

Not Rejected Is Not Proven

Why the absence of evidence against a result is not evidence for it, and what that means for conformity decisions, method comparison, calibration and analyser evaluation in gas analysis

Erik van der Werff · Qualimetric · Version 1.0 · October 2026

Summary

Many decisions in a gas laboratory are framed as yes-or-no questions. Does the gas meet the specification? Do two methods agree? Is the calibration function adequate? Does the analyser perform as required? In practice each question has three possible answers: demonstrated yes, demonstrated no, and not decided. When the third answer is quietly folded into one of the other two, a result that has merely not been rejected is reported as if it had been proven. This paper shows where that happens in four everyday situations, illustrates it with worked numbers, and lists reporting practices that keep the undecided zone visible.

1. Two statements that sound the same

In everyday language, *there is no evidence that A is false* and *there is evidence that A is true* are used interchangeably. In classical probability they would indeed be equal, because the negation of not-A is A itself. They are not, however, statements about the same thing. The first measures the absence of evidence *against* A; the second measures the presence of evidence *for* A. The space between them is the part we do not know.

Evidence theory makes this explicit by giving every proposition two numbers: its *belief*, the support the data provide for A, and its *plausibility*, one minus the support for not-A [8]. With 60 % of the evidence supporting A, 10 % supporting not-A and 30 % undecided, the belief in A is 0.6 and its plausibility 0.9. The gap of 0.3 is quantified ignorance. It shrinks only with better data, never with better wording.

The same structure appears in any survey that reports only the share of respondents who agree: every neutral respondent is silently counted as disagreeing. A fall in agreement from 70 % to 60 % may be nothing more than a shift from agreement to indifference, with the share who actually disagree unchanged. Measurement decisions work the same way:

Evidence supports A	Evidence is inconclusive	Evidence supports not-A
Conformity demonstrated	Undecided at the stated confidence	Non-conformity demonstrated
Agreement demonstrated	Neither agreement nor disagreement shown	Disagreement demonstrated

A binary report has only two boxes, so the middle column must be assigned to one side. Which side it lands on is a decision, not a measurement result, and it should be made explicitly rather than by default. The four sections that follow show what happens when it is not.

2. Conformity: inside the limit is not the same as conforming

Consider a component with an upper specification limit of 0.50 mol%, measured with a standard uncertainty of 0.03 mol% (expanded uncertainty $U = 0.06$ mol%, $k = 2$). Assuming a normal distribution, the probability that the true value is below the limit follows directly from each result [1]:

Result (mol%)	Interval $\pm U$	P(conform)	Simple acceptance	Guarded decision (95 %)
0.40	0.34 – 0.46	99.96 %	Pass	Conformity demonstrated
0.47	0.41 – 0.53	84 %	Pass	Undecided
0.52	0.46 – 0.58	25 %	Fail	Undecided
0.60	0.54 – 0.66	0.04 %	Fail	Non-conformity demonstrated

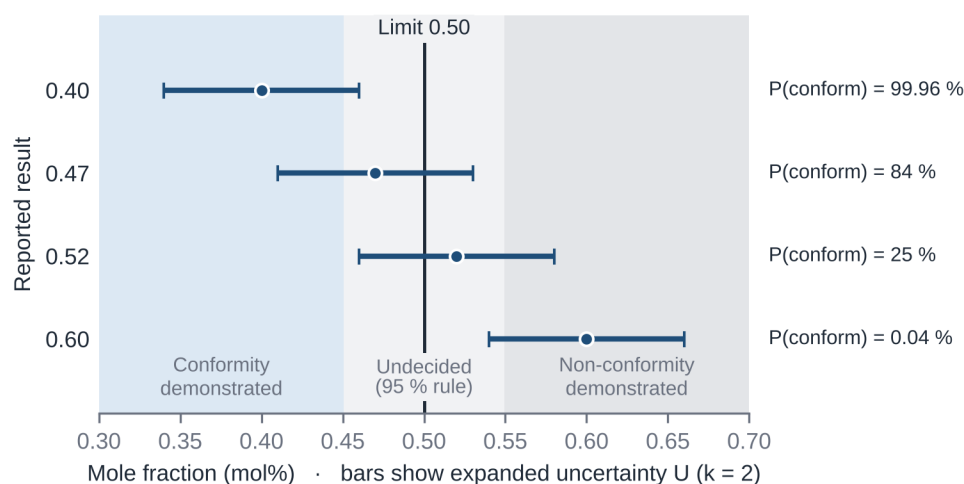


Figure 1. The same four results judged against a single limit. Under simple acceptance only the position of the point matters; with a 95 % decision rule, results between 0.451 and 0.549 mol% are neither accepted nor rejected.

