Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Pharmaceutical and Biomedical Analysis
33 (2003) 7–20
JOURNAL OF PHARMACEUTICAL AND BIOMEDICAL ANALYSIS
www.elsevier.com/locate/jpba

# Application of an improved procedure for testing the linearity of analytical methods to pharmaceutical analysis

Howard Mark *

*Mark Electronics, 21 Terrace Avenue, Suffern, NY 10901, USA*

Received 12 July 2002; accepted 27 March 2003

**Abstract**

Examination of the requirements of the Food and Drug Administration (FDA) for evaluating the linearity of an analytical method reveals them to be unsatisfactory, in both the definition of linearity and in the specifications for testing this property of an analytical method. A new definition for linearity is proposed, along with a new method for evaluating this property of an analytical method, one that is consistent with the definition. The method is tested by re-evaluating the linearity of data collected from a Near-Infrared method of analysis of pharmaceutical preparations.
© 2003 Elsevier B.V. All rights reserved.

*Keywords:* Linearity; Validation; FDA guidelines; Linearity testing; Linearity definition

## 1. Introduction

When the Food and Drug Administration (FDA) evaluates whether an analytical method is ''suitable for its intended purpose'', proof of which is essentially the definition of ''validated'', it considers a large number of characteristics of that method. Not all of those characteristics are necessarily required to be demonstrated in every case. Evaluation of the linearity of the relationship between the actual analyte concentration and the test result from the method, however, is required for quantitation testing for impurities, and for assay methods.

The requirement for linearity is independent of the technology used to ascertain the analyte concentration. In the end, even the most modern instrumental methods that rely on multivariate chemometric computer methods have to produce a number that represents the final answer for that analyte, and that is the test result from that instrument. This term, therefore, holds good for every analytical methodology from manual wet chemistry to the latest high-tech instrument.

Many analytical methods are known where the relationship between the raw measured data and the analyte concentrations are non-linear. Electrochemical measurements, for example, rely on the Nernst equation, which indicates a logarithmic relationship between the cell voltage and the analyte concentration. Spectroscopic measurements rely on Beer's Law, which also expresses a

* Tel.: +1-845-357-9496; fax: +1-845-369-9110.
*E-mail address:* hlmark@prodigy.net (H. Mark).

0731-7085/03/$ - see front matter © 2003 Elsevier B.V. All rights reserved.
doi:10.1016/S0731-7085(03)00346-7

logarithmic relationship between the measured transmittance data and absorbance, absorbance being the quantity that theoretically is proportional to the analyte concentration.

Using spectroscopy as the basis for the discussion here, we find that while Beer's Law shows that the measured absorbance is proportional to concentration, in practice many sources of interference can occur which will cause deviations from the theory. For example, stray light will cause deviations from linearity at low transmittance levels, as will excessive bandwidth of the monochromator. Saturation of the detector, or operation at too high a signal level will make the detector response become non-linear with respect to the optical energy, which will make the computed absorbance non-linear (with respect to concentration) at high energy levels (i.e. high transmittance of the sample). Other effects also exist.

"Linearity" is defined in section 7 of the Glossary to [1] as:

> "The linearity of an analytical procedure is its ability (within a given range) to obtain test results which are directly proportional to the concentration (amount) of analyte in the sample".

The following quote is found in section III of [2]:

> "In some cases, to obtain linearity between assays and sample concentrations, the test data may have to be subjected to a mathematical transformation prior to the regression analysis".

This quote clearly indicates that if the raw data is not itself linearly related to the analyte concentrations, then it may be made linear through a mathematical transformation. As indicated above, any suitable mathematical function may be used for the linearization process.

The guidelines also contain the following passage (in section III of [2]):

> "If there is a linear relationship, test results should be evaluated by appropriate statistical methods, for example, by the calculation of a regression line by the method of least squares".

This passage, then, applies to testing the relationship for linearity. Thus, as currently written, there is a distinction drawn between making the relationship between raw data and concentrations become linear through a mathematical transformation (which may make use of any suitable mathematical function), and testing whether that relationship, after transformation, is linear (which is currently specified to make use only of straight lines). Thus there is a difference between linearizing the relationship, and testing whether the linearization was successful, and it is important to distinguish between the types of mathematical functions that are currently specified to be used for the two activities.

The FDA/ICH guidelines provide a definition of the meaning of the term "linearity", given above. This definition is an extremely strict one, one which is unattainable in practice when noise and error are taken into account. Fig. 1 illustrates the problem with this definition. Fig. 1 shows schematically a plot of a set of hypothetical data that most would agree represents a substantially linear relationship between the test result and the analyte concentration. While there is a line that meets the criterion that "test results are directly proportional to the concentration of analyte in the sample", none of the data points fall on that line. Therefore, in the strict sense of the phrase, none of the data representing the test results can be said to be proportional to the analyte concentration. In the

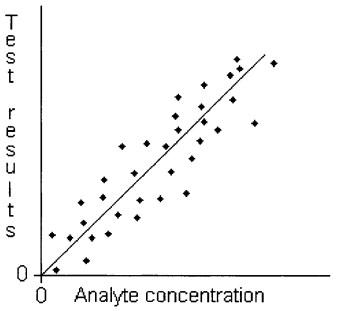


Fig. 1. A representation of linear data.

face of non-linearity of response, there are additional, systematic, departures from the line, as well as random departures, but in neither case is any data point strictly proportional to the concentration.

Less strict descriptions of linearity are also provided. One recommendation is visual examination of a plot (unspecified, but presumably also of the method response vs. the analyte concentration); this method is also recommended by Taylor [3] and by Meier and Zund [4]. This method works fairly well, but is subjective and not amenable to the application of statistical tests, making an objective mathematical evaluation unattainable. It also is open to different interpretations, and is unsuitable for the application of computerized or automated screening methods.

