Postprint of TrAC Trends in Analytical Chemistry Volume 77, March 2016, Pages 167-185
DOI: https://doi.org/10.1016/j.trac.2015.12.006

Evaluation of analytical calibration based on least-squares linear regression for instrumental techniques: A tutorial review

Francisco Raposo
Instituto de la Grasa, Consejo Superior de Investigaciones Científicas (IG-CSIC)
Campus Universitario Pablo de Olavide, Carretera de Utrera km 1, Edificio 46
41013 Seville, Spain

ABSTRACT

The analytical calibration of an instrumental method is very important, being considered as a key point in method validation. There are different validation guidelines; showing that analytical calibration process variety prevails in terms of nomenclature, methodology employed and acceptance criteria. Very common mistakes in the analytical calibration process are the use of correlation and/or determination coefficients as a test for linearity, the negligence in the heteroscedasticity of the experimental data and selection of appropriate weighting factor, misunderstanding about the regression through the origin and using zero-point calibration. Once the calibration function is established, their linearity can be confirmed by using different procedures such as graphical plots, statistical significance tests and numerical parameters. In particular, deviation from back-calculated concentrations expressed in the form of percentage of relative error (*%RE*) can be considered very useful for unambiguous linearity evaluation. Some case studies were included to explain the linearity assessment from a practical viewpoint.

KEYWORDS

Analytical calibration; calibration error; correlation coefficient; determination coefficient; least-squares method; linear regression; linearity assessment;

Abbreviations: AIC, Akaikes information criterion; ADL, average deviation from linearity; AOAC, association of official analytical chemists; ANO, analysis of variance; AUT, automatically; BC, back-calculated; Cal-VG, calibration section of validation guidelines; C.I, confidence interval; CS, case study; CVDL, coefficient of variation of deviations from linearity; DOF, degree of freedom; DEV, deviation from back-calculated concentrations; EMA, European medicines agency; ESTD, external standard; GOF, goodness of fit; GRA, graphical; H_A, alternative hypothesis; H_0, null hypothesis; ICH, international conference of harmonization; ICP, inductive coupled plasma; INAB, Irish national accreditation board; ISTD, internal standard; IUPAC, International union of pure and applied chemistry; *j*, calibration levels; JRC-FCM, Joint research centre-food contact material; *k*, calibration replicates; LIN, linearity; LLQ; lowest limit of quantification; LOF, lack-of-fit; N, number of calibration data; MAN, Mandel; NATA, national association of testing authorities of Australia; OLS, ordinary least-squares; PAR, peak area ratio; PE, pure error; QC, quality coefficient; *r*, correlation coefficient; R^2, determination coefficient; RE, relative error; REG, regression; RES, residuals; RR, relative residual; RSE, residual standard error; RTO, regression through origin; SE, standard error; SSDL, sum of squares of deviations from linearity; SLO, slope; SQT, significance of quadratic term; STA, statistically; SWGTOX, scientific working group for forensic toxicology, TOST, two-one sided test; US FDA, United States food and drug administration; USP, United States pharmacopeia; VFAs, volatile fatty acids; WF, weighting factor ; WLS, weighted least squares; ZPC, zero point calibration.

CONTENTS

1. Introduction

The validation is required in analytical chemistry to demonstrate the performance of the method and the reliability and consistency of the analytical results. Therefore, before an analytical method can be implemented for routine use, it must first be validated to demonstrate that it is suitable for its intended purpose. Several chemical organizations have developed at national or international level different validation guidelines in the field of analytical chemistry. Among them, some international well known validation references are provided by the Association of Official Analytical Chemists (AOAC) **[1]**, the International Union of Pure and Applied Chemistry (IUPAC) **[2]**, the analytical chemistry group EURACHEM **[3,4]** and the European Medicines Agency (EMA) **[5]**. On the other hand, some validation guidelines have been published at national level by regulatory authorities such as US FDA (the United States Food and Drug Administration) **[6]**, USP (the United States Pharmacopeia) **[7]**, ANVISA (National Health Surveillance Agency of Brazil) **[8]**, INAB (Irish National Accreditation Board) **[9]** and NATA (National Association of Testing Authorities of Australia) **[10]**. Also it is possible to find specific validation guidelines published by analytical companies such as Agilent Technologies (authored by Huber) **[11].** Moreover some guidelines are specific for particular research areas such as the food contact materials (FCM) (authored by Bratinova and co-workers belonging to European Commission-EC/Joint Research Center-JRC/Institute for Health and Consumer Protection-IHCP) **[12]** and the scientific working group for forensic toxicology (SWGTOX) **[13]**. Although there is a general agreement among these validation guidelines, diversity prevails in terms of nomenclature, methodology employed and acceptance criteria **[14]**. In addition, analytical chemists mostly are familiar with the validation guidelines relating to their research area but experimental designs and acceptance criteria are different among diverse disciplines **[15]**. Specifically for the pharmaceutical field, representatives from the industry and regulatory agencies from Europe, USA and Japan tried to harmonize the terms and basic requirements for new pharmaceuticals trough the validation guideline so-called International Conference of Harmonization (ICH) **[16]**.

The method validation procedure includes different performance parameters that have been defined and commented in the majority of international and national guidelines. It is important to consider that the analytical calibration of an instrumental method is a very important part of its development, therefore calibration process can be considered as one of the key points in method validation. However, very often reviewing the scientific literature it is possible to conclude that currently there are some problems relating to analytical calibration. Firstly, there is ambiguous terminology (terms and definitions) dealing to calibration practice. Secondly, there is a lack of clear understanding about the analytical calibration procedure, from planning the experiments to fitting the experimental data. Thirdly, the use of correlation (r) and/or determination (R^2) coefficients as a test for linearity, the negligence in the heteroscedasticity of the experimental data, misunderstanding about the regression through the origin (RTO) and using zero-point calibration (ZPC) and finally, the presence of curvilinearity can be considered as very common mistakes in the analytical calibration process.

The objectives of this manuscript are:

1) To summarize the analytical calibration guidelines issued by different chemical institutions for validation of analytical and bioanalytical methods. By this way, it is possible to report the inconsistencies among documents dedicated to calibration as part of the validation of instrumental methods.
2) To explain theoretical and experimental concepts dealing to analytical calibration as a way to perform a suitable and complete calibration procedure, including the selection of the optimal calibration function and the clarification of the linearity concept.
3) To understand the assessment of the linearity for calibration curves by using some examples or case studies.

2. Calibration of analytical methods

2.1. Calibration: general concepts

Calibration is a metrological operation formally defined as the operation that, under specified conditions, in a first step, establishes a relation between the quantity values with measurement uncertainties provided by measurement standards and corresponding indications with associated measurement uncertainties and, in a second step, uses this information to establish a relation for obtaining a measurement result from an indication **[17]**. Other general definition for calibration has been provided by IUPAC as the operation that relates an output quantity to an input quantity for a measuring system under given conditions **[18]**. These generic definitions have given rise to misunderstanding, as it is usual to find in the literature different names for the calibration operation on the one hand, and different types of calibration called by the same name, on the other. By this way, IUPAC distinguish between two different approaches to calibration in the analytical chemistry field: firstly, for species identification or qualitative analyses; and secondly for quantitative analysis **[18]**. In addition, in the last case depending on the relation between those quantities and the measuring system, different classifications of quantitative calibration can be considered such as direct (also known as absolute) and indirect (also known as relative). Additionally, the term calibration has been applied to equipments and processes **[19]**.

This document will be focused in the chemical calibration process, which means the quantitative indirect calibration process, also known as **analytical calibration**. The most frequent analytical calibration operations in chemical analysis are related to spectrophotometric or chromatographic procedures. In these cases, the value of the standard (reference value) is expressed in a quantity different from the output quantity; therefore the measurement and the measurand are different. The analytical calibration is traditionally carried out in order to establish a relationship between the instrumental responses and some different calibration standards characterized by a known value of the measurand. Using IUPAC definition, calibration in analytical chemistry refers to the operation that determines the functional relationship between measured values and analytical quantities, characterizing types of analytes and their amount **[18]**. That is, the estimation of a mathematical function that will relate the instrumental signal to the analytical property to be determined (usually the concentration).

2.2. Analytical calibration: stages

In summary, analytical calibration involves the preparation of a set of standards containing a known amount of the analyte of interest, measuring the instrument response for each standard and establishing the relationship between the instrument response and analyte concentration. Taking into account that measurements are instrument specific, therefore, the analytical calibration process can be separated in two stages. Firstly, planning or designing the experiments for the calibration study; and secondly, fitting the experimental data. In the following, each stage will be described in more detail including some information compiled from 14 validation guidelines with relevance among the scientific community. Calibration information from these validation guidelines (Cal-VG) has been summarized in Table 1 for further evaluation.

2.2.1. Stage 1: Planning the experiments for the calibration study

The most common calibration technique is the multipoint system, in which several different points on the calibration curve are used to calculate the response versus the concentration relationship. However, the design of multipoint calibration experiments strongly depends on the purpose of the experiment and on existing knowledge. It is important to note that the calibration study design must answer the main question: "*how many experiments are needed?".* Due to time and other constrains, this question often translates into *"what is the absolute minimum needed?"*. The idea to suggest a minimum set of standards for calibration must be proposed with awareness, because it is assumed that only this minimum would then be carried out. It must be pointed out that the recommended number of calibration experiments varies in the literature. There is no one formula to fit all answers because every situation for analytical calibration is different. Some practical aspects such as cost and availability of material used to prepare the calibration curve may perhaps also influence the choice of the number of experiments. However, these aspects should not impair the accuracy of the experimental results generated. Analysts should use some form of systematic planning to obtain the experimental data to achieve the goals of the calibration. By this way, they should be able to answer the following key questions that will help to carry out an appropriate experimental calibration design.

2.2.1.1. *“Which range of concentration must be covered by the calibration curve?”*

There is an agreement among Cal-VG that the calibration should cover a working range in which the usual content of real samples is expected. By this way, a working range between 0-150% can be considered as appropriate.

2.2.1.2*. “How many sequences of calibration will be carry out?”*

The information dealing to this question was not included in the majority of Cal-VG. Specifically, EMA **[5]** and NATA **[10]** suggested at least **three** sequences or rounds for calibration. A higher number, six, was suggested by JRC-FCM **[12]** but considering only single measurement, what is apparently equivalent to 2 or 3 sequences using triplicate and duplicate measurements, respectively.

Independently to the number of sequences, two important issues are the independency of standards and the stability of the instrumental signal over time. Firstly, the sequences should ideally be independent from each other, because replicate measurements on the same calibration standard give only partial information about the calibration variability. It would only cover the precision of the instrument used to make the measurements, and not the preparation of the standards. Therefore, each sequence should analyze “genuine” standards. Secondly, there are many factors which affect the magnitude of the instrument response that could vary considerably from day to day. Then, the different sequences must be obtained in a well planned study carried out over at least 2-3 different weeks of analytical work.

2.2.1.3. “How many points (calibration levels) are necessary for a calibration curve?”

A sufficient number of calibration standards is needed to define the response profile in relation to concentration. In general, the more points exist in the calibration curve the better. However, different number of calibration levels can be found in the literature, ranging from 3 to 10. Specifically, for Cal-VG there is an agreement that 5-6 calibration standards are necessary to carry out an appropriate calibration.

2.2.1.4. *“How should the calibration levels be distributed?”*

Another issue related to the number of standards is their distribution over the whole working range. Evenly distributed or equidistant concentrations could be considered as the best option. For short calibration ranges, to distribute the calibration standards to be equidistant is relatively easy to design. However, for a wide calibration range the application of calibration designs based on standard concentrations that correspond to multiples of the next concentration is frequently found in practise. This approach should be strongly discouraged because the relatively broad spacing of the upper standards in such geometric series could mask the situation where the detector is reaching saturation and the instrument’s responses are levelling off somewhere between the last two standards. Therefore, it is preferable to use a **partial arithmetic series**, where the concentrations of the upper standards differ by a constant amount, not by a constant factor. As example, for a calibration range between 10-1000 mg/L, an acceptable 7 point series should be 10, 50, 100, 250, 500, 750 and 1000 mg/L.

2.2.1.5*. “How many replicates are necessary for each calibration point?”*

Replication of calibration standards is an excellent way to minimised the random calibration error and then increase the precision of the values predicted from measurements of real samples. However, only some of the selected Cal-VG recommended carrying out replicate analyses for calibration curves (IUPAC **[2]**, EMA **[5]**, INAB **[9]**, and SWGTOX **[13]**). They suggest a replication pattern of two, three or more. Triplicate measurements of each standard can be found in many experimental calibration curves for research studies. It is important to note that more than six replicates do not provide additional benefit from the statistical point of view **[20]**. Therefore, considering the advantages of replicate measurements against the time necessary and other economic issues, **three** replicates at each concentration level can be considered as an appropriate quantity of replicates.

2.2.1.6. “Which calibration (quantification) mode, external or internal standard (ESTD/ISTD) methodology, *should be used?”*

This subject is directly related to the quantitation principle or methodological calibration approach **[21]**. In ESTD, the response signal of the analyte alone is plotted against concentration to generate a calibration curve. In contrast, ISTD requires a structural analogue of the analyte to be measured which is added to calibration standards and also to samples. In this case, the response signal ratio between analyte/IS is plotted versus analyte/IS concentration or alternatively only versus analyte concentration. Although ISTD is generally beneficial for classical instrumentation techniques, the experimental data should be cautiously checked before to take a choice to the methodology to be used from the quantitative viewpoint **[22]**.

2.2.2. Stage 2: Fitting the experimental data by regression triplet

In statistic science, the term regression is used to describe a group of methods that summarize the degree of association between one variable (or set of variables) and another variable (or set of variables) **[23]**. Regression and correlation play an important part in the interpretation of quantitative analytical methods and also for comparative purposes. The concepts of correlation and regression are intimately related because the calculation and handling of data are similarly based on least-squares method. Nevertheless, correlation and regression must be interpreted totally different **[24]**. Correlation may be described as the degree of association between two random variables, whereas regression expresses the form of the relationship between specified values of one (the independent) variable and the means of all corresponding values of the second (the dependent) variable. An important application of regression is the case of analytical calibration where both variables (instrumental response and concentration of analyte in calibration standards) normally show a direct relation. In summary, linear regression methods try to determine the best linear relationship between experimental data points while correlation assesses the association between them.

An important concept in this manuscript for fitting the calibration data is the so-called regression triplet, which include method, model and fitting technique. The application of calibration functions requires a mathematical equation that relates the instrumental response and the concentration of the calibration standards to predict the concentration of unknown samples. Although this procedure is very common among analysts, a significant source of bias and imprecision in analytical measurements can be caused by the inadequate choice of the regression triplet for the standard curve that finally will transform signal measurements of samples into concentration units. Therefore, it is important to note that the regression triplet must be selected carefully.

