ELSEVIER

Journal of Chromatography A, 850 (1999) 43–51

JOURNAL OF CHROMATOGRAPHY A

Lack-of-fit testing of ion chromatographic calibration curves with inexact replicates

D.E. Coleman[a], L.E. Vanatta[b,*]

[a]*Alcoa Technical Center, AMCT-D-10, 100 Technical Drive, Alcoa Center, PA 15069, USA*
[b]*Air Liquide Electronics Chemicals and Services, Box 650311, M/S 301, Dallas, TX 75265, USA*

Abstract

Calibration studies involve the preparation and analysis of replicates for multiple concentrations of standards. Curves that are fitted through the data are evaluated for their adequacy of fit. A helpful test is a lack-of-fit procedure, which is performed easily by most statistical software. When coupled with R^2_{adj}, the procedure differentiates between data that are not linear and those that are simply noisy. The test requires data from exact replicates of the various standard levels involved. However, in ppt-level ion chromatography, the above condition may be impossible to meet. With the common anions (e.g., chloride, nitrite), the working standards must be prepared by mass and all liquids must be poured; transfer pipets contaminate at these concentrations. Since it is virtually impossible to pour out the desired mass exactly, final concentrations will vary slightly. Consequently, a different approach is needed for lack-of-fit testing. This paper discusses reasonable alternatives and applies them to actual data.

Keywords: Lack-of-fit testing; Calibration; Statistical analysis; Water Analysis; Inorganic anions

1. Introduction

In many fields, trace-level analytical methods are becoming more commonplace. The semiconductor industry is a primary example, where detection limits in the low-ppt range are needed. At these concentrations, contamination of samples and standards becomes a vexing problem.

One analysis that is crucial is the determination of common anions in ultrapure water, the most-used chemical in the manufacture of semiconductors. Because even slight amounts of contaminants (e.g., chloride) can cause device failures, the specifications for these species are 50 to 100 ppt (w/w) [1]. At these levels, preparation of standards must be done with extreme care. These solutions are made by weighing the various constituents. Usually, plastic transfer pipets are utilized to deliver the desired amounts of stock standard and of diluting water. Using this protocol, it is generally possible to obtain target concentrations with high precision. However, all pipets tested to date (by one author, L.E.V.) have severely contaminated the solutions at the ppt level. Consequently, these standards are made strictly by pouring from the source container into the destination vessel.

Under these ''no-pipet'' conditions, target concentrations cannot be achieved precisely. An amount close to the target is poured from the stock standard and the mass recorded to four decimal places. Then, the amount of diluting water needed to obtain the desired final concentration is calculated. That number of grams is poured in as precisely as possible and

*Corresponding author. Tel.: +1-972-995-7541.
E-mail address: lynn.vanatta@airliquide.com (L.E. Vanatta)

0021-9673/99/$ – see front matter
PII: S0021-9673(99)00369-6

the mass recorded, again to four decimal places. The actual concentration can be determined quite accurately from these measurements. However, the variation in concentration suffers among replicates of the same level, since pipets cannot be used.

These actual concentrations, along with the corresponding peak areas (PAs) from the instrument, are used to construct a calibration curve. There is no statistical difficulty in performing the regression analysis on such data; the results are just as valid as those from situations where target concentrations are achieved precisely. A problem arises, though, when the lack-of-fit (LOF) test is performed on the curve. The LOF test requires the existence of exact replicates at each concentration; any LOF result obtained using inexact multiples may be in error. This test is an important one to conduct when assessing the adequacy of a given calibration-curve model. A low *p*-value (e.g., less than 0.05) for the LOF test indicates that at least one term is missing from the regression equation and should be found, if possible (typically, peak area is roughly proportional to concentration and a straight-line regression is used).

