

# Structured heteroscedasticity in errors-in-variables calibration

Research note · version 1.0 · October 2026

**Author:** Erik van der Werff, Qualimetric, Mill, The Netherlands

**About this document.** This is an independent research note by the author, written as background to the author's work on calibration. It has not been peer reviewed and is not an official publication of a journal, standards body or other organisation. The note presents the method and the design of its evaluation; the simulation study (Section 5) and the application to real data (Section 6) are described as planned work, and where the text states expected rather than demonstrated results, this is said explicitly. Comments and questions are welcome.

**Suggested citation:** Van der Werff, E. (2026). Structured heteroscedasticity in errors-in-variables calibration. Research note, version 1.0. Qualimetric, Mill, The Netherlands.

## Abstract

Errors-in-variables (EIV) calibration requires a standard uncertainty for every response value, and in current practice this uncertainty is taken from the replicate standard deviation at each calibration level. With the three to five replicates that are usual, these per-level variances are so imprecise that the weights derived from them are largely noise, and block-wise measurement order reduces them further by hiding drift in the level means. We argue that the two axes of an analytical calibration have fundamentally different error structures. Reference values carry independent, heterogeneous uncertainties that originate from separate preparation routes and are stated per material. Response values come from a single measurement process whose variance is a smooth function of the signal. Exploiting this asymmetry, we model the response variance by a two-component (additive plus proportional) variance function, estimated from all replicates jointly and used as the weight in the EIV fit, while the reference uncertainties are kept as given. We derive the consequences for estimation, goodness of fit, coefficient testing and inverse prediction, and we discuss when a variance function should not be used, in particular when the signals of components interfere. A Monte Carlo study is designed to quantify bias, the realism of the stated uncertainty and the coverage of the predicted value, as well as model selection, intercept testing and the effect of drift, and an application to a real calibration dataset is outlined.

**Keywords:** errors-in-variables regression; variance function; structured heteroscedasticity; analytical calibration; weighted Deming regression; inverse prediction; measurement uncertainty

## In brief

The main message of this note in six points, without formulas:

- **The problem.** Calibrating an instrument with errors-in-variables (EIV) regression requires an uncertainty for every measured response. In practice this uncertainty is taken from three to five repeat measurements per calibration level — far too few to be reliable (Figure 2).
- **The key idea.** The two axes of a calibration are different. The reference values carry independent uncertainties, stated separately for each calibration material. The responses all come from one instrument, whose scatter changes smoothly with the size of the signal (Figure 1).
- **The proposal.** Describe the scatter of the response by a simple variance function with two parameters, estimated from all repeat measurements together, and use it to weight the EIV fit. The reference uncertainties are kept as given (Figure 4).
- **What it brings.** With reliable weights, the goodness-of-fit test, the choice of the calibration model and the test of the intercept become meaningful, and an uncertainty can be given for any measured value, including values between calibration levels.
- **What it does not fix.** Drift combined with a block-wise measurement order, and interference between components, remain problems. Interleaved measurement is still the better design (Figure 3).
- **Status.** The theory and the design of the simulation study are set out; the detailed derivations (Appendix A), the simulation results and an application to real data are outside the scope of this note.

**How to read this note.** Readers mainly interested in the practical message can read Section 1, Section 3.1 and Sections 7 and 8. Sections 2 and 4 give the formal treatment and can be skipped on a first reading. Terms, abbreviations and symbols are explained in the Glossary at the end of the document.

## 1. Introduction

### 1.1 Calibration as an errors-in-variables problem

Calibration of an analytical instrument establishes the relation between a reference quantity, the composition of a calibration material, and the response of the instrument. Both quantities are uncertain (Figure 1). The reference value carries the uncertainty stated by the producer of the material; the response carries the repeatability and stability of the instrument. Ordinary least squares ignores the first of these. Errors-in-variables (EIV) regression, which includes Deming regression and generalized least squares with errors in both variables, accounts for both, but at a price: it requires an uncertainty for every point on both axes, and the quality of the fit and of its uncertainty statement depends directly on the quality of these inputs.

