outcomes (incidents, losses, failures, claims) over a known exposure, graded per VRS-GEN-012.
- **calibration curve** — the relationship between the IR number (or a criterion's awarded points) and the observed outcome rate in the labelled dataset, plotted by band.

- **monotonicity** — the property that a worse input never produces a higher awarded rung and never a higher IR.
- **calibration** — the governed act of moving a threshold to validated maturity and, at the next edition boundary, removing its **(prior)** marker (Clause 10, VRS-GEN-004 §4.5).

54The calibration ladder

4.1 Every threshold that a criterion rung places a number on shall carry exactly one of four **maturity stages** in the calibration record:

STAGE	MEANING	MARKER IN THE RUNG	EDITION DISPOSITION
S0	Arbitrary value with no stated derivation	—	Not permitted in any released edition
S1	Derived from an anchor statement (Clause 5) or elicitation (Clause 6); passes the stress tests (Clause 7)	(prior)	Publishable as a launch-edition threshold
S2	Fitted to the labelled loss dataset (Clause 8)	(prior)	Publishable; remains a prior until validated
S3	Fitted and passing out-of-sample validation (Clause 9)	marker removed at the next edition	Eligible for calibration (Clause 10)

4.2 The stage shall be recorded per threshold as `{ criterion, rung, value, unit, stage, derivation, anchors[], fitted_on, validated_on }` in the calibration record. A threshold without a stage field is S0 and shall not be released.

4.3 A criterion’s maturity is the minimum stage over its thresholds; an edition’s maturity is the minimum over the thresholds it uses. Both shall be stated in the calibration evidence report (Clause 11).

4.4 A threshold below S3 shall carry the **(prior)** marker in its rung text (VRS-GEN-002 §7.5.8), so that any reader of a criterion table sees which numbers are launch-edition values. The marker is a content element of the rung, not a rating attribute; it does not change the points the rung awards.

4.5 No threshold shall be demoted silently. Moving from S2/S3 back to S1 (for example, the dataset was found defective) is an edition act under VRS-GEN-004 and shall be disclosed in the calibration evidence report.

65Anchor derivation (S0 → S1)

5.1 Where a published reference quantity exists for what a threshold expresses, the threshold shall be derived from it by an **anchor statement** of the form: *“An input value X (from anchor A) shall reach rung R of criterion C.”* The threshold is then the value that places rung R at X.

5.2 An anchor statement shall target a **named rung** of a criterion, never an interior value, so the derivation reads as a policy (“the sector benchmark reaches the mid rung”) rather than as a bare number. Where a criterion’s rungs step (best $\rightarrow$ 0), the anchored rung fixes one threshold and the neighbouring thresholds shall be set by a documented ratio convention (for example, successive halving/doubling), each recorded.

5.3 Where two or more anchors constrain one threshold, the derivation shall use the anchor of the highest evidence grade (VRS-GEN-012); the others become **consistency checks** recorded in the report (Clause 11) with the implied value and the discrepancy.

5.4 Every S1 threshold shall list its anchors with a date and a retrievable citation in the calibration record. An anchor that cannot be retrieved is not an anchor; the threshold reverts to S0 and shall not be released.

5.5 Worked derivation — the RL-3 field-failure-rate rungs (VRS-GEN-102). RL-3 *Field failure rate* awards points on a ladder whose rungs are class-weighted failures per 1,000 operating hours: the top rung at ≤ 0.5 (**prior**), the next at ≤ 1 (**prior**), the mid rung at ≤ 2 (**prior**), the low rung at ≤ 4 (**prior**), and 0 above that. To anchor the ladder, adopt ISO 14224:2016 as the reliability-data convention and take the population benchmark failure rate for the comparable equipment class as the anchor A. Anchor statement: “A model whose field failure rate equals the ISO 14224 population benchmark for its equipment class shall reach the mid rung of RL-3.” If that benchmark is ≈ 2 failures per 1,000 h, the mid threshold is fixed at **2 (prior)** and the neighbouring thresholds follow the halving/doubling convention of 5.2 (1, 0.5 above; 4 below). **Consistency check:** a draft that placed the benchmark model at the top rung (≤ 0.5) would rate an average-reliability model as best in class — which the rung-occupancy test (Clause 7) flags as a distribution failure; the derivation therefore holds the mid rung at the benchmark. (Robots are not the ISO 14224 population; the anchor fixes the *scale of consequence a buyer treats as mid* until the fitted value in S2 replaces it.)