Researchers have examined a related, but different, difficulty, the ''errors in *X*'' problem (also known as the ''errors in variables'' problem), where there is considerable uncertainty (either absolute or relative) in the values of the independent variable(s) [2]. In this situation, one may or may not hit target, but there is uncertainty in the actual concentration. In contrast, the context of this paper is that the actual concentrations are accurately known, but they usually vary about target. Consequently, the methodologies (such as model-II regression) developed to solve this related, errors-in-*X* problem do not, unfortunately, apply here.

This paper addresses the LOF-test problem for ppt-level ion chromatographic (IC) data from seven common anions: F, Cl, NO_2, Br, NO_3, SO_4 and PO_4. Eight inexact replicates of each of nine concentrations (blank, 25, 37.5, 50, 62.5, 75, 100, 150 and 200 ppt) were prepared by pouring and then chromatographed. The resulting data were first discussed in a previous paper [3], where the objective was to compare and apply two detection-limit techniques to the IC results. In that work, the actual peak areas were regressed against the actual concentrations, using a straight-line model and the fitting technique of ordinary least squares (OLS). Only the Traditional LOF test was used to evaluate the calibration curves. In retrospect, the results of that procedure were misleading in the original publication. This present paper looks not only at that test, but also at alternative (and possibly more powerful) procedures. Of interest is the ability of each protocol to detect any inadequacy of the straight-line models, given the absence of exact replicates (except for the blanks).

2. Experimental

The experimental protocol was the same as that described in a previous work [3] and is not repeated here.

3. Results and discussion

3.1. Lack-of-fit strategies

Five alternative LOF strategies were devised for this study. Each was designed to address the issue of inexact replicates. The five tests, and the Traditional lack-of-fit procedure, are explained below. In addition, the advantages and disadvantages of each approach are discussed, and are displayed in Table 1.

3.1.1. Traditional lack-of-fit strategy (PA vs. Actual)

The LOF test can provide an indication that a calibration model is inadequate. For example, a straight-line model might have been used when, in actuality, the unknown underlying relationship was, e.g., quadratic, exponential or piecewise-linear. An advantage of the Traditional LOF test is that it provides an objective way to obtain evidence of the model's inadequacy – without forcing the analyst to try a variety of alternative models, and pick and choose among them. A disadvantage is that the LOF test does not provide guidance on how to address any lack of fit that is detected; no superior model is suggested, so one must appeal to first principles, study a plot of the residuals vs. concentration, or resort to trial and error.

Table 1
Comparison of lack-of-fit strategies

Strategy	Advantages	Disadvantages
PA vs. Actual (Traditional)	Is simple	May generate F-test p-values that are too low
	Requires no data adjustments	
PA vs. Target	Is simple	Ignores possibly influential bias and variation in x_{act}[a]
	Requires no data adjustments	
PA vs. Average	Removes bias in x_{act}	Ignores possibly influential variation in x_{act}[a]
Scaled PA vs. Average	Removes bias in x_{act}	May add variation to response[a]
Quadratic Term	Involves no data adjustments	May miss non-quadratic true relationships
	Can detect lack of fit if x^2 term is significant	
	May suggest a better model	
ANOVA-on-Residuals	Can detect lack of fit without having to propose an alternative model	May ignore possibly influential variation in x_{act}[a]
		Does not suggest a better model

[a] Could distort the p-value, making it too high or too low.
Note: x_{act}=actual concentration.

The traditional LOF test is based on the following logical argument: (1) Exact replicates provide the opportunity to estimate ''pure error'' (i.e., what should be expected as random experimental variation), quantified by the pure-error standard deviation, s_e. This estimate provides the yardstick by which one can judge unexplained variation. (2) After a straight-line fit (or any fit), root mean squared error (RMSE) provides an estimate of the ''left-over'' or unexplained variation. (3) If RMSE$>>s_e$, there is evidence that the unexplained variation is in excess of what could reasonably be expected to be due to random variation. (4) Since it is not random, this excess variation must be due to an underlying, systematic deviation from the fit, implying that at least one term is missing from the model.

