# Procedure for the selection and validation of a calibration model
# I —Description and Application

Brigitte Desharnais[a,b,*], Félix Camirand-Lemyre[c], Pascal Mireault[a,b], Cameron D. Skinner[b]

[a]*Department of Toxicology, Laboratoire de sciences judiciaires et de médecine légale*
*1701 Parthenais Street, Montréal, Québec, Canada H2K 3S7*
[b]*Department of Chemistry & Biochemistry, Concordia University*
*7141 Sherbrooke Street West, Montréal, Québec, Canada H4B 1R6*
[c]*Department of Mathematics, Université de Sherbrooke*
*2500 boulevard de l'Université, Sherbrooke, Québec, Canda J1K 2R1*

**Abstract**

Calibration model selection is required for all quantitative methods in toxicology and more broadly in bioanalysis. This typically involves selecting the equation order (quadratic or linear) and weighting factor correctly modelizing the data. A mis-selection of the calibration model will generate lower quality control (QC) accuracy, with an error up to 154%. Unfortunately, simple tools to perform this selection and tests to validate the resulting model are lacking. We present a stepwise, analyst-independent scheme for selection and validation of calibration models. The success rate of this scheme is on average 40% higher than a traditional "fit and check the QCs accuracy" method of selecting the calibration model. Moreover, the process was completely automated through a script (available in Supplemental Data 3) running in RStudio (free, open-source software). The need for weighting was assessed through an $F$-test using the variances of the upper limit of quantification and lower limit of quantification replicate measurements. When weighting was required, the choice between $1/x$ and $1/x^2$ was determined by calculating which option generated the smallest spread of weighted normalized variances. Finally, model order was selected through a partial $F$-test. The chosen calibration model was validated through Cramer–von Mises or Kolmogorov–Smirnov normality testing of the standardized residuals. Performance of the different tests was assessed using 50 simulated data sets per possible calibration model (e.g., linear-no weight, quadratic-no weight, linear-$1/x$, etc.). This first of two papers describes the tests, procedures and outcomes of the developed procedure using real LC-MS/MS results for the quantification of cocaine and naltrexone.

*Author to whom correspondence should be addressed. Email: brigitte.desharnais@msp.gouv.qc.ca

## 1. Introduction

Every toxicologist performing quantitative method development eventually faces the challenge of choosing a calibration model for the analyte. Most data acquisition and processing software (e.g., Agilent's ChemStation or Mass Hunter, AB Sciex's Analyst®) offer options with regards to forcing the calibration equation through the origin, applying a weight and model order (e.g., quadratic or linear). When only the most common weighting (none, $1/x$, $1/x^2$) and model order (linear, quadratic) options are taken into account, there are six possible calibration models per analyte.

Although in principle, all systems should have a linear response to the concentration and generate linear calibration curves, in reality, some physical and chemical phenomenon can create quadratic calibration curves. Processes such as competition in the LC-MS ionization process or saturation of the detector will create saturation phenomenon at high concentrations, even if this is imperceptible to the naked eye. It is important to properly identify occurrences of quadraticity in the data, because this can have a large impact on quality control (QC) accuracy. Simulations using experimentally obtained calibration curves showed a 12% average improvement in QC accuracy when properly using the quadratic calibration model with uniformly weighted data (Supplemental Data 1). In a similar fashion, Gu et al. [1] demonstrated that there is a notable improvement in QC accuracy when the proper weighting is used for the calibration curve. Identifying the correct calibration model is, therefore, a crucial step in method validation that will have impacts on QC accuracy in production.

The Scientific Working Group in Toxicology (SWGTOX) guidelines state that "ultimately, the best approach is to use the simplest calibration model that best fits the concentration response relationship" [2]. The SWGTOX recommends that the fit be evaluated using a standardized residuals plot. Although this type of graph is a very useful tool to roughly estimate the fit, the visual interpretation of the data renders model selection very analyst-dependent and therefore subjective. The SWGTOX validation guidelines also mention that the correlation coefficient (r) alone cannot be used to evaluate the fit, and that other alternatives can be used, such as analysis of variance - lack of fit (ANOVA LOF), significance of the second order term and the coefficient of determination. However, the calculations for these tests are not detailed in the validation guidelines, and there are no recommendations with regards to the circumstances in which they should be applied. Additionally, as is shown in the second paper of this series, the ANOVA LOF and significance of the second order term techniques have significant issues in terms of performance or ease of use.

In order to address these issues, we have developed a stepwise, systematic method to choose and validate the calibration model for an analyte. This method is not biased by the interpretation of the analyst since conclusions are reached by comparing test results to a cut-off. Furthermore, the testing and interpretation has been automated using a script in RStudio, allowing scientists with limited knowledge or comfort in statistics to perform these tests easily, reliably and quickly. As an example, a method validation for 60 analytes required 1 hour of data treatment time to objectively select the calibration model. Using 2610 calibration data sets spread over different calibration models, curva-

ture levels (magnitude of the $x^2$ term) and %RSD values, this automated scheme was shown to have a success rate in average 40% higher than the traditional method of fitting with more complex models until QC accuracy is acceptable (Supplemental Data 1). This vast improvement in the exactness of calibration model selection will ultimately result in higher QC accuracy in production. The selection method was developed by testing different approaches on data sets both from 50 analytes quantified by LC-MS/MS and simulated data sets with varying numbers of replicates. In this paper, we detail the calculations and interpretation steps that constitute the developed process. As practical examples, two different analytes were chosen to demonstrate this protocol: cocaine and naltrexone. The theoretical basis underlying the choice of each test, as well as the mathematical considerations for different aspects of the scheme (including data collection, outliers and forcing calibration through the origin), is covered in the second paper (II —Theoretical Basis).

