October 25, 2021

John Griggs
Environmental Methods Forum
US Environmental Protection Agency
Washington, DC 20460

It is our understanding that you now chair EPA's Environmental Methods Forum whose stated goal is to address issues such as analytical methods for emerging contaminants and issues associated with method development and validation.

The Environmental Monitoring Coalition (EMC) was created in 2020 to address a void created by the dissolution of EPA Environmental Laboratory Advisory Board. Founding EMC partner organizations include:

- American Council of Independent Laboratories,
- Association of Public Health Laboratories,
- The NELAC Institute, and
- Water Environment Federation.

EMC was established in response to the need for the environmental monitoring community to have a mechanism to develop consensus opinions on issues effecting environmental monitoring. One issue brought to EMC's attention relates to the use of correlation coefficient (r) and coefficient of determination (r^2) as measures of calibration quality in EPA methods that that rely on generation of calibration curves.

It is now 40 years since Van Arendonk and Skogerboe stated "One practice that should be discouraged is the use of the correlation coefficient as a means of evaluating goodness of fit of linear models."[1] and 23 years since the International Union of Pure and Applied Chemistry pointed out that "The correlation coefficient, which is a measure of two random variables, has no meaning in calibration…".[2] As well as being technically incorrect, the use of r and r^2 as measures of calibration quality cause many practical problems. Both measures strongly favor reducing relative residuals at the top end of the calibration curve, at the expense of accuracy at the lower end of the curve. It is common to observe calibration curves that pass method criteria for r and r^2, while introducing relative error of over 100% at the low end of the curve. Conversely calibrations that have reasonably low error across the calibration may fail r and r^2 criteria while being perfectly reasonable to use.

Superior alternatives to r and r^2 are readily available and are already included in most EPA methods. Relative Standard Error (RSE) is included in SW-846 method 8000 and in 40 CFR Part 136. The RSE provides a single number to provide a measure of curve quality and is a far superior alternative to r and r^2. Note: A link to an Excel spreadsheet to calculate RSE can be found here: https://nelac-institute.org/docs/comm/emmec/Basic%20RSE%20calculatorv4.xlsx.

[1] Anal. Chem. 53, 1981, 2349-2350
[2] IUPAC, Pure & Appl. Chem. 70(4), 993– 1014 (1998)

Alternatively, Relative Error (RE) can be used to evaluate individual points within the curve. Relative error is included in method 8000 and is the primary method of calibration evaluation in drinking water methods. RSE or RE are required in the laboratory accreditation standards published by The NELAC Institute (TNI). However, the TNI requirements only affect a small population of laboratories as the majority of states do not accredit/certify wastewater or hazardous waste laboratories.

Attachment 1 provides some data showing that curves with a perfect coefficient of determination (1.000) can have errors of over 1000% at low concentrations while calibration curves with r^2 as low as 0.958 can have an RSE of < 20%. "For almost any calibration, the correlation coefficient and coefficient of determination lead us in the direction of choosing the wrong calibration."[3]

Addition of RSE and RE to EPA methods over the last few years is a great improvement. Unfortunately, in SW-846 Method 8000 the language is not very clear regarding the use of r or r^2 in conjunction with RSE/RE. Some people make the interpretation that they are alternatives, others that they are both required. 40 CFR Part 136 has a similar issue. Section 136.6 (b) (4)(x) indicates RSE "may" be used and does not discuss RE.

Ideally, for clarity, and to eliminate the use of outdated and inferior measures of calibration quality, r and r^2 need to be removed from EPA all methods, and in particular EPA approved methods. This is clearly possible, since most drinking water methods currently do not include r or r^2.

For SW-846 method 8000 removing r and r^2 alone is sufficient. There is no need for anything to be added since RSE and RE are already in place. See Attachment 2. It is critically important that the language in the second paragraph of 11.5.6.3 which suggests a calibration curve with an R2 of <0.99 would not be acceptable be removed as the data in Attachment 1 shows very good data can be obtained from curves which do not meet this criterion.

For wastewater methods RSE and RE need to be added to replace the existing language in every EPA method in Part 136. We understand this could be a difficult process, especially for older methods not codified in Part 136. Alternatively, Section 136.6 could be revised in the next Method Update Rule as shown in Attachment 3.

An alternative to the two approaches described above would be for the Environmental Methods Forum to issue a policy statement acknowledging that using correlation coefficient or coefficient of determination as measures of calibration quality in EPA methods that that rely on generation of calibration curves is an outdated concept that should be replaced with RSE or RE. An example of such a statement is provided in Attachment 4.

A more detailed discussion of these issues is available.[4] It is also worth noting that the problems caused by r and r^2 become even more acute with modern instrumentation such as triple quadrupole GCMSMS, because of the wider working range that is possible.

---

[3] Burrows, Richard, Modern Mass Spectrometers and the Correlation Coefficient: Are they Compatible?, National Environmental Monitoring Conference, August, 2021

[4] Evaluating the Goodness of Instrument Calibration for Chromatography Procedures, LCGC, October 2020, Richard Burrows and Jerry Parr. https://www.chromatographyonline.com/view/evaluating-the-goodness-of-instrument-calibration-for-chromatography-procedures

We would like to have the opportunity to discuss this issue further. Please contact either of use to set up a meeting.

Sincerely,

Jerry Parr *David Friedman*

Jerry Parr
EMC Chair
jerry.parr@nelac-institute.org
1-817-308-0449

David Friedman
EMC Vice-chair
friedmanconsulting@outlook.com
1-703-389-3821

CC Adrian Hanley, OW OST
Dan Hautman, OW OGWDW
Kim Kirkland, OLEM ORCR
Robin Segall, EMC OAR

Attachment 1: Comparison of R2, %RE and %RSE for Selected Compounds

Modern Mass Spectrometers and the Correlation Coefficient: Are they Compatible?

Richard Burrows

August 4, 2021 - National Environmental Monitoring Conference

Table 1. Analysis by Time-of-Flight Mass Spectrometry

Analyte	Linear Unweighted			Quadratic Unweighted		
	R2	RE, %	RSE, %	R2	RE, %	RSE, %
Hexadecane	0.998	1109	213	1.000	326	134
2,4,5-Trichlorophenol	0.996	1335	535	1.000	220	90.4
Chrysene	0.999	166	62.4	1.000	142	68.7
Analyte	**Linear Weighted**			**Quadratic Weighted**		
	R2	**RE, %**	**RSE, %**	**R2**	**RE, %**	**RSE, %**
Hexadecane	0.963	<30	18.5	0.986	<30	13.2
2,4,5-Trichlorophenol	0.958	<30	19.8	0.985	<30	13.8
Chrysene	0.985	<30	11.7	0.987	<30	12

Table 2. Analysis by Triple Quad Mass Spectrometry

Analyte	Linear Unweighted			Quadratic Unweighted		
	R2	RE, %		R2	RE, %	
2,4-Dinitrophenol	0.995	186				
Benzo(ghi)perylene	0.999	11260				
Pentachlorophenol				0.998	14638	
Analyte				**Quadratic Weighted**		
				R2	**RE, %**	
2,4-Dinitrophenol				0.993	<30	
Benzo(ghi)perylene				0.981	<30	
Pentachlorophenol				0.98	<30	

Figures 1-4 show linear and quadratic curves with no weighting and 1/concentration^2 weighting for hexadecane and 2,4,5-Trichlorophenol using Time-of-Flight mass spectrometry.