The formal LOF test is conducted by constructing a LOF ANOVA (analysis of variance) table, shown below for the PO_4 data in this study:

Source	Lack of fit			
	DFs	Sum of squares	Mean square	F-ratio
Lack of fit	7	321 371.8	45 910.3	0.4734
Pure error	63	6 109 537.6	96 976.8	
Total error	70	6 430 909.4		p-value= 0.8503

For purposes of discussion, temporarily suppose that the target concentrations were exactly achieved. There were eight replicates at each of nine concentrations, for a total of 8·9=72 observations, hence 72 degrees of freedom (DFs). A straight-line calibration uses two DFs (leaving 70) and gives a sum of squares (SS) of residuals equal to 6 430 909.4. The pure-error sum of squares is the sum of the squared deviations about the mean of the blank

replicates, plus the sum of the squared deviation about the mean of the replicates spiked at 25 ppt, plus etc. Each squared deviation contributes one DF to the pure-error SS; the mean for each concentration that has replicates subtracts one DF. The result is $(8 \cdot 9) - 9 = 63$ DFs for pure error, leaving $70 - 63 = 7$ DFs for LOF. The LOF sum of squares is the total SS minus the pure-error SS: $6\,430\,909.4 - 6\,109\,537.6 = 321\,371.8$.

To assess the statistical significance of the LOF sum of squares, the LOF SS must be standardized and compared to a pure-error yardstick. This evaluation is performed by dividing each SS by its DF, to get the mean squared error (MSE): LOF MSE $= 321\,371.8/7 = 45\,910.3$ and pure error MSE $= 6\,109\,537.6/63 = 96\,976.8$.

For the LOF test, the null (or ''starting'') hypothesis is that there is no difference between the LOF MSE and the pure-error MSE. The statistical significance is assessed by taking the ratio: LOF MSE/pure-error MSE $= 45\,910.3/96\,976.8 = 0.4734$. This ratio is then compared to the F distribution, which characterizes what would be expected under the null hypothesis. An F-table lookup of the ratio, for the associated DF values, will provide a p-value: $F^{-1}(0.4734, \text{DFs}=7, 63) = 0.8503$. The p-value represents the likelihood that, under the null hypothesis, a ratio at least this high would be obtained. In this case, the p-value is high and, therefore, not significant; it provides no evidence of lack of fit. A low p-value (e.g., under 1%) is evidence that the straight-line model is not adequate. If a low p-value is obtained, a plot of the residuals vs. concentration may reveal one or more additional terms to be required to fit the data better.

Note that R^2 (adjusted or not) quantifies the proportion of the total variation that is explained by the model, but does not have an associated p-value to judge statistical significance. Also, a non-significant LOF p-value could (not surprisingly) be accompanied by either a good R^2 statistic (i.e., high R^2, near 1.00), or, if the data are noisy, by a poor (i.e., low) R^2 statistic.

The LOF test, then, is an objective way to assess the adequacy of a model, but it can only be computed if there are exact replicates in the study dataset. Replicates provide the yardstick for comparison (i.e., the MSE for pure error). Additionally, replicates at different concentrations are useful for determining if the measurement precision changes with concentration, thereby requiring regression-fitting techniques more advanced than ordinary least squares (e.g., weighted least squares). Technically, LOF testing can be done even if there are only blank replicates. However, it is dangerous to assume that the variation observed when measuring blanks will be the same encountered when measuring the non-zero standards (a plot of residuals could provide some basis for testing that assumption). Typically, if variation changes, it will increase with concentration. Thus, having only blank replicates will likely result in false detections of lack of fit, since the pure-error MSE will be too small. Such a number in the F-test's denominator will lead to an inflated ratio and perhaps a deceptively low p-value.