## Two axes, two different kinds of uncertainty

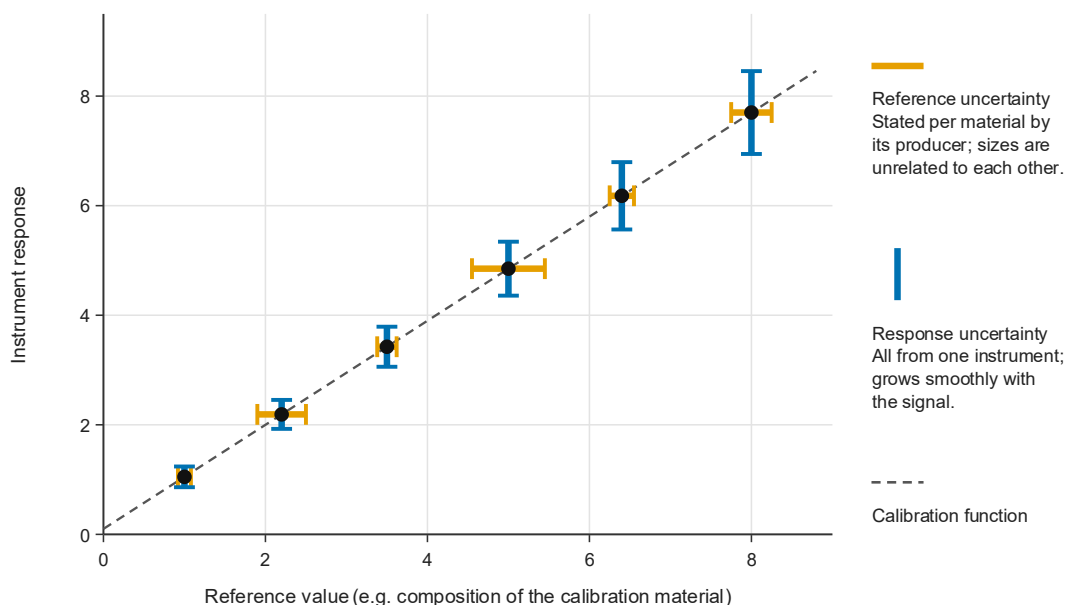


**Figure 1.** The asymmetry at the heart of this note (schematic). Horizontal bars: reference uncertainties, stated separately for each calibration material and unrelated to each other. Vertical bars: response uncertainties, which come from one measurement process and change smoothly with the signal.

Composition as a reference quantity deserves a separate remark. A composition is usually the result of a normalisation, in which the component fractions are scaled to sum to unity (or 100 %). This closure makes the components of one composition mutually dependent: an error in one component is redistributed over all others, which induces correlations, mostly negative, between their values. Treating the components as independent variables is therefore an approximation; its consequences for the independence assumption on the reference axis are discussed in Section 2.3.

### 1.2 Limitations of current practice

The motivation for this work is practical. Five features of the way calibrations are performed in routine laboratories limit the value of current EIV procedures.

**Too few replicates per level.** The response uncertainty at a calibration level is in practice estimated from 3–5 replicate measurements, sometimes 2. A standard deviation from  $n$  replicates has a relative standard uncertainty of approximately  $1/\sqrt{2(n-1)}$ : about 40 % for  $n = 4$ , about 70 % for  $n = 2$ . Weights derived from such variances are largely noise (Figure 2). In an EIV fit this noise propagates into the coefficients, into the goodness-of-fit statistic and into the uncertainty of the predicted value. It has long been known that with weights of this quality, an unweighted fit can outperform the weighted one [Tellinghuisen]. The opposite risk arises when the response uncertainty is taken as the standard deviation of the mean of the replicates,  $s/\sqrt{n}$ . This quantity decreases without limit as  $n$  grows, but only the random, short-term part of the variation averages out; components that are

common to all replicates of a level — drift, between-run effects, the state of the instrument during the block — do not. With many replicates  $s/\sqrt{n}$  therefore underestimates the uncertainty of the mean response, and the calibration point receives a weight it does not deserve. For this reason ISO 6143 limits the number of replicates that may be used in this way to a maximum of 10.

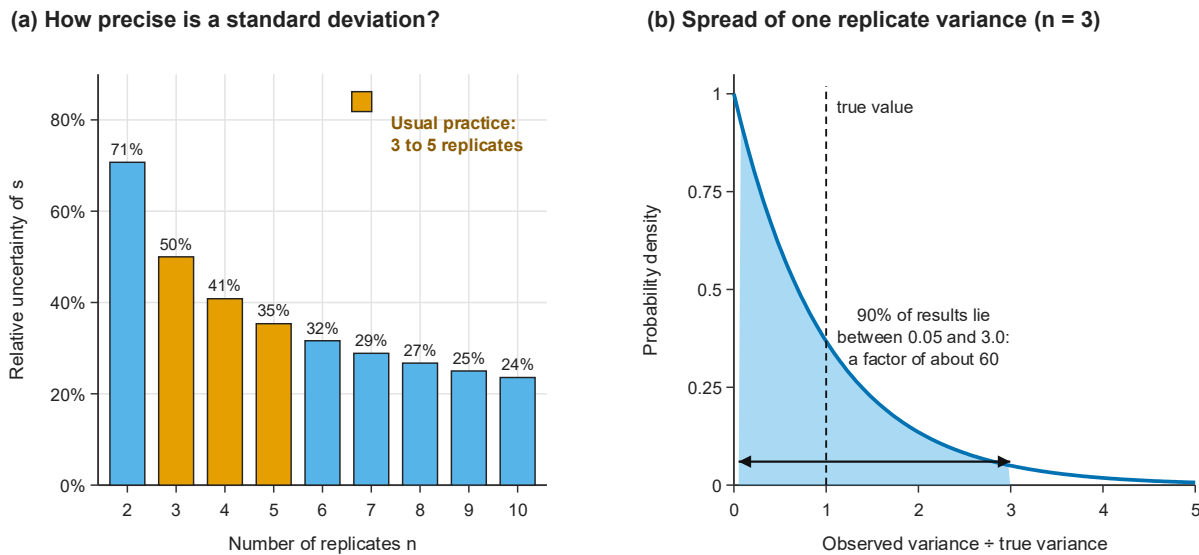


**Figure 2.** Why a handful of replicates is not enough. (a) Approximate relative standard uncertainty of a standard deviation  $s$ ,  $1/\sqrt{2(n-1)}$ , as a function of the number of replicates  $n$ . (b) For  $n = 3$ , the ratio of observed to true variance; the shaded area contains 90 % of the outcomes. Two calibration levels with exactly the same true variance can easily show variances that differ by a factor of ten or more.

