TECHNICAL NOTE TN-003 · CRITERICA INTELLIGENCE RESEARCH · NOVEMBER 2026

# Reading a Band: Probability, Calibration and Interval

Criterica Intelligence Research

Corresponding: Jack Cosentino, founder · research@criterica.com

Published 8 November 2026  Updated 8 November 2026  Version 1.0

12-minute read

ABSTRACT

A diligence memo carries probabilities, intervals and a stated reliability. This note defines each one, shows how calibration is measured with a reliability diagram and expected calibration error, explains the public promotion gate of ECE at or below 0.05 on a temporal holdout, and shows why a stated confidence of 78% carries no information until its calibration has been demonstrated.

## IN ONE PARAGRAPH

A probability in a memo is a claim until its calibration is shown. Calibration is agreement between stated probabilities and observed frequencies, read on a reliability diagram and summarised as expected calibration error on a later time period than the model trained on. Intervals are tested by coverage: an 80% band should contain about 80% of outcomes. Ask for the diagram, the coverage and the holdout date before relying on a number.

SERIES
Technical Note TN-003

VERSION
1.0

PUBLISHED
8 November 2026

STATUS
Published

AUTHORS

Criterica Intelligence Research

Corresponding author: Jack Cosentino, founder

WHO THIS IS FOR AND THE DECISION IT INFORMS

**Reader.** Diligence readers at funders, buyers, insurers and lenders, and the courts and regulators who review the methods they rely on.

**Decision.** Whether a probability or band in a memo can be relied on at face value, and what evidence to ask for before it is.

**Cost of getting it wrong.** An uncalibrated 78% that comes true 66% of the time misprices every position it touches by the same margin, in the same direction.

01

## The numbers in a memo

A diligence memo on a legal asset typically carries three kinds of number: a probability, such as the chance a matter resolves before trial; an interval, such as a duration band from 14 to 41 months; and a statement about how much those numbers can be trusted. The first two drive the decision. The third determines whether the first two should.

This note defines each figure precisely, shows how reliability is measured, and explains the public promotion gate a Criterica model clears before any of its figures reach a reader. It is written for the person who has to sign the memo.

02

## Definitions

| TERM | DEFINITION | WHAT A READER DOES WITH IT |
|---|---|---|
| Probability | A number from 0 to 1 attached to a defined event over a defined horizon, for example resolution before trial within 24 months. | Prices and ranks the event across matters. |
| Interval (band) | A range expected to contain the outcome with a stated frequency, for example a P10 to P90 band that should contain 80% of outcomes. | Sizes on the edge, plans on the middle. |
| Calibration | Agreement between stated probabilities and observed frequencies: events given 70% happen about 70% of the time. | Decides whether a probability can be used at face value. |
| Discrimination (AUC) | How well scores separate matters with the outcome from matters without it. | Decides whether the ranking is informative. |
| Expected calibration error | The weighted average gap between stated probability and observed frequency, across bins of matters. | One number summarizing calibration. |
| Coverage | The share of outcomes that fall inside a stated band. | Tests whether a band is as wide as it claims. |

Calibration and discrimination are different properties and a model needs both.[2] A model can rank matters well and still state probabilities that are systematically too high. Another can be well calibrated on average and rank poorly. A memo that reports only one of them reports half the evidence.

03

## Why “confidence 78%” means nothing until calibration is shown

A figure such as “78% likely” or “confidence 78%” looks like a measurement. On its own, it is a claim. The only way to know what a stated 78% is worth is to take every past case where the system said something near 78% and count how often the event happened.

If it happened about 78% of the time, the number is calibrated and can be used directly: priced, summed across a book and compared with other numbers. If it happened 66% of the time, the

number is overconfident by twelve points, and every decision that used it at face value was mispriced in the same direction. Nothing in the figure itself distinguishes the two cases. The evidence is outside the number, in its track record on matters it did not see.

> A probability without its calibration evidence is a confident-sounding opinion.