Under simple acceptance, the result of 0.47 mol% is reported as compliant, although there is a 16 % probability that the true value exceeds the limit. The result has not been rejected; it has not been proven either. JCGM 106 handles this with an acceptance limit set inside the specification by a guard band. For 95 % confidence of conformity that acceptance limit is $0.50 - 1.645 \times 0.03 = 0.451$ mol%, and a symmetrical rejection limit lies at 0.549 mol%. Between the two, the honest answer is that the measurement cannot decide.

Which decision rule is appropriate is a contractual choice that depends on who carries the risk. ISO/IEC 17025 does not prescribe one, but it does require that the rule be agreed with the customer and stated whenever a statement of conformity is reported [2, 3]. A report that says only “conforms” without stating the rule hides the middle column of the table above.

3. Method comparison: no significant difference is not agreement

A laboratory compares a new method B with an established method A for superior calorific value, using paired measurements on the same samples. Before looking at the data, the users decide that a mean difference within ± 0.10 MJ/m³ is acceptable for their purpose. Two outcomes are shown in Figure 2.

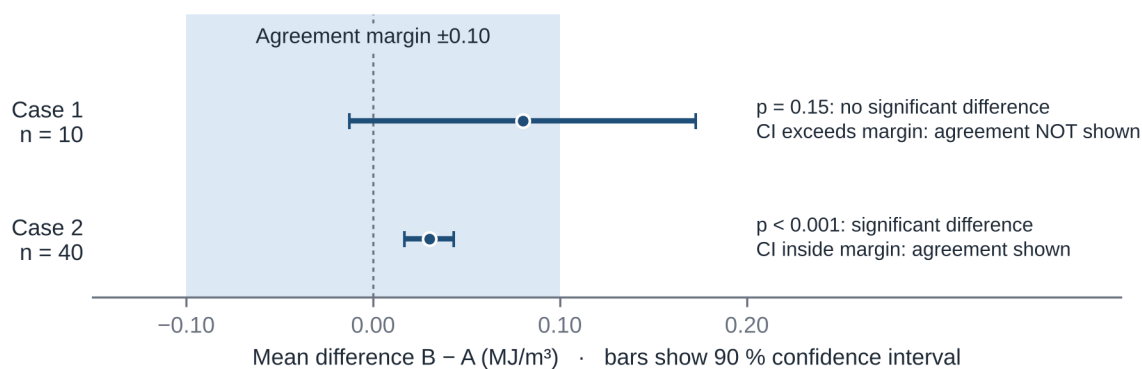


Figure 2. Two hypothetical comparisons. The significance test and the equivalence assessment reach opposite conclusions in both cases.

In case 1 ($n = 10$, mean difference 0.08 MJ/m^3 , standard deviation of the differences 0.16 MJ/m^3) the t-test gives $p = 0.15$: no significant difference. Yet the 90 % confidence interval runs from -0.01 to $+0.17 \text{ MJ/m}^3$, so the data are equally compatible with a bias of 0.17 MJ/m^3 , well outside the margin. Agreement has not been shown; the comparison simply lacked the power to see anything. In case 2 ($n = 40$, mean difference 0.03 MJ/m^3 , standard deviation 0.05 MJ/m^3) the difference is highly significant, $p < 0.001$, but the whole interval from 0.017 to 0.043 MJ/m^3 lies inside the margin. The bias is real, small and correctable, and agreement for this purpose is demonstrated.

A significance test asks whether there is *any* difference. Small, noisy data sets tend to answer no; large, precise ones tend to answer yes, regardless of whether the difference matters. An equivalence assessment reverses the burden of proof: the acceptable margin is fixed first, from the intended use, and agreement is claimed only when the confidence interval lies entirely inside it. Requiring the 90 % interval to fall within the margin is equivalent to the two one-sided tests (TOST) procedure at a 5 % level [7]. ISO 4259-5 addresses the related question of whether two methods that claim to measure the same property agree sufficiently [6]; the underlying principle is the same. Agreement is something to be demonstrated against a predefined criterion, not inferred from a failure to detect a difference.