Another recommendation in the guidelines is to use "statistical methods"; calculation of a linear regression line is advised. This is not so much a definition of linearity as an attempt to evaluate it. If regression is performed, then the correlation coefficient, slope, y-intercept and residual sum of squares are to be reported. There is no indication given, however, as to how these quantities are to be related to linearity, and as Anscombe shows, they are not [5]. Anscombe presents several (synthetic) data sets, to which he then fits a straight line using Least Squares regression, as the guidelines recommend. One data set is substantially linear (much as the plot represented in Fig. 1 is substantially linear), while another is a data set that is obviously very non-linear. Other data sets have other faults. When linear regression is performed on any of these data sets as recommended by the guidelines, all the recommended regression statistics are identical for the different sets of data. It is immediately obvious, therefore, that the regression results cannot distinguish between the different cases, since the regression results are the same for all of them.

Other linearity tests exist, in addition to the ones in the official guidelines. The Durbin–Watson (DW) statistic [6–11] has been proposed [12,13] as a statistically-based test method for evaluating linearity. While a step in the right direction, upon close examination the DW statistic is found to have a fatal flaw: The value that DW should have is two, for residuals from regression data that meet all the theoretical requirements, i.e. that are random, independent, have a Gaussian distribution and represent a linear relation between the two variables. In statistical jargon, they have an expected value of two (see, for example, pages 180–185 in [11]). However, calculating the DW statistic for the data sequence:

$$0,\ 1,\ 0,\ -1,\ 0,\ 1,\ 0,\ -1,\ 0,\ 1,\ 0,\ -1,\ \ldots$$

also results in a computed value of two, despite the fact that this sequence is non-random, non-independent, does not have a Gaussian distribution and is definitely not linear. Thus we expect that any set of residuals showing a similar cyclic behavior will also compute out to a value of DW that will erroneously indicate satisfactory behavior of the residuals, creating a rather insidious fault in this test.

A somewhat similar test is based on a statistical $F$-test, as recommended by Taylor (see page 102 in [3]) and also by Hald [14] and by Dixon and Massey [15]. This $F$-test is based on comparing within-sample estimates of precision to the overall error of the analysis. Ideally they should be the same, but if non-linearity is present, the overall error will be larger (by a statistically significant amount) than the within-sample precision. Unfortunately this test is not only, as Taylor states, insensitive but also suffers from the difficulty that any bias in the estimates of the concentration will inflate the $F$-value and be taken as an indicator of non-linearity when in fact some other phenomenon is affecting the data. Furthermore it requires multiple readings of every sample by both the method under test and the method used to determine the actual concentration of the analyte, making it impractical to apply on a routine basis, and inapplicable to already-existing data.

Hald [14] also recommends testing whether the residuals have a Gaussian distribution, since it is unlikely that the residuals will be so distributed if there is appreciable non-linearity in the relationship between concentration and the test results. While that is true, the test is non-specific; this test is again very insensitive to actual non-linearity (especially for small numbers of samples), and furthermore suffers from the same difficulties as

the preceding test (the $F$-test), in that other types problems with the data may be erroneously called non-linearity.

Therefore, none of these methods are completely satisfactory. In fact, the recommendations of the official guidelines for evaluating linearity, both the definitions and the recommended method(s) for assessing it, while well-intended are themselves not suitable for their intended purpose in this regard. Therefore, let us start by proposing a definition that circumvents the problems of the above definitions and that can serve as a basis for further discussion. We, therefore, propose to define linearity as follows: "Linear data is data where the relationship between analyte concentrations and test results can be fitted (in the Least-Squares sense) as well by a straight line as by any other function".

In this article we present the details concerning a new method of testing the linearity of data. A previous note [16] examined the feasibility of the method by testing it on Anscombe's synthetic data, but it was never used to verify the linearity of real data, neither were any details concerning its operation presented. The application of the method to Anscombe's data shows it to be extremely promising as a replacement for the unsatisfactory methods currently specified. In this paper we, therefore, present the derivation and details of the operation of this new method of evaluating data, and report on it's ability to test linearity by applying it to data from a real analytical method: a method of NIR spectroscopic analysis using diffuse transmittance measurements. This data was collected as part of a validation study and was previously published [13].

## 2. Theory

We propose a method of determining non-linearity (or showing linearity) that bears a close resemblance to the current method of assessing linearity that the FDA and ICH guidelines recommend (that of fitting a straight line to the data, and assessing the goodness of the fit). But as we showed, based on the work of Anscombe [5], the currently recommended method for assessing linearity is faulty because it cannot distinguish linear from non-linear data.

An extension of that method, however, can. As with our definition of linearity (given above), our test almost seems to be the same as the FDA/ICH approach, which we discredited. The difference is that we include the possibility of fitting other functions to the data and comparing the fits, whereas the FDA/ICH guidelines only specify trying to fit a straight line to the data. Our test is also in line with out proposed definition of linearity: conceptually, we can try to fit functions other than a straight line to the data, and if we cannot obtain an improved fit, we can conclude that the data is linear.

It is possible to fit other functions to a set of data using least-squared mathematics. In fact, the well-known Savitzky–Golay (S–G) algorithm is based on fitting polynomials to data [17]. We differ from S–G in that, while S–G fits a polynomial to small sections of the data, we extend it to fit the polynomial to the entire data set at once, rather than a few points at a time. Many texts exist dealing with this subject, but we will follow the presentation of Arden [18]. Arden points out and discusses in detail, many applications of numerical analysis, but they share common characteristics. The data is assumed to be univariate and to follow the form of some arbitrary mathematical function, where the nature of the function may be undetermined. From Taylor's theorem, however, any function can be approximated by a polynomial, although the degree of the polynomial may also not be known a priori (the "degree" of a polynomial being the highest power to which the variable is raised in that polynomial). Through the application of Taylor's theorem, therefore, polynomials become a surrogate for "any mathematical function".

Polynomials have occasionally been used for other aspects of calibration in data analysis, in the manner shown by Mandel (see page 100 in [19], Meier and Zund [4] use this method similarly). Mandel did what we mention elsewhere in this paper: he used an orthogonal quadratic polynomial to improve the fit of a function to a data set, after non-linearity was established. But it is

repetitive to say that this technique has not previously been used as a means of testing for non-linearity, as opposed to reducing or eliminating it.

Our current goal is not to approximate the relationship between test results and analyte concentration as well as possible, but only to ascertain whether a straight line fits the data as well as a polynomial. We will see that for this purpose we need not use polynomials of high degree.