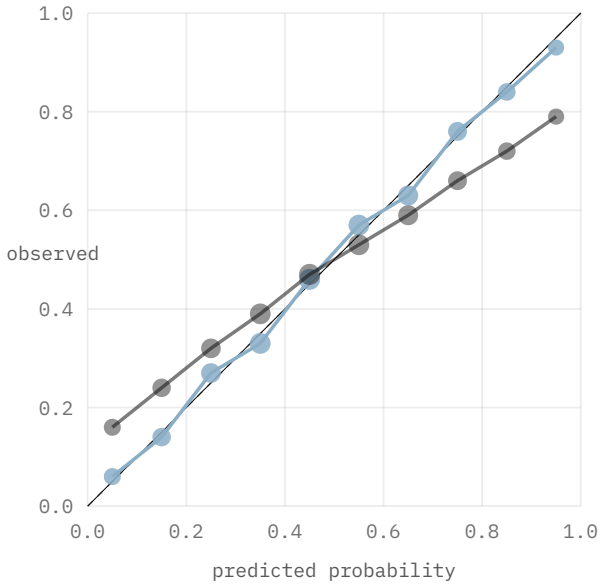
The same applies to labels that are not numbers at all. “High confidence”, “strong case” and a green indicator carry no calibration evidence and cannot be tested. The test that follows can only be run on numbers that were stated in advance and kept.

04

## Reading a reliability diagram

The reliability diagram is the standard picture of calibration. Matters scored on a holdout period are grouped by their stated probability, and for each group the observed frequency of the event is plotted against the average stated probability. A perfectly calibrated model lies on the diagonal.

FIGURE 1 **Reliability diagram for two models on the same holdout** Illustrative




Model A, calibrated
ECE 0.016 · clears the 0.05 gate

Model B, overconfident
ECE 0.064 · does not clear the 0.05 gate

Dot area scales with the number of holdout matters in each bin.

Illustrative dataset: 2,150 fictional matters in ten bins. This figure is evidence of how calibration reads on a diagram. Model A tracks the diagonal. Model B is overconfident: its high scores come true less often than stated and its low scores more often. The diagonal is perfect calibration.

## How to read it

- Points on the diagonal: stated probability matches what happened.
- Points below the diagonal at the high end: the model says more likely than it turns out to be.
- Points above the diagonal at the low end: the model says less likely than it turns out to be.
- A curve flatter than the diagonal, as with Model B, is the signature of overconfidence: the model spreads its scores further from 50% than the evidence supports.

In the illustration, Model B's matters scored near 0.75 resolved the stated way 66% of the time. A reader using Model B's 75% at face value overstates that event by nine points on every matter in the bin. Model A's matters scored near 0.75 resolved that way 76% of the time.

05

# Expected calibration error and the promotion gate

Expected calibration error compresses the diagram into one number. For each bin, take the absolute gap between the average stated probability and the observed frequency, weight it by the share of matters in that bin, and add the bins together.[1]

$$\mathrm{ECE} = \sum_{b} \left(n_b / N\right) \times \left|\, \mathrm{stated}_b - \mathrm{observed}_b \,\right|$$

ILLUSTRATIVE HOLDOUT, BY BIN

| BIN (STATED) | MATTERS | OBSERVED, MODEL A | GAP A | OBSERVED, MODEL B | GAP B |
|---|---|---|---|---|---|
| 0.05 | 120 | 0.06 | 0.01 | 0.16 | 0.11 |
| 0.15 | 180 | 0.14 | 0.01 | 0.24 | 0.09 |
| 0.25 | 240 | 0.27 | 0.02 | 0.32 | 0.07 |
| 0.35 | 300 | 0.33 | 0.02 | 0.39 | 0.04 |
| 0.45 | 320 | 0.46 | 0.01 | 0.47 | 0.02 |
| 0.55 | 300 | 0.57 | 0.02 | 0.53 | 0.02 |
| 0.65 | 260 | 0.63 | 0.02 | 0.59 | 0.06 |
| 0.75 | 200 | 0.76 | 0.01 | 0.66 | 0.09 |
| 0.85 | 140 | 0.84 | 0.01 | 0.72 | 0.13 |
| 0.95 | 90 | 0.93 | 0.02 | 0.79 | 0.16 |

Computed over the illustrative table, Model A has an ECE of 0.016 and Model B 0.064. Criterica's public promotion gate for production classifiers is ECE at or below 0.05 together with AUC at or above 0.70, both measured on a later time period than the model was trained on. Model A would clear the calibration gate; Model B would not, however well it ranked.