2.2.2.1. Regression method: Least-squares

There are various regression methods such as ranked regression, multiple linear regression, nonlinear regression, principal-component regression, partial least-squares regression. The most common statistical method used is the **least-squares** regression, which works by finding the "best curve" through the data that minimizes the residual sum of squares.

2.2.2.2. Regression model: Linear

There are a number of least-squares regression models such as linear, logarithmic, exponential and power. The most common measurement model is the one described by the simple **linear** function because it has many advantages, among them, simplicity and ease of use. Therefore, historically many analytical methods have relied on linear models for the calibration relationship, where calibration data (pairs of analyte concentrations-x_i, and instrument responses-y_i) are used as an input to the least-squares regression.

The literature is plenty of information relating to linear regression. Among the interesting information it is possible to highlight a book fully dedicated to this topic by Montgomery et al. **[25]**. More recently, Andrade-Garda et al. published and excellent chapter devoted to classical linear regression by least-squares method **[26]**. Taking into account the information included in these references and from general texts and books, in this manuscript only fundamental information dealing to linear regression will be covered.

The linear regression analysis yield an equation that gives the best fit to the data points. This **calibration function** can be defined as the mathematical relationship for the chemical measurement process, relating the expected value of the observed signal or response variable to the analyte amount. An observation on the i^{th} unit in the population, denoted by y_i, is:

$$Y_i = \beta_0 + \beta_1 \cdot x_i + \varepsilon_i \tag{1}$$

where the parameters β_1 *and* β_0 are usually called regression coefficients. These coefficients have a simple and often useful interpretation; β_1 is the slope of the regression line, β_0 is the intercept of the regression line and ε_i is the difference between the observed value of Y and the value on the true regression line that corresponds to $x = x_i$.

It is also important to define **the prediction** equation*:*

$$\hat{y}_i = b_0 + b_1 \cdot x_i + e_i \qquad (2)$$

where $\hat{y}_i$ is the predicted value of Y for x_i, b_0 and b_1 are the estimators of regression coefficients and e_i are the residuals $(e_i = y_i - \hat{y}_i)$, representing the difference between the observed and predicted value of *Y*. Each fit to calibration data can be carried out using modern statistical software packages and the typical output from a linear regression analysis includes the necessary statistical parameters for calculation such as slope (b_1), intercept (b_0) and their corresponding standard error (SE) or standard deviation (SD) such as SE_{b1} and SE_{b0}, for the slope value and the y-intercept value, respectively. Additional information such as *r*, R^2, residual standard error (RSE), *F* or Fisher-Snedecor statistic, degrees of freedom (DOF) and different sum of squares (SS) values can be also obtained from a typical regression analysis **[27]**. It is important to note that the estimators of the regression coefficients obtained using the best regression model should be unbiased, but they have confidence intervals (C.I) around their mean values, which vary following the confidence level (normally at 95%). Therefore, a calibration line should be described appropriately using the mean values but also their associated errors **[28]**

2.2.2.3. Regression fitting technique

There are two fitting techniques for linear regression model using least-squares method such as ordinary least-squares (OLS) and weighted least-squares (WLS) **[29]**. This classification is based in the behaviour of the response's standard deviations or variances over the selected working range of the calibration curve. There are several statistical tests that could be used to check for the homogeneity/heterogeneity of standard deviations or variances. They were described in different papers and general textbooks and were reviewed by Sayago et al. **[30]**. In spite that is a normal that most of reports in the literature refer to OLS, it is important to note that OLS should be only used when experimental data have constant variance (homoscedasticity) while WLS is more appropriate when the variance varies (heteroscedasticity). Normally, the fitting technique choice will depend on the range of concentrations of interest. When a narrow range is considered, an ordinary or unweighted linear model is usually adapted, while a larger range may require a more complex or weighted model.

2.2.2.3.1. Ordinary least-squares: assumptions

OLS regression is often selected for a mathematical fit of the relation between concentration and instrumental response. It is usually the first guess and the starting point for the selection of regression curve. There are some basic assumptions necessary for OLS to be valid: (i) relationship between *y* and *x* variables is linear; (ii) *x* is without error or less than one-tenth of the error in *y*; (iii) errors in *y* are normally distributed; (iv) error in *y* is homoscedastic; and (v) errors associated with different observations are independent **[25].** It is important to note that all the assumptions are not routinely checked, but their verification using different statistical tests in flowchart mode was proposed by de **Souza and Junqueira [31]**. Considering that OLS provides statistically accurate estimates only when all assumptions are fulfilled, in any other case, when some assumptions are not satisfied, OLS can not be considered as the best fitting technique. Nevertheless, from a general point of view, violation of the main assumption relating to variance homogeneity is very frequent due to the fact that many analytical methods produce heteroscedastic data, situation in which every calibration point does not have equal impact on the regression **[32].** Then deviations at high concentrations will influence the regression line more than deviations at low concentrations. In fact, the assumption of constant variance for instrumental analysis is normally incorrect. Instead of the variance, the RSD is the constant parameter over a considerable range for many analytical methods. In this case, OLS is not totally adequate for fitting regression and provides biased estimators of regression coefficients (slope and intercept), mainly sensitive to extreme data points **[33]**. In consequence, concentration values for unknown samples could be incorrectly estimated **[34]**.

2.2.2.3.2. Weighted least-squares: weighting factors

Heteroscedasticity must not be neglected in the analytical calibration process. The solution for non-constant variance is to use an alternative regression procedure such as WLS, which is similar to OLS but defines weights to the calibration data **[35]**. Relating to Cal-VG information, the majority of guidelines (9 out of 14) includes some information about WLS in cases of heteroscedasticity. By this way, in spite of their extra complexity, WLS is now becoming rather more common fitting technique for calibration purposes of inductive coupled plasma (ICP) **[36]** and chromatographic techniques **[37, 38]**. The general concept of WLS is well understood; the principle of weighting is to provide more importance to data points with a low variance and less importance to data points with high variance. Therefore, an optimal weighted model will balance the regression line to generate an evenly distributed error throughout the calibration range.

On the other hand, given the evidence of heteroscedasticity, the major problem is the proper assignment of the weighting factors (**WF/w_i**). The question about which weights to apply does not have a simple answer and different WF may be appropriate in different situations depending on the characteristics of the calibration data set. The most reported procedure is to weight according to the inverse of the variance in the response at that point (WF as function of $1/s^2$) **[35]**. However, this WF is generally impractical because it requires several determinations for each calibration point and also because a fresh calibration line should be performed each time the method is used. Alternatively, the variance can be modelled as a direct function of *x* or *y* values. Therefore, WF as function of $1/x^{0.5}$, $1/x$, $1/x^2$ and $1/y^{0.5}$, $1/y$, $1/y^2$ have been evaluated as approximations of variance and used to establish regression functions **[39]**. The selection of WF based in $1/x^2$ as best function was suggested in the literature **[40]** and also has been justified with scientific reasoning **[41]**.

2.2.2.3.3. Common mistakes in regression

There are a couple of very common mistakes in the application of least-squared linear regression. They will be described in detail to avoid them in the routine analytical laboratory work.

2.2.2.3.3.1. Regression using zero-point calibration

A particularized problem in analytical chemistry is to include the ZPC (0, 0) as data on the constructed calibration curve before to apply for the regression fitting technique. Unfortunately, this procedure could be a source of error. Firstly, the majority of analytical instruments have an inputted background signal or noise which is expected to be non-zero. Secondly, the closest way to measure the zero-point calibration is by running a true blank. But even in this case, random noise signal detection is obtained; and only exceptionally, for some chromatographic techniques the blank sample can be considered as a real ZPC. The only way in which ZPC can be added to a regression data set is when a real standard zero-point has been used and the observed response is either zero or too small and then reasonably can be interpreted as zero **[42]**. If the ZPC subject is not considered, the calibration curve used will include a fictitious extra calibration point, being the final effect on reliability results variable depending of the experimental calibration design (levels, replicates, range, and distance of first calibration point to zero). Anyway, the ZPC should not be considered as a true measured data point and therefore, it should be never included falsely to calibration curves **[43]**.

The regression using ZPC is only covered in one guideline of Cal-VG. Specifically, SWGTOX **[13]** states briefly that "the origin shall not be included as a calibration point", although no further explanation is included.

2.2.2.3.3.2. Regression through the origin

RTO is a special case of linear regression where the absence of constant term (intercept) is a simplification of the statistical model **[44]**. Although linear regression is one of the most familiar statistical tools, RTO has been scarcely treated in textbooks. In addition, when this subject has been treated, it was in controversial mode with no clear statement about their use **[45]**. By one hand, it is possible to find some advice that dropping the constant term from a regression could diminish the model's fit to the calibration data. On the other hand, from the theoretical point of view, there are some circumstances under the studies in which RTO looks suitable or even required. For example, in a chemical reaction in which there is not input (x=0) there will not be output (y=0). However, although theory forbids a constant term value in the regression equation careful consideration of data should be necessary. By this way, to know theoretically that y=0 when x=0 can not be considered enough justification to apply RTO. In addition, it is important to note that the majority of instrumental devices include software packages to allow creating a calibration curve that is forced through the origin (for example, by specifying removal of the intercept or selecting a zero constant option) **[42]**. In this case, the regression parameters that will be used to estimate the concentration of unknown samples are obtained using different equations to the best line through the centroid (point which is the average of the given x and y values) **[28]**. However, a calibration curve must not be forced trough the origin unless it is demonstrated that the intercept is not significantly different from zero. Therefore, once it has been established that a linear fit of the data is appropriate, it should be very useful to test whether the intercept significantly differs from zero or contrarily is totally negligible. There are different statistical ways to check the significance of intercept values:

- One approach is by means of Student's test statistic (*t*-test) or the corresponding *p*-value. Considering the null hypothesis b_0=0 (usually at 95% confidence level). If the intercept is not significantly different from zero, then the *t*-experimental value will be lower than the *t*-critical value. In the same line, the *p*-value will be higher that the specified 5% significance level (usually comparing with 0.05) **[42]**.
- Another simplest approach is by means of the confidence interval for the intercept. If this spans zero, the constant value can be considered as statistically not significant **[42]**.
- Alternatively, the SE of the intercept (SE_{b0}) can be compared with the intercept value to know if the curve passes through the origin. If intercept value is higher than SE_{b0}, the constant value can be considered statistically significant. If intercept value is lower or equal than SE_{b0}, the constant value can be considered statistically not significant **[46]**.

Nevertheless, it is important to consider that the statistical justification to drop a non-significant intercept is not enough when the practical relevance of this action is considered. Some important information that should be known about RTO is **[44]**: i) the residuals will not sum to zero; ii) the line may not necessarily pass through the regression centroid and then affecting the confidence band for the regression line (does not expand from the centre of the line).

In summary, RTO is a statistical assumption and not an underlying principle for reporting results. Forcing the regression line through the origin is generally not in agreement with the main objective of the regression, which is assessing the best prediction model to fit the data. Therefore, using RTO either arbitrarily or statistically sounds may unfavourably impact quantitative results, highly in the beginning and softly at the end of the linear calibration range.

The RTO issue is briefly treated in only two of Cal-VG. Firstly, IUPAC guideline suggests that "a test for intercept significantly different from zero can also be made" **[2]**. Secondly, Huber considers helpful RTO. He states the following sentence: *"A linear regression equation applied to the results should have an intercept not significantly different from zero. If a significant nonzero intercept is obtained, it should be demonstrated that this has no effect on the accuracy of the method"* **[11]**.

3. Linearity as validation performance parameter

3.1. General approach

3.1.1. Linear versus non-linear regression models

Linear regression is perhaps the most used statistical model in calibration due to be considered as the easiest way to fit experimental data. Therefore, the preliminary evaluation of one analytical method is routinely considered from the linear calibration point of view. Unfortunately, linear regression can be also considered as an abused statistical method taking into consideration the common mistake to presume linear relationship onto any set of experimental calibration data. However, systematically force a function using the linear regression model to fit any set of calibration standards is not required, often irrelevant and may lead to large errors in the measured results for real samples **[47]**. In fact, the straight-line model is never valid for some instrumental methods while for others only in a limited interval range, and above it a significant departure from linearity is present. Particularly, very often occurred in practice for wide calibration ranges, the calibration plot is linear at low analyte concentrations, but a close examination reveals that curvature exists at higher analyte levels. In looking the use of the right model one important question arise: At what point does curvature become so severe that linear calibration is not useful? **[48]**. On the other hand, fitting by non-linear least squares can be challenging, but quadratic (second degree polynomial) least squares model fit appropriately the experimental data when a calibration curve has a smoothly curving nature. In fact, the majority of Cal-VG includes some information relating to the possibility of applying non-linear regression for calibration curves. Therefore, calibration data must be analyzed with care to evaluate if the non-linear regression model provides a better fit than the linear one **[49]**.

3.1.2. Clarification of linearity concept

The terminology dealing to linearity is somewhat confusing because there are different ways of thinking about linear functions. From the systems theory, the linearity criterion refers to the relationship between the quantity introduced (input) and the quantity back-calculated from the calibration curve (output). The response function term looks more suitable because it is used to describe the relationship between the instrumental response and the concentration. As it was explained before, a linear dependence of the signal and the analyte concentration is certainly the most convenient case and commonly used in chemical analysis. However, there are some analytical techniques with a clear non-linear response **[50]**. Therefore, the terms analytical response or response function or standard curve would have been more appropriate for this validation characteristic **[47]**.

The same confusion in concept between response function and linearity of results can be found in the scientific literature, with the majority of guidelines, text and papers dealing to validation including the term linearity. This should be the reason why many analysts still continue to use the linearity terminology when referring to the response function of calibration curves. Exceptionally, some Cal-VG such as EMA **[5]**, FDA **[6]** and ANVISA **[8]** do not contain the word linearity, it was replaced by calibration or standard curve. Similarly, Eurachem **[4]** and SWGTOX **[13]** guidelines used alternative terms such as working range and calibration model, respectively.

On the other hand, another source of confusion dealing to linearity is the different meaning from graphical against mathematical points of view. By this way, graphically should be possible to distinguish two shapes such as straight-line and non-linear. However, a quadratic model showing clearly a parabola or curved graph may be really a linear calibration. This is due to algebraically speaking only the coefficients of the mathematical function are considered.

3.2. Evaluation of linearity

Considering that the scope of this manuscript is restricted to analytical methods dealing to real linear relationship and associated plots, the linearity concept will be maintained. Once the calibration function is established, it should be tested (validated) for total conformance to the regression model. Therefore, the linearity assessment of the calibration function should be checked before its use. As can be seen from Cal-VG, the linearity assessment, as a required performance in method validation, has always been subject to different definitions and interpretations. The same problematic is present in published papers **[49, 51-54]**. This is due to the recommendations that are sometimes complicated or controversial and do not detail the experimental designs, the statistical calculations and the respective assumptions that need to be checked. In addition, there are very limited applicable implementation procedures that can be followed by analysts for assessing linearity. This manuscript was planned as a way to establish a practical approach to evaluate the linearity range applied to in-house validated methods. Therefore, linearity can be further confirmed by using alternative procedures such as graphical plots, statistical tests and numerical parameters that were previously reported in the analytical literature. They have been compiled to facilitate comparisons of analytical methodologies from the linearity assessment point of view and, some of them, should be proposed as formal acceptance criteria. Table 1 summarizes the different linearity tests included in the Cal-VG for decision making. Following the different linearity procedures will be described in detail.