Figures 5-10 show linear and quadratic curves with no weighting and 1/concentration^2 weighting for 2,4-Dinitrophenol, Benzo(ghi)perylene, and Pentachlorophenol using triple-quad mass spectrometry.

The table above and the calibration curves clearly demonstrated that weighting is better than no weighting and that a quadratic fit is better than a linear fit for both polar and non-polar compounds for different technologies.

Attachment 2: Language for Method 8000D on Initial Calibration with Suggested Changes

11.5.1 Linear calibration using average calibration or response factor

As calculated in Sec 11.4, each CF or RF represents the slope of the line between the origin and the given standard response. If the relative standard deviation (RSD) of variation in the factors is ≤ 20%, the linear model is generally representative over the range of calibration standards.

11.5.1.1 If the RSD is ≤20% over the calibration range, the slopes of the lines for each standard are sufficiently close to one another that the use of the linear model is generally appropriate over the range of standards that are analyzed; or may be used to determine sample concentrations. Alternatively, either of the two methods described in 11.5.4 may be used to determine calibration function acceptability.

NOTE: The RSD approach is equivalent to a 1/x2 weighted linear least square regression line that is forced through the origin. 11.5.1.2

Given the potentially large numbers of analytes that may be analyzed in some methods, it is likely that some analytes may exceed the acceptance limit for the RSD for a given calibration. In those instances, it is recommended, but not required, that corrective actions as described in Sec. 11.5.6.1 be followed. Sec. 11.5.6.1 also provides alternative uses for initial calibrations that do not meet their criteria of acceptability.

11.5.2 Linear calibration using a least squares regression

11.5.2.2 ~~In the specific case of an unweighted linear least squares regression (i.e., a regression that varies both a and b), the correlation coefficient (r) can be used to measure the "goodness of fit."~~

~~The instrument data system will typically calculate r. An r value of +1.00 indicates a positive perfect correlation; an r value of -1.00 indicates a negative perfect correlation; an r value of 0 indicates no correlation.~~

~~However, if the regression line is forced through the origin or the weighting factor is variable, then the coefficient of determination, more often termed r 2 , should be used to measure the "goodness of fit", such that $0 \leq r^2 \leq 1$. This shows the strength of the association between x and y. The r 2 value allows the analyst to determine the percent of the data closest to the line of best fit. For consistency, it is acceptable to use r 2 for linear unweighted curves as well. An r 2 value of 1.00 indicates that all variability in response is due to variation in concentration.~~

~~In order for the linear regression model to be used for quantitative purposes, r or r 2 should be ≥ 0.995 or 0.99, respectively. Alternatively, either of the two methods described in Sec. 11.5.4 may be used to determine whether the calibration function meets acceptance criteria.~~ It is recommended that the resulting calibration curve be inspected by the analyst as described in Sec. 11.5.4.1. 11.5.2.3.

11.5.3.1 ~~Linear and non-linear least squares regressions are mathematical methods that minimize differences (the residuals) between observed instrument response, yi, and calculated response, yi', by adjusting coefficients of the polynomial (a, b, c, and d) to obtain the polynomial best fitting the data.~~

~~The coefficient of determination (r 2) may be used as a measure of goodness of fit. See Sec. 11.5.2.2 for the definition of r 2 .~~

~~11.5.3.2 Under ideal conditions (i.e., a "perfect" fit of the model to the data), the r 2 will equal 1.00. In order to be an acceptable non-linear calibration, the r 2 must be ≥ 0.99. Alternatively, either of the two methods described in 11.5.4 may be used to determine calibration function acceptability. It is recommended that the resulting calibration curve be inspected by the analyst, as described in Sec. 11.5.4.1.~~

11.5.4 Acceptance criteria independent of calibration model

Either of the two procedures described in Secs. 11.5.4.1 and 11.5.4.2 may be used to determine calibration function acceptability for linear and non-linear curves. These include refitting the calibration data back to the model. Both % Error and Relative Standard Error (RSE) evaluate the difference between the measured and the true amounts or concentrations used to create the model.

Percent error between the calculated and expected amounts of an analyte should be ≤ 30% for all standards. For some data uses, ≤50% may be acceptable for the lowest calibration point.

The RSE acceptance limit criterion for the calibration model is the same as the RSD limit for or in the determinative method. If the RSD limit is not defined in the determinative method, the limit should be set at ≤20% for good performing compounds and ≤30% for poor performing compounds. A list of known poorly performing compounds can be found in Sec. 16 of this document.

11.5.6.1 Corrective action may be needed if the calibration criteria (RSD~~/r 2~~ and %Error/RSE) are not met. If any analyte for any calibration standard has a percent error > ±30% as described in Section11.5.4.1, corrective action may be needed. Some recommended courses of action and additional options for modifying the calibration ranges follow. More specific corrective actions that are provided in the applicable determinative methods will supersede those noted in Method 8000. Generally, the calibration should not be used for quantitative analyses of that analyte when the calibration criteria (RSD~~/r 2~~ and % Error/RSE) are not met.

11.5.6.2 For all calibration models the following options are allowed. However, if none result in an acceptable calibration, a new initial calibration must be performed.

11.5.6.3 Generally, the first option is to check the instrument operating conditions. The suggested maintenance procedures in Sec. 11.11 may be useful in guiding such adjustments. This option will apply in those instances where a linear instrument response is expected. It may involve some trade-offs to optimize performance across all target analytes. For instance, changes to the operating conditions necessary to achieve linearity for problem compounds may cause the RSD for other compounds to increase, but as long as all analytes meet the RSD limits for linearity, the calibration is acceptable. If the initial calibration for any analyte does not meet the acceptance criteria (e.g., RSD/RSE > 20%~~ or r 2 < 0.99~~), the analyst may wish to review the results (proper identification, area counts, calibration or RFs, and RSD/RSE) for those analytes to ensure that the problem is not associated with just one of the initial calibration standards.

If criteria for RSD/RSE~~/r 2~~ has been met for the calibration model but the % error of one or more of the individual calibration points at the extreme ends of the calibration range exceeds the criteria described in Sec. 11.5.4.1, the usable range of the calibration may be narrowed to the standards that meet the % error criteria, but the calibration points used to generate the initial curve are retained. The LLOQ becomes the lowest end of the adjusted calibration range. The calibration model should meet the RSD/RSE~~/r 2~~ criteria (Secs. 11.5.1 – 11.5.3) and the minimum number of data points (Sec. 11.5.3.1) before this option can be used.

Attachment 3: Suggested Changes for Instrument Calibration for EPA Wastewater Methods

Suggested Preamble Language

Most of the wastewater methods developed by EPA in the last 40 years, including those promulgated in Part 136, contained a general statement such as this language from section 7.2.2 of Method 625:

> Calculate response factors for each compound using equation 1.. If the RF value over the working range is constant (< 35%), the RF can be assumed to be invariant and the average RF can be used for calculations. Alternatively, the results can be used to plot a calibration curve of response ratios, As/Ais, vs. concentration ratios Cs/Cis.