## 2. Materials and methods

### *2.1. LC-MS/MS quantification*

Cocaine and naltrexone (Cerilliant, Round Rock, TX, USA) were spiked in bovine blood at concentrations of 5, 10, 15, 50, 75, 100, 400, 500 and 1000 ng/mL to produce calibration standards. Considering that most of the samples analyzed with this method fall in the low concentration range (i.e., therapeutic concentrations), it is appropriate to place more calibration levels at the lower end of the working range. Cocaine-$D_3$ and codeine-$D_3$ (Cerilliant) were used as internal standards (IS, 5 and 100 ng/mL, respectively). Solid phase extraction of the standards was performed using Oasis cartridges (HLB 3cc, product WAT094226, Waters, Mississauga, ON, Canada). A 2 mL volume of blood was extracted and reconstituted in 100 $\mu$L of 15:85 methanol:ammonium formate (10 $mM$). The samples were analyzed on an Agilent 1200 HPLC equipped with an AB Sciex 4000 QTrap mass spectrometer. An aliquot (5 $\mu$L) was injected and separated on an Agilent Zorbax Eclipse C18 column (100 $\times$ 2.1 mm, 3.5 $\mu m$) using a 25 minute step/ramp gradient from 10 $mM$ ammonium formate + 0.2% formic acid to methanol. Quantitative analysis was performed with m/z transition 305.2/183.0 Da for cocaine (the $^{13}$C-containing species was used to reduce the signal and remove saturation at the upper levels of the working range) and m/z transition 342.1/212.0 Da for naltrexone. The peak area ratio of the analyte to the IS was used as the response. This method has been validated according to ISO 17025 and CAN-P-1578 guidelines and is currently used as a routine quantification method. Five injections of each extracted standard were performed in order to create measurements replicates on which to base the statistical analysis. The selection of this experimental setup to obtain replicate measurements is explained in the "Results and discussion" section of the second paper (Raw data necessary and replicate analysis). Chromatographic data analysis was performed with Multiquant$^{\text{TM}}$ (AB Sciex, Framingham, MA, USA).

### *2.2. Simulated data sets*

To validate the calibration model selection accuracy for all six types of possible models (linear-no weight, quadratic-no weight, linear-$1/x$, etc.), simulated data were produced

using a script written and run in RStudio (RStudio, Boston, MA, USA). R (programming environment, `https://www.r-project.org/`) and RStudio (graphical interface, `https://www.rstudio.com/`) are free open-source statistical software tools. The script for simulated data generation is available in Supplemental Data 2.

Using experimental LC-MS/MS calibration data for 50 analytes, intervals spanning the maximum and minimum calibration parameter values for $b_0$, $b_1$ and $b_2$ for quadratic models were established. Synthetic calibration data were generated using calibration parameters chosen at random from within these intervals. For the present study, interval boundaries were $9 \times 10^{-3}$ to $5 \times 10^{-1}$ for $b_0$, $3 \times 10^{-3}$ to $8 \times 10^{-1}$ for $b_1$ and $-7 \times 10^{-5}$ to $-7 \times 10^{-8}$ for $b_2$. Using these parameters, the predicted signal ($y_i$) for each concentration level was calculated.

For every weighting scheme (none, $1/x$, $1/x^2$), each data set was assigned a maximal %RSD value at random between 1% and 20%. This 20% upper boundary was chosen by keeping in mind the SWGTOX guidelines that state precision values should not be higher than 20% [2]. From the randomly assigned %RSD at the lower limit of quantification (LLOQ), the standard deviations for other concentration levels were calculated according to the chosen weighting pattern.

Using the calibration parameters and calculated standard deviations, 50 data sets, each with 5, 7 or 10 normally distributed replicate measurements, were generated at each concentration level for each of the six calibration models tested here.

*2.3. Heteroscedasticity testing*

Only the description of the calculations and/or R functions used to carry out tests will be described in this section. The purpose of the tests and interpretation of the results will be described in Section 3. All calculations required to choose and validate a calibration model were performed using an R script running in RStudio. A compressed file containing all required R scripts as well as instructions for their use, including a video tutorial, is available as Supplemental Data 3.

The presence of heteroscedasticity (a change in variance across concentration levels) was determined by calculating the probability that the variance of measurements at the upper limit of quantification (ULOQ) was equal to or smaller than the variance of measurements at the LLOQ using an $F$-test. This unilateral $F$-test was performed using the following RStudio formula:

$$var.test(MeasurementsLLOQ, MeasurementsULOQ, alternative = \text{“}less\text{”}) \quad (1)$$

with the probability being stored in the $P$ value element of the output list.