3.2.1. Graphical mode

The first step for linearity assessment should be plotting the paired data and to carry out the visual inspection of the experimental data provided by the calibration plot. By this way, it is possible to quickly identify problems with the data showing the general shape of the curve and how well the curve fits the calibration standards. This suggestion was included in the majority of Cal-VG. More specific plots to check linearity are detailed following.

3.2.1.1. Residuals plot

Linearity can be tested informally by examination the plot of residuals produced by linear regression of the responses on the concentrations in an appropriate calibration set. A visual evaluation of the pattern of the residuals plot is very simple and straightforward. In spite of their usefulness, a residual plot can not be considered as a potent tool to identify deviations from the linear regression model. This is due to no statistical test is involved and some experience may be necessary for the interpretation of these plots. The residual plot is useful for the judgement of linearity as per the sign sequence. A residual plot shown random behaviour in a constant range, without systematic pattern is indicative of the correct linearity. However, residuals values fairly distributed between positive and negative values have been suggested as indicator of the deviation in relation to the linearity assumption **[25, 31]**. The majority of Val-VG includes implicitly the residuals plot as a way to check the linearity of calibration curves.

3.2.1.2. Linearity plot

The sensitivity of the analytical instrumental was suggested by Dorschel et al. to be useful to check the linearity of the calibration curves **[55].** This approach was also included in a validation paper published by Huber **[56]**, and then sometimes is termed in the literature as Huber's linearity test **[57]**. In this method, the sensitivity is easy to calculate as the ratio between the individual response values and the corresponding concentrations. Specifically for chromatographic techniques, the linear range was defined as the range of concentration over which the sensitivity (slope of calibration curve) in the form of response factor (RF) for ESTD or relative response factor (RRF) for ISTD can be considered as constant within a defined tolerance. A tolerance limit of ±5% was suggested by IUPAC to evaluate the linearity of chromatographic calibration curves **[58]**. Considering that visual examination can help to evaluate the data, usually the assessment is carried out by figures referred as **linearity or sensitivity plots**, representing the sensitivity versus concentration (expressed using logarithmic scale for wide calibration ranges). By this way, the median value of slope is selected instead the average value taking into account that the median value is an estimator robust to outliers. Tolerance limits are determined multiplying the median value by constant factors of 0.95 and 1.05 for lower and upper limits, respectively. The calibration range is considered to be within the linear range if no results are outside of the tolerance limits. In the same way, the data points should form a straight line with very low or zero slope.

This plot methodology was only included in two guidelines of Cal-VG, INAB **[9]** and Agilent **[11]**, as a way to check appropriately the linearity of calibration curves.

3.2.2. Non-graphical mode

Unfortunately, linearity cannot be demonstrated over a given working range by simple visual observation of calibration and residuals plots because both processes are subjective. The US FDA validation guideline includes that "*the simplest model that adequately describes the concentration–response relationship should be used*" **[6]**.

3.2.2.1. Statistical significance tests

The quality of linear fit must be evaluated using **significance tests** to check if the equation defining a calibration standard is appropriate. Although the calculations differ, the significance tests are carried out in similar way stating the null (H_0) and alternative hypotheses (H_A). The H_0 is usually that there no difference between the values being compared. There are many types of significance tests useful to examine different statistical parameters, being very common the *t*-test (comparing mean values) and the *F*-test (comparing variances). Four significance tests have been reported in the literature for linearity testing, all of then based on Fisher-Snedecor or *F*-distribution corresponding to the comparison between the tabulated (F_{crit}) and the experimental (F_{exp}) *F*-values.

3.2.2.1.1. Analysis of variance (ANOVA) test

This routine statistical test was suggested by Danzer and Currie **[18]**. The experimental value is given by:

$$F_{\text{exp}} = (s_{\text{y/x}} / s_{\text{y}})^2 \qquad (3)$$

where $s_{y/x}$ is RSE and s_y is the pure error (PE). Therefore, this test considers useful to check whether the residual variance is larger than the squared PE from the study of genuine replicate samples.

In spite that this test is an IUPAC recommendation, it was not considered by none of Cal-VG.

3.2.2.1.2. Analysis of Variance lack-of-fit test (ANOVA-LOF)

The ANOVA-LOF or simply LOF test has been recommended previously as a sound test applied to calibration experiments as sole basis **[59, 60]**. The LOF is a statistical test derived from the analysis of variance applied to regression. This test is based on the comparison of the variability of a set of calibration measurements and is relatively simple. It can be easily implemented on much spreadsheet software. Specifically, RSE can be divided in two components: i) PE or random experimental error, corresponding to the variability of replicates around their common means; ii) LOF error, which is the variability of the group means around the regression line. In this case the null hypothesis means that no significant LOF is present.

This test can be carried out in two ways **[61]**:

- Fisher-Snedecor test. This is a conventional *F*-test between the two components of RSE, specifically the ratio between the so-called LOF and PE variances. The model is adequate (assumption of absence of LOF) if F_{exp} *is* lower than F_{crit}.
- *p*-value test. This is a statistical parameter that represents the probability. It is important to note that a high *p*-value means that the starting assumption (absence of LOF) cannot be rejected. Therefore, a *p*-value greater than 0.05 (at

95% confidence level) must be obtained to fail reject (accept) the starting assumption about the LOF for the linearity of the calibration curve.

Unfortunately, this test has some drawbacks that make it not necessarily conclusive, fail to be quantitative and then **affecting its usefulness** in general application **[62, 63]**:

- Sufficient replicates of calibration during the validation of the analytical method are required. It exist the possibility that data may be judged nonlinear if an inadequate number of replicates are employed in the data analysis. However, replicate requirement should not be a practical limitation in most cases. Most important is the fact that true replicates must be used to avoid overestimates failing the test due to using simple replication of the calibration standard.
- A major problem of this test is that it depends on the precision of the method. By this way, for precise data the test is very sensitive and less likely to pass it. Similarly, when imprecise data are processed, the test can fail to detect apparently large deviations from linearity.
- If the null hypothesis is not rejected, it does not mean that the linear model is correct, only that the model is not contradicted by the data or that insufficient data exist to detect the inadequacies of the model. In addition, there are causes of LOF other than non-linearity that can arise in calibration curves.

This statistical test was included in 4 of the Cal-VG, but in a different form. By short way, IUPAC **[2]** included that a test of significance can be undertaken by comparing LOF variance with that due to PE, including the explanation that non-linearity is not the only reason for LOF. INAB **[9]** stated that *“a LOF test may be utilised to underpin the visual assessment”*. Also, SWGTOX **[13]** briefly suggested to apply LOF as alternatively evaluate unweighted calibration model. On the other hand, JRC-FCM **[12]** explained widely the LOF test following the ISO 11095 approach **[60]**.

3.2.2.1.3. Mandel´s test

This test is useful to make a decision about linearity by seeing whether a straight line or quadratic curve fits the calibration data better **[64]**. Mandel´s test compares the RSE of both models by an *F*-test with a null hypothesis that there are not significant differences between the residual variances of both models. For this test, if F_{exp} *is* lower than F_{crit} means that the quadratic model did not provide a significantly better fit than the straight line model. Therefore, in this case the quadratic or second degree polynomial model cannot be selected on statistical grounds.

It is important to note that this test was also suggested by IUPAC **[18]**, but this approach included a simplified formula when compared to the original one given by Mandel:

$$F_{IUPAC} = [(S_{y/x})^2_{LIN} - (S_{y/x})^2_{NON\text{-}LIN})] / (S_{y/x})^2_{NON\text{-}LIN} \qquad (4)$$

$$F_{Mandel} = [(N\text{-}2)\cdot(S_{y/x})^2_{LIN} - (N\text{-}3)\cdot (S_{y/x})^2_{NON\text{-}LIN})] / (S_{y/x})^2_{NON\text{-}LIN} \qquad (5)$$

where N is the overall number of standards necessary to calculate DOF. It was found that the IUPAC simplification is not valid in general although it can be used safely when the variances of the linear and alternative models are very similar, typically when they differ by less than 10%. Another drawback of Mandel´s test is that a higher number of calibration standards, compared to routine calibration, are necessary to detect non-linearity correctly **[65]**.

This test was suggested only in one of Cal-VG. By this way, JRC-FCM **[12]** includes a wide explanation according to IUPAC suggestion **[18]**.

3.2.2.1.4. Significance of the quadratic term (SQT) test

This is another approach to evaluate the presence of systematic curvature for calibration data. Similarly to Mandel´s test, SQT compares a straight-line regression model with a second-order regression model based on comparison of mean squares (MS) values of quadratic term and residuals. In this case, deviation from linearity is detected when MS are different, and then F_{exp} exceeds the F_{crit} value and the QT can be considered as significant **[42]**. Alternatively, deviation from linearity is detected when the second-order regression coefficient (b_2) is significantly different from zero, than means that b_2 interval does not include the zero value **[66]**.

This test has a low prevalence in Cal-VG. Only SWGTOX **[13]** cited the use of SQT as alternative test for calibration model assessment.

3.2.2.2. Numerical parameters

This section includes different numerical parameters used for evaluating the goodness of fit (GOF) of the calibration curves, which means how well a linear regression model fits the calibration standards.

3.2.2.2.1. *r* and/or *R^2*

Both coefficients are regression parameters commonly used when performing analytical calibration assessments as indicators for GOF. These coefficients are very popular among scientific community, probably due to the simple communication and suggestive interpretation of concepts and also quickly availability from statistical software and calculators. Considering their wide use it is important to define briefly both coefficients: *r* is an indicator of the degree of correlation between both variables, while R^2 is an indicator of the proportion of variability in the response explained by the regression. In practice, calibration curves should be dealt with R^2 rather than *r* since R^2 shows the proportion by which the variance of the dependent variable is reduced by knowledge of the corresponding independent variable. However, *r* values have been extensively reported for GOF of calibration curves. In addition, erroneously both coefficients have been used interchangeably **[12, 34]**. In any case, the use of *r* and/or R^2 is a subject of much controversy and it is clearly reflected in the information of Cal-VG dealing to this issue. Firstly, the linearity evaluation of analytic methods reported in some guidelines such as USP **[7]**, Agilent **[11]** and ICH **[16]** suggest that *r* values should be reported. Secondly, ANVISA **[8]** includes an acceptance criterion that relies in *r*. Finally, IUPAC [**2**], Eurachem **[3,4]**, INAB **[9]**, NATA **[10]** and SWGTOX **[13]** criticize those analytical methods reporting high values of *r* and/or R^2, close to unity, that considering linearity as appropriate without further evaluation. It is important to consider the meaning of both coefficients to state that correlation or response variability and linearity are only loosely related. Therefore, these statistical parameters should be considered misleading in the context of testing the linearity. In fact, the use of *r* and/or R^2 as linearity test can be considered as the most frequently reported misconception about linear regression. Unfortunately, this is a topic of high relevance over the time with no definitive solution among scientific community. Maximum relevance should be provided by analysts engaged in the validation of analytical methods to definitively avoid the use of *r* and/or R^2 as proof of linearity. To highlight this idea, Table 3 compiles some published articles dealing to calibration curves in which previously was included some information about the misleading use of *r* and/or R^2 for linearity evaluation **[34, 51, 53, 59, 67-90]**.

3.2.2.2.2. The residual standard error (RSE)

This statistical parameter is also known as the standard error of regression (SER) or residual standard deviation (SD_{res}) and represented by $s_{y/x}$. It is a measure of the average residual or deviation of the data from the fitted regression line. It is used to calculate many useful regression statistics including significance test of the intercept and the slope, some outlier test and the C.I. values. In addition, it should also be used as another way to check the GOF of the calibration, considering that it provides a quantification of the spread of the data around the regression line **[91]**. The smaller the RSE, the closer are the measured data point to the calculated calibration curve. Unfortunately, it is affected by the choice of units and their value can only be used to compare the results obtained from the same instrument system. Thus, it is not totally appropriate for comparison because results are depending on the arbitrary units for signal measurement. However, NATA guideline **[10]** suggests their use as a measure of GOF considering the repeatability precision of *y* values.

3.2.2.2.3 The relative standard deviation of the slope (%RSD_{SLOPE})

This mathematical measure of linearity was suggested by Cuadros-Rodriguez et al. **[92]**. The $\%RSD_{SLOPE}$ value is characteristic for the analytical calibration method and independent of the instrument used. It should be used as a mathematical measure of GOF and as a comparison criterion for the dispersion of the experimental data around the regression or calibration line. Expressed as percentage, the equation to calculate this parameter is:

$$\%RSD_{SLOPE} = (SE_{b1} / b_1) *100 \qquad (6)$$

This parameter is scarcely used in Cal-VG, being only reported in JRC-FCM **[12]**. This guideline suggests high tolerance values, 5% for classical chromatography techniques and 8% for more specialised techniques (mass spectroscopy detection). These values contrast with RSD_{SLOPE} values for an appropriate GOF of experimental results around 1-2% as showed the examples included in the reference paper **[92]**.

3.2.2.2.4. Quality Coefficients (QC)

Different quality coefficients (QC) were reported in the literature for the GOF assessment of calibration lines based on calibration curve deviation from either *x* or y values **[93]**. Hu et al. **[94]** defined the QC using the relative residual (RR) values as:

$$QC\ (Hu) = [(\Sigma\ \%\ RR)^2 / (n-1)]^{1/2} \qquad (7)$$

$$\%RR = [\text{residuals} / y_{\text{meas}}] = [(y_{meas}\text{-}y_{pred}) / y_{meas}]\cdot 100 \qquad (8)$$

where measurement deviations or residuals are calculated from the difference between y_{meas} (the measurement at each data point) and y_{pred} (the measurement predicted by regression model). This parameter can be expressed more appropriately as percentage to comparative purpose. The limit of acceptable quality for calibration curves using atomic spectroscopy was set at 5%. These QCs accepts that RSD is constant and measure the average relative deviation of the measurements from the model: the better the model fit to the experimental points, the smaller the QC. Unfortunately, little deviations in the lower sector of the calibration line may result in high relative errors and thus in elevated QC values. As consequence, calibration lines with problems at the highest concentrations may possibly pass unobserved, while small deviations at the lowest concentrations may be considered as unacceptable. In order to deal with this issue an alternative QC using instrumental signal average (ȳ) and based on the acceptance of a constant variance was proposed **[93]**:

$$\text{QC (mean)} = [(\Sigma\ \%\ RR)^2 / (n\text{-}1)]^{½} \qquad (9)$$

$$\%RR = [\text{residuals} / \bar{y}] = [(y_{meas}\text{-}y_{pred}) / \bar{y}]\cdot 100 \qquad (10)$$

The alternative QC eliminated the effect of a different location of a similar residual in the calibration curve and then being more robust because relates each residual with the mean signal response value.