5.6 Availability, security and support thresholds shall be anchored the same way: RL-4 *Availability under duty* (VRS-GEN-102) against the IEC 61703:2016 availability model; cyber thresholds against IEC 62443-3-3:2013 security levels; each recorded as an anchor statement naming the rung it fixes.

7 6 Structured elicitation (S0 $\rightarrow$ S1 where no anchor exists)

6.1 Where no citable anchor exists for a threshold, the threshold shall be set by **structured expert judgement** following the classical model: a panel of ≥ 5 experts drawn from at least three stakeholder classes (risk-bearers, operators, independent engineers), calibration questions with known answers to weight the experts, and performance-weighted aggregation (Cooke, 1991).

6.2 Manufacturers of rated products may participate as **respondents** to seed questions but shall not be **weighted panel members** for thresholds that affect their own products (VRS-GEN-003 independence).

6.3 The elicitation shall record, per threshold: the question as posed, each expert’s quantiles (5/50/95), the calibration score, the aggregated distribution, and the value adopted (the aggregated median unless the report states otherwise).

6.4 The elicitation question shall be framed as loss-relevance (“at what value of this input does expected loss cross the level this rung represents?”), not as importance, so that the resulting threshold is answerable from evidence an underwriter can obtain.

6.5 Until an elicitation has been run, a threshold set by the drafting team shall carry stage S1 with **derivation: "drafting-team judgement, elicitation pending"**, the **(prior)** marker in its rung, and a note to that effect in the report. This is the honest description of the current launch-edition thresholds.

8 7 Synthetic stress tests (required for every edition, pre- and post-data)

7.1 Before an edition is released, the methodology shall pass the following tests on synthetic cohorts generated by the toolkit. Failure of any test blocks release; the report records the results and the cohort specification.

7.2 Monotonicity. For every criterion and every input, holding all else fixed, a worse input value shall never reach a higher rung, and a higher awarded rung shall never yield a lower IR. (Tested by finite-difference sweeps over each input's range.)

7.3 No-evidence handling. A criterion whose evidence bar is not met shall award 0 points — never a null, never a discard, and never a re-scaling of the remaining criteria (VRS-GEN-005). The reachable maximum shall reduce only for criteria not applicable by class (VRS-GEN-002 §7.5.6).

7.4 Rung occupancy. On a synthetic cohort drawn from documented input distributions, no rung of any criterion shall be unreachable, and the top rung of a criterion shall not be reached by more than the share stated in the report. A calibration that places the sector benchmark model at a criterion's top or bottom rung has mis-set that threshold.

7.5 Boundary sanity. The anchor statements of Clause 5 shall reproduce their target rung when the anchor input is fed to the implementation.

7.6 Prior sensitivity. Perturbing every **(prior)** threshold by $\pm 20\%$ shall change the awarded rung for no more than the share stated in the report and shall preserve the IR rank order of the cohort at a Kendall τ of no less than 0.85. Larger movement indicates the IR is an artefact of the threshold rather than of the evidence, and the threshold shall be flagged for priority fitting.

7.7 Evidence-grade honesty. For the same inputs, lowering the evidence grade of every record by one step shall never raise the awarded points and shall never raise the tier.

7.8 Determinism. The same inputs and calibration record shall produce byte-identical outputs (reproducibility requirement of VRS-GEN-005).

9 8 Fitting (S1 → S2)

8.1 Fitting shall begin only when the labelled loss dataset for a criterion meets the **minimum fitting set**: ≥ 30 rated models with $\geq 10,000$ pooled operating hours each and ≥ 1 outcome event per 5 models, all records graded B or better (VRS-GEN-012). Below this, the threshold remains S1.