*2.4. Variance evaluation for weight selection*

Variance evaluation was performed by first applying each weighting scheme ($W_i$) to the measurements to calculate the concentration levels' normalized weighted variances

($V_W^i$), which were used to calculate the total normalized weighted variance ($V_W$) [3].

$$S = \sum_i \sqrt{W_i} \tag{2}$$

$$V_W^i = \frac{Var\left\{y_{i1}, y_{i2}, \ldots, y_{ij}\right\} \times W_i}{S^2} \tag{3}$$

$$V_W = Var\left\{V_W^1, V_W^2, \ldots, V_W^i\right\} \tag{4}$$

where $S$ is the scaling factor, $W_i$ was the weighting applied at the $i^{th}$ concentration level (e.g., for a $1/x$ weighting at the 5 ng/mL concentration level $W_5$ will be $1/5 = 0.2$), $V_W^i$ is the weighted and normalized variance at the $i^{th}$ concentration level for weighting scheme $W$ and concentration level $i$, $Var$ is the variance operator, which calculates the variance of the elements inside the braces, $y_{ij}$ is the measurement at the $i^{th}$ concentration level and the $j^{th}$ replicate and $V_W$ is the total normalized weighted variance for weight $W$. Three values of $V_W$ should be obtained using this calculation, one for each possible weight (uniform or "no weight" ($W_i = 1$), $1/x_i$ and $1/x_i^2$).

*2.5. Partial F-test for model order selection*

To perform the partial $F$-test, the sum of squares for the linear ($y_L = b_1 \cdot x + b_0$) and quadratic ($y_Q = b_2 \cdot x^2 + b_1 \cdot x + b_0$) models were calculated by [4]

$$SS_{reg,Q} = \sum_{n=1}^{i} W_i \times n_j \times \left(\hat{y}_i - \bar{y}\right)^2 = \sum_{n=1}^{i} W_i \times n_j \times \left(\left\{b_2 \cdot {x_i}^2 + b_1 \cdot x_i + b_0\right\} - \bar{y}\right)^2 \tag{5}$$

$$SS_{reg,L} = \sum_{n=1}^{i} W_i \times n_j \times \left(\hat{y}_i - \bar{y}\right)^2 = \sum_{n=1}^{i} W_i \times n_j \times \left(\left\{b_1 \cdot x_i + b_0\right\} - \bar{y}\right)^2 \tag{6}$$

where $SS_{reg,Q}$ and $SS_{reg,L}$ are the sum of squares of the regression of the quadratic and linear models, respectively, $W_i$ is the weighting applied at the $i^{th}$ concentration level, $n_j$ is the number of measurement replicates (here, 5) per concentration level, $\hat{y}_i$ is the predicted measurement at the $i^{th}$ concentration level (obtained by inserting the value of $x$ in the calibration equations, which need to be previously determined) and $\bar{y}$ is the average of measurements over all samples analyzed (nine concentration levels × five replicates = 45 measurements).

The sum of the residuals squared ($SS_{res,Q}$) was calculated for the quadratic model from

$$SS_{res,Q} = \sum_{n=1}^{i \times j} W_i \times \left(y_{ij} - \hat{y}_i\right)^2 \tag{7}$$

where $y_{ij}$ was the $j^{th}$ measurement at the $i^{th}$ concentration level.

The $F$ statistic was then calculated by

$$F_{calc} = \frac{SS_{reg,Q} - SS_{reg,L}}{\left(\frac{SS_{res,Q}}{n-3}\right)} \tag{8}$$

where $n$ is the total number of measurements ($i \times j$, here 45).

The probability ($P$) associated with the calculated $F$ statistic was found using the RStudio command

$$1 - pf\left(F_{calc}, 1, (n-3)\right) \tag{9}$$

*2.6. Normality of the residuals*

Normality of the standardized residuals was evaluated through the Kolmogorov–Smirnov (KS) and Cramer-von Mises (CVM) tests. The calculations necessary for these tests are fairly complicated and are covered in the accompanying paper ((II —Theoretical basis of the developped procedure). The probability values (output of the tests) were collected in the .txt file created by the script.

## 3. Results and discussion

*3.1. Raw data*

Raw data resulting from the replicate analysis ($j = 5$) of the nine calibration standards for cocaine and naltrexone are presented in Table 1. Calibration curves and variance graphs for both analytes are shown in Figure 1.

*3.2. Heteroscedasticity testing*

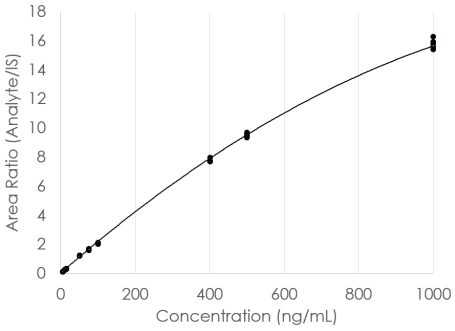
The purpose of testing for heteroscedasticity was to determine if weighted least-squares regression was necessary. Data are heteroscedastic if the absolute error (standard deviation of the replicates) varies systematically across concentration levels. Figure 2 shows the calibration curve of simulated homoscedastic (a) and heteroscedastic (b and c) data sets. In least-squares regression, the best model parameters (e.g., slope and intercept for linear models) are found by minimization of the sum of the squared error between the measured values and the values predicted by the model (squared residuals). In unweighted (also called uniformly weighted) least-squares regression, the default regression, all squared errors are treated equally in the summation. On the other hand, when data are heteroscedastic, there is greater confidence in the measured values that have the smallest error. This greater certainty should be used advantageously by giving a greater weight to the values with the smallest error in the summation of errors and therefore a greater influence in fixing the calibration parameters [5].