No criteria were given as to how to evaluate such a curve. In the 2017 Method Update Rule, EPA promulgated Methods 608.3, 624.1 and 625.1 and the language was revised to read:

> Calculate the mean (average) and relative standard deviation (RSD) of the response factors. If the RSD is less than 35%, the RF can be assumed to be invariant and the average RF can be used for calculations. Alternatively, the results can be used to fit a linear or quadratic regression of response ratios, As/Ais, vs. concentration ratios Cs/Cis. If used, the regression must be weighted inversely proportional to concentration. The coefficient of determination (R^2; Reference 10) of the weighted regression must be greater than 0.920 (this value roughly corresponds to the RSD limit of 35%). Alternatively, the relative standard error (Reference 11) may be used as an acceptance criterion. As with the RSD, the RSE must be less than 35%. If an RSE less than 35% cannot be achieved for a quadratic regression, system performance is unacceptable and the system must be adjusted and re-calibrated.

Reference 10. http://en.wikipedia.org/wiki/Coefficient_of_determination (accessed on 09/10/2013)
Reference 11. 40 Code of Federal Regulations 136.6(b)(4)(x)

Because the older methods do not address the evaluation of calibration curves that do not use the average response factor approach and because having two very different criteria in the 2017 methods created confusion in the laboratory community, EPA is revising Section 136.6(b)(4)(x) to clearly indicate RSE (or RE) is the preferred approach.

Suggested Changes to the Test of 136.6 (b)(4)(x)

Changes in calibration model.
(A) Linear calibration models do not adequately fit calibration data with one or two inflection points. For example, vendor-supplied data acquisition and processing software on some instruments may provide quadratic fitting functions to handle such situations. If the calibration data for a particular analytical method routinely display quadratic character, using quadratic fitting functions may be acceptable. In such cases, the minimum number of calibrators for second order fits should be six, and in no case should concentrations be extrapolated for instrument responses that exceed that of the most concentrated calibrator. Examples of methods with nonlinear calibration functions include chloride by SM4500-Cl-E-1997, hardness by EPA Method 130.1, cyanide by ASTM D6888 or OIA1677, Kjeldahl nitrogen by PAI-DK03, and anions by EPA Method 300.0.

(B) As an alternative to using the average response factor, the quality of the calibration ~~may be~~must be evaluated using ~~the~~ Relative Standard Error (RSE) or Relative Error (RE). The acceptance criterion for the RSE/RE is the same as the acceptance criterion for Relative Standard Deviation (RSD), in the method.

RSE is calculated as:

$$\%\,\mathrm{RSE} = 100 \times \sqrt{\frac{\sum_{i=1}^{n}\left[\frac{x_i' - x_i}{x_i}\right]^2}{(n-p)}}$$

where:
x'i = Calculated concentration at level i
xi = Actual concentration of the calibration level i
n = Number of calibration points
p = Number of terms in the fitting equation (average = 1, linear = 2, quadratic = 3)

Relative Error (RE) is calculated using the following equation:

$$\%\ Relative\ Error = \frac{x'_i - x_i}{x_i} \times 100$$

x_i = True value for the calibration standard
x'_i = Measured concentration of the calibration standard

This calculation must be performed for two (2) calibration levels: the standard at or near the mid-point of the initial calibration and the standard at the lowest level.

(C) Using the RSE/RE as a metric has the added advantage of allowing the same numerical standard to be applied to the calibration model, regardless of the form of the model. Thus, if a method states that the RSD should be ≤ 20% for the traditional linear model through the origin, then the RSE/RE acceptance limit can remain ≤ 20% as well. Similarly, if a method provides an RSD acceptance limit of ≤ 15%, then that same figure can be used as the acceptance limit for the RSE. The RSE ~~may~~ is to be used ~~as an alternative to~~instead of correlation coefficients and coefficients of determination for evaluating calibration curves for any of the methods at Part 136. If the method includes a numerical criterion for the RSD, then the same numerical value is used for the RSE/RE. Some older methods do not include any criterion for the calibration curve – for these methods, if RSE/RE is used the value should be ≤ 20%. ~~Note that the use of the RSE is included as an alternative to the use of the correlation coefficient as a measure of the suitability of a calibration curve. It is not necessary to evaluate both the RSE and the correlation coefficient.~~

Attachment 4: Suggested Policy Memo

MEMORANDUM

Subject: Use of correlation coefficient (r) and coefficient of determination (r^2) as measures of calibration quality

Some older EPA methods use a correlation coefficient (r) and coefficient of determination (r^2) as measures of calibration quality. Such measures are now considered inappropriate and the Environmental Methods Forum recommends these measures not be used and instead use Relative Standard Error (RSE) or Relative Error (RE) to evaluate calibration curves as an alternative to using the average response factor.

RSE is calculated as:

$$\%\,\mathrm{RSE} = 100 \times \sqrt{\frac{\sum_{i=1}^{n}\left[\frac{x_i' - x_i}{x_i}\right]^2}{(n-p)}}$$

where:
x'i = Calculated concentration at level i
xi = Actual concentration of the calibration level i
n = Number of calibration points
p = Number of terms in the fitting equation (average = 1, linear = 2, quadratic = 3)

Relative Error (RE) is calculated as:

$$\%\ Relative\ Error = \frac{x'_i - x_i}{x_i} \times 100$$

x_i = True value for the calibration standard
x'_i = Measured concentration of the calibration standard

This calculation must be performed for two (2) calibration levels: the standard at or near the mid-point of the initial calibration and the standard at the lowest level.

Using the RSE/RE as a metric has the added advantage of allowing the same numerical standard to be applied to the calibration model, regardless of the form of the model. Thus, if a method states that the RSD should be ≤ 20% for the traditional linear model through the origin, then the RSE/RE acceptance limit can remain ≤ 20% as well. Similarly, if a method provides an RSD acceptance limit of ≤ 15%, then that same figure can be used as the acceptance limit for the RSE. The RSE is to be used instead of correlation coefficients and coefficients of determination for evaluating calibration curves for any of the methods at Part 136. If the method includes a numerical criterion for the RSD, then the same numerical value is used for the RSE/RE. Some older methods do not include any criterion for the calibration curve – for these methods, if RSE/RE is used the value should be ≤ 20%.

Figure 1. Calibration Curve for Hexadecane with Quadratic Curve Fit and 1/Conc^2 Weighting, 0.005 to 0.5 ng (GC/TOFMS)