The QC factor has null influence in the information compiled from Cal-VG, because it was never reported as useful parameter to check for linearity criterion.

3.2.2.2.5. Deviation from back-calculated concentrations (DEV)

An aspect very important in analytical chemistry is the fit-for-purpose principle **[3, 4, 48, 85]**. The primary goal of an analytical procedure is to give accurate measurements in the future for real unknown samples; so deviations of model fit with regard to standard calibration concentrations are of primary interest and they must be evaluated. For each standard curve there is an optimal design to obtain the most accurate back-calculated results. By this way, the models should be retained or rejected based on the accuracy of the back-calculated (BC) results regardless of other statistical properties or numerical parameters that could be considered only informative and barely relevant for the objective of the assay **[47]**. In order to assess the linearity, comparing directly the back-calculated concentrations with the theoretical or nominal values of the calibration standards is a simple way to become aware of the error contribution from the whole regression range **[39, 95]**. The deviations from the proposed linear calibration model can be expressed as relative error (RE) of the estimated regression line:

$$\%RE= [(x_{\mathrm{meas}}-x_{theo})/x_{theo}]\cdot 100 \quad (11)$$

where x_{meas} is the measured or experimental value and x_{theo} is the theoretical or nominal concentration. For each calibration point, a negative or positive deviation can be obtained by simple calculation. Σ*%RE is an* overall parameter defined as the sum of absolute values that was proposed to select the best fitting technique for regression **[39, 96]**. The mean disadvantage of this sum parameter is similar to RSE, it can only be used to compare results obtained from the same instrument. On the contrary, one advantage of *%RE* parameter is their usefulness to check individually if appropriate fitting for all the calibration points is carried out, at both high and low concentration levels. Alternatively, this parameter should be useful as a plot. *%RE* versus concentration values using log units has been previously reported as valuable indicator of linearity **[97]**. As novelty in the present manuscript, with the purpose of further accuracy evaluation using this numerical parameter, two different complementary parameters have been proposed. By one hand, $\%RE_{AVER}$, as the overall error from the whole calibration curve. On the other hand, $\%RE_{MAX}$*,* as the maximum specific error at any point of the calibration curve also should be considered.

Some of Cal-VG included acceptability criterion for linearity relating to deviation from regression line. By one hand, Agilent **[11]** and ICH **[16]** only states that analysis of deviation should be helpful for evaluating linearity. On the other hand, US FDA **[6]**, ANVISA **[8]**, JRC-FCM **[12]** clearly specify the conditions to be met in developing a calibration curve such as 20% to the low limit of quantification (LLQ) and 15% to the rest of nominal concentrations. In addition, EMA **[5]** suggests the same deviations limits of 15-20%, but clearly specifies that deviation should be presented as back-calculated concentrations of the calibration standards.

3.2.2.3. Deviation from linearity tests

Recently quadratic and higher-order polynomial models have been used more frequently as functional relationship for calibration curves. As a result, more sophisticated parameters to evaluate the linearity have been described in the literature. Among them, Akaikes information criterion (AIC), average deviation from linearity (ADL), sum of squares of deviations from linearity (SSDL), coefficient of variation of deviations from linearity (CVDL) and two-one sided test (TOST) are the most used. As it was previously stated, this manuscript is dealing specifically to linear regression and further explanation about the new parameters is out of the scope of this manuscript and can be found elsewhere **[98-104]**.

4. Evaluation of linearity: Case studies

The literature is plenty of theoretical information dealing to validation procedures. However, there is in general a lack of practical information, particularly to calibration practice and linearity assessment. Employing some of the criterion for assessing linearity in isolation can be a risk because they are only partly addressed to real linear calibration. To increase the knowledge about this important issue, this section provides some case studies (CS) to explain in detail different ways to check linearity from the practical analytical viewpoint. Therefore, illustrative examples were selected from the literature as approaches that best suited the typical problems dealing to linearity. Table 4 includes CS 1-7 in which the regression data output were obtained using the software package Statgraphic® Plus 5.0. Table 5 summarizes the same regression parameters and additionally the *%RE* information for CS 8-10.

To understand the experimental designs of the different CS is necessary to define some abbreviations: *N* as the total number of calibration data; *j* as the number of calibration levels; *k* as the number of replicates for each calibration level; and PAR as peak area ratio.

4.1. Linearity Plots

Case study 1. Determination of Ibuprofen (1A) and Biochanin (1B) by HPLC [49]

4.1.1. Experimental design

CS 1A: Matrix = mix of water-acetonitrile (50:50); N=14, j=7; k=2

CS 1B: Matrix = mix of water-acetonitrile (67:33); N=14, j=7; k=2

4.1.2. Calibration data: X (amount-mg/tablet) versus Y (signal-arbitrary units)

CS 1A: 103.9/265053; 103.9/261357; 139.3/345915; 139.3/345669; 180.1/445684; 180.1/445753; 200.3/494700; 200.3/493846; 219.9/540221; 219.9/539610 278.1/683881; 278.1/ 683991; 305.7/755890; 305.7/754901.

CS 1B: 0.158/0.1212; 0.158/0.1211; 0.315/0.4036; 0.315/0.4152; 0.631/1.8395; 0.631/1.8351; 1.261/3.8405; 1.261/3.8461; 2.522/8.5235; 2.522/8.5399; 5.045/16.8070; 5.045/16.6986; 10.090/34.0687; 10.090/33.9168

4.1.3. Sensitivity values:

CS 1A: Median = 2471; Tolerance range (±5%) = 2348-2595

CS 1B: Median = 3.0478; Tolerance range (±5%) = 2.8954-3.2002

4.1.4. Linearity evaluation

Linearity plots derived from experimental data are represented in Figure 1. It is important to note that although the same high values of R^2 (0.9998) were obtained, the Figures 1a&b noticeably show two different trends. Then confirming one more time the lack of reliability of R^2 as a proof of linearity. By this way:

- CS 1A represents the ideal case where all the sensitivity values (response factors) are between ±5% tolerance limits. According to the trend of Figure 1c the relationship between drug amount and instrumental signal can be considered as a linear regression in the whole calibration range.
- CS 1B represents an unfavourable example where the majority of sensitivity values (response factors) are outside of ±5% tolerance limits. Therefore, the trend of Figure 1d shows that regression clearly deviates from a straight-line model.

4.2. ANOVA test

Case study 2. Determination of iron by Spectrophotometric method [59, 105]

4.2.1. Experimental design

Matrix= not specified; *N*=12, *j*=6, *k*=2

4.2.2. Calibration data: X (mg/mL) versus Y (absorbance units)

64/138; 64/142; 128/280; 128/282; 192/423; 192/425; 256/565; 256/567; 320/720; 320/725; 384/870; 384/872

4.2.3. Calculation of variances

$(S_{y/x})^2 = 28.05$; $(S_y)^2 = 4.75$

4.2.4. F-test values

F_{exp}= 5.91 (Equation 1); F_{cri} (0.05, 10, 6) = 4.06

4.2.5. Evaluation of linearity

Taking into account that F_{exp} is higher than F_{crit} in this CS the statistical H_0 (there is no difference between variances) would be rejected at 95% confidence level and regression should be considered as non-linear.

4.3. LOF test

Case study 3. Serotonin measurement by reverse-phase HPLC and fluorescence detection [53]

4.3.1. Experimental design

Matrix = *Planaria* sp.; *N*=18, *j*=6, *k*=3

4.3.2. Calibration data: X (amount-ng) versus Y (PAR of serotonin/IS)

CS 3A: 0.35/0.0956; 0.35/0.0948; 0.35/0.0934; 0.5/0.1356; 0.5/0.1393; 0.5/0.1361; 1.0/0.2575; 1.0/0.2551; 1.0/0.2535; 2.0/0.5028; 2.0/0.4962; 2.0/0.4940; 5.0/1.2605; 5.0/1.2534; 5.0/1.2516; 7.0/1.6706; 7.0/1.6950; 7.0/1.6928.

CS 3B: 0.35/0.0914; 0.35/0.0948; 0.35/0.0976; 0.5/0.1356; 0.5/0.1361; 0.5/0.1423; 1.0/0.2475; 1.0/0.2551; 1.0/0.2635; 2.0/0.4820; 2.0/0.4962; 2.0/0.5148; 5.0/1.2216; 5.0/1.2534; 5.0/1.2905; 7.0/1.6206; 7.0/1.6928; 7.0/1.7450.

4.3.3. Error sum squares calculation (see reference 53)

CS 3A: Residual (SS_{RES}) = $4.84 \cdot 10^{-3}$; Pure error (SS_{PE}) = $0.47 \cdot 10^{-3}$; Lack-of-fit (SS_{LOF}) = $4.37 \cdot 10^{-3}$

CS 3B: Residual (SS_{RES}) = 0.0153; Pure error (SS_{PE}) = 0.0109; Lack-of-fit (SS_{LOF}) = 0.0044

4.3.4. Degrees of freedom calculation (see reference 53)

CS 3A&B: Residual (DOF_{RES}) = 16; Pure error (DOF_{PE}) = 12; Lack- of-fit (DOF_{LOF}) = 4

4.3.5. Calculation of associated variances (σ^2=SS/DOF)

CS 3A: Residual (σ^2_{RES})=$3.03 \cdot 10^{-4}$; Pure error (σ^2_{PE}) = $0.38 \cdot 10^{-4}$; Lack-of-fit (σ^2_{LOF}) = $1.09 \cdot 10^{-3}$

CS 3B: Residual (σ^2_{RES})=$9.6 \cdot 10^{-4}$; Pure error (σ^2_{PE}) = $9.1 \cdot 10^{-4}$; Lack-of-fit (σ^2_{LOF}) = $1.1 \cdot 10^{-3}$

4.3.6. F-test values

CS 3A: Fisher ratio (σ^2_{LOF}/ σ^2_{PE}): F_{exp} = 27.97; F_{crit} (0.05, 4, 12) = 3.259

CS 3B: Fisher ratio (σ^2_{LOF}/ σ^2_{PE}): F_{exp} = 1.202; F_{crit} (0.05, 4, 12) = 3.259

4.3.7. Linearity evaluation

In this CS the statistical hypotheses are the null (*H_0*) if there is not LOF (regression is linear), whereas the alternative (*H_A*) means that LOF is present and therefore non-linear model should be selected. To check hypotheses, both *F* values, one obtained experimentally versus the critical one, are compared to obtain a statistical decision about linearity of calibration curves. Unfortunately, LOF is a statistical test that depends of instrumental signal precision. To evaluate the influence of the precision, the calibration data of CS 3A were fictitiously modified in such way that CS 3B are three replications providing the same average value but higher SD and RSD. Specifically, average RSD of true data was lower than 1.0% while for the modified data increased to 3.3%. The results obtained demonstrated that precision is a key factor in the LOF test. By one hand, for original precise data F_{crit} is lower than F_{exp}, which means than the LOF term is highly significant and consequently a different non-linear model should be proposed to describe the relationship between *y* and *x*. On the other hand, the imprecision in the instrumental signal measurements surprisingly provided a lower F_{exp}, and a contradictory conclusion about the positive acceptance of calibration curve linearity can be achieved.

4.4. Mandel´s test

Case study 4. Fluorescence measurement of unspecified analyte [65]

4.4.1. Experimental design

N=11, j=11; k=1

4.4.2. Calibration data: X (concentration-mM) versus Y (signal-arbitrary units)

0/0.10; 1/3.8; 2/7.50; 3/10.0; 4/14.4; 5/17.0; 6/20.7; 7/22.7; 8/25.9; 9/27.5; 10/30.0.

4.4.3. Quadratic regression output:

$Y = (-0.0594\pm0.3344) + (3.8990\pm0.1556)\cdot x + (-0.0888\pm0.0149)\cdot x^2$; R^2= 0.9985; RSE= 0.4389

4.4.4. Calculation of variances

$(S_{y/x})_{LIN}{}^2 = 0.9232$; $(S_{y/X})_{NON}{}^2 = 0.1927$;

4.4.5. F-test values (see reference 65 for calculation)

F_{IUPAC} =3.79 (Equation 4); F_{Mandel} = 35.13 (Equation 5); F_{crit} (0.05, 1, 8) = 5.32

4.4.6. Linearity evaluation

Mandel´s test can be summarized as a comparison of RSE between linear and non-linear models using the well-know conceptual F-test. In this case, the statistical hypotheses are the null (H_0) if there is not difference between RSE values and then the quadratic model does not improve the linear one. It is important to note that the original formulation of Mandel, including DOF to avoid erroneous conclusion, should be considered. For this example, using Mandel definition the F_{exp} is higher than F_{crit} and then H_0 can be rejected and the quadratic model considered as appropriate. On the other hand, IUPAC definition provides a contrary conclusion ($F_{exp} < F_{crit}$) considering that H_0 cannot be rejected and then a linear model could be erroneously accepted because the alternative second order polynomial model does not improve the fit from the statistical point of view.

4.5. Significance of quadratic term (SQT)

<u>Case study 5.</u> No information was reported about the source of measurement [42]

4.5.1. Experimental design

N=11, *j*=11; *k*=1

4.5.2. Calibration data: X (concentration) versus Y (response)

0/0.10; 2/19; 4/40; 6/71; 8/116; 10/164; 12/225; 14/299; 16/376; 18/466; 20/566

4.5.3. Quadratic regression output

Y = (2.3916±1.47059) + (4.5104±0.342103)·x + (1.1815±0.01647)·x^2; R^2= 0.9999; RSE= 1.9303

4.5.4. Mean squares calculation (Statgraphic® Plus 5.0 software was used)

Quadratic term (MS_{QT})=19164; Mean square of residuals (MS_{RES})=3.726;

4.5.5. *F*-test values

F_{exp} (MS_{QT}/ MS_{RES}) = 5143; F_{crit} (0.05, 1, 8) = 5.32

4.5.6. Linearity evaluation

For this statistical test the hypotheses are similar that for Mandel´s test but using MS values. The experimental value of *F*, coming from the comparison of the quadratic MS term with the residual MS, exceeds the critical value. Similarly, the interval limit of the quadratic coefficient ranges 1.165<b_2<1.198, excluding the zero value. Therefore, the QT is significant and the second-order polynomial is a better representation of the data compared with a linear model.

4.6. %RSD_{slope}

Case study 6. Determination of caffeine by HPLC [92]

4.6.1. Experimental design

N=12, j=4; k=3

4.6.2. Calibration data: X (concentration-μg/mL) versus Y (PA signal-arbitrary units) 5/1181; 5/1124; 5/1123; 10/2302; 10/2184; 10/2357; 15/3463; 15/3304; 15/3253; 20/4527; 20/4476; 20/4424

4.6.3. %RSD_{slope} calculation

%RSD_{slope} = SE_{b1}/b_1*100 (Equation 4) = (3.68/221.16)*100 = 1.66%

4.6.4. Linearity evaluation

The error in the slope of the calibration curve can be related with the quality of analytical calibration. In this case, the error value obtained is lower than the suggested experimental value of 2% and then the regression could be considered as linear.