The presentation in Arden's book, which we follow, makes the assumption that there is a single (univariate) mathematical system (corresponding to "analyte concentration" and "test reading"), and that there is a functional relationship between these two variables of interest, although again, the nature of the relationship may be unknown. The function is approximated by a polynomial, and any given polynomial must minimize the sum of the squares of the differences between each datum and the corresponding point of the polynomial.

By far the easiest type of polynomial to deal with, and therefore, the most widely used approximating functions are simple polynomials; these are also convenient in that they are the direct result of applying Taylor's theorem, since Taylor's theorem produces a description of a polynomial that estimates the function being reproduced:

$$Y = a_0 + a_1X + a_2X^2 + a_3X^3 + \ldots + a_nX^n \tag{1}$$

where X and Y correspond to the test results and the analyte concentrations. Often a polynomial of degree 2 (quadratic) can provide a satisfactory fit to the data. Polynomials of higher degree may provide a better fit, if the data requires it.

The mathematics of fitting a polynomial by least squares are relatively straightforward, and we sketch the derivation here following Arden, but as we shall see is rather generic: Starting from equation 1, we want to find coefficients (the $a_i$) that minimize the sum-squared difference between the data and the function's estimate of that data, given a set of values of X. Therefore, we first form the desired differences:

$$D = a_0 + a_1X + a_2X^2 + a_3X^3 + \ldots + a_nX^n - Y \tag{2}$$

Then we square those differences and sum those squares over all the sets of data (corresponding to the samples used to generate the data);

$$\sum_i D^2 = \sum_i (a_0 + a_1X + a_2X^2 + a_3X^3 + \ldots + a_nX^n - Y)^2 \tag{3}$$

The problem now is to find a set of values for the $a_i$ that minimizes $\Sigma D^2$. We do this by the usual procedure of taking the derivative and setting it equal to zero. In this case, we take the derivative of $\Sigma D^2$ with respect to each $a_i$ and set each of those derivatives equal to zero. Since there are multiple $a_i$, this creates not one, but a whole set of equations. There are $n+1$ different $a_i$ (including $a_0$), therefore, we wind up with $n+1$ equations, although here we only show the first three of the set for our exposition:

$$\begin{aligned}&\partial\left(\sum_i D^2\right)/\partial a_0 \\ &= \partial\left(\sum (a_0 + a_1X + a_2X^2 + a_3X^3 + \ldots + a_nX^n - Y)^2\right)/\partial a_0 \\ &= 0\end{aligned} \tag{4a}$$

$$\begin{aligned}&\partial\left(\sum_i D^2\right)/\partial a_1 \\ &= \partial\left(\sum (a_0 + a_1X + a_2X^2 + a_3X^3 + \ldots + a_nX^n - Y)^2\right)/\partial a_1 \\ &= 0\end{aligned} \tag{4b}$$

$$\partial\left(\sum_i D^2\right)/\partial a_2 = \partial\left(\sum(a_0 + a_1X + a_2X^2 + a_3X^3 + \ldots + a_nX^n - Y)^2\right)/\partial a_2 = 0 \quad (4c)$$

etc.

The details consist of taking the indicated derivative of each term (while noting that $\partial(\Sigma_i F^2) = 2\,\Sigma_i F\,\partial F$ (where F is the inner summation of the $a_i X$)), separating the summations, dividing by two to eliminate the constant term and subtracting the term involving Y from each side of the resulting equation. All this puts the equations into their final form. The details can be found in chapter 9 in Arden [18], here we show only the result:

$$a_0\sum(1) + a_1\sum_i X + a_2\sum_i X^2 + a_3\sum_i X^3 + \ldots + a_n\sum_i X^n = \sum_i Y \quad (5a)$$

$$a_0\sum_i X + a_1\sum_i X^2 + a_2\sum_i X^3 + a_3\sum_i X^4 + \ldots + a_n\sum_i X^{n+1} = \sum_i XY \quad (5b)$$

$$a_0\sum_i X^2 + a_1\sum_i X^3 + a_2\sum_i X^4 + a_3\sum_i X^5 + \ldots + a_n\sum_i X^{n+2} = \sum_i X^2Y \quad (5c)$$

etc.

The values of X and Y are known since they constitute the data, and therefore, the summations, once evaluated, are constants. Therefore, Eqs. (5a), (5b) and (5c) comprise a set of n+1 equations in n+1 unknowns, the unknowns being the various values of the $a_i$. Therefore, solving Eqs. (5a), (5b) and (5c) as simultaneous equations for the $a_i$ results in the calculation of the coefficients that describe the polynomial (of degree n) that best fits (in the least squares sense) the data.

In principle, the relationships described by Eqs. (5a), (5b) and (5c) could be used directly to construct a function that relates test results to sample concentrations. In practice we find that correlation between the various powers of X is an important consideration that must be taken into account. We find, for example, that if we square each of the numbers in the sequence ⟨1 ... 10⟩, creating the corresponding sequence ⟨1, 4, 9, 16, ... 100⟩, and then apply the formula for calculating the correlation coefficient to the two sequences, we find that the correlation coefficient of the integers from 1 to10 with their squares is 0.974—a rather high value. The formula for calculating correlation coefficient is [11]:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{X})(y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{X})^2\sum_{i=1}^{n}(y_i - \bar{Y})^2}}$$

where: r represents the correlation coefficient; $X_i$, $Y_i$ represent individual data points; n represents the number of x, y pairs; $\bar{X}$, $\bar{Y}$ represent the means of the x and y values, respectively.

Correlation effects are of concern for us. Our goal is to formulate a method of testing linearity in such a way that the results can be justified statistically. Ultimately we will perform statistical testing on the coefficients of the fitting function that we use. We will use a *t*-test to see whether any given coefficient is statistically significant, compared with the standard error of that coefficient. We do not need to solve the general problem, however, just as we do not need to create the general solution implied by equation 1. In the broadest sense, equation 1 is the basis for computing the best-fitting function to a given set of data, but that is not our goal. Our goal is only to determine whether the data represent a linear function or not. To this end it suffices to ascertain only whether the data can be fit better by any polynomial of degree greater than 1, than it can by a straight line (which itself is a polynomial of degree 1). To this end we need to test a polynomial of degree higher than 1. While in some cases, the use of more terms may be warranted, as we shall see, we need test only the ability to fit the data using only one term of degree greater than one.