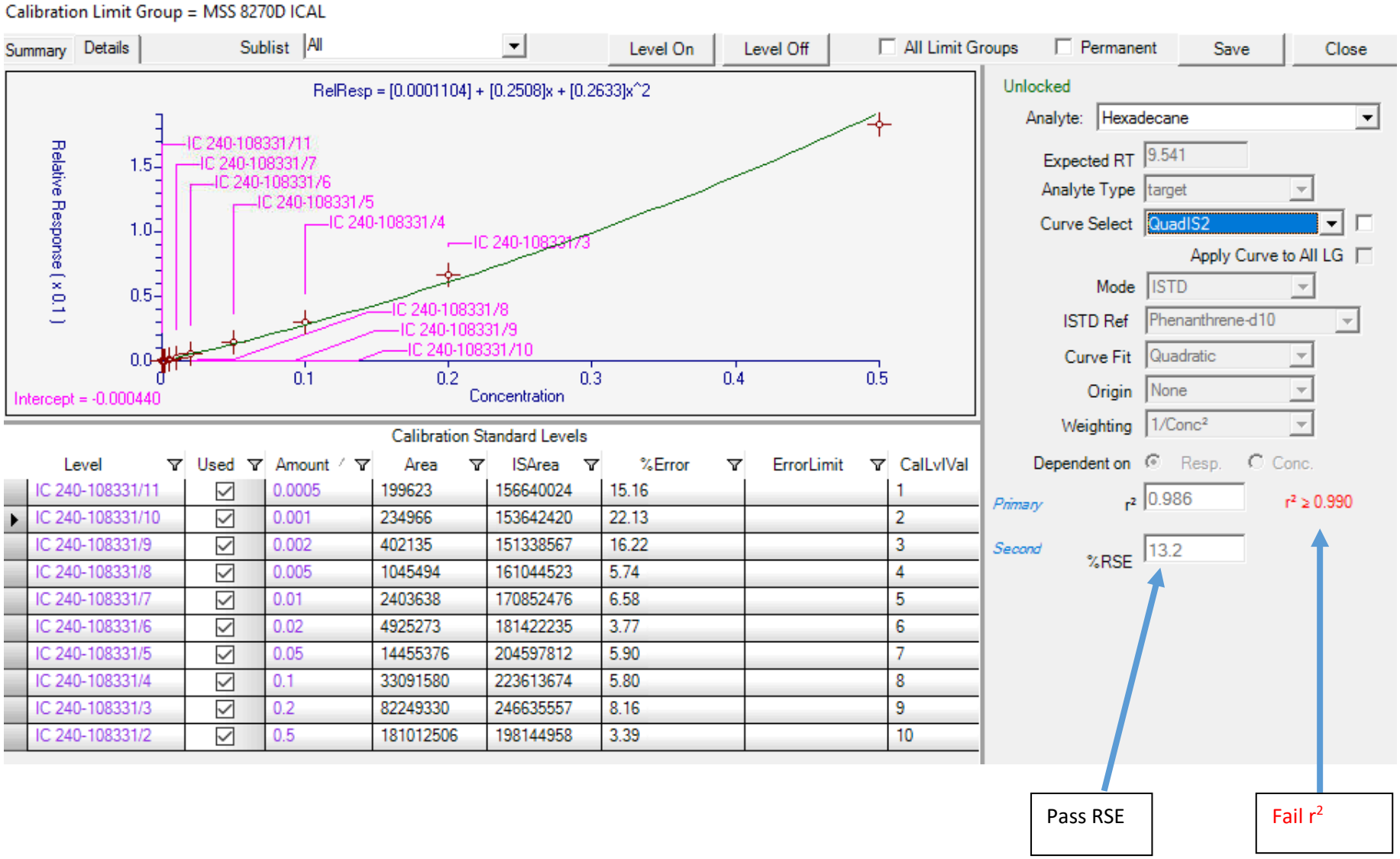


Figure 2. Calibration Curve for Hexadecane with Linear Curve Fit and No Weighting, 0.005 to 0.5 ng (GC/TOFMS)

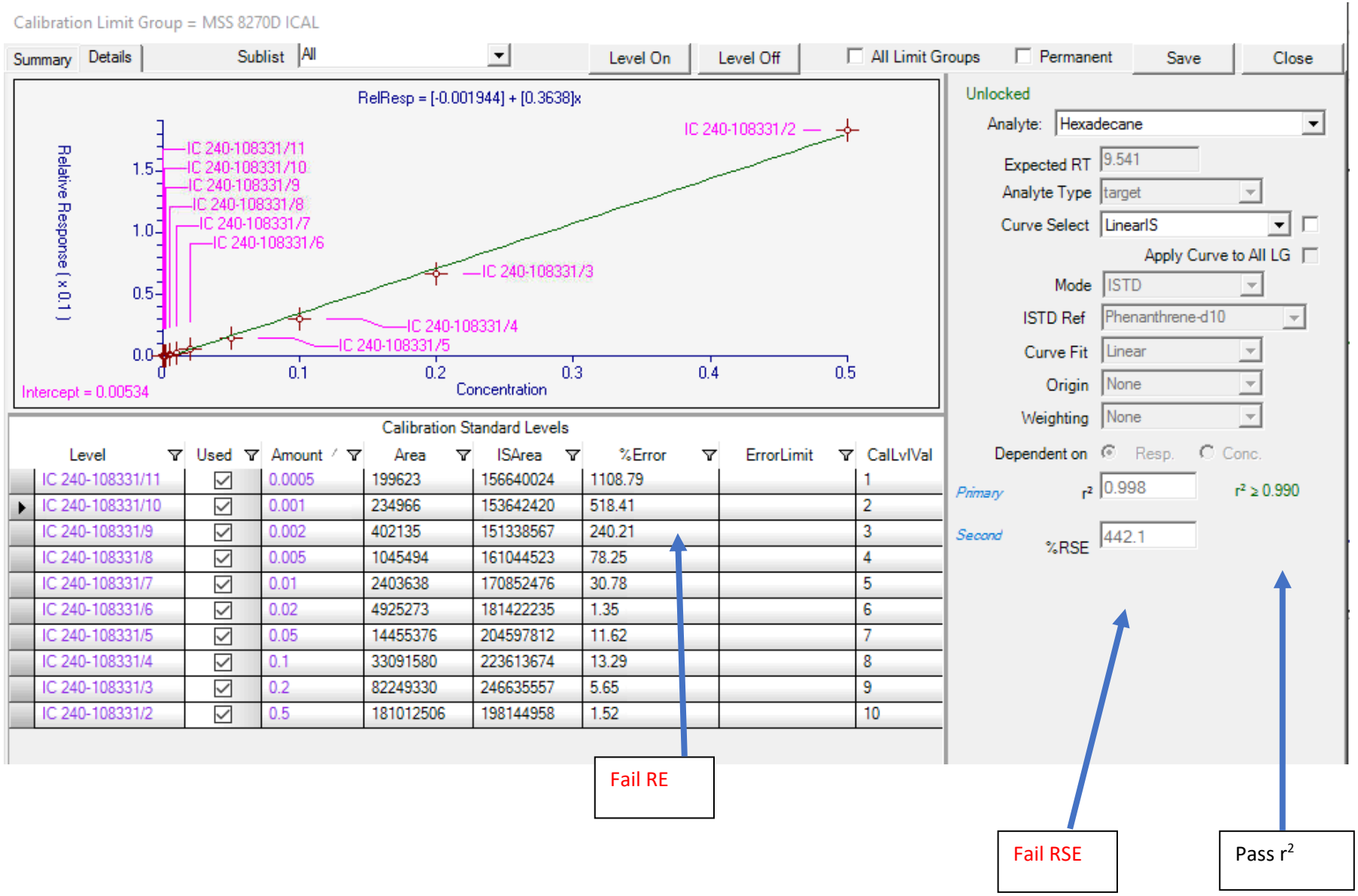


Figure 3. Calibration Curve for 2,4,5-Trichlorophenol with Quadratic Curve Fit and 1/Conc^2 Weighting, 0.005 to 0.5 ng (GC/TOFMS)