**Non-optimal measurement order.** Calibration materials are typically measured block-wise — material 1 five times, then material 2 five times, and so on — because switching is inconvenient and requires purging. Replicates measured within a few minutes capture only very-short-term repeatability. Slow drift (detector, temperature, baseline) does not appear in the within-level variance but is confounded with the level: each block sees a different state of the instrument. The consequences are (i) per-level variances that are too small, (ii) an apparent curvature or offset in the calibration function that is really a time trend, and (iii) an over-confident uncertainty statement. The last point deserves emphasis: the uncertainty is underestimated, not merely misstated. The variation between blocks is a genuine part of the variation of a calibration point — an unknown is measured at another moment, in another state of the instrument — but under block-wise order it is never observed as variance: it is absorbed into the level means, where it appears as apparent structure in the calibration function instead of as uncertainty. The weights are then too large, the goodness-of-fit statistic is inflated and invites a higher-order model, and the uncertainty of the predicted value is too small. Taking the standard deviation of the mean,  $s/\sqrt{n}$ , as the response uncertainty compounds this, because it further reduces only the within-block part. Interleaved measurement (1, 2, ...,  $n$ , 1, 2, ...) distributes drift over all levels and yields realistic within-level variances, but is rarely done (Figure 3). It is

nevertheless the better design: block-wise order is common practice for reasons of convenience, not because it is optimal.

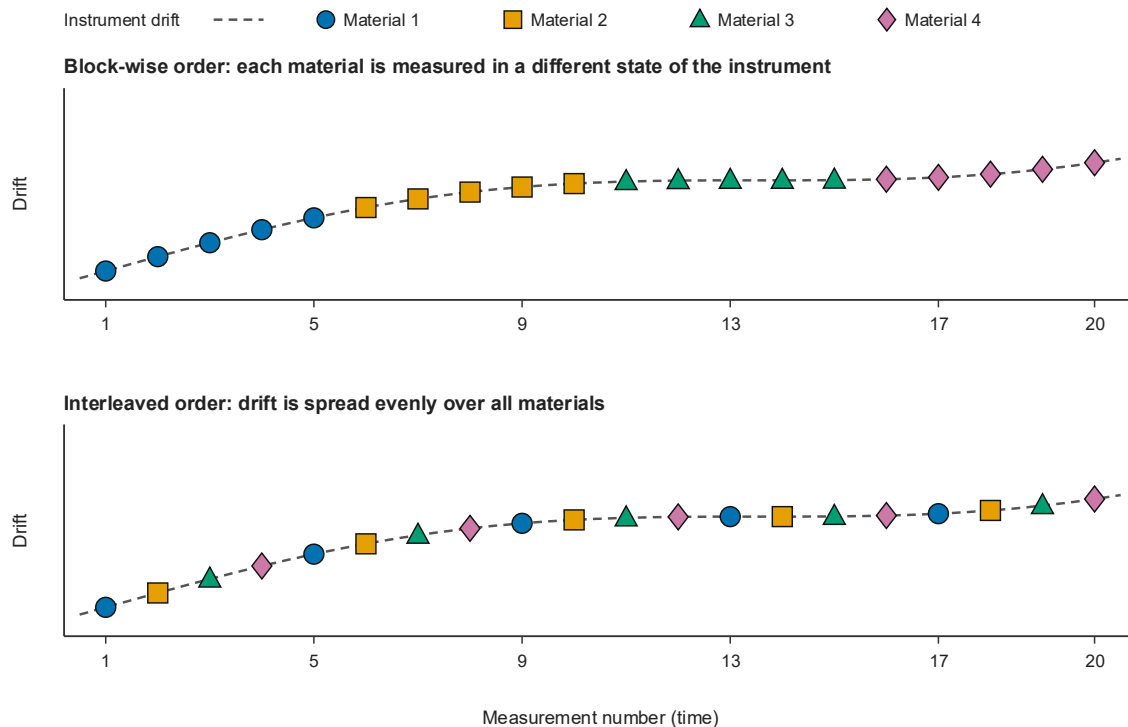


**Figure 3.** Measurement order and drift (schematic). With block-wise order (top) the drift is never seen as scatter within a material; it shifts the level means instead and can mimic curvature or an offset in the calibration function. With interleaved order (bottom) each material samples the whole drift, so the replicate scatter is realistic.

**Heterogeneous reference uncertainties.** Calibration materials come from different producers, batches and preparation routes, with uncertainties that differ per material and are stated per material. There is no natural model that ties them together, and there is no reason to want one.

**Per-point uncertainties as a formal requirement.** Standards and procedures for EIV calibration require a standard uncertainty for each response value and treat these as given inputs, independent per point. The procedure is silent on how reliable those inputs are, and the practitioner has no diagnostic to check them.

**Model selection and coefficient testing built on noisy weights.** The usual rule — take the lowest polynomial order for which the goodness-of-fit test passes, and drop coefficients (in particular the intercept) that are not significant — is sound in principle but rests entirely on the weights. With replicate variances from 3–4 measurements the  $\chi^2$  statistic and the coefficient covariance matrix are themselves noise: a second-order term is added because one level happened to show a small variance, or an intercept is retained or discarded on the strength of a t-value that means nothing. Model selection then models noise. This limitation is rarely named, because the tests are formally available; they are just not informative.