4.7. Quality coefficient (QC)

Case study 7. Determination of copper by AAS-Graphite furnace [93]

4.7.1. Experimental design

N=6; j=6; k=1

4.7.2. Calibration data: X (concentration-ng/mL) versus Y (absorb.-arbitrary units) 0/0.0000; 10/0.0410; 40/0.1240; 60/0.1890; 80/0.2460; 100/0.3050

4.7.3. QCs calculation

QC (Hu) = 6.2 (Equations 5-6); QC (mean) = 2.0 (Equations 7-8)

4.7.4. Linearity evaluation

The decision support based in different QC shows controversial results. The proposal of Hu et al. gave a value higher than 5% suggested as limit of acceptable quality. On the contrary, the QC (mean) provided a value than can be considered as satisfactory. This is due that QC (Hu) is more severe for deviations in the low concentration range and too tolerant for deviations in the high concentration range.

4.8. RE as function of RTO and ZPC

Case study 8. Determination of unspecified drug by HPLC [96]

4.8.1. Experimental data

N=10; j=10; k=1

4.8.2. Calibration data: X (concentration-ng/mL) vs Y (signal-arbitrary units·10^{-3})

1/1.4; 2/2.4; 5/5.6; 10/10.6; 20/20.5; 50/5.9; 100/99.9; 200/199.7; 500/502.5; 1000/999.5

4.8.3. Evaluation of linearity

In this CS, a comparison has been made between regression models with or without intercept and also including the ZPC. The regression statistics have been summarized in Table 5. Calibration data were considered following three possible regression options:

i) The normal OLS regression gives $\%RE_{\mathrm{MAX}}$ values below 15% for all the calibration points.

ii) To study the influence of ZPC, the origin as additional calibration point was included in the experimental calibration data. In this case, the statistical regression output was improved, giving lower average and extreme error values. However, it is suggested do not include fictitious points in the calibration curves to avoid erroneous confidence intervals of regression coefficients that could be affect the accuracy of prediction interval for real samples.

iii) To study OLS-RTO, two different regression curves have been evaluated such as automatically (clicking the software option) or statistically sound. In the statistical case, there are three possibilities to evaluate the intercept significance:

- The experimental *t*-value (1.66) is lower that the critical value (2.26). Correspondingly, the *p*-value (0.14) is higher that the critical value (0.05). That means that intercept should be removed.
- The intercept interval value (-0.2/1.3) includes zero. That means that the RTO is a reliable option justified from this statistical point of view.
- The standard error of the intercept (SE_{b0} = 0.3255) is lower than the intercept value (b_0=0.5397). That means that intercept should be maintained. Therefore, this way to evaluate the intercept significance looks less reliable.

The decision to force RTO can be considered as an inappropriate analytical decision taking into account the values of *%RE* parameter in both options (automatic and statistical). The calibration curves without intercept resulted in higher *Σ%RE* and *%RE$_{AVER}$* values that the calibration curve with intercept. In the same way, extreme *%RE$_{MAX}$* values around 40% were obtained forcing RTO. Therefore, bias was considerably higher than 15-20% suggested as limit for the calibration points.

4.9. RE as function of regression fitting technique (OLS/WLS)

Case study 9. Determination of unspecified drug by HPLC [96]

4.9.1. Experimental data

N=14; *j*=7; *k*=2

4.9.2. Calibration data: X (concentration-ng/mL) versus Y (signal-arbitrary units)

5/0.0632; 5/0.0725; 10/0.1126; 10/0.1344; 50/0.6078; 50/0.5830; 100/1.0714; 100/1.1227; 500/5.1290; 500/5.4232; 1000/10.3892; 1000/10.5105; 5000/46.7262; 5000/51.1182

4.9.3. Evaluation of linearity

In the present CS, a comparison has been made between OLS and WLS regression fitting techniques. In addition, the influence of different WF in the GOF of regression triplet has been analyzed. The preliminary analysis of regression statistics showed that:

- R^2 values were very similar and always higher than 0.99. Anyway, this parameter has no significance for testing linearity.
- The slope values (b_1) were marginally affected by weighting, ranging from 0.00977 to 0.01055.
- The intercept values (b_0) were different among the diverse regression models. As usual, for WLS the intercept value was lower, specifically 8-35% that values for OLS regression, looking that intercept is relatively poorly estimated for unweighted fitting technique. In addition, although with less difference against OLS, b_0 values ranged between 0.01636 and 0.07074 when different WF were used. Similarly, the standard error (*SE$_{b0}$*) dropped considerably, by this way decreasing %RSD from 125% to 21%. Therefore, the influence of the intercept value can be considered as a key factor for the GOF of the different regression models.

RSE and *%RE* parameters can be useful to evaluate the GOF of different linear regression models. In theory, calibration regressions should be selected taking into account if adequately fit to the experimental data, and then providing accurate (lowest RSE and *%RE*) calibration curves. The results of CS 9 (Table 5) showed that $1/x^2$ and $1/y^2$ are the best WF considering the lowest values of RSE, Σ*%RE*, $\%RE_{AVE}$ and $\%RE_{MAX}$. Relating to $\%RE_{\mathrm{MAX}}$, the value drops from no weighting (393%) to $1/y^{0.5}$ (115%), $1/x^{0.5}$ (108%), $1/y$ (36%), $1/x$ (33%), $1/x^2$ (13%) and $1/y^2$ (12%) when WF were used. Then, bias was considerably greater that acceptable limits (20%) at the low concentration section of the calibration range, except in the case of using $1/x^2$ and $1/y^2$ as WF.

5. Calibration tutorial: full regression analysis and linearity evaluation

This evaluation involves 4 different steps, all of which are easy to perform using statistical software. The basic steps are displayed in Figure 2 and explained following.

- Step 1. Plot of calibration data: instrument response versus concentration

As can be expected for calibration curves, the first step is to plot the *y* (instrument response) versus the *x* values (normally concentration). Suspect points than can be considered as outliers or leverage data should be investigated to see if they are correct and they belong to data set. If there is some degree of curvature, this situation may also be detectable.

- Step 2. Determination of instrumental response behaviour

The importance of this step must be emphasized considering that one assumption behind OLS is that the SD of the instrument response does not change (homoscedasticity) over the full range of *x* values for which the model will be applied. A first evaluation about constant SD can be carried out looking at the spread of data from the calibration plot. If it is appear that the spread increases, then the assumption of homoscedasticity will be probably not correct. Further evaluation of instrumental response behaviour can be carried out statistically to check homo/heteroscedasticity but is out of the scope of this manuscript additional explanation. Ignoring heteroscedasticity situation have negative impact in the regression, firstly in the estimation of coefficients and secondly in the prediction intervals for real samples.

- Step 3. Postulation of the regression triplet: method, model and fitting technique

A common starting point for analytical calibration is least-squared method, linear model and OLS as the fitting technique. However, the WLS fitting technique must be used in analytical calibration so frequently to avoid an erroneous straight-line model. In this case, an important issue is the selection of WF.

Once the appropriate regression triplet has been selected, it can be used to fit the calibration data and regression analysis to obtain the usual statistical parameters (b_1, b_0, RE_{b1}, RE_{b0}, RSE, R^2)

- Step 4. Evaluation of linearity

As it was previously stated the linearity of calibration data is very important only to strictly linear regression models. This evaluation can be carried out in different ways that can be grouped into two general modes such as graphical and non-graphical. The graphical assessment includes residuals and linearity plots. In addition, the non-graphical evaluation could be carried out by different statistical significance tests and also numerical parameters that were previously explained from the theoretical viewpoint.

Following a new experimental case study is presented as an example to carry out an overall calibration diagnostic.

Case study 10. Determination of volatile fatty acids (VFAs) by GC [106]

5.10.1. Experimental data

Matrix: aqueous samples; N=21; j=7; k=3

5.10.2. Calibration data: X (concentration-ng/mL) versus Y (PAR of n-C5/IS)

10/0.04465; 10/0.04322; 10/0.04435; 50/0.23650; 50/0.23678; 50/0.23592; 100/0.48307; 100/0.48556; 100/0.48876; 250/1.28845; 250/1.29006; 250/1.26648; 5000/2.60038; 500/2.52749; 500/2.58343; 750/3.70000; 750/3.83859; 750/3.80000; 1000/4.90000; 1000/5.05255; 1000/5.10000

5.10.3. Evaluation of calibration curve

- Step 1. Plot of calibration data: instrument response versus concentration

Figure 3a shows the scatter-plot of 21 PAR versus their respective calibration standard concentrations. All points appear to be well-behaved and there is no evidence of curvature.

- Step 2. Determination of instrumental response behaviour

For this example the SD (and variance) of instrument response values clearly increase with the concentration of calibration standards. This concentration dependent trend confirms the heteroscedasticity of *y* measurements.

- Step 3. Postulation of the regression triplet: method, model and fitting technique

Visualizing the calibration plot is perceptible that least-squares and linear model should be an appropriate selection. About the fitting technique, the practical effect of heteroscedasticity is that WLS should be chosen instead of OLS. As it was previously suggested, it looks that $1/x^2$ can be considered as the best option for weighting.

- Step 4. Evaluation of linearity

Table 6 summarizes the results obtained after apply the different linearity tests. Following further explanation is provided.

A) Graphically

A1. Residuals plot

When the proposed regression triplet is fit to the data, the corresponding residuals plot should be generated as well. Residual plot can be considered as one of the most useful tools in the calibration-diagnosis process but some previous experience in judgment is necessary. For this example, the residuals plot (Fig 3b) does not show an ideal random patter, but a clear non-random patter (parabola or sinusoidal) is neither showed. On the other hand, the graph exhibits a trumpet shape what confirms that WLS should be applied as fitting technique.

A2. Linearity plot

The method can be considered as linear in the full calibration range if none of the sensitivity points (relative response factors) intersects or are beyond the ±5% tolerance limits. For this example, Figure 3c shows that there is a deviation from linearity because the sensitivity values (0.00432; 0.00443; 0.00446) corresponding to the replicate calibration standards of 10 mg/L are of all them below the low tolerance limit (median value $_x$ 0.95= 0.00469).

B) Statistical significance tests

The ANOVA and LOF tests showed that deviation from linearity is present. However, Mandel´s and SQT tests demonstrated that better fit to quadratic model was not the reason to the lack of linearity. By this way, confirming that there are reasons for LOF different to curvature presence.

C) Numerical parameters

- R^2 was near 1, but this parameter is not proof of GOF to linear model.
- %RSD_{slope} was lower that the suggested limit of 2%, giving evidence of correct linearity.
- QC (mean) was lower than the 5% limit, indicating good linearity. However, this test has poor significance in this case because it was designed for homoscedastic calibration curves.
- $\%RE_{MAX}$ was lower than 15-20% considered as acceptance limit to bias deviation from calibration curves.

D) Final decision about linearity

This example shows what frequently happens for real experimental data, there are contradictory results when the linearity of a calibration curve is evaluated using different viewpoints [107]. In this CS, linearity plot and some statistical significance tests showed that calibration data could be treated as a non-linear model. Although it is important to specify that linearity graph was linear except for the lowest calibration point, and some statistical tests such as ANOVA and LOF are not totally reliable due to be precision dependent. On the other hand, some of the numerical parameters used to evaluate the linearity suggested that regression follows a linear model. Therefore, the important question is which decision to take considering the contrary results. To reply appropriately, *%RE* can be considered as a key numerical parameter because in analytical chemistry the selected calibration curve will be used to predict the *x* values of real samples from experimental *y* values. In this CS, the deviations calculated from back-concentrated concentrations were 3.6% in the worst case. This result is very far from 15-20% suggested as acceptance criterion. By this way, it is possible to conclude that calibration data could be considered as linear for the whole calibration range considered (10-1000 mg/L).

6. Conclusions

Based on the theoretical aspects of analytical calibration and the results reported in the selected CS, the following conclusions can be drawn.

- A very good option for analytical calibration should be using triplicate measurements of six calibration standards covering a working range in which the usual content of real samples is expected. The calibrations levels should be evenly distributed or alternatively for wide calibration ranges partial arithmetic series should be suggested. The quantification calibration mode (ESTD/ISTD) should be select according the specific characteristic of each analytical determination. In any case, three sequences or rounds for calibration over at least 2-3 different weeks are suggested to check the independency of standards and the stability of the instrument signal over time.
- The regression triplet including method, model and fitting technique should be selected appropriately to fit the experimental calibration data.
- The common mistakes in analytical calibration should be avoided following the next suggestions. Firstly, although analytical software will probably giving information of r and/ or R^2, both parameters will be never considered as indicators for evaluating the linearity of calibration curves. Secondly, heteroscedasticity must not be neglected in the calibration process. WLS should be considered as the best fitting technique option to regression curves. Relating to WF, they should be calculated as function of $1/x^2$. Thirdly, ZPC and RTO can be considered as two common mistakes for regression providing a significant bias for quantitative results. Therefore, they should be not considered.
- There are different ways to assess the linearity of calibration curves by using graphical plots, statistical significance test and some numerical parameters. The final decision obtained using some of them should be contradictory. Specially, *%RE* should be considered as helpful to appropriately evaluate the linearity of calibration curves considering its unambiguous conclusion when compared to well establish acceptance limits.
- A straight-line model is commonly preferred for calibration curves but deviation from linearity is a very common situation in several analytical techniques. This situation should be resolved applying polynomial models as the best response function for the accurate calibration curves.

Acknowledgements

The author would like to thanks to Dr. Dámaso Hornero and Dr. María Roca for the critical reviews and helpful discussions about the manuscript.

References

[1] AOAC peer-verified methods, program manual on policies and procedures, 1993.

[2] R. W. M. Thompson, S.L. Ellison, “IUPAC Technical Report. Harmonized Guidelines for Single-Laboratory Validation of Methods of Analysis”, *Pure Appl. Chem* 74 (2002) 835–855.

[3] Eurachem Guide. The Fitness for Purpose of Analytical Methods. A Laboratory Guide to Method Validation and Related Topics. First edition, 1998.

[4] Eurachem Guide. The Fitness for Purpose of Analytical Methods. A laboratory Guide to Method Validation and related topics. Second edition, 2014.

[5] European Medicines Agency (EMA). Committee for Medicinal Product for Human Use (CHMP). Guideline on bioanalytical method validation Guideline on bioanalytical method validation Table of contents,” 2012.

[6] US FDA. United States Department of Health and Human Services Food and Drug Administration. Guidance for Industry. Bioanalytical Method Validation,” no. May, 2001.

[7] Unites States Pharmacopeia (USP) XXV. General Information /〈1225〉 Validation of compendial procedures,” 2003.

[8] Agência Nacional de Vigilância Sanitária (ANVISA). Guide for validation of analytical and bioanalytical methods. Resolution-RE n. 899, 2003.