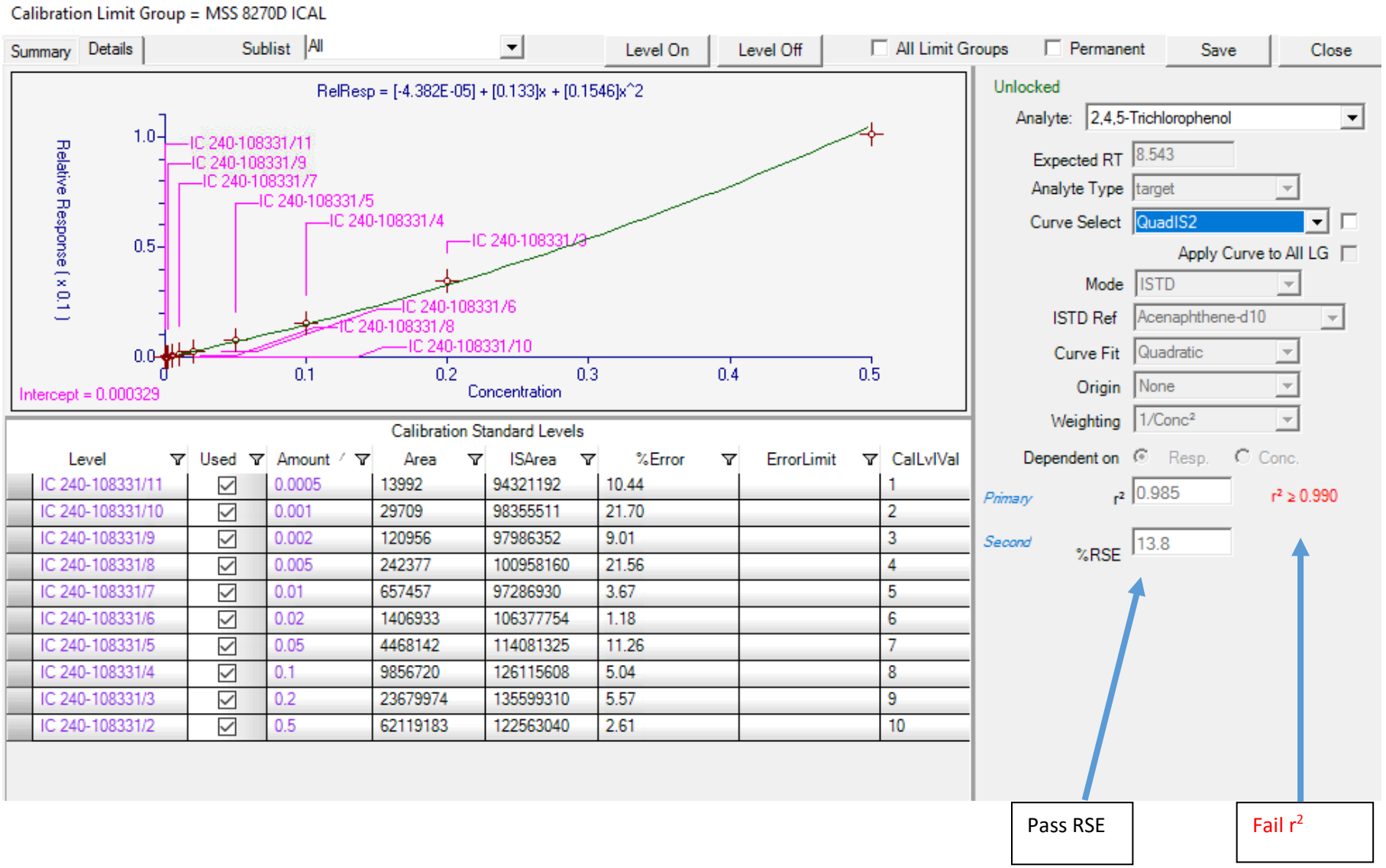


Figure 4. Calibration Curve for 2,4,5-Trichlorophenol with Linear Curve Fit and No Weighting, 0.005 to 0.5 ng (GC/TOFMS)

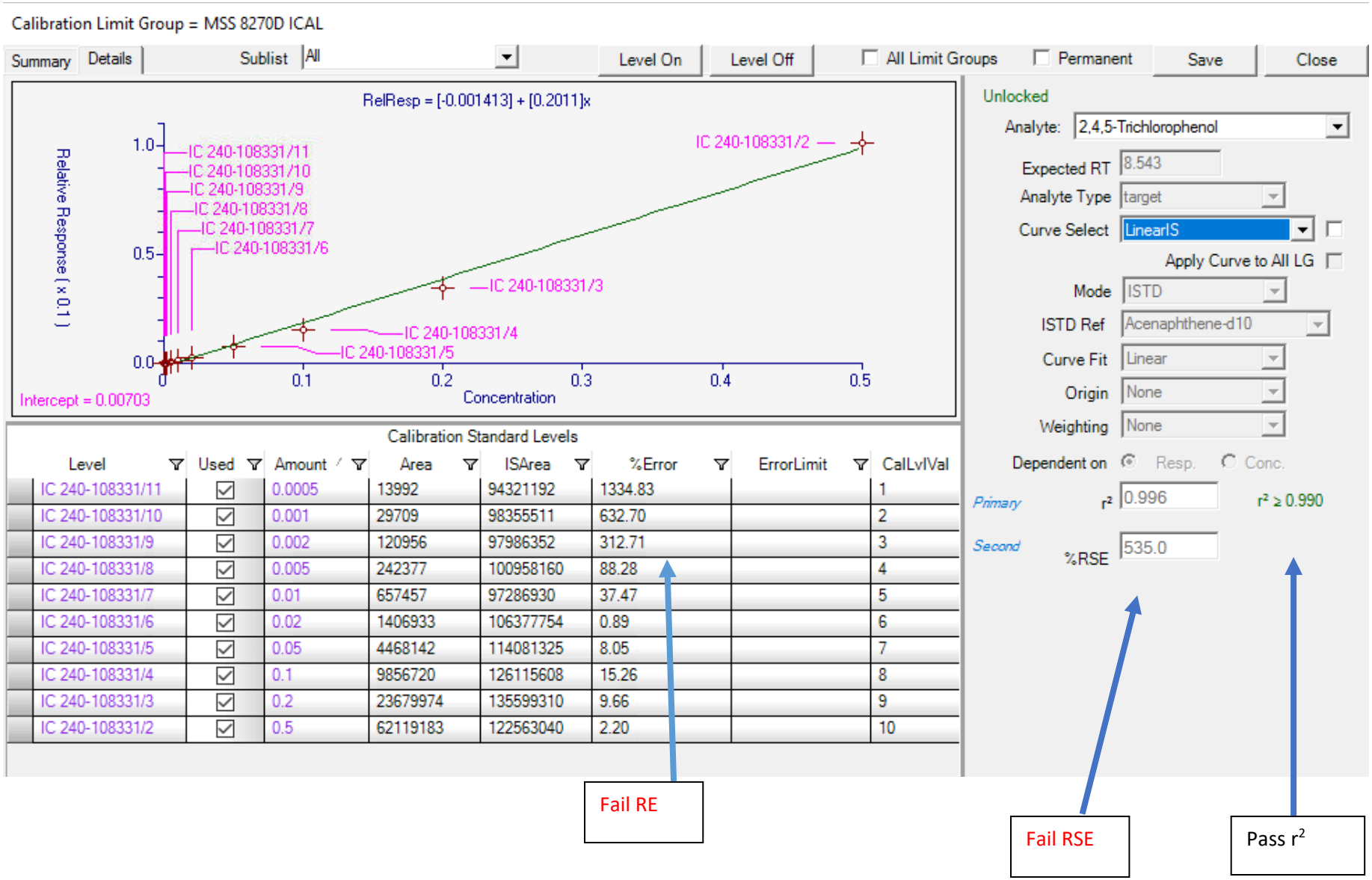


Figure 5. Calibration Curve for 2,4-Dinitrophenol with Quadratic Curve Fit and 1/Conc^2 Weighting, 0.8 to 048 ng (GC/MS/MS)