### 1.3 Existing approaches

Three lines of work are relevant. In ordinary calibration, where the reference axis is taken as error-free, variance-function weighting is well established: replicates are pooled to estimate the two or three parameters of a smooth variance function, so that the weights rest on many degrees of freedom instead of a few per level [Carroll & Ruppert; Davidian & Carroll; Tellinghuisen 2004, 2026]. The two-component model of Rocke and Lorenzato, which combines an additive and a proportional error, is essentially the variance function adopted here. In method comparison, weighted Deming regression with precision profiles uses variance functions on both axes [Hawkins & Kraker 2026]; there both axes are measurement methods and the error structure is symmetric. In metrological calibration, finally, EIV regression as described in ISO 6143, ISO/TS 28037 and by Fuller uses per-point uncertainties on both axes and takes them as given inputs.

### 1.4 Contribution and outline

The calibration situation is a third case, covered by neither line of work: the reference axis has independent, heterogeneous and externally given uncertainties, whereas the variance on the response axis is structured by the measurement process. In this paper we (i) make this asymmetry explicit in the measurement model, (ii) replace the per-level replicate variances on the response axis by a variance function while keeping the reference uncertainties as given, (iii) work out the consequences for estimation, covariance, model selection and inverse prediction, and (iv) compare the method with current practice in a simulation study that includes the effect of measurement order, and illustrate it on a real calibration dataset. Figure 4 gives an overview of the procedure. Section 2 introduces the measurement model and its assumptions, Section 3 the variance function and the conditions under which it may be used, and Section 4 the estimation procedure. Sections 5 and 6 present the simulation study and the application, and Section 7 discusses limitations and the relation to existing standards.

#### The proposed procedure at a glance

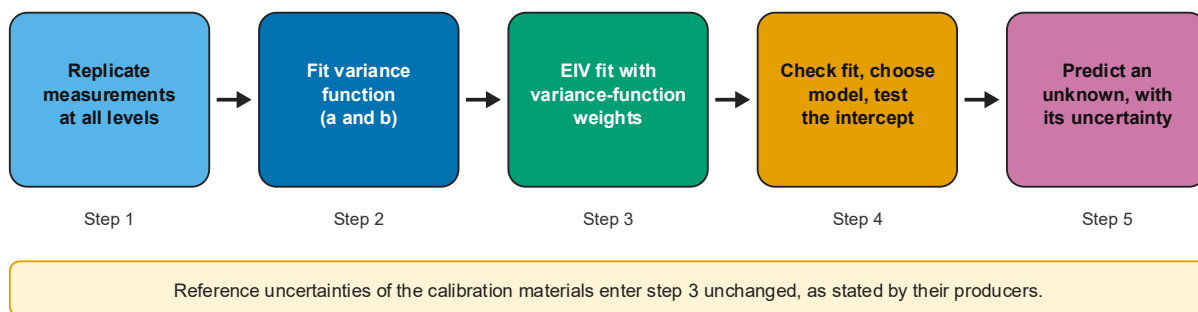


**Figure 4.** Overview of the proposed procedure. The replicates of all calibration levels are pooled to estimate one variance function (Sections 3.2 and 4.3), which supplies the weights for the EIV fit (Sections 4.1–4.5). Model checks and tests (Sections 4.6–4.7) and the prediction of unknowns (Section 4.8) all build on these weights.

## 2. Measurement model and notation

### 2.1 Observations, true values and errors

Let  $i = 1, \dots, m$  index the calibration levels and  $j = 1, \dots, n_i$  the replicate measurements at level  $i$ .

The reference value supplied with calibration material  $i$  is an observation of a true but unknown value  $\xi_i$ :

$$x_i = \xi_i + \delta_i, \quad \delta_i \sim (0, \sigma_{\delta,i}^2)$$

The uncertainty  $\sigma_{\delta,i}$  is stated by the producer of the material; it is not estimated from the calibration data. Different materials have different  $\sigma_{\delta,i}$ , and the  $\delta_i$  are taken as mutually independent (see 2.3).

The instrument response is an observation of a true response  $\eta_i$ :

$$y_{ij} = \eta_i + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim (0, \sigma_\varepsilon^2(\eta_i))$$

The  $\varepsilon_{ij}$  are independent across  $i$  and  $j$ , with a variance that depends on the true response through a variance function  $\sigma_\varepsilon^2(\cdot)$  — this is the structured heteroscedasticity of the title, developed in Section 3.

Normality of the errors is not required for the estimator; it is assumed only in the simulation study (Section 5).

### 2.2 Functional relation

The true values are linked by a deterministic relation with parameter vector  $\beta$ :

$$\eta_i = f(\xi_i; \beta) \quad (\text{response function})$$

or, equivalently for inverse use,

$$\xi_i = G(\eta_i; \beta) \quad (\text{analysis function})$$

The practically relevant output of a calibration is the analysis function  $G$ , applied to a measured response  $y_0$  of an unknown to obtain an estimate of its true reference value  $\xi_0$ , together with an uncertainty. Both  $f$  and  $G$  are treated as low-order polynomials or other parametric forms; the method does not depend on the specific form. Notation follows Fuller ( $\xi, \eta$  for true values;  $\delta, \varepsilon$  for errors), which is the common usage in the EIV literature.

### 2.3 Assumptions and their justification

The model rests on four assumptions.