3.1.2. PA vs. Target strategy

A simple, but naive, strategy is to ignore the fact that the target concentration was not obtained in the study samples, and simply to use the target concentration in the LOF analysis. Clearly, the danger of obtaining a misleading LOF p-value depends on how much the actual concentrations deviate from the target, in both mean deviation and variation about that mean. At most, mean absolute deviation should be no more than a few percent of the target concentration. (See Table 2).

3.1.3. PA vs. Average strategy

An improved, but simple, strategy is to replace each target concentration with the mean of the actual concentrations obtained. This approach has an advantage over the PA vs. Target protocol; the former eliminates the effect of systematic bias between actual and target. However, the problem of any excessive variation in the actual concentrations remains. (See Table 2, and compare results at targets 2 and 4 ppt for the two strategies.)

3.1.4. Scaled PA vs. Average strategy

A more sophisticated approach that builds upon the Average strategy involves adjusting the peak-area results. For many analytes and matrices, PA might be assumed to be roughly proportional to concentration. In other words, a straight-line model, with intercept of zero, is used to relate PA to concentration. Thus,

Table 2
Simulated data to illustrate three strategies (PA vs. Target, PA vs. Average and Scaled PA vs. Average) for determining lack of fit[a]

Target concentrations, X_T	1.00	2.00	4.00
Actual concentrations, X_A	0.99, 1.02, 1.00, 1.01	1.70, 2.10, 2.30, 1.80	4.50, 4.70, 4.60, 4.60
$\bar{X}_A$	1.005	1.975	4.600
$\overline{\lvert X_T - X_A \rvert}$	0.01	0.23	0.60
(% of X_T)	(1%)	(11%)	(15%)
$\overline{\lvert \bar{X}_A - X_A \rvert}$	0.010	0.225	0.050
(% of X_T)	(1%)	(11%)	(1%)
Actual peak area, Y_A	220, 178, 178, 177	280, 301, 290, 293	589, 613, 548, 608
Scaled PA, Y_S	223, 175, 179, 176	326, 283, 249, 321	602, 600, 548, 608
	Adequacy of strategies at each concentration		
PA vs. Target strategy ($x = X_T$, $y = Y_A$)	(OK)	$\overline{\lvert X_T - X_A \rvert}$ large (excess variation)	$\overline{\lvert X_T - X_A \rvert}$ large (biased mean)
PA vs. Average strategy ($X = \bar{X}_A$, $y = Y_A$)	(OK)	$\overline{\lvert \bar{X}_A - X_A \rvert}$ large (excess variation)	(OK)
Scaled PA vs. Average strategy ($X = \bar{X}_a$, $y = Y_S$)		Scaling may add variation	

[a] All concentrations in ppt units; all peak areas in PA units. See text for discussion. Note: $\bar{X}_A$ = mean of actual concentrations; $\lvert X_T - X_A \rvert$ = mean absolute deviation of $(X_T - X_A)$; $\lvert \bar{X}_A - X_A \rvert$ = mean absolute deviation of $(\bar{X}_A - X_A)$.

independent of the slope of such a model (provided that slope>0), a first-order multiplicative adjustment to each PA value is (average/actual). In other words, PA is replaced with: PA·(average/actual). Actual concentrations that are below average will have their PA values scaled higher, and those that are above average will be scaled lower (see Table 2). The advantage of this approach is that it reduces bias, but at the expense of possibly increasing variation, since the adjustment is approximate.

3.1.5. Quadratic Term strategy

Another strategy for LOF does not involve adjusting analytical data, but instead tries a more complex calibration model. One then assesses the statistical significance of the higher-order term(s) by examining the *p*-value of the additional coefficient(s). (The null hypothesis: regardless of the other terms, the coefficient of the new term is zero.) This strategy is easy to try, but can be subtle with respect to selecting an additional term. If possible, the term (e.g., quadratic, exponential) should be dictated by the chemistry or physics of the method. Otherwise, selection should be based on any clear systematic pattern in a plot of the residuals (from a straight-line fit) vs. concentration. Typically, if the plot's form is neither known nor discernible, a quadratic is added. This choice loosely depends on the fact that for many functions, their Taylor-series approximations have terms that successively decline in magnitude. Finding a statistically significant higher-order term is evidence of LOF, and, furthermore, may suggest an improved calibration model. Unfortunately, however, the choice of term constrains the power of the strategy. For example, LOF may be missed if a quadratic term is added when the true relationship is a trending sinusoid.