Hence, while in general we may wish to try fitting equations of degrees 3, 4, ... m (where m is some upper limit less than n), we need begin by using only polynomials of degree 2, i.e. quadratic fits.

A complication arises, caused by the correlation effects. We learn from the theory of multiple regression analysis, that when two (or more) variables are correlated, the standard error of both variables is increased over what would be obtained if equivalent but uncorrelated variables are used. This is discussed by Daniel and Wood (see page 55 in [20]), who show that the variance of the estimates of coefficients (from their standard errors) is increased by a factor of;

$$\mathrm{VIF} = 1/(1 - \mathrm{R}^2) \tag{6}$$

when there is correlation between the independent (X) variables, where R represents the correlation coefficient between the variables and we use the abbreviation VIF, as is sometimes done, to mean Variance Inflation Factor. Arden describes a general method for removing the correlation between the various powers of X in a polynomial, based on the use of orthogonal Chebyshev polynomials. Other types of orthogonal polynomials also exist and could be used, such as Legendre polynomials, Jacobi polynomials, and others.

But this method is unnecessarily complicated for our current purposes, and in any case has a severe limitation of its own; when applied to actual data. Chebyshev and other types of orthogonal polynomials are orthogonal only if the data is uniformly, or at least symmetrically, distributed along the X-axis; in practical applications, real data will seldom meet that requirement.

Since we do not need to deal with the general case, we can use a simpler method to orthogonalize the variables, one also based on the presentation of Daniel and Wood, who describe a transformation that makes the square of that variable uncorrelated with the variable itself. This is done by computing a new variable Z with the property that for the given data set, $(\mathrm{X}-\mathrm{Z})^2$ is uncorrelated with X. Thus, once Z is computed, it is subtracted from each of the original values of X and the result is squared. A symmetric distribution of the data is not required since the data distribution is taken into account in the formula. Z is calculated, by imposing the condition that $(\mathrm{X}-\mathrm{Z})^2$ is to be uncorrelated with X, this requires that the condition:

$$\sum\nolimits_i (\mathrm{X_i} - \bar{X})(\mathrm{X_i} - \mathrm{Z})^2 = 0 \tag{7}$$

must be met (where the summation is taken over all the samples in the set). The formula on the left-hand side of equation 7 is essentially the numerator of the formula for the correlation coefficient between X and $(\mathrm{X}-\mathrm{Z})^2$. Solving equation 7 for Z is not obvious, therefore, we will show how to solve equation 7 for Z. First expand the square term in equation 7;

$$\sum\nolimits_i (\mathrm{X_i} - \bar{X})(\mathrm{X_i^2} - 2\mathrm{X_i Z} + \mathrm{Z}^2) = 0 \tag{8}$$

Then multiply through and collect terms:

$$\sum\nolimits_i (\mathrm{X_i^2}(\mathrm{X_i} - \bar{X}) - 2\mathrm{X_i Z}(\mathrm{X_i} - \bar{X}) + \mathrm{Z}^2(\mathrm{X_i} - \bar{X})) = 0 \tag{9}$$

Separate the summations and bring constants outside the summations:

$$\sum\nolimits_i \mathrm{X_i^2}(\mathrm{X_i} - \bar{X}) - 2\mathrm{Z} \sum\nolimits_i \mathrm{X_i}(\mathrm{X_i} - \bar{X}) + \mathrm{Z}^2 \sum\nolimits_i (\mathrm{X_i} - \bar{X}) = 0 \tag{10}$$

Since $\Sigma_i(\mathrm{X_i} - \bar{X}) = 0$, the last term in equation 10 vanishes, leaving:

$$\sum\nolimits_i \mathrm{X_i^2}(\mathrm{X_i} - \bar{X}) - 2\mathrm{Z} \sum\nolimits_i \mathrm{X_i}(\mathrm{X_i} - \bar{X}) = 0 \tag{11}$$

Equation 11 is now easily rearranged explicitly for Z:

$$\mathrm{Z} = \frac{\sum_{i=1}^{N} \mathrm{X_i^2}(\mathrm{X_i} - \bar{X})}{2\sum_{i=1}^{N} \mathrm{X_i}(\mathrm{X_i} - \bar{X})} \tag{12}$$

It is also relatively straightforward to show that equation 12 is equivalent to the expression on page 121 in [20]. Thus we see that equation 12 (or the one in [20]) provides the value of Z that causes $(\mathrm{X}-\mathrm{Z})^2$ to be uncorrelated with X. Z will equal $\bar{X}$ if the data are symmetrically (or uniformly) distributed as is the case of Mandel's data [19]. but in the general case will not equal $\bar{X}$.

Creating an orthogonal variable using equation 12 provides the advantage that the data in the resulting variable is orthogonal to the original X data regardless of the distribution of the X values. This procedure can, therefore, be applied to a set of real data without concern for the distribution of that data.

It is also possible to set up expressions corresponding to equation 7, for representing terms that are the numerators of the correlation coefficients between X and the third, fourth, and even higher powers of (X−Z):

$$\sum_i (X_i - \bar{X})(X_i - Z_3)^3 = 0 \tag{15A}$$

$$\sum_i (X_i - \bar{X})(X_i - Z_4)^4 = 0 \tag{15B}$$

where the various $Z_i$ represent the value of Z needed to make the variable for the corresponding power of X uncorrelated with X itself. Solving each of the Eqs. (15A) and (15B) for the corresponding $Z_i$ will provide a value that creates a term for the corresponding power of X that is uncorrelated with X. From Eq. (15A) we obtain the expression;

$$Z_3^2 \sum_i X_i(X_i - \bar{X}) - Z_3 \sum_i X_i^2(X_i - \bar{X}) + \frac{1}{3} \times \sum_i X_i^3(X_i - \bar{X}) = 0 \tag{16A}$$

which is quadratic in $Z_3$ and may be solved by the usual formula, or by an approximation method (also discussed by Arden [18]). Application of this formula to several sets of test data followed by further study of the behavior of polynomials of odd degree (specifically, the relation between X and $X^3$) reveals that for any data that could represent actual validation data, no real roots of the equation exist; the roots of Eq. (16A) are complex (in the sense of being of the mathematical form a+bi, where i represents the square root of −1.

