

The acceptable r^2 for a calibration curve, and why it is not enough.

A coefficient of determination above 0.995 tells you the points line up. It does not tell you whether the curve will report an accurate result at the low end of the range, which is where most regulatory decisions are made.

SUMMARY

Many calibration acceptance criteria still rest on r or r^2 . This paper shows, with a worked eight-level example, why the highest r^2 can belong to the least accurate curve, and sets out the metrics that measure calibration error directly: RSE, RSD, relative error at the low and mid points, residual plots, and a documented weighting rationale. It closes with a seven-step ICAL review checklist.

THE SHORT VERSION

- 01 Most methods accept $r^2 \geq 0.99$, the same requirement as $r \geq 0.995$. Some labs set a stricter internal limit such as 0.999.
- 02 A passing r^2 shows the points line up. It cannot show the curve reports accurate results near your reporting limit.
- 03 Judge the curve by back-calculated percent error at each standard and by RSE, and record why you chose the model and weighting.

The number every analyst checks first.

Open almost any initial calibration (ICAL) report and the first value a reviewer looks for is the correlation coefficient. Many methods and SOPs still set acceptance at $r \geq 0.995$, which is the same requirement as $r^2 \geq 0.990$. If the curve clears that line, it is accepted, and every sample in the batch is quantitated against it.

The habit is understandable. r^2 is a single number, every chromatography data system reports it, and a value of 0.999 looks like proof of a good fit. The problem is that r^2 answers a different question from the one the lab needs answered. The lab needs to know how accurately the curve will report a concentration, especially near the reporting limit. r^2 describes how well the points line up across the whole range.

What r^2 actually measures.

The coefficient of determination is the fraction of the variance in instrument response that the regression line explains. An unweighted least-squares fit places the line to minimize the sum of squared absolute residuals. Detector response grows with concentration, so the absolute residuals at the top of the range are far larger than those at the bottom, and they dominate both the fit and the r^2 value.

Take a residual of 200 area counts. At a 100 ppb standard with a response near 100,000, that is a 0.2% miss. At a 0.5 ppb standard with a response near 500, the same 200 counts is an error of roughly 40%. The regression treats those two misses as equal. The analyst reporting a result near the low standard does not.

This is why the International Union of Pure and Applied Chemistry has described the correlation coefficient as having “no meaning in calibration”, a position Burrows and Parr cite in their review of calibration acceptance practice ^[1].

A worked example.

The table below is an eight-level calibration from 0.5 to 100 ppb, with the kind of proportional scatter typical of chromatographic detectors. The same data were fit three ways: unweighted linear regression, linear regression weighted $1/x$, and linear regression weighted $1/x^2$. Each standard was then back-calculated against its own curve and compared with its true concentration.

LEVEL	TRUE CONC. (PPB)	RESPONSE (AREA)	UNWEIGHTED RE	$1/X$ RE	$1/X^2$ RE
1	0.5	540	-36.7%	-0.5%	+0.8%
2	1	1,010	-20.8%	-3.0%	-2.6%
3	2	2,090	-5.7%	+2.7%	+2.7%
4	5	4,880	-5.8%	-2.9%	-3.1%
5	10	10,350	+2.5%	+3.5%	+3.2%
6	25	24,600	-1.3%	-1.3%	-1.7%
7	50	51,200	+3.2%	+2.8%	+2.4%
8	100	98,300	-0.7%	-1.3%	-1.7%
R^2			0.99960	0.99954	0.99945
RSE			17.6%	2.8%	2.8%

Illustrative data set constructed for this article. RE is the relative error of each back-calculated standard. Every derived value is computed directly from the responses shown.