The $F$-test was applied to the measurements at the LLOQ and the ULOQ, where the difference in variance (error) is the largest inside the calibration range for heteroscedastic data sets of the $1/x$ and $1/x^2$ type. The $P$-value obtained represents the probability that variance at the ULOQ is smaller than or equal to the variance at the LLOQ (null

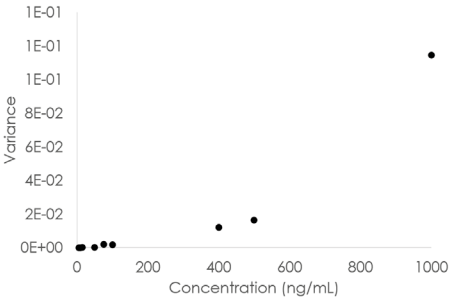
Table 1: Area ratio (analyte/IS) at nine concentration levels with five measurement replicates for cocaine and naltrexone

| **Concentration (ng/mL)** | **Area ratio (analyte/internal standard) Cocaine** | | | | |
|---|---|---|---|---|---|
| | **j = 1** | **j = 2** | **j = 3** | **j = 4** | **j = 5** |
| 5 | 0.131 | 0.127 | 0.131 | 0.126 | 0.130 |
| 10 | 0.256 | 0.249 | 0.244 | 0.249 | 0.249 |
| 15 | 0.340 | 0.333 | 0.328 | 0.331 | 0.311 |
| 50 | 1.235 | 1.257 | 1.224 | 1.234 | 1.225 |
| 75 | 1.596 | 1.663 | 1.710 | 1.656 | 1.613 |
| 100 | 2.055 | 2.046 | 2.109 | 2.033 | 2.127 |
| 400 | 7.733 | 7.727 | 7.964 | 7.687 | 7.747 |
| 500 | 9.688 | 9.447 | 9.557 | 9.476 | 9.346 |
| 1000 | 16.298 | 15.575 | 15.807 | 15.926 | 15.420 |

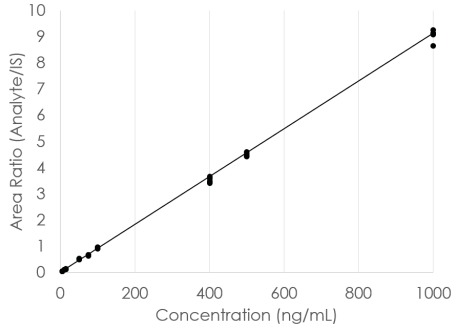
| **Concentration (ng/mL)** | **Area ratio (analyte/internal standard) Naltrexone** | | | | |
|---|---|---|---|---|---|
| | **j = 1** | **j = 2** | **j = 3** | **j = 4** | **j = 5** |
| 5 | 0.052 | 0.054 | 0.047 | 0.049 | 0.053 |
| 10 | 0.101 | 0.104 | 0.103 | 0.099 | 0.102 |
| 15 | 0.133 | 0.132 | 0.136 | 0.131 | 0.135 |
| 50 | 0.528 | 0.510 | 0.515 | 0.497 | 0.503 |
| 75 | 0.669 | 0.676 | 0.649 | 0.682 | 0.639 |
| 100 | 0.923 | 0.909 | 0.924 | 0.964 | 0.964 |
| 400 | 3.419 | 3.451 | 3.497 | 3.673 | 3.595 |
| 500 | 4.426 | 4.455 | 4.458 | 4.529 | 4.600 |
| 1000 | 8.656 | 9.092 | 9.139 | 9.110 | 9.269 |



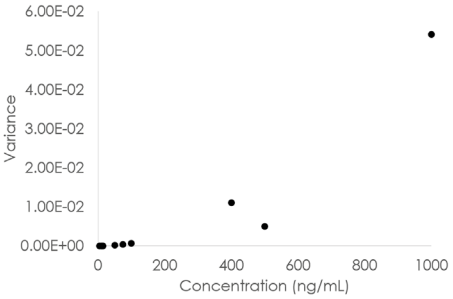

(a) Calibration curve of cocaine
Calibration equation:
$y = 7 \times 10^{-6}x^2 + 0.0224x + 0.0174$,
where $y$ is the area ratio and $x$ is the concentration in ng/mL.




(b) Variance graph of cocaine

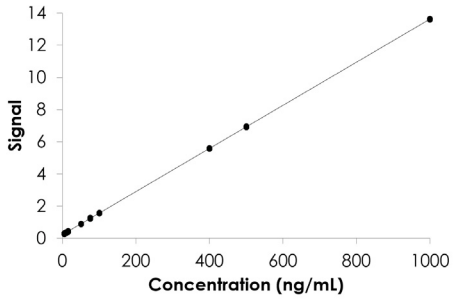



(c) Calibration curve of naltrexone
Calibration equation:
$y = 0.0091x + 0.0059$,
where $y$ is the area ratio and $x$ is the concentration in ng/mL.