Similarly, Eq. (15B) results in the following expression, which is cubic in $Z_4$, and which can also be solved using either known algebraic methods [21], or approximation methods:

$$Z_4^3 \sum_i X_i(X_i - \bar{X}) - \frac{6}{4} Z_4^2 \sum_i X_i^2(X_i - \bar{X}) + Z_4 \times \sum_i X_i^3(X_i - \bar{X}) - \frac{1}{4} \sum_i X_i^4(X_i - \bar{X}) = 0 \tag{16B}$$

Since Eq. (16B) is cubic in $Z_4$, it is guaranteed to have at least one real root, and linearity testing can proceed. Arguing by induction, we conclude that polynomials of even degree are amenable to this procedure, while polynomials of odd degree are not. We will shortly see, however, that this consideration is moot, even though similar expressions can be generated to correspond to higher powers of X, to create corresponding variables for powers of X that are uncorrelated to X.

While $Z_4$ is not necessarily orthogonal to Z, it is orthogonal to the data (X), and so will all powers of $(X-Z_i)$ be orthogonal to X. Therefore, each one could be tested separately, for as many terms as are needed to make up a polynomial of the desired degree, if this were necessary. Should it become necessary to evaluate non-linearity terms that are represented by higher powers of $X_i$ they need not be evaluated simultaneously, each variable: $(X-Z)^2$, $(X-Z_4)^4$, etc. can each be evaluated separately, preventing possible intercorrelations between the $Z_i$ from influencing the results.

Taylor's theorem tells us that, while any function can be approximated by a polynomial the terms of a Taylor expansion results in coefficients of the polynomials that necessarily decrease for higher powers of the polynomial, due to the presence of n! in the denominator of the Taylor formula (where n represents the power of any given term). Therefore, Taylor's theorem tells us that we will rarely, if ever, have to go beyond the quadratic term, so the issue of orthogonality of terms, as well as the problem of polynomials of odd degree; all become moot. Testing data for quadratic non-linearity will suffice to reveal the presence of any nonlinearity in the data.

At this point we note that equations 7, and indeed the whole derivation leading to them is familiar to us, in a different context. In using spectroscopy to do quantitative analysis, we use an equation for a calibration model similar to equa-

tion 1 to express Beer's Law; one of the representations of the equation involved is [22]:

$$C = b_0 + b_1X_1 + b_2X_2 + \ldots + b_nX_n \quad (17)$$

Equation 17 is commonly used to represent the equations needed for doing quantitative spectroscopic analysis using what is called the MLR algorithm (also sometimes called P-matrix or Inverse Least Squares). The various $X_i$ in equation 17 represent entirely different variables, in spectroscopic analysis they are absorbances at different wavelengths. Nevertheless, starting from equation 17, we can derive the set of equations for calculating the MLR calibration coefficients, in exactly the same way we derived equations 7 from equation 1. This derivation is found in [11] and also in [22]. Comparison of those derivations with equations 1–7 is instructive, they are exactly parallel. Because of this parallelism we can set up the following equivalencies:

Coefficient equivalences

| MLR coefficient | Corresponding polynomial coefficient |
|---|---|
| $b_0$ | $a_0$ |
| $b_1$ | $a_1$ |
| $b_2$ | $a_2$ |
| $b_4$ | $a_4$ |
| etc. | etc. |

Data equivalences

| MLR variable | Corresponding polynomial coefficient |
|---|---|
| $X_1$ | $X$ |
| $X_2$ | $(X-Z)^2$ |
| $X_4$ | $(X-Z_4)^4$ |
| etc. | etc. |

and we see that by using $X_i$ $(X-Z)^2$, $(X-Z_4)^4$, etc. as the MLR variables $X_1$, $X_2$, $X_4$, etc. respectively, we can use the common and well-understood statistical methods (and computer programs) of multiple regression analysis to perform the necessary calculations. A consideration of key importance is that, along with the values of the coefficients of the various powers of X, we can obtain all the common statistical estimates of variances, standard errors, goodness of fit, etc. that these computer programs produce for us, along with the ones specified by the FDA. Of special interest is the fact that many programs compute estimates of the standard errors of the coefficients, as described by Draper and Smith (see, for example, page 129 in [11]). This allows testing the statistical significance of each of the coefficients, which, as we recall, are now the coefficients of the various powers of X that comprise the polynomial we are fitting to the data.

This is the basis of the new test for non-linearity. We need not use polynomials of high degree since our goal is not necessarily to fit the data as well as possible. Especially since we expect that well-behaved methods of chemical analysis will produce results that are already close to being linearly related to the analyte concentrations, we expect non-linear terms to decrease as the power of X increases. Thus we need only test the fit of a quadratic equation to the data to test for linearity, although there is nothing to stop us from testing equations of higher degree if we choose. Data well-described by a linear equation will produce a set of coefficients with a statistically-significant value for the term $X^1$ (which is X, of course) and small, statistically non-significant values for the coefficients of the variables representing $X^2$ or higher powers of X.

One "recipe" for performing the test is, therefore, as follows:

1) Ascertain the actual concentration (Y) of the analyte and measure the test result (X).
2) Compute Z from the test results according to equation 12.
3) Compute the new variable $(X-Z)^2$ from each value of X.
4) Regress X and $(X-Z)^2$ against Y, using an MLR program that computes the desired

statistics (it is required that the t-value for the coefficients is included among these statistics).

5) Inspect the t statistic of the coefficients of X and $(X-Z)^2$, to determine if the linear term is statistically significant and whether the t value for the coefficient of $(X-Z)^2$ indicates statistical significance; if so, that indicates that statistically significant nonlinearity exists.