8.2 The **outcome variable** per criterion shall be pre-registered in the report before fitting (for example, severity-weighted safety events per 1,000 operating hours in the 12 months *following* the rating window). The rating shall be computed from data **before** the outcome window; look-ahead is prohibited.

8.3 Fitting shall minimise a pre-registered loss (rank loss for ordering claims; Brier score for tier-as-probability claims) subject to the monotonicity constraints of 7.2.

8.4 Fitting shall adjust **only** the numeric (**prior**) thresholds inside rungs. It shall not change the points a rung awards, add or remove a rung, or alter a criterion's point range (VRS-GEN-004 §4.5, §4.6). A fitted threshold shall stay within a stated distance of its S1 value (shrinkage — recorded), so that fitting corrects a prior rather than inventing one.

8.5 Fitting shall be performed on a **training partition** only; the held-out partition is reserved for Clause 9 and shall not be inspected during fitting.

10 9 Out-of-sample validation (S2 → S3)

9.1 Validation shall use a held-out partition of $\geq 25\%$ of models, stratified by sector and manufacturer, or a temporal split (rate on year t , outcomes in $t+1$). Temporal splits are preferred when the dataset spans ≥ 2 years.

9.2 Acceptance criteria (all required):

- **Discrimination:** Spearman ρ between the IR number and held-out loss ≤ -0.40 (higher IR, lower loss) with 95 % CI excluding zero; per-criterion $\rho \leq -0.25$ against that criterion's own outcome.
- **Calibration curve:** observed loss rate is monotone across IR bands (a worse band shall show no lower loss) with no adjacent-band inversion.
- **Tier honesty:** discrimination at the Certified tier shall be no worse than at the Unverified tier for the same models; more evidence shall not degrade prediction.
- **Stability:** re-fitting on bootstrap resamples changes no (**prior**) threshold beyond the tolerance stated in the report, and changes the awarded rung for $< 10\%$ of models.

9.3 A threshold that fails validation stays S2; the report shall state which criterion failed and the remediation (more data, a revised rung via amendment under VRS-GEN-004, or a retired input).

11 10 Calibration — dropping the (prior) marker

10.1 A threshold is **calibrated** — its (**prior**) marker removed from the rung — only when it is S3 (validated) and the removal is published as an edition act (amendment or revision, VRS-GEN-004 §4.5). Calibration produces an edition change to a threshold; it shall never adjust a rating between editions.

10.2 Until a threshold reaches S3 it stays a (**prior**) and is displayed as such in the criterion table. A rating that depends on a (**prior**) threshold is unaffected in its points and continues to carry its tier and edition stamp on its face (VRS-GEN-004 §6.3); the launch-edition status of the threshold is disclosed in the calibration evidence report, not on the rating.

10.3 Sector AR thresholds (the **-201** protocols) are calibrated per sector when they are S3 against sector-specific outcomes; the sector's data twin (**-501**) is regenerated, never hand-edited.

10.4 Calibration is an edition act under VRS-GEN-004, announced in the calibration evidence report, and reversible under 4.5. A calibrated threshold that a later dataset contradicts is returned to S1/S2 by a new edition, with its derivation restated.

12 11 Calibration evidence report

11.1 Every methodology edition shall publish a **Calibration Evidence Report** alongside the regenerated data twins, containing at minimum: the maturity table (every threshold, its stage S1 (prior) / S2 fitted / S3 validated / calibrated, derivation and anchors); the elicitation records (Clause 6); the stress-test results (Clause 7) with the cohort specification; for S2/S3 thresholds the dataset description (n models, exposure, outcome counts, grade mix, partition scheme), the fitting objective and result, the validation statistics with confidence intervals, and calibration-curve tables; the list of demotions and their causes; and the maturity of every criterion and sector.

11.2 The report shall be machine-generated by the toolkit from the calibration record and the test outputs so that report and implementation cannot diverge, and shall carry the edition identifier and the record hash.

11.3 The report is a **public document** from the first published edition onward; the underlying dataset is not (VRS-GEN-012 provenance and consent rules govern its release).

13 12 Independence and anti-tuning controls

12.1 Thresholds shall be fitted to **observed loss** and to nothing else. It is prohibited to adjust any threshold with reference to the result it would produce for an identified model, manufacturer, sector participant, client, sponsor, or data partner.