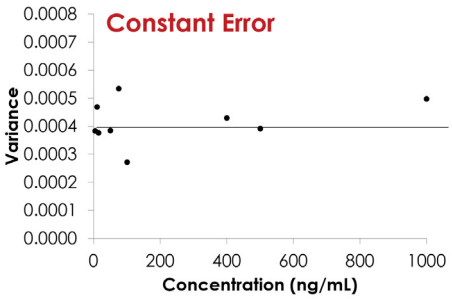



(d) Variance graph of naltrexone

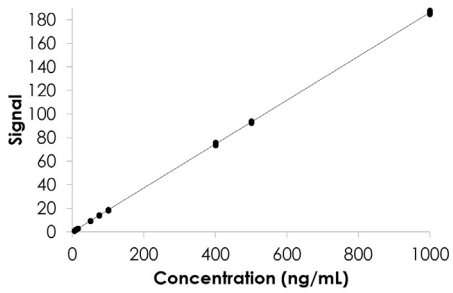
Figure 1: Calibration curves and variance graphs of cocaine and naltrexone



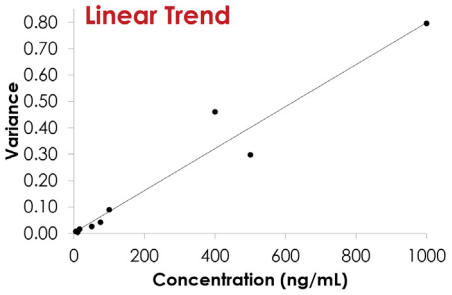

(a) Calibration curve with homoscedastic data



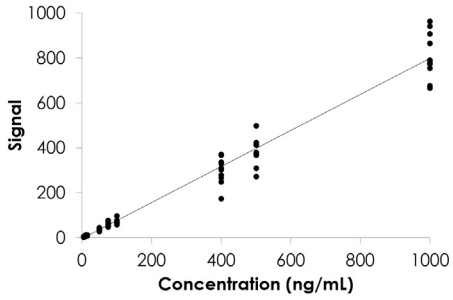

(b) Variance graph with homoscedastic data



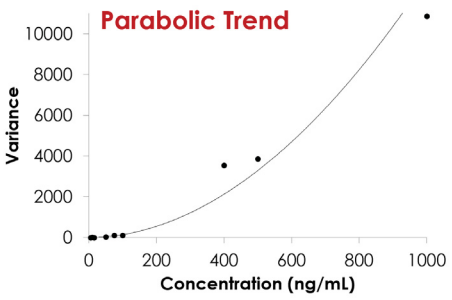

(c) Calibration curve with heteroscedastic data, $1/x$ weight




(d) Variance graph with heteroscedastic data, $1/x$ weight




(e) Calibration curve with heteroscedastic data, $1/x^2$ weight




(f) Variance graph with heteroscedastic data, $1/x^2$ weight

Figure 2: Calibration curves and variance graphs of linear models, created to show as well as possible the patterns of changing variance

hypothesis). If $P > 0.05$, this null hypothesis is accepted and data are considered to be homoscedastic (constant variance across concentrations), therefore no weighting is required. On the other hand, if $P < 0.05$, the null hypothesis is rejected and we accept the alternative hypothesis, which states that the variance at the ULOQ is larger than the variance at the LLOQ. This means that data are heteroscedastic and a weighting factor, which will be decided using the variance evaluation, should be used.

For cocaine and naltrexone, the $F$-test yielded $P$-values of $7 \times 10^{-9}$ and $7 \times 10^{-8}$, respectively, indicating that the data sets were heteroscedastic and weighting should be applied in both calibration procedures.

Application to simulated data showed this test is robust, with an average success rate of 98% for all types of calibration model utilizing five replicate measurements (Table 2). The success rate represents the percentage of data sets that was correctly classified (e.g., declared homoscedastic when it was indeed homoscedastic). Because the $P$-value threshold is set at 0.05 (5%), a 95% success rate is expected. Since the observed success rates for all models are near the expected rate, as is the average rate, this test is considered robust.

*3.3. Variance evaluation for weight selection*

Most data analysis software offers unweighted or uniform regression (weighting factor = 1) as the default as well as weighted regression using $1/x$ and $1/x^2$ weighting factors. The theoretical basis for these three common weighting factors is beyond the scope of this paper, but is a result of the type of noise that dominates over the calibration range [6, 7, 8]. Examination of the variance plot (variance of replicates vs. concentration) provides confirmation of the heteroscedasticity test results and is also suggestive of the appropriate weighting factor. The variance plot is provided as a PDF output when the R script is executed. Constant error across the calibration range from the LLOQ to the ULOQ is indicative that unweighted regression was appropriate (see Figure 2 (d)). Weighted regression was justified through heteroscedasticity testing and is apparent as increasing error across the variance plot. Plots which exhibited a linear trend, where variance increased proportionally to the concentration, indicated that $1/x$ weighting should be selected (see Figure 2(e)). If a parabolic trend was found, where variance increased proportionally to the square of the concentration, a $1/x^2$ weighting factor should be used [1] (see Figure 2(f)).