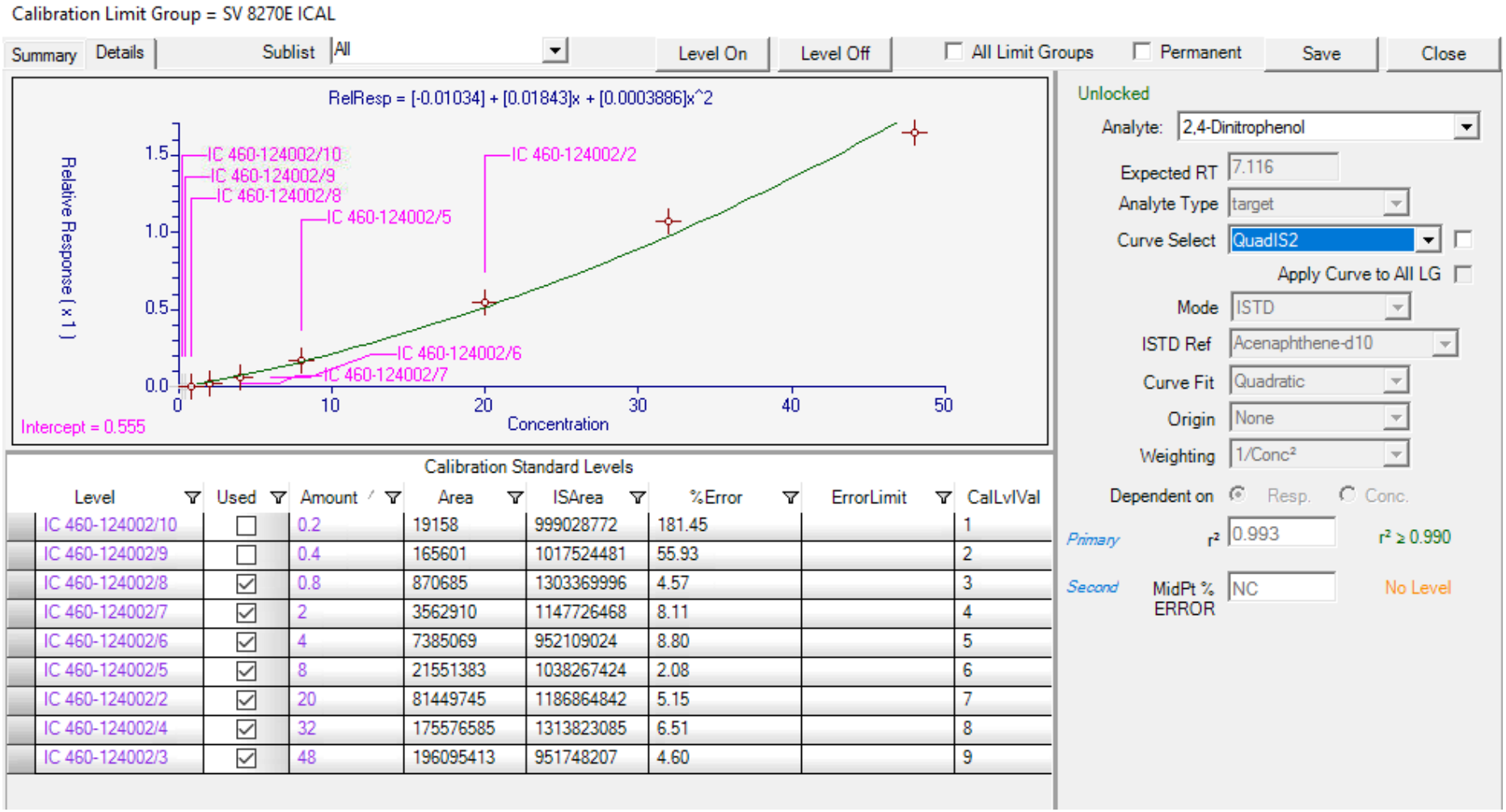


Figure 6. Calibration Curve for 2,4-Dinitrophenol with Linear Curve Fit and No Weighting, 0.8 to 048 ng (GC/MS/MS)