**Independence of the  $\delta_i$ .** This assumption is justified when the calibration materials come from separate preparations. It is violated when materials share a parent mixture or a common purity assessment, in which case  $\text{cov}(\delta_i, \delta_k) \neq 0$ . Normalisation is a related case. When the composition of a material is obtained by normalising the component fractions to

unity (Section 1.1), the components within that material are correlated through the closure constraint; when the materials share inputs to the normalisation, such as a common analysis or a common purity value of the matrix component, this correlation extends to the  $\delta_i$  of one component across materials. The method does not address these cases. The assumption is stated explicitly so that the asymmetry argument is not misread as a claim that reference uncertainties are always independent.

**Independence of the  $\varepsilon_{ij}$ .** Random repeatability errors are independent between replicates and between levels. Shared systematic components, such as gain drift, baseline drift or a common integration setting, violate this assumption; their treatment is discussed in Section 3.3.

**Correct functional form of  $f$ .** The calibration function is assumed to be correctly specified. As usual, this is checked with the goodness-of-fit statistic and the residuals (Section 4.6).

**Variance function.** The response variance is assumed to depend on the calibration material only through the true response  $\eta_i$ . The variance function is introduced in Section 3; its justification and its limits, in particular interference between components of the material, are discussed in Section 3.1.

### 3. Structured heteroscedasticity of the response

#### 3.1 The asymmetry argument

The two axes differ not in the size of their uncertainties but in their *origin*. Reference uncertainties arise from  $m$  separate processes; there is no information in one that helps estimate another, and the producer has already done the estimation. Response uncertainties arise from one process — one detector, one integration, one operator, one day — and it is unreasonable to treat  $m$  separate variances as free parameters when they are known to be tied to the signal by the physics of the detector. Estimating them separately throws away that knowledge and, with few replicates, replaces it by noise.

Can the response variance be modelled? Modelling the response variance as a smooth function of the signal is an assumption, not a fact, and it has to be defended. The argument in favour is physical: the sources of repeatability — sample introduction, detector noise, integration of the signal — are the same at every calibration level and vary smoothly with the size of the signal. It is implausible that the true variance jumps back and forth between levels without a cause. The observed per-level variances do jump, but that is exactly what sampling noise predicts: with  $n = 3$  the ratio  $s_i^2/\sigma^2$  follows  $\chi^2_2/2$ , whose central 90 % interval runs from about 0.05 to 3.0 — a factor of about 60 between two levels with identical true variance. Erratic per-level variances are therefore no evidence against a variance function; with few replicates they are what a variance function predicts (Figure 5).

### Per-level estimates scatter; the true variance is smooth

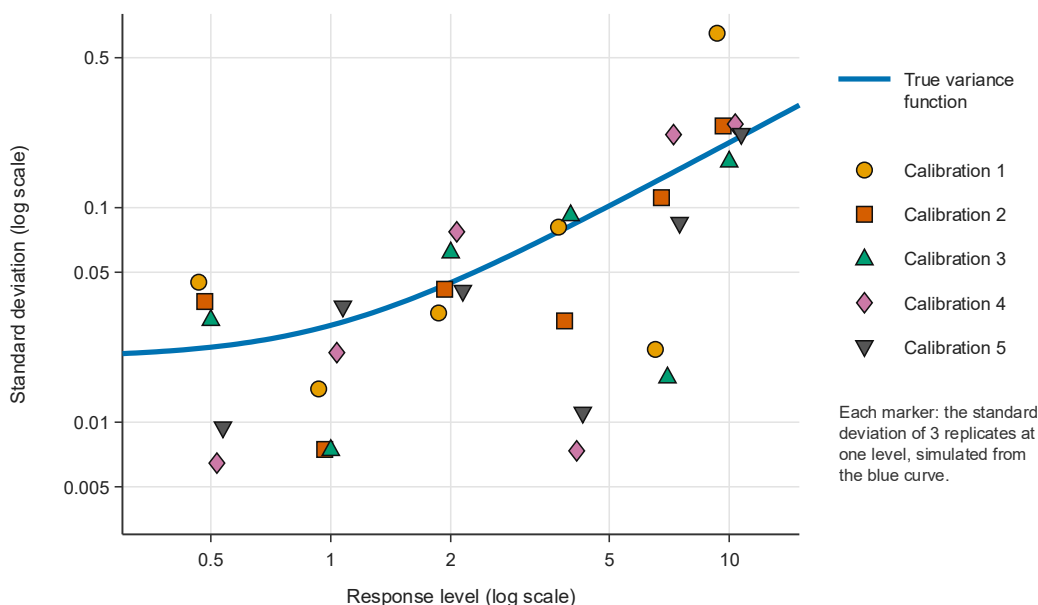


**Figure 5.** Simulated illustration, not measured data. Five calibrations were simulated from the same smooth variance function (blue curve), with three replicates at each of six levels. The per-level standard deviations jump up and down without any real cause: erratic per-level variances are what sampling noise predicts, not evidence against a variance function.

The argument against is that the response to a calibration material is not determined by the analyte alone. The materials differ in composition, and the signal of one component can be affected by the signals of the others: overlapping or partially resolved signals, a large neighbouring signal that perturbs the baseline, a matrix effect on sensitivity. Such contributions depend on the material, not on  $\eta$ , and are not captured by  $\sigma^2_{\varepsilon}(\eta)$ . They act in two ways. Their random part adds a material-specific variance component, so that the variance at level  $i$  becomes  $\sigma^2_{\varepsilon}(\eta_i) + \tau_i^2$ . Their systematic part — for example a consistent integration error caused by a neighbouring signal — shifts the level mean and does not appear in the replicates at all.