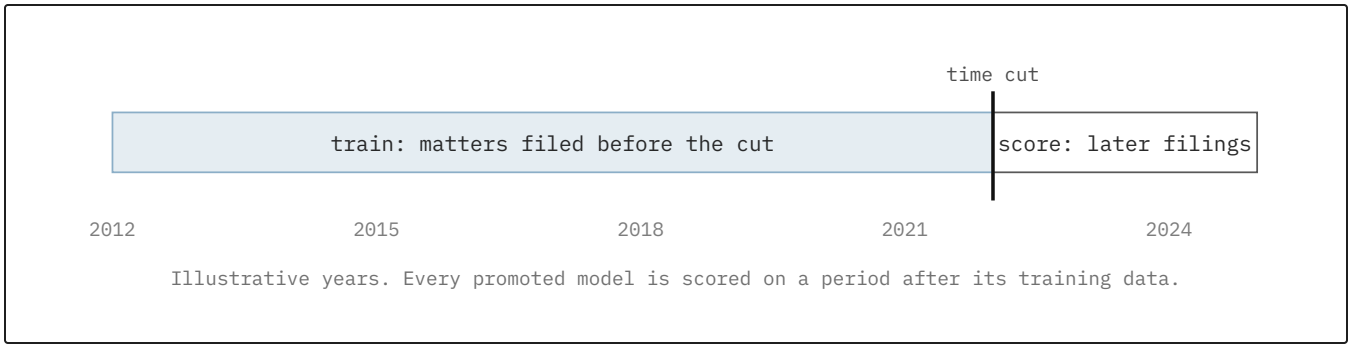
A model that fails the calibration gate is not promoted. Recalibration on held-out data is a standard remedy and is applied where it is appropriate, and the recalibrated model is then scored again on a period it has not seen. The result, pass or fail, is recorded in the model registry.

06

## Temporal holdout: scoring on a period the model never saw

Calibration measured on the same matters a model was trained on is meaningless: a model can memorize its training data. Calibration measured on a random sample of held-back matters is better, but it still mixes eras. Matters from the same year share judges, rules and economic conditions, so a random split lets a model learn the era and then be tested on the same era.

FIGURE 2 **Temporal holdout** Illustrative




The model is trained on matters filed before a time cut and scored only on matters filed after it, the way it will be used. This figure is evidence of the order of operations: train before the cut, score after it.

Criterica validates on a later time period. The model trains on matters filed before a cut and is scored on matters filed after it, which is how it will actually be used: estimating matters that have not resolved yet, in conditions that may have shifted. A model that cannot survive that test does not reach production. It is the closest a validation can come to the decision a reader is making.

07

## Calibrating an interval

Bands are calibrated the same way probabilities are. A P10 to P90 band claims to contain 80% of outcomes. On a holdout period, the test is direct: count the share of resolved matters whose actual outcome fell inside their stated band. A coverage near 80% means the band is as wide as it should be. Coverage well below 80% means the bands are too narrow and the model is overconfident about timing or value; coverage well above means they are wider than necessary.

For a reader, coverage answers the practical question behind every band: if I size on the edge, how often will the matter land beyond it? A calibrated P90 duration means roughly one matter in ten runs longer. That is a reserve assumption that can be written down and tested against the book later.

08

## Four common misreadings

### Reading discrimination as calibration

A high AUC says a model ranks matters well. It says nothing about whether its probabilities are the right size. A memo that cites AUC alone has shown that the ordering is informative, not that a stated 60% means 60%.

### Reading an average as a guarantee

A low ECE is an average over bins. A model can be well calibrated overall and less so in a thin region, often at the extremes where few matters fall. The reliability diagram shows where the evidence is dense and where it is sparse, which is why it should accompany the single number.

### Reading a band as a promise about one matter

An 80% band is calibrated across many matters: about one in five should land outside it. A single matter landing outside its band is not evidence the model failed. A book in which far more than one in five land outside is.

### Reading old validation as current

Calibration is measured on a period. When rules, courts or case mix change, a model validated on an earlier period may drift. The date of the holdout belongs beside every calibration figure, and the figure should be refreshed as new resolutions arrive.