This test procedure has several advantages: it gives an objective, unambiguous determination, based on standard statistical methodology, of whether any non-linearity is present in the relationship between the test results and analyte concentration. Since it is based on regression analysis, it is a straightforward extension of the method currently specified by the FDA. It provides a means of distinguishing between different types of non-linearity (i.e. the need for polynomials of various degrees, if necessary in different situations), if they are present, since only those that have statistically-significant coefficients are active. It is also more sensitive than the DW statistic as well as being immune to the "fatal flaw" that afflicts DW. Because of the extreme variability of DW for small numbers of samples, the tables in Draper and Smith for the thresholds of the DW statistic only give the values for more than ten samples. Since this new method of linearity testing depends on calculating the t value rather than comparing variances, it is applicable to data from smaller numbers of samples.

## 3. Experimental

Two groups of workers independently devised nearly identical measurement protocols to validate analytical methods for similar sample types; the details are described in [13]. Briefly, FOSS/NIR-Systems model 6500 NIR spectrometers, each fitted with Intact™ tablet analyzer modules were used to collect transmittance spectra of the samples. One group was measuring tablets, the other group was measuring capsules, and the main difference between the experimental setups was that each group used a sample mask specific to their samples. The data spectra used were each the result of averaging together 32 scans, the default value for the instrument.

Except for a few very rare situations, NIR analysis requires that the instrument and calibration algorithm be "trained" using actual samples of the same type that are to be measured. The vast majority of attempts to calibrate the instruments through the use of standard samples have failed, resulting in inaccurate answers when applied to "real" samples. For this reason, many books exist that discuss the principles of NIR analysis and how to implement proper calibration methodology [23–27]. The requirement to use actual samples for the calibration exercise means that the calibration process is considerably more extensive than is needed for other technologies. The reward for going through that exercise is an analytical method that is rapid (usually $< 1$ min), requires no chemicals (with the concomitant advantage of not having to dispose of them) and, being computer-controlled, is easily interfaced to both process control and corporate data-management computers. For these reasons, NIR is likely to play a major role in the FDA's current PAT initiative. It is not feasible to reproduce in this article the extensive discussions found in the books; all of the listed books are recommended to the reader interested in further study. Since analysis using NIR spectroscopy requires that the samples have their concentrations measured using a method of known accuracy to provide reference values, the reference values for the samples used in this study were measured using already-validated HPLC methods appropriate for each sample type.

All calculations were performed using programs written in MATLAB™.

## 4. Results and discussion

Details about the samples used are presented in [13]. For the tablet study 96 samples were used for the calibration, and 42 samples were used as an independent test set. Plots of the NIR (test method) versus HPLC (reference method) values is presented in Fig. 2.

Similarly, for the capsule study 70 samples were used for the calibration, and 210 samples were

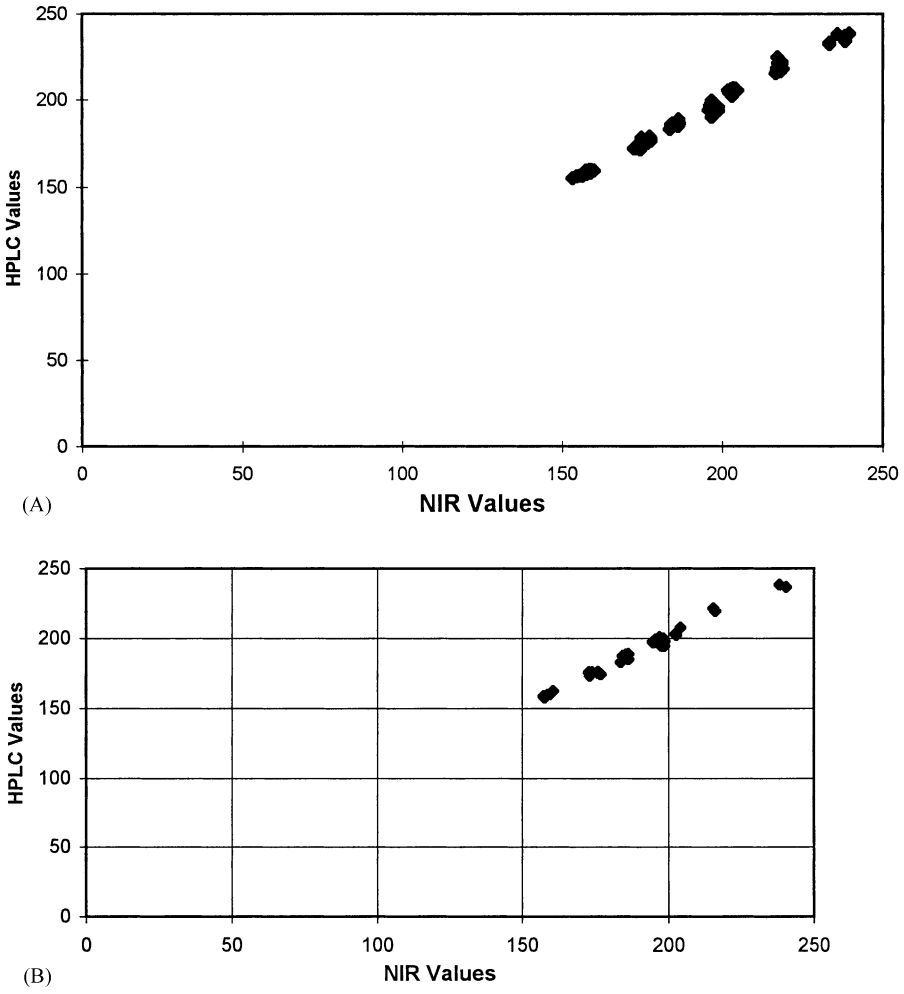


Fig. 2. Plots of the NIR vs. the HPLC values for the tablet product. (A) calibration data. (B) test data.

used as an independent test set. Plots of the NIR (test method) versus HPLC (reference method) values is presented in Fig. 3.

To the eye, all sets of data appear satisfactorily linear. The plot of the test data for the capsule product reveals, not surprisingly for process samples, that the range of the values for the test samples is extremely limited. As we will see, this affects the statistics that are computed for this data, especially the correlation coefficient. In the previous study [13] the DW statistic also was computed to assess the linearity of these data, the conclusion of that test also was that these data show no evidence of non-linearity.