3.1.6. ANOVA-on-Residuals strategy

A final sophisticated strategy for LOF testing exploits the fact that the residuals from the fit should be completely random (assuming that the straight-line calibration model is correct, and that the usual calibration and regression assumptions hold). If the PA–concentration relationship is more complex, one would expect residuals to be at different mean levels for different concentrations. Such a variation in means might give a noticeable pattern (e.g., sinusoidal, parabolic) to the residuals plot, and patterns are a clear indication that something is missing from the

chosen model. The natural strategy to follow, if it appears that the pattern can be easily ''matched'' by adding one or more terms to the model, is to follow the Quadratic Term strategy. Otherwise, the ANOVA-on-Residuals strategy could be considered.

One-way ANOVA is a formal statistical test of equality of means. The procedure can be applied to the residuals, using the target concentrations to define the groups (i.e., target concentration is no longer treated as a continuous variable, but as a categorical variable). Thus, the steps are: (a) using a straight-line model and OLS, fit PA vs. actual; (b) compute residuals from the fit; (c) do one-way ANOVA of residuals vs. target concentration; (d) note the overall *p*-value from the ANOVA. The *p*-value from the ANOVA indicates the weight of evidence for there being non-straight-line behavior; i.e., a low *p*-value indicates that some of the residuals are ''too high'' or ''too low'', or both. (Technically, this *p*-value should be adjusted for the degrees of freedom used in the straight-line fit.) An advantage of this approach is that there is no presumption of (and hence no dependence upon) the nature of the departure from a straight line. This asset is shared by the Traditional LOF test, but is missing in the Quadratic Term strategy. Naturally, the ANOVA procedure also carries the disadvantage of being non-constructive when a low *p*-value is obtained; the test does not suggest a more complex function for the calibration.

3.2. Results from lack-of-fit analysis

3.2.1. Data analysis for day-to-day biases

Prior to analyzing the IC data for LOF, an analysis was done for each anion, to ensure that there was no evidence of outliers (none were found), or day-to-day biases. The latter analysis is similar to the procedure for the ANOVA-on-Residuals strategy: (a) using a straight-line model and OLS, fit PA vs. actual; (b) compute residuals from the fit; (c) do one-way ANOVA of residuals vs. day (as a categorical variable); (d) note the overall *p*-value from the ANOVA; (e) if the *p*-value is significant (i.e., less than 0.05), identify days on which measurements were unusually low or high (graphically, or by employing the formal Tukey–Kramer test); (f) eliminate from further analysis all measurements from identified days; (g) repeat steps (a) and (b), using only the remaining data. (See Table 3 for the number of days and number of measurements remaining.)

3.2.2. LOF results

The various LOF strategies were applied to the remaining anion data. Overall results are explained below; *p*-values are given in Table 3.

Residual plots (which are among the most informative plots in calibration analysis) show little evidence of LOF for all anions except F and SO_4 (see Fig. 1). There appears to be slight convex curvature in the F residuals. An alternative explana-

Table 3
The *p*-values (in %) from LOF strategies; biased data have been removed[a]