09

## What to ask for in diligence

1. For every probability in the memo, the reliability diagram and ECE of the model that produced it, on a later time period.
2. For every band, its stated coverage and its measured coverage on that period.
3. The AUC of each model, so ranking and calibration are both evidenced.
4. The model version behind each figure, so the figure can be regenerated and later scored against the outcome.
5. The status of each model and, for any stub, the comparison supplied instead.

These are the items a Criterica model card is designed to carry, and cards are shared with institutional reviewers under NDA. MC-002 sets out what a card contains and publishes a sample built from fleet-level figures. WP-006 explains how to act on each figure once its calibration is established.

10

## The question, as the reader has it

A diligence reader at a funder, a buyer, an insurer or a lender is handed a probability and asked to sign. The decision they face is whether to price on it. Their problem is not that probabilities are uncertain; it is that a bare probability gives them no way to find out whether it was earned. A point estimate such as 78% fails this reader because it carries no track record: two systems can both say 78% and mean very different things, and nothing in the figure separates them.

What the reader needs is the evidence beside the number: how often events given roughly 78% actually happened, on matters the system had not seen, and how wide a band must be to contain the outcome with the frequency it claims.

11

## What is measured

| QUANTITY | UNIT | POPULATION |
|---|---|---|
| Probability of an event | A number from 0 to 1 over a stated horizon | Closed matters that could have had the event, from filed court records with a recorded termination |
| Interval (band) | The unit of the outcome, for example months or US dollars, with a stated coverage | Closed matters of the same case type, venue and stage |
| Calibration error | Weighted average gap between stated and observed frequency | Matters scored on a later period than the model was trained on |
| Coverage | Share of outcomes inside the stated band | Matters scored on that later period |

Inclusion is at the level of a recorded termination: a matter counts as an observed outcome only when its record shows how it ended. Matters still open are not outcomes, which is a form of censoring that the duration models handle explicitly and that the event-probability evaluation handles by scoring each event only over its stated horizon.

12

## How a number travels from the record to the band

1. **Record to structured event.** Filings and orders are read into dated events, and parties, counsel and judges are matched across records by entity resolution on a knowledge graph.

2. **Event to model class.** Probabilities come from probabilistic classification; durations from survival analysis in the Kaplan-Meier and proportional-hazards tradition, which uses open matters correctly. Sparse venues borrow strength through hierarchical partial pooling.
3. **Model class to calibrated output.** Classifier scores are recalibrated on held-out matters, because high-capacity models often rank well and state probabilities that are too extreme (see the references).
4. **Output to band shown.** Intervals are fitted as quantile or conformal intervals, the class of method that gives a band a coverage guarantee under a stated exchangeability assumption and a way to test it.

This is the reason the note spends its length on calibration: it is the step that turns a score into a probability a reader may price on.

13

## Validation design, in full

Validation is a design before it is a number. Four choices define it.

1. **What is frozen.** The model, its recalibration and its version are fixed at the time cut. Nothing is adjusted after the later period is examined.
2. **What is compared.** Each stated probability or band is compared with how the matter actually closed in the later period.
3. **How leakage is prevented.** The point-in-time boundary applies to every input, not only the outcome. A field that was recorded after the cut, or that restates an earlier one, is not visible to the model for an earlier estimate.
4. **What is reported.** Three metric families: discrimination, calibration and interval coverage, with the date of the holdout beside each.

For calibration, two instruments are used together. The reliability diagram shows where the model is over- or under-confident. The Brier score, a proper scoring rule, summarises probability accuracy

and decomposes into reliability, resolution and uncertainty components, which separates how honest the probabilities are from how informative the ranking is. A proper rule cannot be improved by shading probabilities away from the model's honest belief, which is why it is the right companion to a binned error.

Figure 1 is evidence of how calibration reads on a diagram: one illustrative model tracks the diagonal and one is overconfident. Figure 2 is evidence of the order of operations in validation: train before the cut, score after it.

14

## Scope and precision

A calibrated probability is a statement about frequencies across many comparable matters. It is not a forecast of the realized dollars of one matter, and a probability of resolution is not a statement about who prevails; direction of outcome is read only from sources that record it.