Tables 1 and 2 present the results of applying the new linearity test to the tablet product and the capsule product, respectively. The test was applied separately to the calibration data, and to the test data for each product. For comparison purposes, a straight line, as recommended by the current guidelines, was also fitted to each data set.

From Table 1 of [13] we find that there were 96 samples in the calibration set for tablets, 70 samples in the calibration set for capsules, 42 samples in the validation set for tablets and 210 samples in the validation set for capsules. The corresponding critical values for the t statistic with those numbers of samples, for a two-tailed test at

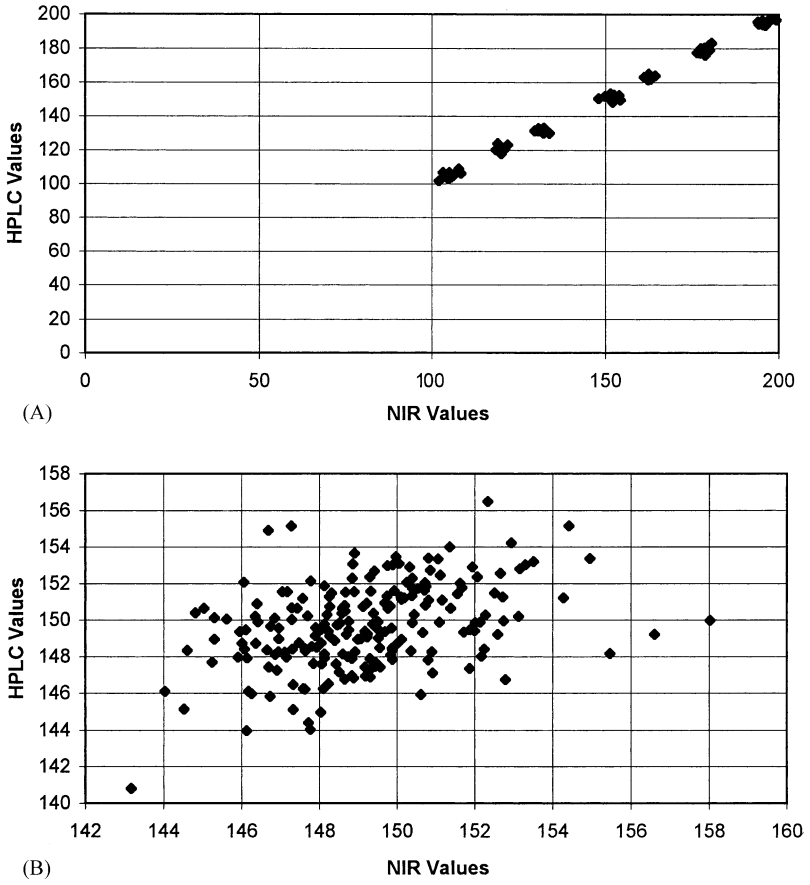


Fig. 3. Plots of the NIR vs. the HPLC values for the capsule product. (A) Calibration data. (B) Test data.

99% confidence, are, from [28]: 2.6291, 2.6501, 2.6981 and 2.6006, respectively. Comparing the t-values in Tables 1 and 2 to these critical values, we find that the t-values for the linear term of the regression is statistically significant in all cases, and except for the test set for the capsule product, where the limited range affected the results, they were all highly significant. The low value of correlation coefficient for the test set from the capsule product indicates that the limited range is the cause of the low values for all the statistics. Use of the t-value for evaluating the linear term is superior to the use of the correlation coefficient (specified by the guidelines), since tables of critical values of t are more common and easier to evaluate than are tables of critical values for the correlation coefficient. Furthermore, having a known statistical value for testing the significance of the linear term provides an objective test for whether there is indeed sufficient data for making the evaluation; from Fig. 3C alone it is not at all clear whether this is the case, due to the limited range of the data.

Similarly, we find that the quadratic terms are non-significant, consistent with and confirming the previous results, but through the use of a test statistic that is more specific, more easily interpreted and in more common use. Having two coefficients with their corresponding t-values separates the linear from the non-linear contributions to the relationship, and yet as a multivariate method allows both pieces of the relationship to be tested separately but simultaneously. In this case there is no reason to suspect higher-order non-

Table 1
The results of testing the linearity of the data from the tablet product

| Parameter | Coefficient when using only linear term | t-value when using only linear term | Coefficient including quadratic term | t-value including quadratic term |
|---|---|---|---|---|
| *Results for calibration data* | | | | |
| Constant | 0.000 | | −0.3376 | |
| Linear term | 1.0000 | 85.62 | 1.0000 | 86.4 |
| Square term | – | – | 0.0007 | 1.67 |
| S.E.E.[a] | 2.42 | | 2.39 | |
| R[b] | 0.9937 | | 0.9938 | |
| *Results for test data* | | | | |
| Constant | 2.37 | | 2.53 | |
| Linear term | 0.9917 | 52.3 | 0.9917 | 51.92 |
| Square term | – | – | −0.0004 | −0.693 |
| S.E.E.[a] | 2.24 | | 2.26 | |
| R[b] | 0.9928 | | 0.9928 | |

[a] S.E.E., standard error of estimate.
[b] R, correlation coefficient.

linearity, but if there were, these could also be tested by including the variables corresponding to the higher-degree polynomials, as indicated by the expressions presented in Eqs. (16A) and (16B), or their obvious extensions.