Anion	PA vs. Actual (%)	PA vs. Target (%)	PA vs. Average (%)	Scaled PA vs. Average (%)	Quadratic Term (%)	ANOVA-on-Residuals (%)	Days	*n*	*p* for Δ prec. (%)
F	15.8	8.4	8.7	11.1	**1.3**	15.6[b]	5	45	**1.1**
Cl	7.0	88.1	88.3	86.6	92.1	92.0	7	63	**1.7**
NO_2	82.7	62.2	62.2	63.0	82.7	72.0	5	45	55
Br	**0.01**	72.2	88.3	80.0	16.9	74.0	8	71[c]	**0.5**
NO_3	49.9	64.2	63.6	62.7	6.9	72.0	5	45	62
SO_4	NA; no blank response	**5.0**	**4.9**	**4.7**	**4.1**	7.2	6	54	**0.3**
PO_4	NA; no blank response	85.0	84.0	85.0	73.4	90.8	8	72	**2.7**

[a] Note that the PA vs. Actual LOF test is suspect because precision may change with concentration. This suspicion is indicated by low *p*-values in the ''*p* for Δ prec.'' column, which is from straight-line fits of sample SD of PA (grouped by average concentration) vs. average concentration (low *p*-values indicate statistically significant slope). Also, all *p*-values ≤5% are emboldened.

[b] A *p*-value of 5.9% was obtained using a Welch's test (which may be applicable) for comparison of means for groups with non-constant variance.

[c] One missing value. All *p*-values ≤5% are emboldened.