Three consequences follow. First, the variance function is a hypothesis to be checked per component: under the model the standardised replicate variances  $(n_i - 1) \cdot s_i^2 / \sigma^2_{\varepsilon}(\hat{\eta}_i)$  follow  $\chi^2$  with  $n_i - 1$  degrees of freedom, a likelihood-ratio test of the variance function against  $m$  free per-level variances ( $m - 2$  degrees of freedom) tests it jointly, and individual levels outside the  $\chi^2$  range are flagged. With few replicates this test has little power, so it must be supplemented by knowledge of the measurement: a material with an interference is identified from the raw signal (for example its separation from neighbouring signals), not from its replicate variance. Second, where an interference is known, the model is extended for that material only, by a term  $\tau_i^2$  that is estimated or assessed separately, rather than abandoned for all materials. Third, per-level weighting is not the safe alternative it appears to be: with three replicates, a variance increase by a factor of two or three caused by interference cannot be distinguished from sampling noise, so per-level variances do not

detect the problem either; they merely replace a model that may be wrong at a few levels by estimates that are imprecise at all levels. The systematic part of an interference is invisible to both approaches; it appears only in the residuals of the fit, where the residual-versus-replicate comparison of 4.3 provides a diagnostic.

### 3.2 Variance functions

A single variance function is adopted throughout — the two-component (additive plus proportional) form of Rocke and Lorenzato:

$$\sigma_{\varepsilon}^2(\eta) = a^2 + (b \cdot \eta)^2$$

Here  $a$  is the noise floor of detector and integration at zero signal,  $b$  the relative standard deviation at high signal, and  $a/b$  the response at which the two contributions are equal. This is the behaviour of practically every analytical detector: constant absolute spread at the low end, constant relative spread at the high end (Figure 6). The form is not selected from the data but imposed from prior knowledge of the measurement process; the data estimate only  $a$  and  $b$ . This is the statistically appropriate response to the replicate problem: when the data are too sparse to select a variance model, the model must come from outside the data.

**The two-component variance function**

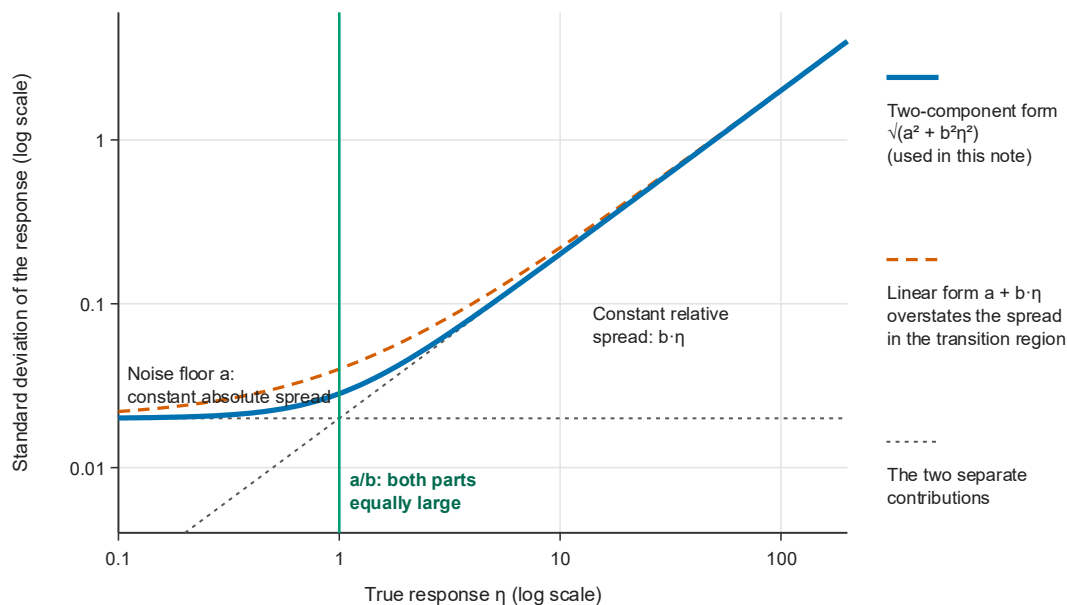


**Figure 6.** The two-component variance function of Section 3.2 (illustrative values  $a = b = 0.02$ ). At low signal the standard deviation levels off at the noise floor  $a$ ; at high signal it is proportional to the signal. The linear form  $a + b \cdot \eta$ , which corresponds to fully correlated error sources, overstates the standard deviation where the two contributions are of similar size.