These characteristics were the basis for an automated, analyst independent selection of the required weighting using a variance evaluation. Indeed, properly weighted variances should be constant across the calibration range. For example, if the raw variances increase linearly with concentration ($x$), multiplying all variances by the appropriate weighting, $1/x$, will result in constant weighted variances across all the calibration range. Conversely, multiplying by an inappropriate weighting factor, say $1/x^2$, will produce changing weighted variances across the calibration range. Therefore, the weighting factor producing the most uniform set of weighted variances, as evaluated by taking the variance of the weighted and normalized variances for the different concentration levels, is the closest to the proper weight and should be used. Variance evaluation also acts as a double-check of the $F$-test result, building a healthy redundancy in the weighting

Table 2: Success rate of different tests in the process of calibration model selection and validation for 5, 7 or 10 simulated measurement replicates; 50 data sets were generated for each weighting/order combination

| | 5 replicates | | | | | |
|---|---|---|---|---|---|---|
| Model order | Linear | Quad. | Linear | Quad. | Linear | Quad. |
| Weighting | 1 | 1 | $1/x$ | $1/x$ | $1/x^2$ | $1/x^2$ |
| $F$-test (Heteroscedasticity) (%) | 98 | 92 | 100 | 98 | 100 | 100 |
| Variance test (Weight selection) (%) | 100 | 98 | 58 | 70 | 100 | 100 |
| Partial $F$-test (Order selection) (%) | 96 | 96 | 98 | 88 | 90 | 50 |
| Validation (CVM) (%) | 100 | 100 | 100 | 100 | 100 | 100 |
| | 7 replicates | | | | | |
| Model order | Linear | Quad. | Linear | Quad. | Linear | Quad. |
| Weighting | 1 | 1 | $1/x$ | $1/x$ | $1/x^2$ | $1/x^2$ |
| $F$-test (Heteroscedasticity) (%) | 94 | 98 | 100 | 100 | 100 | 100 |
| Variance test (Weight selection) (%) | 100 | 100 | 86 | 90 | 100 | 100 |
| Partial $F$-test (Order selection) (%) | 98 | 100 | 100 | 94 | 98 | 48 |
| Validation (CVM) (%) | 100 | 100 | 98 | 100 | 100 | 100 |
| | 10 replicates | | | | | |
| Model order | Linear | Quad. | Linear | Quad. | Linear | Quad. |
| Weighting | 1 | 1 | $1/x$ | $1/x$ | $1/x^2$ | $1/x^2$ |
| $F$-test (Heteroscedasticity) (%) | 100 | 98 | 100 | 100 | 100 | 100 |
| Variance test (Weight selection) (%) | 100 | 100 | 98 | 92 | 98 | 100 |
| Partial $F$-test (Order selection) (%) | 98 | 100 | 88 | 90 | 94 | 58 |
| Validation (CVM) (%) | 100 | 100 | 100 | 98 | 98 | 98 |

selection for the calibration curve.

Calibration data and variance plots obtained for cocaine and naltrexone are shown in Figure 1. Both variance plots show a parabolic pattern, although this pattern was subjectively less clear for naltrexone. Variance test scores for cocaine ($V_{W_1} = 1.7 \times 10^{-6}$; $V_{W_{1/x}} = 2.0 \times 10^{-9}$; $V_{W_{1/x^2}} = 6.2 \times 10^{-12}$) and naltrexone ($V_{W_1} = 3.8 \times 10^{-7}$; $V_{W_{1/x}} = 5.0 \times 10^{-10}$; $V_{W_{1/x^2}} = 2.6 \times 10^{-12}$) confirmed that a $1/x^2$ weighting factor should be used to build calibration models for both analytes, since this weight produced the smallest spread of weighted variances. Both the plot and weighted variance evaluation provide confirmation of the heteroscedasticity $F$-test results. Nearly, all LC-MS/MS analyses spanning a few concentration orders of magnitude can be expected to produce data with this weighting [1].

It is important to note here that sampling statistics govern the variance estimation at each concentration level. Thus, the smaller the number of replicates, the more likely the variance estimation is to be erroneously large or small. This estimation error propagates into the weighted variances as a bias toward an erroneous weighting and can result in incorrect selection of the weight. Tests with simulated data show that this happens up to 42% of the time for $1/x$ data with five replicates (Table 2). To overcome this fundamental limitation in the data requires increasing the number of replicates, which increased the success rate in identifying the proper weighting factor to 86% for 7 replicates and 92% for 10 replicates. For this reason, the authors suggest that the use of seven measurement replicates when selecting and validating the calibration model, which provides improved performance with the tests compared to the five measurement replicates suggested by the SWGTOX guidelines. In general, improved performance occurs for all tests with increased replicates, but it is the most marked in the weight selection step. For diverse practical reasons, analysts may justifiably use five measurement replicates and, with the aid of the calibration model selection scheme presented here, produce validated calibration models. However, they need to realize that the trade-off will be an increased frequency of incorrect weight and/or order selection that can ripple through to lower accuracy and precisions in the results.