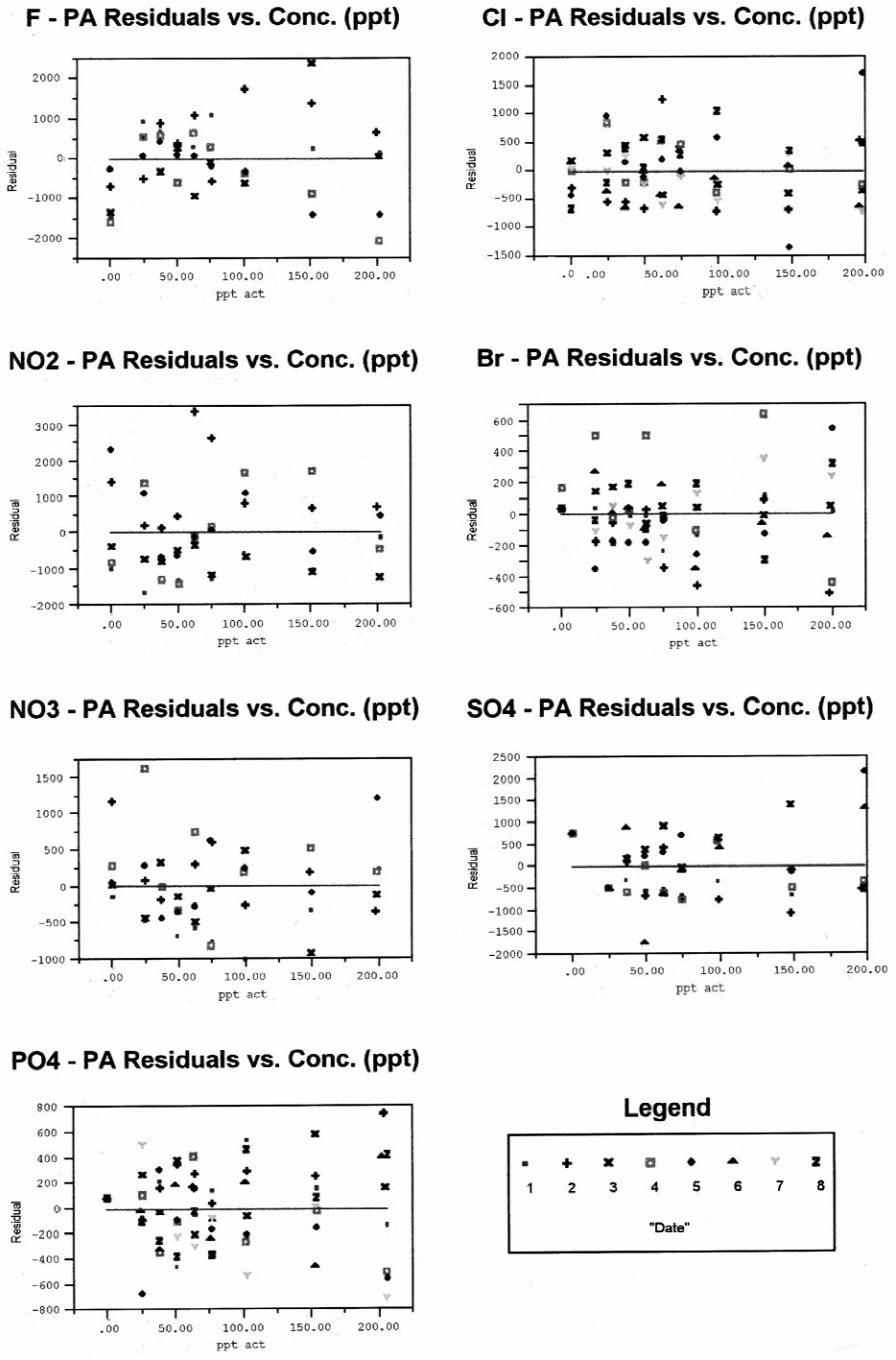

Fig. 1. Plots of residuals from straight-line fits: PA (in PA units) vs. actual concentration. Type of symbols indicates day of experiment; all data on days with unusually low or high mean PA values have been eliminated.

tion for fluoride's pattern is that the blanks simply measured ''too low'' for a straight-line fit.

The qualitative impressions from the residual plots are largely born out by the results of the approximate LOF tests. (Items A–C disregard the suspect ''PA vs. Actual'' column): (A) Five of the seven anions (i.e., all but F and SO_4) have no statistically significant *p*-values. Fluoride appears to show curvature in the residual plots (admittedly, this evaluation of the residuals graphs is subjective). (B) The remaining two anions (F and SO_4) have at least one statistically significant *p*-value. (C) For SO_4, the non-ANOVA tests show excellent agreement in *p*-values. For F, only the Quadratic Term test is statistically significant.

For the two anions with curvature (F and SO_4), the quadratic term test has the lowest *p*-value and also shows excellent power to detect curvature. Recall its two advantages and one disadvantage: (A) Involves no approximation resulting from data adjustment and no expenditure of degrees of freedom for group. (B) Is constructive; may suggest a better model. (C) May miss LOF if a quadratic fit is not the true relationship.

There may still be some unresolved day-to-day bias issues with the data. In the F and NO_2 residuals plots, tracking the various symbols still reveals tendencies for certain days to read high, and certain days to read low.

4. Conclusions

From an LOF perspective, it is always preferable to have exact replicates. Having exact replicates also facilitates precision modeling, which involves developing an empirical model (e.g., straight-line) that relates measurement standard deviation to concentration. The remainder of this section assumes that exact replicates are not available (due, perhaps, to ''dilution by pouring'').

Two overall conclusions emerged from this study. First, if the residuals suggest a quadratic term (or some other simple term), the Quadratic Term strategy should be followed, using the *p*-value of that term to judge lack of fit. Second, if another term is not warranted, the PA vs. Average strategy should be implemented. Other conclusions are below.

The Traditional (PA vs. Actual) LOF test for a straight line should not be relied on if there are many inexact replicates, since there is no guarantee of the representativeness of the estimated pure (i.e., experimental) error from whatever exact replicates are available (possibly only blank samples). The estimated pure-error standard deviation is the key ''yardstick'' for assessing the magnitude of deviations (i.e., lack of fit) from the model.

Thus, the best first step in an LOF strategy involving a potentially straight-line calibration is to fit the line, compute the residuals, plot the residuals vs. the actual concentration, and look for the following: (1) outliers; (2) bias or variation that depends on any other factor, such as day; (3) variation that changes with concentration; (4) mean level that changes with concentration.