[9] Irish National Accreditation Board (INAB). Guide to Method Validation for Quantitative Analysis in Chemical Testing Laboratories PS 15,” 2012.

[10] National Association of Testing Authorities, Australia (NATA). Technical Note 17, Guidelines for the validation and verification of quantitative and qualitative test methods, 2013.

[11] L. Huber, “Validation of Analytical Methods,” Agilent technologies, 2010

[12] C. S. S. Bratinova, R. Barbara, “Guidelines for performance criteria and validation procedures of analytical methods used in controls of food contact materials”, Publication office of the European Union, Luxembourg, 2009.

[13] Scientific Working Group for Forensic Toxicology (SWGTOX). Standard practices for method validation in forensic toxicology, J. Anal. Toxicol. 37 (2013) 452-474.

[14] S. Chandran and R. S. P. Singh, “Comparison of various international guidelines for analytical method validation,” *Pharmazie* 62 (2007) 4–14.

[15] D. Stöckl, H. D’Hondt, and L. M. Thienpont, “Method validation across the disciplines-Critical investigation of major validation criteria and associated experimental protocols,” *J. Chromatogr. B 877* (2009) 2180–2190.

[16] International conference on harmonisation (ICH) of technical requirements for registration of pharmaceuticals for human use. Harmonised tripartite guideline. Validation of analytical procedures: text and methodology Q2(R1) 2005”, 1994.

[17] International vocabulary of metrology (VIM)-. Basic and general concepts and associated terms, 3rd edition. Joint Committee for Guides in Metrology (JCGM) 200:2008..

[18] K. Danzer and L.A. Currie, “Guidelines for calibration in analytical chemistry. Part 1: Fundamentals and single component calibration,” *Pure Appl. Chem.* 70 (1998) 993–1014.

[19] L. Cuadros-Rodríguez, L. Gámiz-Gracia, E. Almansa-López, and J. Laso-Sánchez, “Calibration in chemical measurement processes: I. A metrological approach,” *TrAC - Trends Anal. Chem.* 20 (2001) 195–206.

[20] C. Burgess, “Is a sample size of n=6 a magic number,” *Pharmaceutical Technology* 38 (6) 2014

[21] L. Cuadros-Rodríguez, L. Gámiz-Gracia, E.M. Almansa-López, and J.M. Bosque-Sendra, “Calibration in chemical measurement processes: II. A methodological approach,” *TrAC - Trends Anal. Chem.* 20 (2001) 620–636.

[22] J. W. Dolan, “When should an internal standard be used?,” *LC-GC Eur.* 25 (2012) 316–322.

[23] S. Burke, “Regression and calibration”, *LC-GC Eur.*, Online Supplement, Statistics and data analysis, (2001) 13–18.

[24] W.G. Warren, “Correlation or regression: Bias or precision”, *Appl. Stat.* 20 (1971) 148–164.

[25] D. Montgomery, E.A. Peck and G.G. Vining. “Introduction to linear regression analysis”, Fourth edition, John Wiley & Sons, 2006.

[26] J.M. Andrade-Garda, A. Carlosena-Zubieta, R.M. Soto-Ferreiro, J. Teran-Baamonde, Michael Thompson, “Classical linear regression by least squares method”, in: J.M Andrade-Garda (Ed.), Basic chemometric techniques in atomic spectroscopy, 2nd edition, The Royal Society of Chemistry, 2013, pp. 52-122

[27] D. Theodorou, Y. Zannikou, and F. Zannikos, “Estimation of the standard uncertainty of a calibration curve: Application to sulfur mass concentration determination in fuels”, *Accredit. Qual. Assur.* 17 (2012) 275–281.

[28] R. Caulcutt, R. Boddy, Statistics for analytical chemists, 2nd edition, Chapman and Hall Ltd, London, 1983.

[29] L. E. Vanatta and D. E. Coleman, “Calibration, uncertainty, and recovery in the chromatographic sciences”, *J. Chromatogr. A* 1158 (2007) 47–60.

[30] A. Sayago and A. G. Asuero, “Fitting straight lines with replicated observations by linear regression: Part II. Testing for homogeneity of variances”, *Crit. Rev. Anal. Chem.* 34 (2004) 133–146.

[31] S. V. C. De Souza and R. G. Junqueira, “A procedure to assess linearity by ordinary least squares method”, *Anal. Chim. Acta* 552 (2005) 23–35.

[32] J. S. Garden, “Nonconstant variance regression techniques for calibration-curve-based analysis”, *Anal. Chem.* 52 (1980) 2310–2315.

[33] K. Baumann and H. Wätzig, “Appropriate calibration functions for capillary electrophoresis II. Heteroscedasticity and its consequences”, *J. Chromatogr. A* 700 (1995) 9–20.

[34] M. Mulholland and D. B. Hibbert, “Linearity and the limitations of least squares calibration,” *J. Chromatogr. A* 762 (1997) 73–82.

[35] J. N. Miller and J.C. Miller, Statistic and Chemometrics for Analytical Chemistry, 5th Edition 2005

[36] R. S. Nascimento, R. E. S. Froes, N. O. C. e Silva, R. L. P. Naveira, D. B. C. Mendes, W. B. Neto, and J. B. B. Silva, “Comparison between ordinary least squares regression and weighted least squares regression in the calibration of metals present in human milk determined by ICP-OES,” *Talanta* 80 (2010) 1102–1109.

[37] C. Mansilha, a. Melo, H. Rebelo, I. M. P. L. V. O. Ferreira, O. Pinho, V. Domingues, C. Pinho, and P. Gameiro, “Quantification of endocrine disruptors and pesticides in water by gas chromatography-tandem mass spectrometry. Method validation using weighted linear regression schemes,” *J. Chromatogr. A* 1217 (2010) 6681–6691.

[38] C. P. Da Silva, E. S. Emídio, and M. R. R. De Marchi, “Method validation using weighted linear regression models for quantification of UV filters in water samples,” *Talanta* 131 (2015) 221–227.

[39] A. M. Almeida, M. M. Castel-Branco, and A. C. Falcão, “Linear regression for calibration lines revisited: Weighting schemes for bioanalytical methods,” *J. Chromatogr. B* 774 (2002) 215–222.

[40] R. B. Jain, “Comparison of three weighting schemes in weighted regression analysis for use in a chemistry laboratory,” *Clin. Chim. Acta* 411(2010) 270–279.

[41] H. Gu, G. Liu, J. Wang, A.-F. Aubry, and M. E. Arnold, “Selecting the correct weighting factors for linear and quadratic calibration curves with least-squares regression algorithm in bioanalytical LC-MS/MS assays and impacts of using incorrect weighting factors on curve stability, data quality, and assay performance,” *Anal. Chem.* 86 (2014) 8959–8966.

[42] S.L.R. Ellison, V.J. Barwick and T.J. Duguid, Practical Statisticcs for the analytical scientist, The Royal Society of Chemistry, Cambridge, 2009.

[43] J.J. McShane, “Zero is not a valid data for calibration purposes”, 2012 http://www.thetruthaboutforensicscience.com

[44] G. J. Hahn, “Fitting regression models with no intercept term.,” *J. Qual. Technol.* 9 (1977) 56–61.

[45] J. G. Eisenhauer, “Regression through the origin,” *Teach. Stat.* 25 (2003) 76–80.

[46] J.W. Dolan, “Calibration curves, Part I: To be or not to be”, LG-GC Eur. 22 (2009), 190-194.

[47] E. Rozet, A. Ceccato, C. Hubert, E. Ziemons, R. Oprean, S. Rudaz, B. Boulanger, and P. Hubert, “Analysis of recent pharmaceutical regulatory documents on analytical method validation,” *J. Chromatogr. A* 1158 (2007) 111–125.

[48] N. J. Miller-Ihli, T. C. O’Haver, and J. M. Harnly, “Calibration and curve fitting for extended range AAS,” *Spectrochim. Acta Part B At. Spectrosc.* 39 (1984) 1603–1614.

[49] L. Kirkup and M. Mulholland, “Comparison of linear and non-linear equations for univariate calibration,” *J. Chromatogr. A*, 1029 (2004) 1–11.

[50] J. Burrows and K. Watson, “Linearity of chromatographic systems in drug analysis part III: Examples of nonlinear drug assays,” *Bioanalysis* 7 (2015) 1763–1774.

[51] H. T. Karnes and C. March, “Calibration and validation of linearity in chromatographic biopharmaceutical analysis,” *J. Pharm. Biomed. Anal.* 9 (1991) 911–918.

[52] D. Tholen, Evaluation of the linearity of quantitative measurement procedures: A statistical approach; Approved Guideline, 2003.

[53] P. Araujo, “Key aspects of analytical method validation and linearity evaluation”, J. Chromatogr. B 877 (2009) 2224–2234.

[54] M. Sanagi, Z. Nasir, S. L. Ling, D. Hermawan, W. A. Wan Ibrahim, and A. A. Naim, “A practical approach for linearity assessment of calibration curves under the international union of pure and applied chemistry (IUPAC) guidelines for an in-house validation of method of analysis”, *J. AOAC Int.* 93 (2010) 1322–1330.

[55] C. A. Dorschel, J. L. Ekmanis, J. E. Oberholtzer, F. V. Warren Jr., and B. A. Bidlingmeyer, “LC detectors: Evaluation and practical implications of linearity”, *Anal. Chem*. 61 (1989) 951A–968A.

[56] L. Huber, “Validation of analytical methods: Review and strategy,” *LC-GC Eur.*11 (1998) 96–105.

[57] J. Cristale, D. M. Dos Santos, B. S. Sant’Anna, D. C. Sandron, S. Cardoso, A. Turra, and M. R. R. De Marchi, “Tributyltin in crustacean tissues: Analytical performance and validation of method,” *J. Braz. Chem. Soc.* 23 (2012) 39–45.

[58] L.S. Ettre, “Nomenclature for Chromatography“, *Pure & Appl. Chem*. 65 (1993) 819-872.

[59] Analytical Methods Committee, “Is my calibration linear?”, Analyst 119 (1994) 2363–2366.

[60] ISO 11095. Linear calibration using reference materials, 1996.

[61] D. E. Coleman and L. E. Vanatta, “Lack-of-fit testing of ion chromatographic calibration curves with inexact replicates,” *J. Chromatogr. A* 850 (1999) 43–51.

[62] G. Tetrault, “Evaluation of assay linearity (I),” *Clin. Chem.* 36 (1990) 585–586.

[63] M. H. Kroll and K. Emancipator, “A theoretical evaluation of linearity”, *Clin. Chem.* 39 (1993) 405–413.

[64] J. Mandel. “The statistical analysis of experimental data”, Dover Publications, New York, 1964.

[65] J.M Andrade, M.P. Gómez-Carracedo, “Notes on the use of Mandel´s test to check for nonlinearity in laboratory calibrations”, *Anal. Methods* 5 (2013) 1145-1149.

[66] D.L. Massart, B.G.M. Vandegiste, L.M.C. Buydens, S. de Jong, P.J.Lewi, J.Smeyers-Verbeke, “Handbook of chemometrics and qualimetrics”, Elsevier, Amsterdam, 1997.

[67] W. H. Davis Jr. and W. A. Pryor, “Measures of goodness of fit in linear free energy relationships,” *J. Chem. Educ.* 53 (1976) 285–287.

[68] J.S. Hunter, “Calibration and the straight line: current statistical practices”, *J. Assoc. Anal. Chem*. 64 (1981) 574-583.

[69] M. D. Van Arendonk, R. K. Skogerboe, and C. L. Grant, “Correlation coefficients for evaluation of analytical calibration curves,” *Analytical Chemistry* 53 (19781) 2349–2350.

[70] D. G. Mitchell and J. S. Garden, “Measuring and maximizing precision in analyses based on use of calibration graphs,” *Talanta* 29 (1982) 921–929.

[71] Analytical Methods Committee, “Uses (Proper and improper) of correlation coefficients,” *Analyst* 113 (1988) 1469–1471.

[72] H. Sahai, R.P. Singh, “The use of R^2 as a measure of goodness of fit: an overview”, *Va J Sci* 40 (1989) 5-9.

[73] M. Thompson, “Statistics. Abuse of statistics software packages,” *Anal. Proc.* 27 (1990) 142–144.

[74] J.N. Miller, “Is it a straight line?,” *Spectrosc. Int.* 3 (1991) 41–43.

[75] J. N. Miller, “Basic statistical methods for analytical chemistry. Part 2. Calibration and regression methods. A review”, *Analyst* 116 (1991) 3–14.

[76] R. Cassidy; M. Janoski, “Is your calibration curve linear?,” *LC-GC* 10 (1992) 692–695.

[77] D.L. MacTaggart, S.O. Farwell, Analytical use of linear regression. Part I: regression procedures for calibration and quantitation, *J. of AOAC Int* 75 (1992) 594-608.

[78] J. Van Loco, M. Elskens, C. Croux, and H. Beernaert, “Linearity of calibration curves: Use and misuse of the correlation coefficient,” *Accredit. Qual. Assur.* 7 (2002) 281–285.

[79] R. De Levie, “Two linear correlation coefficients,” *J. Chem. Educ.* 80 (2003) 1030–1032.

[80] W. Huber, “On the use of the correlation coefficient r for testing the linearity of calibration functions,” *Accredit. Qual. Assur.* 9 (2004) 726.

[81] M. M. Kiser and J. W. Dolan, “Selecting the best curve fit,” *LC-GC North Am.* 22 (2004) 112–117.

[82] J. Ermer and H. J. Ploss, “Validation in pharmaceutical analysis: Part II: Central importance of precision to establish acceptance criteria and for verifying and improving the quality of analytical data,” *J. Pharm. Biomed. Anal.* 37 (2005) 859–870.

[83] D. B. Hibbert, “Further comments on the (miss-)use of r for testing the linearity of calibration functions,” *Accredit. Qual. Assur.* 10 (2005) 300–301.

[84] A. G. Asuero, A. Sayago, and A. G. González, “The correlation coefficient: An overview,” *Crit. Rev. Anal. Chem.* 36 (2007) 41–59.

[85] J. W. Lee, V. Devanarayan, Y. C. Barrett, R. Weiner, J. Allinson, S. Fountain, S. Keller, I. Weinryb, M. Green, L. Duan, J. a. Rogers, R. Millham, P. J. O’Brien, J. Sailstad, M. Khan, C. Ray, and J. a. Wagner, “Fit-for-purpose method development and validation for successful biomarker measurement,” *Pharm. Res.* 23 (2006) 312–328.

[86] J.M. Sonnergaard, “On the misinterpretation of the correlation coefficient in pharmaceutical sciences,” *Int. J. Pharm.* 321 (2006) 12–17.

[87] T. Singtoroj, J. Tarning, a. Annerberg, M. Ashton, Y. Bergqvist, N. J. White, N. Lindegardh, and N. P. J. Day, “A new approach to evaluate regression models during validation of bioanalytical assays,” *J. Pharm. Biomed. Anal.* 41 (2006) 219–227.