- Binning: ECE depends on the number of bins, so comparisons use the same binning, and the diagram accompanies the number.
- Thin regions: calibration is least certain where few matters fall, often at the extremes of the probability scale.
- Data lag and venue sparsity: the record reflects filings up to its as-of date, and thin venues widen the band.

15

## What a reader can reproduce

The illustrative dataset behind Figure 1 is published in the table in section 5: ten bins, the matters in each and the observed frequencies for both models. A reader can recompute the expected calibration error for both with a spreadsheet and check it against the figures stated in the text, and can build a reliability diagram of their own from the same columns. Reproducing the calibration of

a production model requires the platform and its held-out matters; the model card for that model states its results and is shared with institutional reviewers under NDA.

## What this number means

The reliability diagram and calibration figures in this note are computed from an illustrative dataset of 2,150 fictional matters scored by two fictional models, chosen to show the method clearly. The promotion gates stated here, AUC at least 0.70 and ECE at most 0.05 on a later time period, are the public gates every Criterica production classifier clears. Calibration results for individual models are provided in their model cards under NDA.


REFERENCES

1. Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. Monthly Weather Review, 78(1), 1–3. doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2
2. Murphy, A. H. (1973). A new vector partition of the probability score. Journal of Applied Meteorology, 12(4), 595–600. doi.org/10.1175/1520-0450(1973)012<0595:ANVPOT>2.0.CO;2
3. DeGroot, M. H., & Fienberg, S. E. (1983). The comparison and evaluation of forecasters. The Statistician, 32(1–2), 12–22. doi.org/10.2307/2987588
4. Gneiting, T., & Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. Journal of the American Statistical Association, 102(477), 359–378. doi.org/10.1198/016214506000001437
5. Niculescu-Mizil, A., & Caruana, R. (2005). Predicting good probabilities with supervised learning. Proceedings of the 22nd International Conference on Machine Learning, 625–632. doi.org/10.1145/1102351.1102430
6. Naeini, M. P., Cooper, G. F., & Hauskrecht, M. (2015). Obtaining well calibrated probabilities using Bayesian binning. Proceedings of the 29th AAAI Conference on Artificial Intelligence, 2901–2907.
7. Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. Proceedings of the 34th International Conference on Machine Learning, PMLR 70, 1321–1330. arxiv.org/abs/1706.04599
8. Vovk, V., Gammerman, A., & Shafer, G. (2005). Algorithmic Learning in a Random World. Springer.
9. Romano, Y., Patterson, E., & Candès, E. J. (2019). Conformalized quantile regression. Advances in Neural Information Processing Systems, 32. arxiv.org/abs/1905.03222
10. Angelopoulos, A. N., & Bates, S. (2023). Conformal prediction: a gentle introduction. Foundations and Trends in Machine Learning, 16(4), 494–591. arxiv.org/abs/2107.07511

11. Kaplan, E. L., & Meier, P. (1958). Nonparametric estimation from incomplete observations. Journal of the American Statistical Association, 53(282), 457–481. doi.org/10.1080/01621459.1958.10501452

12. Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting. Proceedings of the Conference on Fairness, Accountability, and Transparency, 220–229. doi.org/10.1145/3287560.3287596

## NOTES

1. Binned ECE depends on the number of bins. Ten equal-width bins are used throughout this note; any comparison between models should use the same binning.
2. AUC, the area under the receiver operating characteristic curve, is the probability that a randomly chosen matter with the outcome is scored above a randomly chosen matter without it.

Prepared by Criterica Intelligence Research with data engineering, validation and editorial review. Criterica Intelligence Research publishes numbered documents in five series; each is versioned, dated and issued as a new version when it changes, with the change recorded.

## CITE THIS DOCUMENT

Criterica Intelligence Research. (2026). Reading a Band: Probability, Calibration and Interval (Technical Note TN-003, version 1.0). Criterica. https://critericaintelligence.com/research/reading-a-band

```
@techreport{criterica_tn003_2026,
  author = {{Criterica Intelligence Research}},
  title = {Reading a Band: Probability, Calibration and Interval},
  institution = {Criterica},
  type = {Technical Note},
  number = {TN-003},
  year = {2026},
  month = nov,
  note = {Version 1.0},
  url = {https://critericaintelligence.com/research/reading-a-band}
}
```

DOI: pending