Discussion of the first three items is beyond the scope of this paper. However, the existence of issue (4) may reveal a systematic residual pattern that suggests an additional model term, such as quadratic or exponential. If such a term is suggested, the Quadratic Term strategy should be followed, using that term (quadratic or not). The term should be added to the straight-line model, regression should be done, and the *p*-value for the term's estimated coefficient should be used as the LOF *p*-value. A small *p*-value is evidence of lack of fit; the term then should be added to the calibration model, and the residual analysis described above should once again be carried out.

If the residuals appear to have no systematic pattern, or one that is too complex to be modeled by a simple additional term (excluding, e.g., sinusoidal, piecewise, or clearly non-linear functions), one of the other LOF strategies should be followed. For each target concentration, if the absolute and relative amounts of variation about the average are not too great, the PA vs. Average strategy is recommended. It: (1) has the benefits of the PA vs. Target approach, but also eliminates the effects of bias in actual concentration, (2) avoids the risk of adding variation (unlike the Scaled PA vs. Average technique) and (3) generally will have greater power to detect lack of fit than the ANOVA-on-Residuals protocol.

These strategies, then, provide the chromatographer with alternate LOF tests when exact replicates are not possible. As the desired concentrations for

standards becomes lower and lower, ''dilution by pouring'' will continue to be required in order to avoid contamination. Consequently, LOF testing of calibration curves will have to rely more and more heavily on non-traditional techniques.

5. Symbols and abbreviations

5.1. Mathematical symbols used

MSE: mean squared error.
R^2_{adj}: R^2, ''penalized'' for each independent variable used in the regression. (R^2 measures the amount of total variation in the response ''explained'' by the dependent variable.)
RMSE: root mean square error (often used for sample standard deviation).
SS: sum of squares.

5.2. Terms used

ANOVA: analysis of variance.
Degrees of freedom (DFs): the number of observations in a study minus the number of parameters estimated using those observations.
F-test: a test that compares the ratio of two chi-squared-distributed statistics (with known degrees of freedom) to the *F*-distribution with the same DF values.
Lack-of-fit sum of squares: total-error sum of squares minus pure-error sum of squares.
Lack-of-fit (LOF) test: a test of the statistical significance of the residual variation that is above and beyond that attributable to pure error.
Mean absolute deviation: the mean of the absolute value of the quantity ''true minus predicted''.
Mean squared error: sum of squares divided by its degrees of freedom.
Null hypothesis: the ''starting assertion'' that is assumed true unless evidence casts sufficient doubt upon it.
OLS: ordinary least squares. A fitting technique that minimizes the sum of squares of the residuals.
p-value: the probability value associated with a statistical test, representing the likelihood that a test statistic would assume or exceed a certain value, if the null hypothesis is true.
Pure error (also called experimental error): unexplained variation that occurs when experimental conditions are replicated and repeated experimental runs are performed.
Pure-error sum of squares: sum of the squares of the quantities: peak areas at a concentration minus the mean response at that concentration.
Residual: the actual (measured) value minus the predicted value.
Total-error sum of squares: sum of the squares of all the peak-area residuals.
Tukey–Kramer test: a formal test for multiple comparison of means. Controls the experiment-wise false-positive rate.
Welch's test: a weighted analysis of variance test (analogous to WLS) for equality of means of different groups. The test is appropriate when there are statistically significant differences in the standard deviations of observations from the different groups.
WLS: weighted least squares. Same as OLS, except weights are added (typically to account for non-constant response variation).

References

[1] Specification for Deionized Water, Texas Instruments.
[2] N. Draper, J. Smith, in: Applied Regression Analysis, 2nd ed, Wiley, New York, 1981, p. 678.
[3] L.E. Vanatta, D.E. Coleman, J. Chromatogr. A 770 (1997) 105.