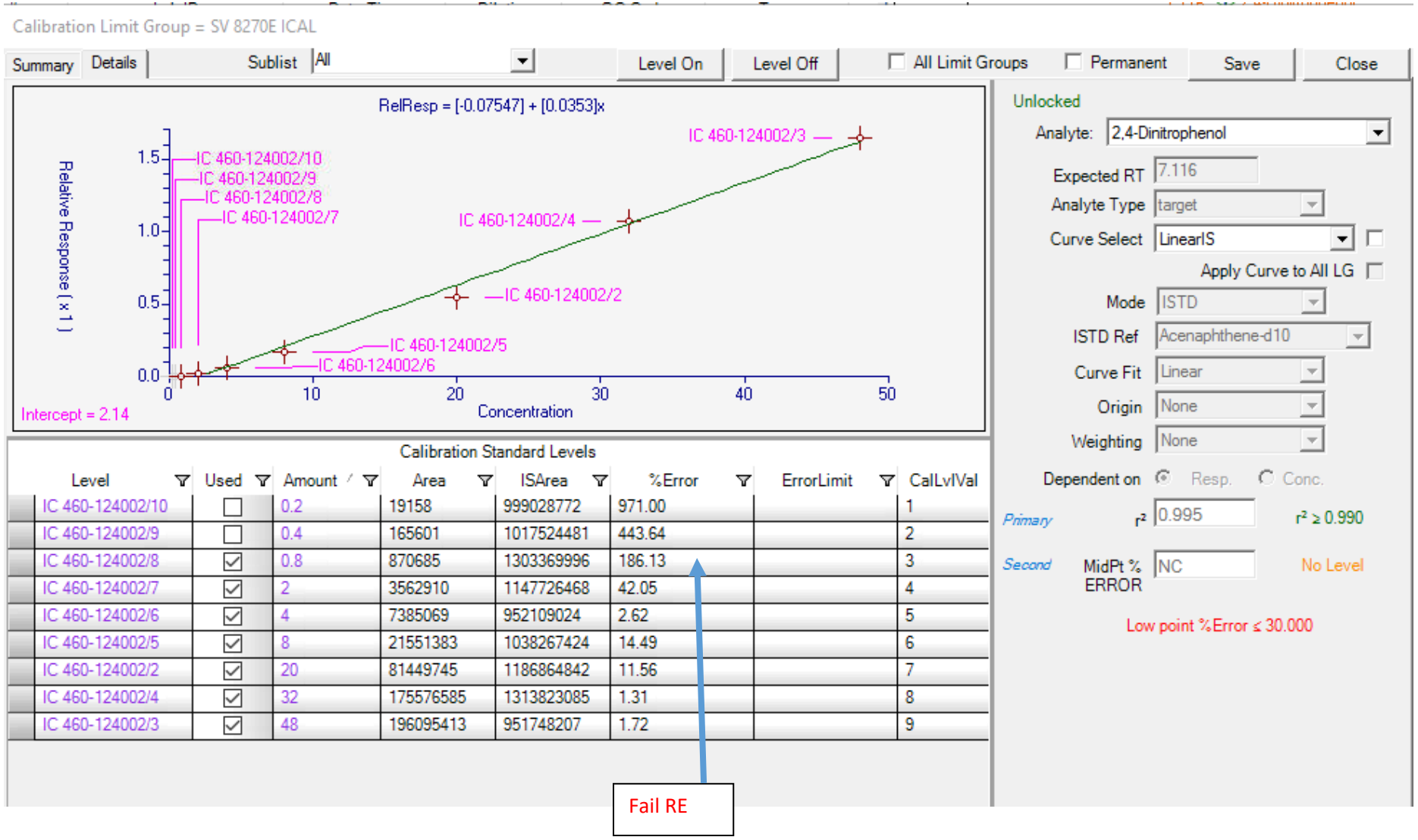
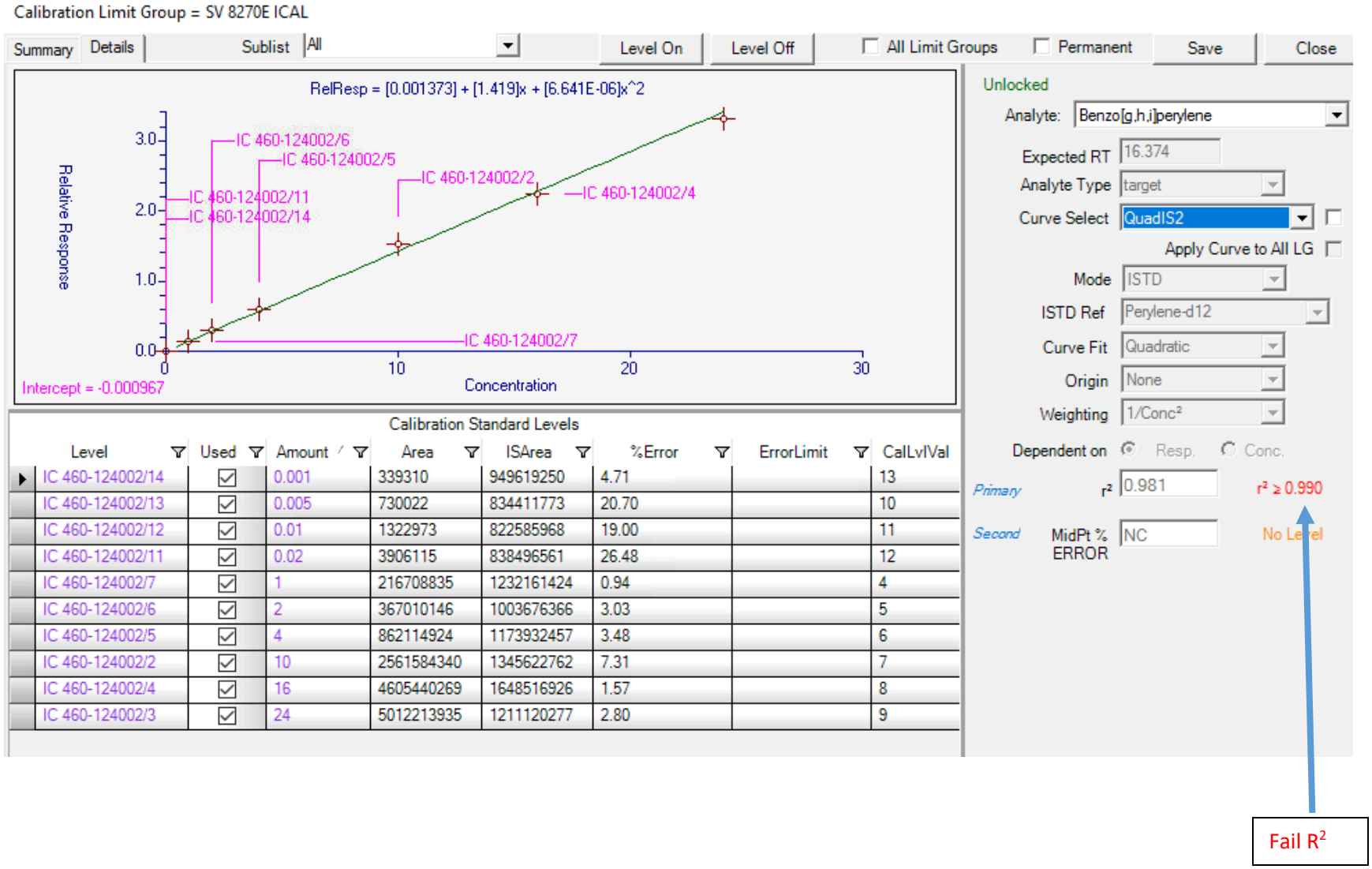


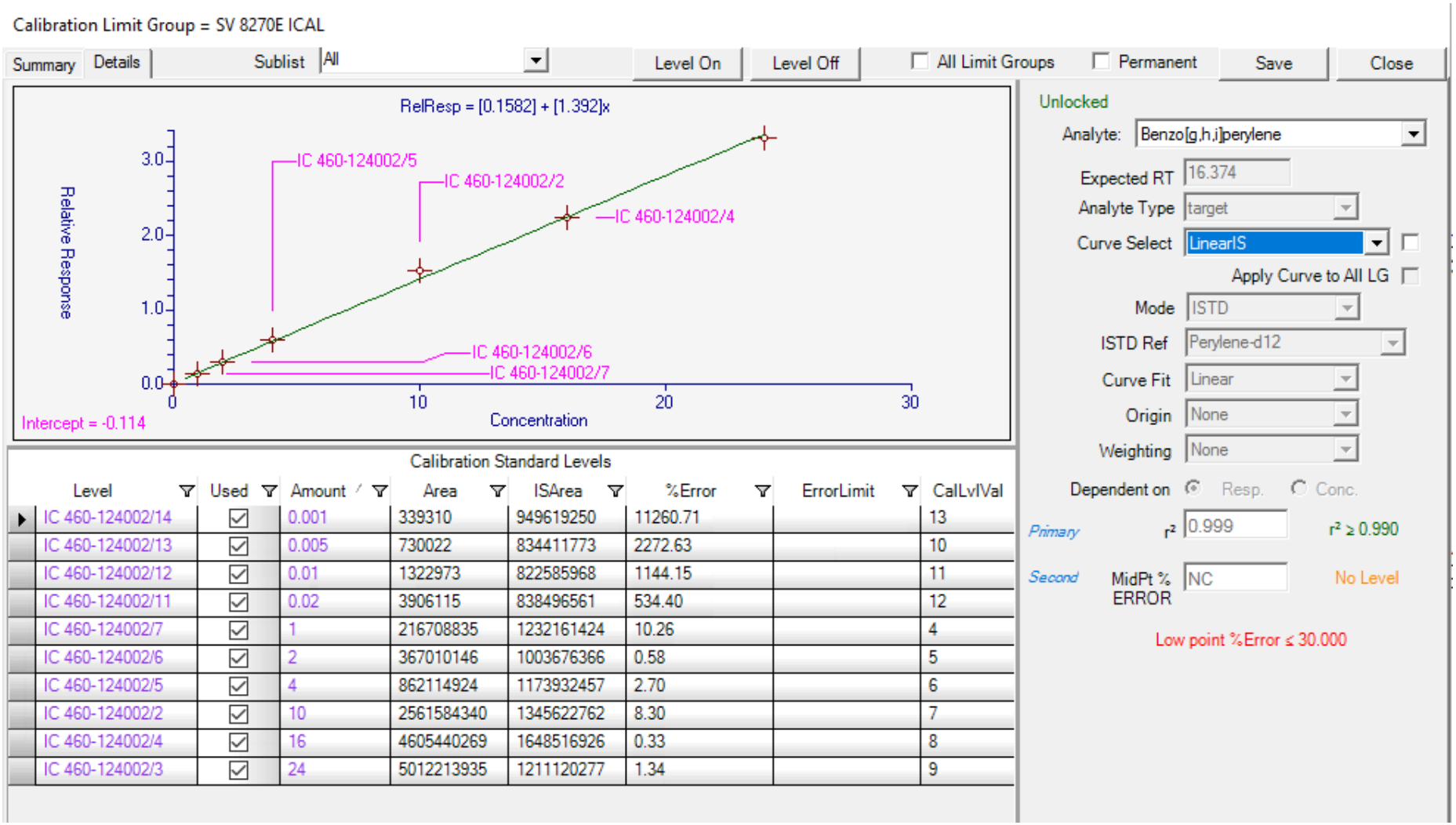
Figure 7. Calibration Curve for Benzo (ghi)peryene with Quadratic Curve Fit and 1/Conc^2 Weighting, 0.001 to 24 ng (GC/MS/MS)