## 5. Conclusions

The new test of linearity can provide an objective, unambiguous decision tool as to whether a given data set exhibits non-linearity in the relationship between the test results and the analyte concentration. It also provides all the statistical results that the current FDA/ICH test procedure recommends, in a context that makes those statistics more meaningful. Through the computation of auxiliary diagnostic statistics, such as the Standard Error of Estimate (S.E.E.) and the Correlation Coefficient, it also provides information as to whether, and how well, an analytical method gives a good fit of the test

Table 2
The results of testing the linearity of the data from the capsule product

| Parameter | Coefficient when using only linear term | t-value when using only linear term | Coefficient including quadratic term | t-value when including quadratic term |
|---|---|---|---|---|
| *Results for calibration data* | | | | |
| Constant | −0.0022 | | 0.0434 | |
| Linear term | 1.0000 | 129.7 | 1.0000 | 128.8 |
| Square term | – | – | −0.0001 | −0.175 |
| S.E.E.[a] | 1.90 | | 1.92 | |
| R[b] | 0.9980 | | 0.9980 | |
| *Results for test data* | | | | |
| Constant | 90.19 | | 90.39 | |
| Linear term | 0.3986 | 6.26 | 0.3988 | 6.3213 |
| Square term | – | – | −0.0359 | −2.14 |
| S.E.E[a] | 2.11 | | 2.09 | |
| R[b] | 0.3986 | | 0.4209 | |

[a] S.E.E., standard error of estimate.
[b] R, correlation coefficient.

results to the actual concentration values. It can distinguish between different types of non-linearities, if necessary, while simultaneously evaluating the overall goodness of the fitting function.

In application to the current set of test data, we find that the results of using an NIR analytical method provides a linear relationship between the test result (from the NIR method) and the actual concentration of the analyte (as measured by the validated reference HPLC method).

In Section 2, we discuss the fact that polynomials may be used as surrogates for "any mathematical function" since by virtue of Taylor's theorem, any mathematical function can be approximated to any desired degree of accuracy by including sufficiently many terms of the Taylor expansion of the function. FDA guidelines recommend that "... the simplest model that adequately describes the relationship..." should be used. This implies that the polynomial of lowest degree that can be used for the given purpose should be used. At the other end of the scale, it is well-known that if n data pairs are available, they can always be fit exactly by a polynomial of degree $(n-1)$, but such fits are virtually always spurious. Thus a polynomial of intermediate degree must be selected for the test. Arguably, there is room for subjectivity in the selection of the degree of the polynomial. The lowest-degree polynomial that can be used for testing linearity is a quadratic. In the absence of evidence to the contrary, and in the light of the FDA guideline, therefore, this is what should be used, and so I recommend.

Since the fitting of the polynomial is the surrogate for fitting "any other function", and the use of the t-test from a multivariate least-squares calculation is a means of assessing the contribution of a term of degree greater than unity to the fitting function over that of the straight line alone, therefore, this method of testing linearity conforms to the definition of "Linearity" as stated in Section 1.

## References

[1] ICH-Q2A, "Guideline for Industry: ICH-Q2A, Text on Validation of Analytical Procedures", Food and Drug Administration, Rockville, Maryland, March 1, 1995.
[2] ICH-Q2B; "Guidance for Industry: ICH-Q2B, Validation of Analytical Procedures: Methodology", Food and Drug Administration, Rockville, Maryland, May 19, 1997.
[3] J.K. Taylor, Quality Assurance of Chemical Measurements, Lewis Publishers, Chelsea, MI, 1987.
[4] P.C. Meier, R.E. Zund, Statistical Methods in Analytical Chemistry, Wiley, New York, 1993.
[5] F.J. Anscombe, Am. Stat. 27 (1973) 17–21.
[6] J. Durbin, G.S. Watson, Biometrika 37 (1950) 409–428.
[7] J. Durbin, G.S. Watson, Biometrika 38 (1951) 159–178.
[8] J. Durbin, Biometrika 56 (1989) 1–15.
[9] J. Durbin, Econometrica 38 (1970) 422–429.
[10] J. Durbin, G.S. Watson, Biometrika 58 (1971) 1–19.
[11] N. Draper, H. Smith, Applied Regression Analysis, third ed, Wiley, New York, 1998.
[12] G.E. Ritchie, R.W. Roller, E.W. Ciurczak, H. Mark, C. Tso, S.A. MacDonald, J. Pharm. Biomed. Anal. 28 (2) (2002) 251–260.
[13] G.E. Ritchie, R.W. Roller, E.W. Ciurczak, H. Mark, C. Tso, S.A. MacDonald, J. Pharm. Biomed. Anal. 29 (1–2) (2002) 159–171.
[14] A. Hald, Statistical Theory with Engineering Applications, Wiley, New York, 1952.
[15] W.J. Dixon, F.J. Massey, Jr., in: Introduction to Statistical Analysis, fourth ed., McGraw-Hill, New York, 1983.
[16] H. Mark, NIR News 13 (3) (2002) 12–14.
[17] A. Savitzky, M.J.E. Golay, Anal. Chem. 36 (8) (1964) 1627–1639.
[18] B.W. Arden, An Introduction to Digital Computing, first ed, Addison-Wesley, Reading, MA, 1963.
[19] J. Mandel, Evaluation and Control of Measurements, Marcel Dekker, New York, 1991.
[20] C. Daniel, F. Wood, Fitting Equations to Data—Computer Analysis of Multifactor Data for Scientists and Engineers, first ed, Wiley, 1971.
[21] W.L. Hart, Brief College Algebra, D.C. Heath & Co, Boston, 1947.
[22] H. Mark, Principles and Practice of Spectroscopic Calibration, Wiley, New York, 1991.
[23] D. Burns, E. Ciurczak, Handbook of near-infrared analysis, in: Practical Spectroscopy Series, second ed, Marcel Dekker, New York, 2001, p. 814.
[24] P. Williams, K. Norris, Near-Infrared Technology in the Agricultural and Food Sciences, American Association of Cereal Chemists, St. Paul, MN, 2001.
[25] B. Osborne, T. Fearn, P.H. Hindle, Practical NIR Spectroscopy, second ed, Longman Scientific and Technical, Essex, UK, 1993.
[26] R. Raghavachan, Near-infrared applications in biotechnology, in: Practical Spectroscopy, Marcel Dekker, New York, 2001, p. 382.
[27] H.W. Siesler, Y. Ozaki, S. Kawata, H.M. Heise, Near Infrared Spectroscopy: Principles, Instruments, Applications, first ed, Wiley-VCH, Weinheim, Germany, 2002.
[28] D.B. Owen, Handbook of Statistical Tables, Addison-Wesley, Reading, MA, 1962.