### *3.4. Partial F-test for model order selection*

With the weighting factor chosen, the next step was to select the model order (i.e., linear or quadratic). The recommended practice by the SWGTOX and the FDA is to choose the model with the lowest order that adequately describes the calibration system under study [2, 9]. Often times in bioanalysis laboratories, this is done using the "Test and Fit" strategy [1], meaning that the lowest order yielding standard and QC accuracies below the 15% or 20% bar is chosen. However, rather than choosing the model which is "good enough", the partial $F$-test can be used to improve the likelihood of selecting the true model order underlying the measurements.

Selection of the appropriate model was done by performing a partial $F$-test. Here, the test was applied to establish if the quadratic calibration model significantly improved the captured variance of the data compared to a linear model [4]. Linear or quadratic calibration responses, which are typically encountered in toxicology validation work, were

the two models compared. However, it is noteworthy that the partial $F$-test allows alternate calibration models to be compared. This test compares the improvement in the sum of squares of the regression when switching from a linear to a quadratic calibration model ($SS_{reg,Q} - SS_{reg,L}$) to the sum of squares of the residuals in the quadratic model ($^{SS_{res,Q}}/_{n-3}$). If there is a significant increase in the variance explained by the quadratic regression, then using a quadratic model is justified [4].

For cocaine and naltrexone, the $P$-values obtained from the partial $F$-test were $1 \times 10^{-13}$ and 0.20, respectively. In the case of cocaine, since $P < 0.05$, the increase in the sum of squares of the regression when switching to a quadratic model was significant, therefore a quadratic (second order) calibration model should be used for this analyte. On the other hand, the $P$-value for naltrexone was $> 0.05$, which means the quadratic model does not capture a significantly greater portion of the measurements' variance. Therefore, a linear model should be used for naltrexone.

Tests with simulated data showed that the erroneous outcome of selecting a quadratic model when in fact the underlying data model was linear happens in 4% of the cases on average (Table 2), near the expected value of 5%. The opposite error (selecting linear when quadratic is the correct model) happens far more often, with an average success rate of 80% (Table 2). Erroneous selection of a linear model mainly happened when the second order term was small relative to the error at the upper concentration levels, for example when increasing variance (heteroscedastic data, especially $^1/_{x^2}$) masked the curvature. Indeed, quadratic, $^1/_{x^2}$ data with $n = 7$ show a 78% success rate for model order selection when %RSD = 2.5%, but the success rate decreased to 26% when %RSD = 20% (Supplemental Data 4). Large and/or increasing variance can mask curvature present in the data and result in an undetectable improvement in the fit obtained when using a quadratic model. Unfortunately, increasing the number of measurement replicates will not have a marked effect in this case. Satisfyingly, when the partial $F$-test fails, the result is to err on the side of caution advocated by the SWGTOX and the FDA: the lowest order model fitting the data (linear) is selected. Ultimately, when the curvature is masked, using a linear model instead of a quadratic one will not have an appreciable impact on the accuracy of the results.

### *3.5. Normality of the residuals*

After having chosen the calibration model that best represents the data (weight and order), its validation was required. In principle, the correct model should describe all the systematic trends in the data with only random error remaining in the residuals [5]. Therefore, the residual errors are expected to follow a normal distribution. Both the CVM and KS procedures can be used to test whether the standardized residual distribution is significantly different from a normal distribution. In practice, to adhere correctly to statistical procedures and for clarity of decision, the user is expected to choose only one of them as a validation test. The authors favor the use of the CVM normality test, because the results obtained in simulations demonstrate that it is stricter than the KS test (lower $P$-values obtained) and therefore has greater ability to detect departure from normality.

The CVM test produced $P$-values of 0.865 and 0.992 for cocaine and naltrexone, respectively. In both cases, KS and CVM produced $P$-values $> 0.05$ suggesting that the standardized residuals did not depart significantly from a normal distribution. The calibration model was, therefore, considered validated. When $P < 0.05$, the distribution of the standardized residuals is significantly different from a normal distribution. This indicates that the calibration model chosen did not accurately account for all systematic trends in the data and therefore should not be validated.

When the CVM test was applied to the simulated data, the success rate was more than 98% across all six calibration models and higher than the expected rate of 95% (Table 2). Again, low numbers of measurement replicates and/or high variance will negatively impact the ability of the tests to detect departure from normality (i.e., inappropriate models producing non-normally distributed residuals). This too points toward the benefits of using a higher number of measurement replicates as a better practice.

When a model fails to pass the validation step, the analyst should attempt to understand why, so that the fundamental problem can be addressed. A detailed exploration of all possible problems is well beyond the scope of this paper but certainly systematic errors or instrument drift should be investigated. Where appropriate, the method should be modified so an adequate model for the data can be found. This might involve a change of IS, a modification of the MS–MS transition(s), a reduction of the dynamic range or a move toward more exotic calibration models (e.g., logarithmic) when justified by the expected analyte/instrument response. The analyst should also be wary of methods with excessively high %RSD since, paradoxically, these methods are easier to validate but are inherently less precise and potentially less accurate.