Why this form, and why only this form. The two terms correspond to two classes of error source that are physically independent: sources that do not depend on the signal (detector

noise, baseline fluctuation, integration noise) and sources that scale with it (the amount of sample introduced — loop volume, loop pressure and temperature, flow and split — and short-term sensitivity changes). Because the sources are independent, their variances add; the standard deviation is  $\sqrt{a^2 + b^2\eta^2}$ , which tends to  $a$  at low signal and to  $b\cdot\eta$  at high signal. The linear form  $\sigma = a + b \cdot \eta$ , sometimes used in practice, would correspond to fully correlated sources and overstates the standard deviation in the transition region. Restricting the analysis to this form is a deliberate choice for parsimony: the argument of this paper is that the data are too sparse to select a variance model, and every additional parameter erodes the gain from pooling. With  $m = 4\text{--}8$  levels and a few replicates, two parameters can be estimated; three can hardly be. Over the narrow ranges typical of calibrations — often a factor of 2 to 10 between the lowest and highest level — the two-component form is moreover practically indistinguishable from other smooth forms such as a power of the mean; within the range it is the smoothness of the function that matters, not its exact shape.

The form does matter in three situations, which are therefore stated as limits rather than ignored. First, detectors governed by counting statistics have a variance term proportional to  $\eta$ , which the two-component form does not contain. Second, over ranges of several decades the relative standard deviation is often not constant at the high end (detector non-linearity, overload), and a power of the mean with an exponent different from 2 may fit better. Third, when the function is extrapolated — to a response  $y_0$  outside the calibration range, or to zero for limits of detection and quantification — the result depends on the assumed form, and  $a$  in particular is then often poorly determined by the data. Material-specific interference (3.1) is a different matter: it is not a function of  $\eta$  at all and is not remedied by any choice of variance function. The robustness of the conclusions to a misspecified form is tested in the simulation (5.1, 5.4).

What can be tested is whether both terms are supported.  $a = 0$  and  $b = 0$  are nested cases; a likelihood-ratio test on the pooled replicate variances (each  $s_i^2 \sim \sigma_\varepsilon^2(\eta_i) \cdot \chi_{n_i-1}^2/(n_i - 1)$  under normality) tells whether the data carry each term. When no calibration level lies near zero,  $a$  will often not be significant; it is then retained on physical grounds, with the statement that it is not determined by the data — see the lower-limit remark in Section 7.

Two diagnostics accompany the variance function: a log-log plot of the replicate standard deviation against the level mean, and the standardised residuals after the fit. A separate variance function is estimated for each component and each detector; pooling across components or instruments is not assumed.

Other forms — power of the mean  $a \cdot \eta^c$ ; a quadratic  $a + b \cdot \eta + c \cdot \eta^2$  with a linear term for counting-statistics detectors — are mentioned in the Discussion as extensions; the estimation framework does not change, only  $\sigma_\varepsilon^2(\cdot)$ .

### 3.3 Heteroscedasticity versus correlation

A variance function describes the diagonal of the covariance matrix of the response errors; it does not imply any off-diagonal terms. Random repeatability is uncorrelated between points, and correlation arises only from shared systematic components. Two of these

deserve comment. Components that are common to the calibration and to the measurement of the unknown largely cancel in the analysis function when the unknown is measured close in time and under the same conditions; what remains is a stability term that belongs in the uncertainty budget of the measurement, not in the calibration fit. Drift under block-wise measurement order, on the other hand, is not correlation in the statistical sense but confounding: it is absorbed into the level means. It is therefore addressed by the design of the measurement sequence (Section 5) and not by the covariance matrix.

### 3.4 Consequences for the covariance structure

With  $\sigma_\varepsilon^2(\eta)$  evaluated at the fitted  $\eta_i$ , the weight matrix of the response axis becomes a smooth function of the fit itself, and the variance-function parameters (a, b, ...) become part of the estimation. The resulting covariance matrix of  $(\hat{\beta}, \hat{a}, \hat{b})$  and the covariance between  $\hat{\beta}$  and the weights will be derived in Appendix A.

## 4. Estimation

### 4.1 Objective function

The EIV estimator minimises, over  $\beta$  and the true values  $\xi_i$ ,

$$S = \sum_i \left[ \frac{(x_i - \xi_i)^2}{\sigma_{\delta,i}^2} + \sum_j \frac{(y_{ij} - f(\xi_i; \beta))^2}{\sigma_\varepsilon^2(f(\xi_i; \beta))} \right]$$

or its equivalent using level means and  $n_i$ . Compare with current practice, where  $\sigma_\varepsilon^2(f(\xi_i; \beta))$  is replaced by the per-level sample variance  $s_i^2$ .

### 4.2 Evaluating the variance at the fitted response

The variance must be evaluated at the fitted response and not at the observed one. Weighting by the observed value gives observations that happen to lie low a high weight and biases the fit downward. Two implementations are possible: iteratively reweighted least squares, in which the fit and the weights evaluated at the fitted values are updated alternately until convergence, or joint maximum-likelihood estimation of  $\beta$  and the variance-function parameters (Section 4.5).

### 4.3 Estimating the variance function

The variance function can be estimated by two routes.

- **Replicate route:** fit  $\sigma_\varepsilon^2(\cdot)$  to the per-level sample variances  $s_i^2$ , weighted by their own uncertainty (relative variance  $2/(n_i - 1)$ ), or fit  $\log s_i^2$  unweighted.
- **Residual route:** fit  $\sigma_\varepsilon^2(\cdot)$  to the squared (studentised) residuals of the calibration fit, iterating with the fit itself. Available even with  $n_i = 1$ .

Comparison of the two routes is a diagnostic: if residuals are systematically larger than the replicate route predicts, a component is missing from the replicates — typically drift under block-wise order.