4. Calibration: an accepted fit is not a proven model

ISO 6143 fits the analytical function by generalized least squares, taking the uncertainties of both the calibration gas compositions and the instrument responses into account [4]. The fit is then judged by a goodness-of-fit parameter, the largest weighted residual, which should not exceed 2. Passing that criterion means that the calibration data do not contradict the chosen function within the stated uncertainties. It does not mean that the function is correct.

The strength of that statement depends entirely on the uncertainties the fit is tested against and on the number of calibration points. With generous uncertainties or only three or four points, a straight line may pass while the detector response is in fact slightly curved. The error then appears where it is hardest to see: in interpolation between the calibration points. Comparing candidate functions, inspecting the pattern of the residuals rather than only their maximum, and using realistic rather than conservative uncertainties all turn “not rejected” into something closer to “demonstrated”.

5. Analyser evaluation: no detected error is not a small error

ISO 10723 evaluates the performance of a natural gas analyser by measuring a set of reference gases that spans the operating range, determining the errors and uncertainties of the analyser and their effect on calculated properties such as calorific value and density [5]. Two statements must be kept apart here. *The evaluation found no error larger than its own uncertainty* is not the same as *the evaluation demonstrates that the error is within the maximum permissible error*.

The first can be true while the second is not, whenever the uncertainty of the reference gases or the repeatability of the analyser is large compared with the requirement. A performance evaluation is only informative when its own uncertainty is small relative to the permissible error it is meant to verify. This becomes more pressing as analysers are extended to biomethane and hydrogen-enriched gas, where additional components and wider ranges are often covered by fewer and less accurate reference gases.

6. Keeping the undecided zone visible

None of this requires new statistics. It requires reporting that does not hide the middle column. In practice:

- **Define the criterion before measuring.** Specification limits, agreement margins and maximum permissible errors follow from the intended use, not from the data.
- **State the decision rule.** Simple acceptance, a guard band or a shared-risk rule are all legitimate choices, provided the report says which one was applied.
- **Allow three outcomes.** Report “not decided at the stated confidence” where that is the truth, rather than forcing the result into pass or fail.
- **Prefer intervals to p-values.** A confidence interval shows both whether there is a difference and whether it matters; a p-value shows only the first.
- **Check that the test can see what matters.** If the uncertainty of the evaluation is comparable with the criterion, a pass carries little information.

7. How NexChrom applies this

NexChrom, Qualimetric’s software for handling gas chromatography data, is built on the standards discussed in this paper, including generalized least-squares calibration according to ISO 6143 and performance evaluation according to ISO 10723, alongside the property calculations of ISO 6976. Its guiding principle is the one set out above: a result used for a decision is reported together with its uncertainty and with the basis on which the decision was made, so that the difference between “not rejected” and “demonstrated” stays visible to the user and to the auditor. More information and demonstrations are available at nexchrom.nl.

References

- [1] JCGM 106:2012, *Evaluation of measurement data – The role of measurement uncertainty in conformity assessment*.
- [2] ISO/IEC 17025:2017, *General requirements for the competence of testing and calibration laboratories*, clauses 7.1.3 and 7.8.6.
- [3] ILAC-G8:09/2019, *Guidelines on Decision Rules and Statements of Conformity*.
- [4] ISO 6143:2001, *Gas analysis – Comparison methods for determining and checking the composition of calibration gas mixtures*.
- [5] ISO 10723:2012, *Natural gas – Performance evaluation for analytical systems*.
- [6] ISO 4259-5, *Petroleum and related products – Precision of measurement methods and results – Part 5: Statistical assessment of agreement between two different measurement methods that claim to measure the same property*.
- [7] Schuirmann, D.J. (1987). A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics*, 15(6), 657–680.
- [8] Shafer, G. (1976). *A Mathematical Theory of Evidence*. Princeton University Press.

About the author

Erik van der Werff has worked for 25 years as an independent consultant in gas analysis, chemometrics and measurement uncertainty through Qualimetric. He is a member of ISO/TC 193/SC 1 (natural gas analysis), ISO/TC 28/WG 2 (precision of test methods) and ASTM D02, and the author of ISO 4259-5. He develops NexChrom. · qualimetric.eu · nexchrom.nl