## 4. Conclusions

We developed a general procedure to select and validate quantitative calibration models (Figure 3). The two model analytes, cocaine and naltrexone, were quantified by LC-MS/MS. The $F$-test demonstrated that both data sets were heteroscedastic and required weighting in the calibration process. Variance evaluation indicated that the spread of weighed normalized variances was the lowest for $1/x^2$ weighting. Visual examination of the variance graph and evaluation of the variance confirmed the $F$-test results. A weight of $1/x^2$ was, therefore, chosen for both analytes. A partial $F$-test demonstrated a significant increase in the sum of squares of the regression when switching from a linear to a quadratic model for cocaine, but not for naltrexone. Therefore, a quadratic calibration model was adopted for cocaine but a linear model was retained for naltrexone. Both calibration models were validated through CVM normality testing of the residuals. Analysis of simulated data sets showed good performance level of all tests; but it also pointed to benefits of increased replicate analysis ($n = 7$) in accurate selection of the calibration model.

Choosing the correct calibration model can have tremendous impact on the accuracy of the QCs. The process of selection and validation of a calibration model explained here is a stepwise, biasfree alternative to other less rigorous methods such as visual inspection of the standardized residuals graph. Simulations using experimentally determined

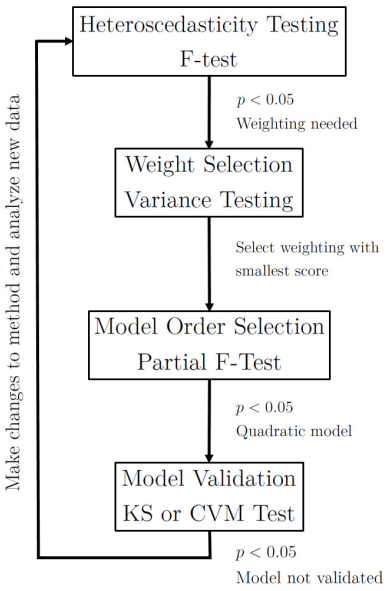



Figure 3: Flowchart for the selection and validation of the calibration model

calibration curves have shown that this approach performs much better than a more traditional approach of fitting increasingly complex models until QC accuracy is satisfying. Additionally, the calculations and interpretation of tests results have been automated through the use of RStudio scripts made available to all readers in Supplemental Data 3. Experimental workload is not modified by the use of this scheme, and only a minute or two per analyte are added to the data treatment time, making this a very efficient option to remove the subjectivity in calibration model selection. This tool is intended to aid analysts in better calibration model selection in toxicology and bioanalysis.

## 5. Funding

Brigitte Desharnais, Félix Camirand-Lemyre and Cameron D. Skinner gratefully acknowledge support of the National Sciences and Engineering Research Council of Canada. Brigitte Desharnais also gratefully acknowledges the support of the Fonds de recherche du Québec – Nature et technologies.

## 6. Acknowledgements

The authors wish to thank Cynthia Côté for her thorough review and comments as well as Sue-Lan Pearring and Chris H. House for their opinion on key matters. The authors are also grateful to Gabrielle Daigneault, Lucie Vaillancourt, Julie Laquerre and Marc-André Morel for their technical work.

## References


[1] H. Gu, G. Liu, J. Wang, A.-F. Aubry, 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, Analytical Chemistry 86 (2014) 8959–8966.

[2] Scientific Working Group for Forensic Toxicology, Scientific Working Group for Forensic Toxicology (SWGTOX) Standard Practices for Method Validation in Forensic Toxicology, Journal of Analytical Toxicology 37 (2013) 452–474.

[3] Davidian, Marie and Haaland, Perry D, Regression and calibration with nonconstant error variance, Chemometrics and Intelligent Laboratory Systems 9 (1990) 231–248.

[4] D.L. Massart, B.G.M. Vandeginste, L.M.C Buydens, S. De Jong, P.J. Lewi, J. Smeyers-Verbeke, Multiple and Polynomial Regression, in: Handbook of Chemometrics and Qualimetrics: Part A, volume 20A of *Data Handling in Science and Technology*, Elsevier, Amsterdam, Netherlands, 1997, pp. 263–303.

[5] D.L. Massart, B.G.M. Vandeginste, L.M.C Buydens, S. De Jong, P.J. Lewi, J. Smeyers-Verbeke, Straight Line Regression and Calibration, in: Handbook of Chemometrics and Qualimetrics: Part A, volume 20A of *Data Handling in Science and Technology*, Elsevier, Amsterdam, Netherlands, 1997, pp. 171–230.

[6] Karnes, H Thomas and Shiu, Gerald and Shah, Vinod P, Validation of bioanalytical methods, Pharmaceutical Research 8 (1991) 421–426.

[7] Hubert, Ph and Chiap, Patrice and Crommen, Jacques and Boulanger, Bruno and Chapuzet, E and Mercier, N and Bervoas-Martin, S and Chevalier, P and Grandjean, D and Lagorce, Ph and others, The SFSTP guide on the validation of chromatographic methods for drug bioanalysis: from the Washington Conference to the laboratory, Analytica Chimica Acta 391 (1999) 135–148.

[8] Ingle, James D. and Crouch, Stanley R., Signal-to-Noise Ratio Considerations, in: Spectrochemical Analysis, Prentice Hall, Englewood Cliffs, United States of America, 1988, pp. 135–163.

[9] Food and Drug Administration, Bioanalytical Method Validation – Guidance for Industry, Technical Report, Silver Springs, United States of America, 2001.