[88] “AMC Technical Brief 3,” *Committee, Anal. Methods* 3(2006) 1–2.

[89] L. Komsta, “Chemometric and statistical evaluation of calibration curves in pharmaceutical analysis - a short review on trends and recommendations,” *J. AOAC Int.* 95 (2012) 669–672.

[90] E. Rozet, E. Ziemons, R. D. Marini, and P. Hubert, “Usefulness of information criteria for the selection of calibration curves”, *Anal. Chem.* 85 (2013) 6327–6335.

[91] D. Stöckl, K. Dewitte, and L. M. Thienpont, “Validity of linear regression in method comparison studies: Is it limited by the statistical model or the quality of the analytical input data?,” *Clin. Chem.* 44 (1998) 2340–2346.

[92] L. Cuadros-Rodríguez, A. M. García Campaña, C. Jimenez Linares, and M. Roman Ceba, “Estimation of performance characteristics of an analytical method using the data set of the calibration experiment,” *Anal. Lett.* 26 (1993) 1243–1258.

[93] P. Vankeerberghen and J. Smeyers-Verbeke, “The quality coefficient as a tool in decisions about the quality of calibration in graphite furnace atomic absorption spectrometry”, *Chemom. Intell. Lab. Syst.* 15 (1992) 195–202.

[94] Y. Hu, J. Smeyers-Verbeke, and D. L. Massart, “Exploratory study on median-based robust regression methods for linear calibration in atomic absorption spectrometric analysis,” *J. Anal. At. Spectrom.* 4 (1989) 605–611.

[95] E. L. Johnson, D. L. Reynolds, D. S. Wright, and L. A. Pachla, “Biological sample preparation and data reduction concepts in pharmaceutical analysis,” *J. Chromatogr. Sci.* 26 (1988) 372–379.

[96] J. W. Dolan, “Calibration curves, part 5: Curve weighting,” *LC-GC Eur.* 22 (2009).

[97] J. W. Dolan, “Calibration curves, part 3: A different view”, LC-GC Eur. 22 (2009) 304-308.

[98] M. H. Kroll, J. Praestgaard, E. Michaliszyn, and P. E. Styer, “Evaluation of the extent of nonlinearity in reportable range studies,” *Arch. Pathol. Lab. Med.* 124 (2000) 1331–1338.

[99] E. Hsieh and J.-P. Liu, “On statistical evaluation of the linearity in assay validation,” *J. Biopharm. Stat.* 18 (2008) 677–690.

[100] E. Hsieh, C.-F. Hsiao, and J.-P. Liu, “Statistical methods for evaluating the linearity in assay validation”, *J. Chemom.* 23 (2009) 56–63.

[101] J.-P. Liu, S.-C. Chow, and T.-C. Hsieh, “Deviations from linearity in statistical evaluation of linearity in assay validation,” *J. Chemom.* 23 (2009) 487–494.

[102] S. J. Novick and H. Yang, “Directly testing the linearity assumption for assay validation,” *J. Chemom.* 27 (2013) 117–125.

[103] D. LeBlond, C. Y. Tan, H. Yang, and H. Pappa, “Confirmation of analytical method calibration linearity,” *Pharmacopeial Forum*, 39 (2013).

[104] H. Yang, S. J. Novick, and D. Leblond, “Testing assay linearity over a pre-specified range”, *J. Biopharm. Stat.* 25 (2015) 339–350.

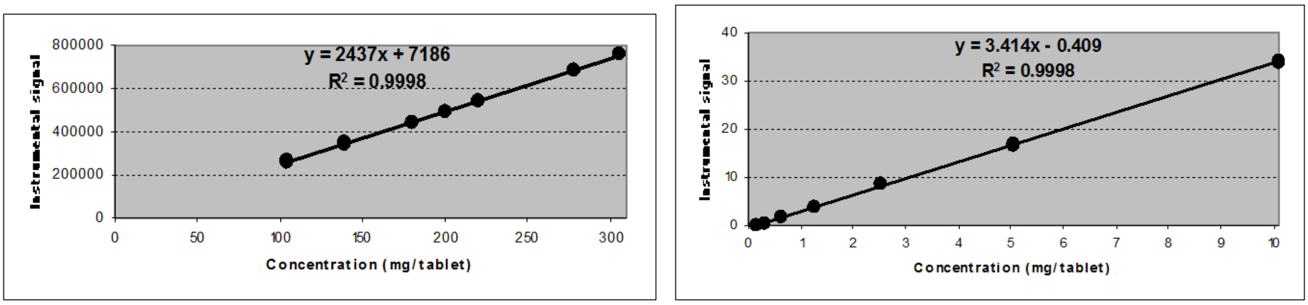
[105] A. C. Olivieri, “Practical guidelines for reporting results in single- and multi-component analytical calibration: A tutorial,” *Anal. Chim. Acta* 868 (2015) 10–22.

[106] F. Raposo, R. Borja, J. A. Cacho, J. Mumme, T. F. Mohedano, A. Battimelli, D. Bolzonella, A. D. Schuit, J. Noguerol-Arias, J.-C. Frigon, G. A. Peñuela, J. Muehlenberg, and C. Sambusiti, “Harmonization of the quantitative determination of volatile fatty acids profile in aqueous matrix samples by direct injection using gas chromatography and high-performance liquid chromatography techniques: Multi-laboratory validation study,” *J. Chromatogr. A* 1413 (2015) 94–106.

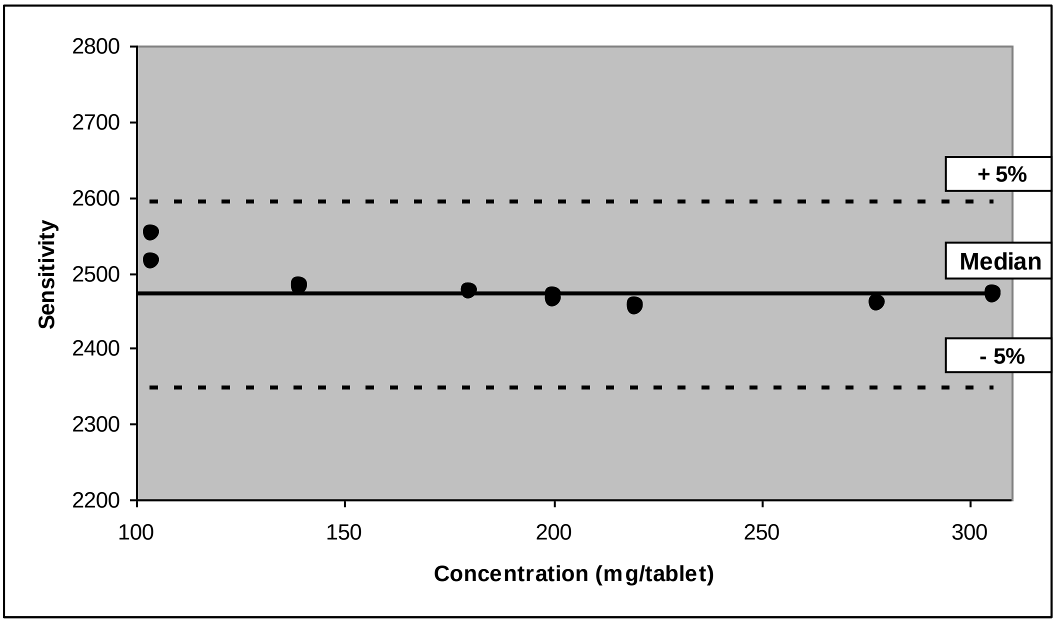
[107] J.M. Mermet, “Calibration in atomic spectrometry: A tutorial review dealing with quality criteria, weighting procedures and possible curvatures”, *Spect. Acta* 65 (2010) 509-523.