0.99960

R^2 OF THE UNWEIGHTED FIT, THE HIGHEST OF THE THREE

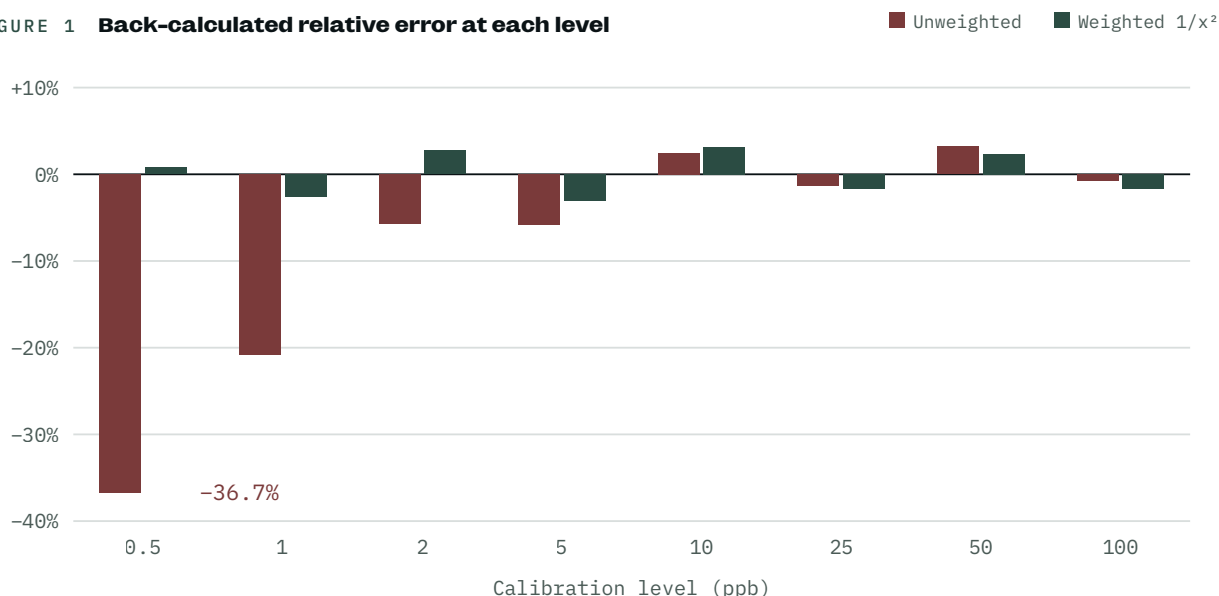
-36.7%

BACK-CALCULATED ERROR AT 0.5 PPB, UNWEIGHTED

3.2%

LARGEST ERROR AT ANY LEVEL, WEIGHTED $1/X^2$

FIGURE 1 **Back-calculated relative error at each level**



The unweighted curve has the highest r^2 of the three, and it is the only one that would bias a low-level result by more than a third. A sample that truly contains 0.5 ppb would be reported near 0.32 ppb. If the reporting limit sat at 0.5 ppb, that sample would be reported as not detected. The weighted curves have a marginally lower r^2 , and they hold every standard within 3.5% of its true value.

Ranking these three fits by r^2 puts them in exactly the wrong order. Burrows and Parr report the same pattern in real environmental data: of the curves they compared, the unweighted regression was the only one that passed an $r = 0.995$ criterion, and it carried the largest error at the low end ^[1].

What a reviewer should look at instead.

None of the alternatives is exotic. Each one measures error in the units the lab reports, as concentration relative to the true value.

METRIC	WHAT IT MEASURES	WHERE IT APPLIES
RSD of response factors	Spread of the response factor across levels	Average response factor calibrations. For these, RSD and RSE are the same value.
Relative standard error (RSE)	Pooled relative error of every back-calculated standard, adjusted for degrees of freedom	Linear and quadratic regression, weighted or unweighted. Not usable when a zero-concentration point is included.
Relative error at the low and mid points	Accuracy at the levels where reporting decisions are made	Any model. Used in recent EPA 500-series drinking water methods.
Relative residual plot	Whether error is random, or shows a trend, curvature, or a fan shape	Any model. Shows when a straight line is the wrong shape for the data.