12.2 The fitting partition shall be **blinded** to manufacturer identity during fitting (identifiers replaced by opaque keys). Unblinding occurs only in the validation report, after thresholds are frozen for the edition.

12.3 Any person who has received compensation from a rated party in the preceding 24 months shall not set, fit, or approve a threshold affecting that party's products (VRS-GEN-003).

12.4 Every change to a threshold shall be a versioned edit to the calibration record with a recorded reason and a Clause 5, 6, 8 or 9 derivation. Changes without a derivation are prohibited and shall be reverted.

12.5 A discrepancy between an assessor's expectation and the rung the evidence awards is **calibration evidence**, not grounds to change that model's points; it enters the next fitting cycle as an observation.

14

13 Toolkit mapping (informative)

CLAUSE	TOOLKIT COMPONENT (<code>STAGE CAPTURE / DATA CALIBRATION</code>)
5 Anchor derivation	<code>priors.mjs</code> — sets thresholds from anchor statements; emits derivation records
7 Stress tests	<code>stress.mjs</code> — monotonicity, no-evidence handling, rung occupancy, boundary sanity, prior sensitivity, grade honesty, determinism
8–9 Fitting/validation	<code>fit.mjs</code> (skeleton until the minimum fitting set exists)
11 Report	<code>report.mjs</code> — renders the Calibration Evidence Report from the record + test outputs
4 Stage ledger	<code>stages.mjs</code> — validates every threshold carries a stage; blocks S0 in released editions

15

Bibliography

- US Bureau of Labor Statistics, *Survey of Occupational Injuries and Illnesses* — incidence rate formula (cases × 200,000 ÷ hours worked).
- OSHA, 29 CFR Part 1904 — Recording and Reporting Occupational Injuries and Illnesses.
- ISO 14224:2016, *Petroleum, petrochemical and natural gas industries — Collection and exchange of reliability and maintenance data for equipment*.
- IEC 61703:2016, *Mathematical expressions for reliability, availability, maintainability and maintenance support terms*.
- IEC 62443-3-3:2013, *Industrial communication networks — Network and system security — System security requirements and security levels*.
- ISO/IEC Guide 98-3:2008, *Uncertainty of measurement — Part 3: Guide to the expression of uncertainty in measurement (GUM)*.
- Cooke, R. M. (1991). *Experts in Uncertainty: Opinion and Subjective Probability in Science*. Oxford University Press.
- Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1), 1–3.

16 Change history

DATE	VERSION	STATUS	CHANGE	AUTHORITY
2026-09-16	1.0.0 (draft)	Draft	VRS 2026 rewrite: recast so calibration targets only the (prior) thresholds inside criterion rungs and produces edition changes to priors (VRS-GEN-004 §4.5), never live adjustments; the four-stage maturity ladder retained (S0–S3) with the S3→calibrated step defined as removal of the (prior) marker at an edition boundary; anchor derivation retargeted from tier boundaries to named rungs (worked derivation re-run on RL-3, VRS-GEN-102); weight-fitting, confidence machinery, renormalisation, qualitative levels, class/dataset medians and the bare calibration marker removed; validation protocol kept as the first-edition review method; see VRS-GEN-005 §6 and VRS-GEN-002 §7.5.8.	VRS 2026 rewrite Stage 4
2026-09-10	0.1.0	Draft	Created: four-stage ladder, anchor derivation, elicitation, stress tests, fitting, validation, graduation, evidence report, independence controls; toolkit mapping	CEO direction 2026-09-10 (IR/AR calibration)

About Veyrum Research Institute

Veyrum Research Institute is the research arm of Veyrum — the independent intelligence and trust layer for the robotics industry (research, ratings and registry, advisory, data, and specialist insurance & finance introductions).

Disclaimer

The opinions in this document are independent and current as of the date of publication, based on information believed reliable; they are not guarantees or warranties of safety, fitness, or compliance, and are not engineering, legal, or financial advice. This document may be revised. © 2026 Veyrum.

Contact: hello@veyrum.com · VRS-GEN-202 · Version 1.0.0