Figure 1. Calibration versus linearity plots for CS 1

1a &1b. Calibration plots for Ibuprofen and Biochanin



1c. Linearity plot for Ibuprofen



1d. Linearity plot for Biochanin

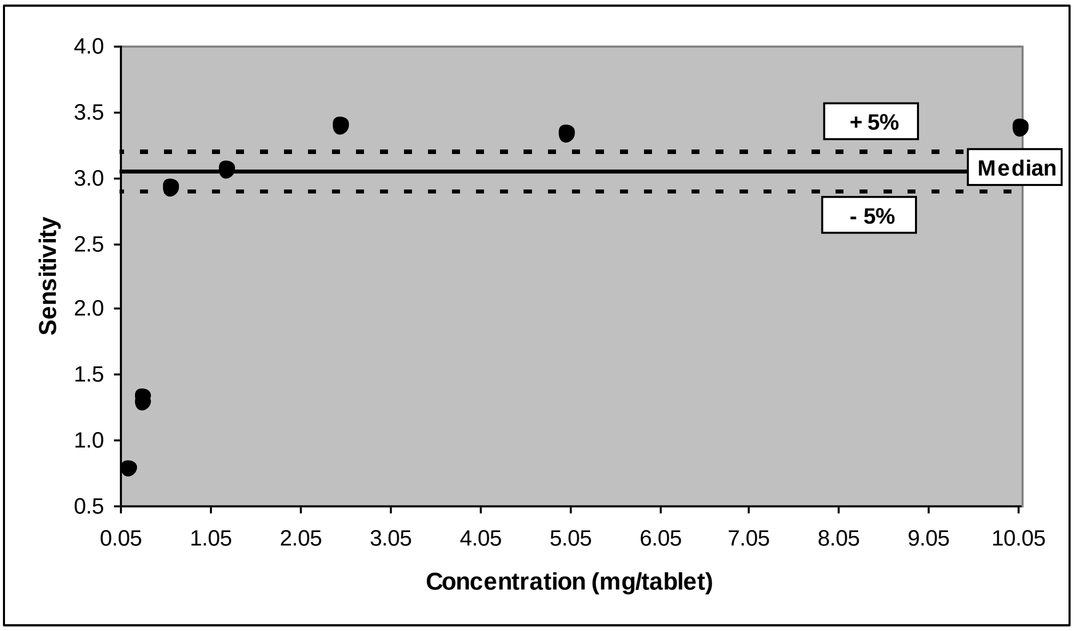


Figure 2. Different steps for analytical calibration

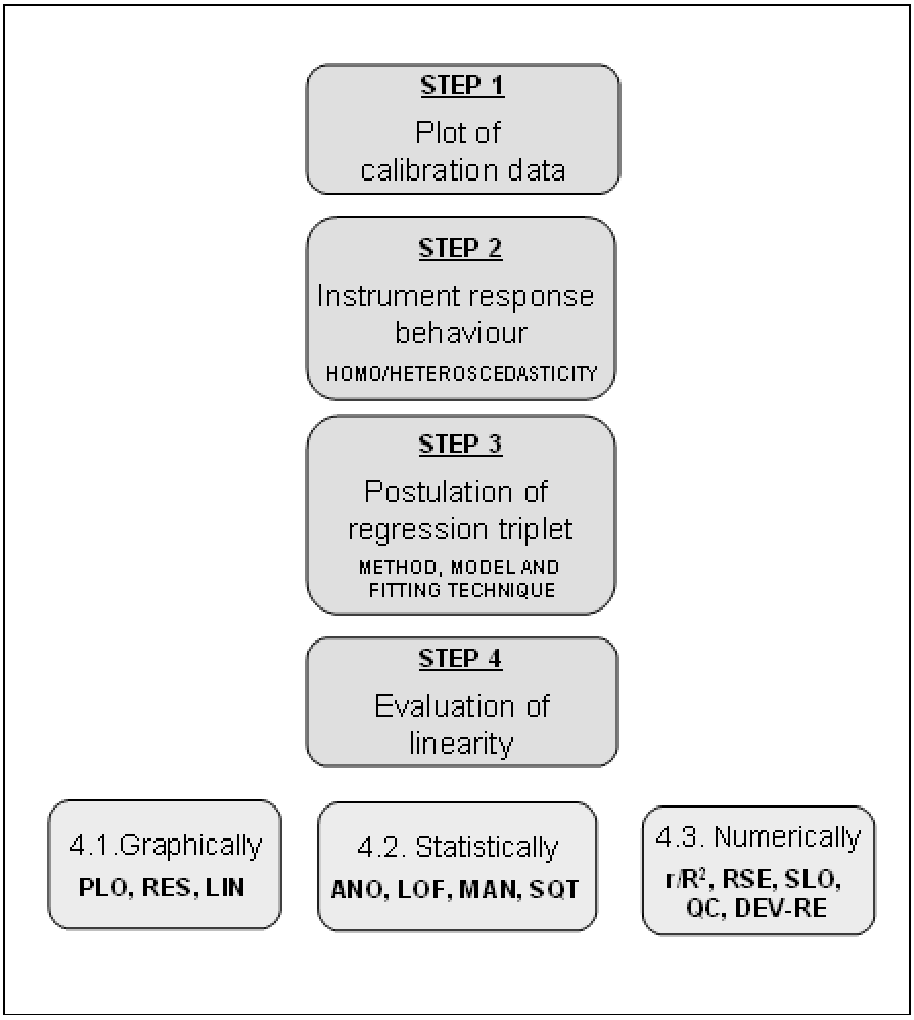


Figure 3a. Calibration plot for CS 10

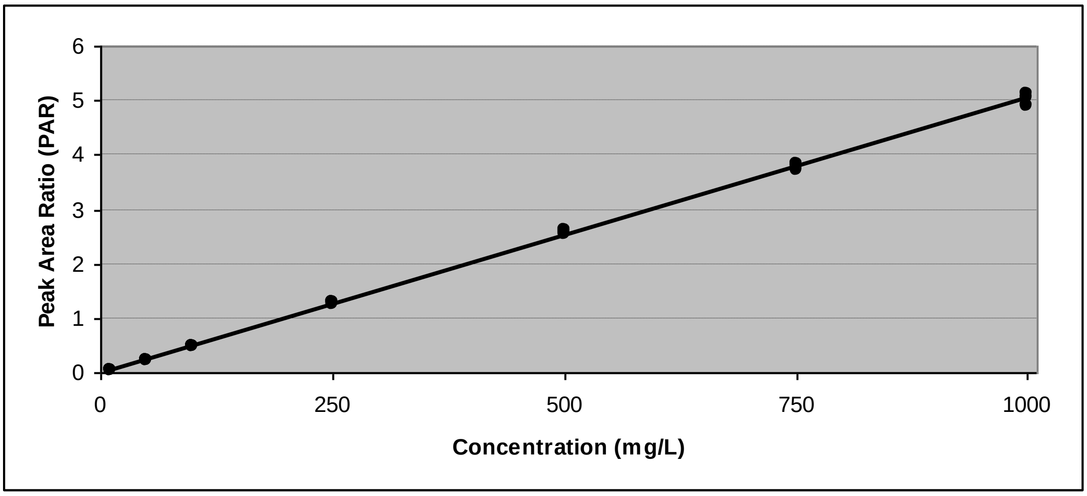


Figure 3b. Residuals plot for CS 10

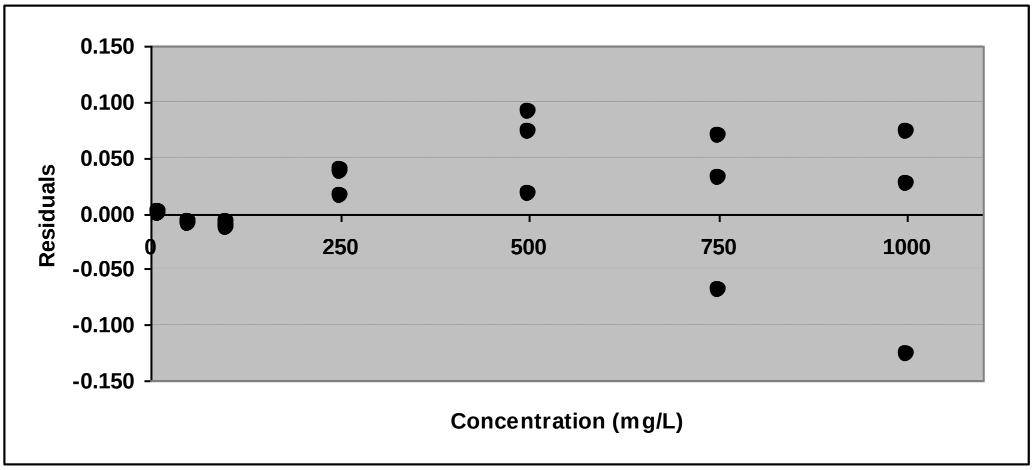


Figure 3c. Linearity plot for CS 10

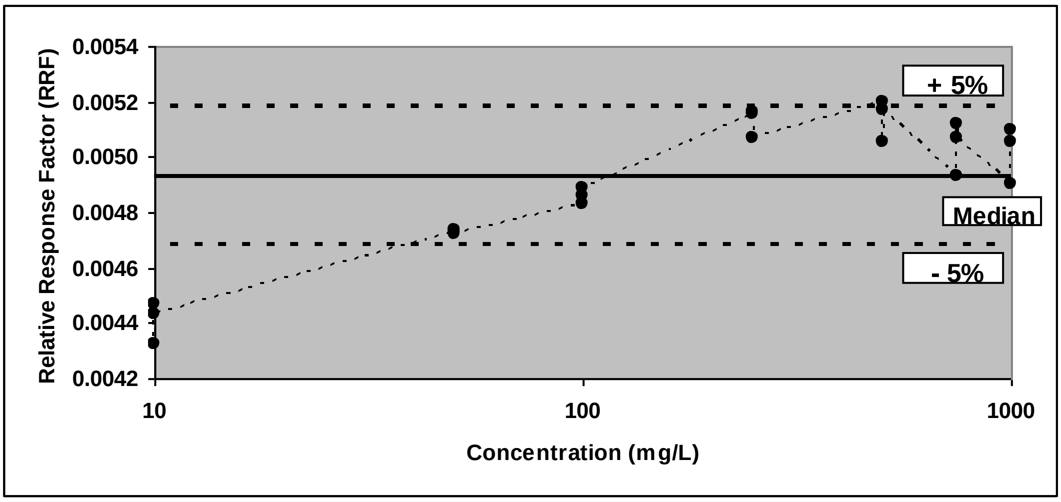


Table 1. Summary of validation guidelines: Calibration process

NAME	AOAC (1993) [1]	ICH (1994) [16]	EUR-1 (1998) [3]	FDA (2001) [6]	IUPAC (2002) [2]	ANVI (2003) [8]	USP (2003) [7]	JRC (2009) [12]	AGILE (2010) [11]	EMA (2011) [5]	INAB (2012) [9]	NATA (2013) [10]	SWGTOX (2013) [13]	EUR-2 (2014) [4]
FITTING EXPERIMENTAL DATA														
REG		X	X	X	X	X	X			X		X	X	X
Non-LIN	X	X	X	X	X	X	X	X	X	X	X	X	X	X
WLS			X	X	X		X	X			X	X	X	X
ZPC/RTO					RTO^T				RTO$^+$				ZPC$^-$	
EVALUATION OF LINEARITY														
A) GRAPHYCALLY														
PLO		X	X			X	X		X		X	X	X	X
RES			X		X			X	X		X	X	X	X
LIN/SEN									X		X			
B) STATISTICALLY														
ANO														
LOF					X			X			X		X	
MAN								X						
SQT													X	
C) NUMERICAL PARAMETERS														
r/R^2		X			X$^-$	X$^+$	X		X		X$^-$	X$^-$	X$^-$	
RSE												X		
SLO								X						
QCs														
DEV		X		X		X		X	X	X$_{BC}$				

Table 2. Advantages and disadvantages of different procedures to check linearity

PROCEDURE	ADVANTAGES	DISADVANTAGES
A) Graphically		
GRA	Helpful to overall view of experimental calibration data	Do not detect lack of linearity in all cases
RES	Very useful for calibration diagnosis	Do not detect lack of linearity for all degrees of curvature. Necessary previous knowledge about regression trends.
LIN	Graphs including acceptance limits are very illustrative for easy understanding	Some terms such as sensitivity, response factors or relative response factors could be confusing Not totally conclusive
B) Statistically		
ANO	Easy to calculate Acceptance limit based in *F* test	Test precision dependent and not conclusive
LOF	Acceptance limit based in *F* test	Calculations are complicate without software program Test precision dependent and not conclusive
MAN	Easy to calculate Acceptance limit based in *F* test	Erroneous versus correct equation for calculation. Explain lack of linearity only in the case of quadratic model
SQT	Acceptance limit based in *F* test	Calculations are complicate without software program Explain lack of linearity only in the case of quadratic model
C) Numerical parameters		
r/R^2	Widely use due to be simple concepts and statistical availability	Misleading, therefore not never useful as linearity test
RSE	Simple concept	Results are instrument dependent
SLO	Easy to calculate	Not conclusive
QC	Acceptance limit based in %	Different equations were provided in the literature Not conclusive for all the calibration range
%RE	Simple calculation Acceptance limit based in % Conclusive from statistical viewpoint	Acceptance criteria of 15-20% should be sometimes considered as excessive high

Table 3. Literature relating to r/R^2 as misleading linearity criterion

Reference	Authors	Date	Coefficients	Statements
[67]	Davis & Pryor	1976	r	"*r*, although widely used as a measure of GOF, does not accurately reflect the deviations of points from the line"
[68]	Hunter	1981	r&R^2	"In fitting functional models values of *r* and R^2 close to ±1 do provide an aura of respectability, but not much else"
[69]	Van Arendonk et al.	1981	r	" A practice that should be discouraged is the use of r as a means of evaluating goodness of fit of linear models"
[70]	Mitchell & Garden	1982	r	"The *r* value does not indicate whether the chosen mathematical model adequately fits the data"
[71]	Analytical Methods Committee	1988	r	"A large value of *r* does not indicate a linear relationship between two measurements" "The r does not indicate linearity or the lack thereof"
[72]	Sahai and Singh	1989	R^2	"A large value of R^2 does not insure a good fit neither the model predict well"
[73]	Thompson	1990	r	"*r* is often mis-applied to calibration data in an attempt to support the presumption of linearity" "$r\approx1$ does not necessarily imply an underlying linear relationship"
[74]	Miller	1991	r	" The magnitude of *r*, considered alone, is a poor guide of linearity"
[51]	Karnes and March	1991	r	"*r* is a poor indicator of how well a linear regression equation fits the linear model" "*r* is of little value in documenting adherence to the linear model"
[75]	Miller	1991	r	"A high value of *r* is thus seen to be no guarantee at all that a straight line rather than a curve, is appropriate for a given calibration plot"
[76]	Cassidy & Janosky	1992	r& R^2	"Values of *r* and R^2 tell us whether there is a reasonable probability that x and y are directly related. They are not intended to measure the degree of linearity of the line of best fit. Consequently, neither *r* nor R^2 should be used the linearity of a calibration curve"
[77]	MacTaggart & Farwell	1992	r	"*r* gives only a relative idea of the linearity inherent in a particular data set"
[59]	Analytical Methods Committee	1994	r	"Hence, *r* is misleading in the context of testing for linearity" "It is better used for correlation, not for quantify linearity"
[34]	Mulholland & Hibbert	1997	r& R^2	"Many analysts depend entirely on the use of R^2 (or *r*) value between 0.999 and 1.000 as an acceptability criterion. This is well known to be inadequate" "*r* does not give any indication of the errors associated with an individual measurement"

Table 3. Literature relating to r/R^2 as misleading linearity criterion (continuation)

Reference	Authors	Date	Coefficients	Statements
[78]	Van Loco et al.	2002	r	“r is not useful indicator of linearity in the calibration model, even for r>0.997” “r is not suitable for assessing the linearity of calibration curves”
[79]	De Levie	2003	r	“r can easily be misinterpreted by chemists as a measure of GOF which it is not”
[80]	Huber	2004	r	“r describes the quality of the fit only poorly and its linearity not at all” “If r is used for testing the quality of the fit with subsequent proof of linearity, it is severely biased”
[81]	Kiser & Dolan	2004	R^2	“Even if the standard curve has R^2>0.9990, the fit will not necessarily be very good” “R^2 is a poor measure of the curve fit quality”
[82]	Emer	2005	r	“r is neither a proof or linearity, nor a suitable quantitative parameter”
[83]	Hibbert	2005	r	“r is not the statistic of choice to determine the extent of linearity”
[31]	De Souza & Junqueira	2006	r &R^2	“the improper recommendation to establish linearity that is most frequently written into protocols and papers is the use of r or R^2”
[84]	Asuero et al.	2006	r	“r close to unity does not necessarily indicate a linear calibration function” “Analyst should avoid being misled by r” “It is surprising that r had been used so frequently to assess the linearity of calibration graphs” “In short, r value is in reality not a measure of model adequacy”
[85]	Lee et al.	2006	R^2	“R^2 is not useful for evaluating the quality of a calibration curve model because it does not penalize model complexity and consequently encourages overfitting”
[86]	Sonnergaard	2006	r	“r is often misused as a universal parameter expressing the quality in linear regression analysis”
[87]	Singtoroj et al.	2006	R^2	“R^2 alone is not adequate to demonstrate linearity since values above 0.999 can be achieved even when the data shows signs of curvature”
[88]	Analytical Methods Committee	2006	r	“Given the importance of linear calibration, it is strange that most analytical chemists are willing to use r as an indicator of linearity” “r in the context of linearity testing is potentially misleading, and should be avoided.
[53]	Araujo	2009	r	“It is extremely important to emphasize that an r-test to check the linearity does not exist. We cannot say that r=0.999 is more linear that r= 0.997”
[89]	Komsta	2012	r &R^2	“r and R^2 are completely unrelated to several phenomena that can occur during calibration. Very high values can be obtained for curves with significant curvilinearity”
[90]	Rozet et al.	2013	R^2	“R^2 do not allow to properly select an adequate response function for the calibration curve”

Table 4. Linear regression output data for CS 1-7

Case Study	SLOPE [b_1]	STD. ERROR [SE_{b1}]	INTERCEPT [b_0]	STD. ERROR [SE_{b0}]	DET. COEF. [R^2]	RSE [$S_{y/x}$]
1A	2437	10	7186	2090	0,9998	2421
1B	3.414	0.015	-0.409	0.064	0,9998	0.183
2	2.286	0.014	-11.4	3.5	0.9996	5.296
3A	0.2416	0.0016	0.016	0.006	0.9993	0.017
3B	0.2416	0.0029	0.016	0.011	0.9977	0.031
4	3.0109	0.0916	1.2727	0.0542	0,9917	0.961
5	28.141	2.202	-68.5	26.050	0.9478	46.18
6	221.16	3.68	45.33	50.39	0.9972	71.26
7	0.0030	0.00005	0.0052	0.0029	0.9989	0.004

Table 5. Linear regression output data for CS 8-10

CASE STUDY	FITTING TECHN.	SLOPE (ST. ERROR)	INTERCEPT (ST .ERROR)	DET. COE.	RSE	RELATIVE ERROR (%RE)		
		[b_1/SE_{b1}]	[b_0/SE_{b0}]	[R^2]	[$S_{y/x}$]	[Σ%*RE*]	%RE_{AVER}	%RE_{MAX}
8[a]	OLS	0.9998 (0.0009)	0.5397 (0.3255)	0.9999	0.8774	25.6	2.6	14.0
8[b]	OLS ZPC	0.9999 (0.0009)	0.4744 (0.2937)	0.9999	0.8442	17.8	1.8	7.4
8[c]	OLS RTO/STA	0.9998 (0.0009)	-	0.9999	0.8774	83.6	8.4	40.0
8[d]	OLS RTO/AUT	1.006 (0.0008)	-	0.9999	0.9588	83.2	8.3	39.9
9[a]	OLS	0,00977 (0.00015)	0,20658 (0.28597)	0.9973	0.9318	1223	58	393
9[b]	WLS ($1/y^{0.5}$)	0.00984 (0.00014)	0.07074 (0.08796)	0.9976	0.3202	358	17	115
9[c]	WLS ($1/x^{0.5}$)	0.00984 (0.00014)	0.06719 (0.08305)	0.9976	0.3029	339	16	108
9[d]	WLS ($1/y$)	0.00994 (0.00015)	0.03160 (0.02102)	0.9974	0.0878	147	7	36
9[e]	WLS ($1/x$)	0.00995 (0.00015)	0.02980 (0.02102)	0.9973	0.0777	140	7	33
9[f]	WLS ($1/y^2$)	0.01055 (0.00025)	0.01636 (0.00347)	0.9934	0.0112	85	4	12
9[g]	WLS ($1/x^2$)	0.01050 (0.00027)	0.01644 (0.00336)	0.9921	0.0106	85	4	13
10	WLS ($1/x^2$)	0.00503 (0.00003)	-0.00665 (0.00080)	0.9992	0.0032	44	2	3.6

Table 6. Case study 10: evaluation of linearity

Linearity Test	**Acceptance Criteria**	**Result**
A. Graphical mode		
Residuals plot	No trend	No trend
Linearity plot	≤ ±5% tolerance limits	Out of tolerance limits
B. Statistical significance tests		
ANOVA	H_0: $F_{\mathrm{exp}} < F_{\mathrm{crit}}$	902926 > 2.4
ANOVA-LOF	H_0: $F_{\mathrm{exp}} < F_{\mathrm{crit}}$	8.65 > 2.96
Mandel	H_0: $F_{\mathrm{exp}} < F_{\mathrm{crit}}$	1.74 < 3.16
SQT	H_0: $F_{\mathrm{exp}} < F_{\mathrm{crit}}$	1.73 < 3.16
C. Numerical parameters		
R^2	None	0.9993
$\%RSD_{\mathrm{slope}}$	< 2%	0.6%
QC (mean)	< 5%	2.6%
$\%RE_{MAX}$	< 15-20%	3.6%