METRIC	WHAT IT MEASURES	WHERE IT APPLIES
Weighting rationale	Whether the weighting matches how variance changes across the range	Any regression. $1/x^2$ gives each level equal relative weight; $1/x$ sits between that and unweighted.

RELATIVE STANDARD ERROR

$$RSE = 100 \times \sqrt{(\sum [(x'_i - x_i) / x_i]^2 \div (n - p))}$$

x'_i is the back-calculated concentration of standard i , x_i its true concentration, n the number of calibration levels, and p the number of terms in the model: 2 for linear, 3 for quadratic.

Why this matters in an audit.

Accreditation requirements have already moved. The TNI 2016 standard for environmental laboratories requires calibration curves generated by regression to be evaluated by RSE or by the relative error of the low and mid calibration points ^[2]. Recent revisions of EPA SW-846 methods, ASTM methods, and the wastewater methods under 40 CFR Part 136 include RSE, relative error, or both as acceptance options ^{[1][3]}.

The practical consequence is that an auditor can ask two questions a correlation coefficient cannot answer. First, what error does this curve introduce at the reporting limit? Second, why was this model and this weighting chosen over the alternatives? A lab that accepted the curve on r^2 alone has no documented answer to either. In a data integrity review, an undocumented model choice is a weakness even when the reported results happen to be correct.

A review checklist for every ICAL.

- 01 **Back-calculate every standard.**
Record the relative error at each calibration level against the curve it belongs to.
- 02 **Report the right pooled statistic.**
RSE for regression curves, RSD for average response factor curves, against the limit in your method or SOP.
- 03 **Check the low and mid standards on their own.**
A good pooled statistic can hide one bad level.
- 04 **Plot relative residuals against concentration.**
Look for a trend, curvature, or a fan shape.

05 Compare every model your method permits.

Average RF, linear and quadratic, each unweighted and weighted, judged on the same error metrics.

06 Document the choice.

Record why the selected model and weighting were chosen, so the rationale is in the record before an auditor asks.

07 Treat r^2 as a supporting value.

It can flag a gross problem. It cannot confirm a good curve.

Where WLTR fits.

Step five is where manual review usually stops short. Fitting and comparing every permitted model for every analyte is slow work in a spreadsheet, so most labs evaluate only the model their chromatography data system applies by default.

WLTR is calibration data integrity infrastructure for regulated analytical laboratories. For each analyte, it derives nine calibration regression models (average response factor, plus linear and quadratic fits across four weightings each) and evaluates them against fifteen ICAL parameters, including RSE, computed orthogonally to the CDS. The methodology is covered by U.S. Patent 12,038,448 B2 ^[4].

CALIBRATION VERIFICATION

See how your current ICAL acceptance criteria hold up against RSE, low and mid point error, and a documented model rationale.

GET COMPLIANCE GAP ASSESSMENT

WLTR.TECH

SOURCES

1. R. Burrows and J. Parr, "Evaluating the Goodness of Instrument Calibration for Chromatography Procedures," *LCGC Supplements* 38(11), November 2020. www.chromatographyonline.com/view/evaluating-the-goodness-of-instrument-calibration-for-chromatography-procedures
2. The NELAC Institute (TNI), 2016 Environmental Laboratory Sector Standard, Volume 1, Module 4: Quality Systems for Chemical Testing. Calibration evaluation requirements as summarized in [1]. nelac-institute.org
3. U.S. EPA, SW-846 Test Method 8000D, Determinative Chromatographic Separations. www.epa.gov/hw-sw846/sw-846-test-method-8000d-determinative-chromatographic-separations
4. U.S. Patent 12,038,448 B2, inventor Walter "Chip" Hinchman Sr. patents.google.com/patent/US12038448B2

Walter Hinchman, MBA

WLTR, INC. • NEVADA C-CORP