Calibration Standard Levels

Level	Used	Amount	Area	ISArea	%Error	ErrorLimit	CalLvlVal
IC 460-124002/14	☑	0.001	339310	949619250	4.71		13
IC 460-124002/13	☑	0.005	730022	834411773	20.70		10
IC 460-124002/12	☑	0.01	1322973	822585968	19.00		11
IC 460-124002/11	☑	0.02	3906115	838496561	26.48		12
IC 460-124002/7	☑	1	216708835	1232161424	0.94		4
IC 460-124002/6	☑	2	367010146	1003676366	3.03		5
IC 460-124002/5	☑	4	862114924	1173932457	3.48		6
IC 460-124002/2	☑	10	2561584340	1345622762	7.31		7
IC 460-124002/4	☑	16	4605440269	1648516926	1.57		8
IC 460-124002/3	☑	24	5012213935	1211120277	2.80		9

Figure 8. Calibration Curve for Benzo (ghi)peryene with Linear Curve Fit and No Weighting, 0.001 to 24ng (GC/MS/MS)



Calibration Standard Levels

Level	Used	Amount	Area	ISArea	%Error	ErrorLimit	CalLvlVal
IC 460-124002/14	☑	0.001	339310	949619250	11260.71		13
IC 460-124002/13	☑	0.005	730022	834411773	2272.63		10
IC 460-124002/12	☑	0.01	1322973	822585968	1144.15		11
IC 460-124002/11	☑	0.02	3906115	838496561	534.40		12
IC 460-124002/7	☑	1	216708835	1232161424	10.26		4
IC 460-124002/6	☑	2	367010146	1003676366	0.58		5
IC 460-124002/5	☑	4	862114924	1173932457	2.70		6
IC 460-124002/2	☑	10	2561584340	1345622762	8.30		7
IC 460-124002/4	☑	16	4605440269	1648516926	0.33		8
IC 460-124002/3	☑	24	5012213935	1211120277	1.34		9

Figure 9. Calibration Curve for Pentachlorophenol with Quadratic Curve Fit and 1/Conc² Weighting, 0.05 to 48 ng (GC/MS/MS)

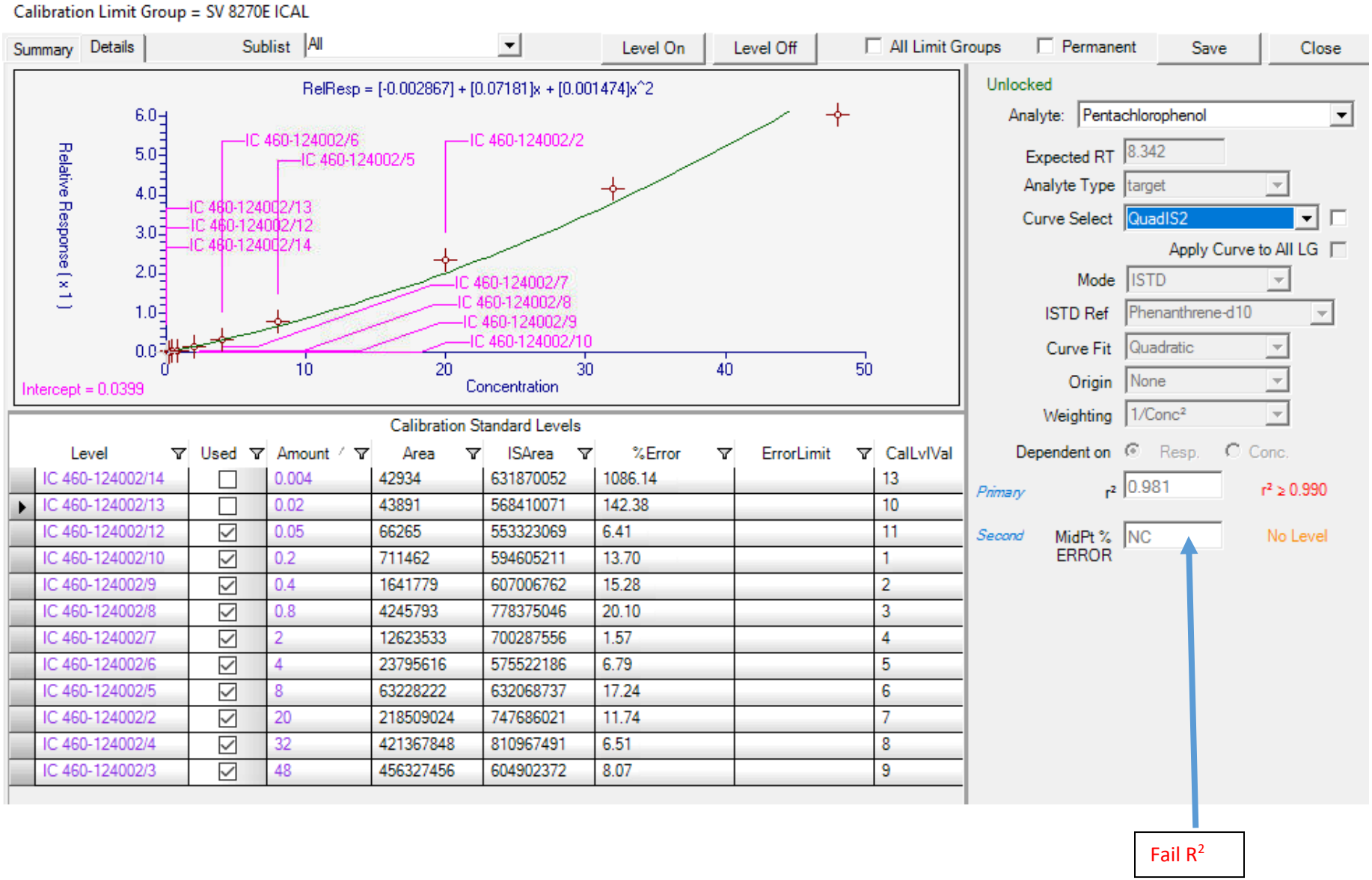


Figure 10. Calibration Curve for Pentachlorophenol with Quadratic Curve Fit and No Weighting, 0.05 to 48 ng (GC/MS/MS)