#### 4.4 Propagation of variance-function uncertainty

The variance-function parameters are estimated from the same data as  $\beta$ , and their uncertainty propagates into  $\hat{\beta}$  and into the predicted value. This contribution can be evaluated analytically from the joint covariance matrix of the likelihood (Section 4.5) or numerically by parametric bootstrap. Which of the two is used, and under which conditions the contribution is negligible (for example as a function of the degrees of freedom  $\nu$ ), is left for further work.

#### 4.5 Joint likelihood

Under normal errors the model of Section 2 has a single likelihood in  $(\beta, \xi_1 \dots \xi_m, a, b)$ . Writing it out gives one consistent basis for estimation (4.2), for the variance-function estimate (4.3), for the propagation of 4.4, and for all tests below. Degrees of freedom:  $\nu = N - p - q$ , with  $N$  the total number of response observations,  $p$  the number of  $\beta$ 's and  $q = 2$  the variance-function parameters; the  $m$  reference values and the  $m$  latent  $\xi_i$  cancel.

#### 4.6 Goodness of fit and model selection

The minimum of  $S$  (equivalently  $-2 \cdot \log L$  up to a constant) is approximately  $\chi^2_\nu$ ; with variance-function weights this statistic recovers its distribution because the weights are no longer noise, and the existing rule “lowest order that passes” becomes usable. Contrast with its behaviour under per-level weights (Section 5).

Three criteria for choosing the order of  $f$  are derived from the same likelihood. The first is the likelihood-ratio (or  $F$ ) test between nested polynomial orders. The second is AICc, the Akaike information criterion with the small-sample correction, which is needed because  $m$  is small. The third is predictive: in a leave-one-level-out procedure, each calibration level is predicted as an unknown from the remaining levels. This criterion measures directly what the user of the calibration needs, and it penalises overfitting most heavily at the ends of the range, where it does most harm.

#### 4.7 Significance testing of individual coefficients

Each coefficient is tested by a Wald  $t$ -statistic,  $t = \hat{\beta}_k / u(\hat{\beta}_k)$ , with  $\nu$  degrees of freedom, where  $u(\hat{\beta}_k)$  comes from the covariance matrix of the joint fit including the uncertainty of  $a$  and  $b$  (4.4) — omitting that contribution makes  $u(\hat{\beta}_k)$  too small and the test too liberal. The likelihood-ratio test between the models with and without  $\beta_k$  is the reference; for linear  $f$  and normal errors the two coincide, and the simulation checks the size of the  $t$ -test.

The test is only as informative as  $u(\hat{\beta}_k)$ . Under per-level weighting the covariance matrix is as noisy as the weights, and the  $t$ -value of, for example, the intercept is essentially arbitrary; under variance-function weighting the same test becomes informative. This is the most direct practical benefit of the method.

Four rules of procedure follow.

- polynomial coefficients are strongly correlated, in particular when the range lies far from zero; test hierarchically from the highest order downward (backward elimination) and never drop a lower-order term while a higher one remains;
- the intercept  $\beta_0$  is the exception to the hierarchy: it has its own physical meaning (blank signal) and may be tested regardless of the order;
- centring or orthogonalising the polynomial reduces the correlation and makes the individual tests more interpretable; report the coefficients back in the original basis;
- a coefficient found not significant is removed by refitting the reduced model, never by setting its estimate to zero and keeping the others. The coefficients are correlated, so the remaining estimates shift; the reduced model is a different constrained optimisation with its own solution, its own covariance matrix and — since the variance-function weights depend on the fitted responses — its own weights. The LR test performs this refit by construction. After each removal the remaining coefficients are re-tested, since their t-values change: one coefficient per step.

**Order of elimination and non-nested alternatives.** Start from the highest plausible model (e.g. third order with intercept) and eliminate downward, the intercept last. Going down first and then up again can end in two non-nested models with the same number of parameters — second order with intercept versus third order without — between which nested tests and AICc cannot decide. This situation means the data cannot separate an offset from curvature, typically because the low end of the range is poorly covered. An intercept that is not significant but whose removal makes the goodness-of-fit test fail is conflicting evidence, and conflicting evidence means insufficient power; the remedy is a level near zero or a blank measurement, not a higher-order term. If a choice must be made without new data: the predictive criterion of 4.6 plus prior knowledge, which for practically every detector favours the low order with a small intercept over a higher order without one.

**The intercept, metrologically.** Dropping  $\beta_0$  when not significant introduces a bias at the low end of the range if the true value is small but non-zero; retaining it when it is noise inflates the uncertainty there. The test alone therefore does not decide; the decision combines the test with prior knowledge of the detector's blank behaviour.

## 4.8 Inverse prediction

For an unknown with measured response  $y_0$  (mean of  $n_0$  replicates), the estimate is  $\hat{\xi}_0 = G(y_0; \hat{\beta})$  with uncertainty combining (i) the response uncertainty  $\sigma_\varepsilon^2(y_0)/n_0$  — available directly from the variance function, where per-level weighting has nothing to offer at a response between calibration levels — and (ii) the uncertainty of  $\hat{\beta}$ . To first order,  $u^2(\hat{\xi}_0) = (\partial G/\partial y)^2 \cdot \sigma_\varepsilon^2(y_0)/n_0 + \mathbf{g}^T \cdot \mathbf{V}(\hat{\beta}) \cdot \mathbf{g}$ , with  $\mathbf{g} = \partial G/\partial \beta$  evaluated at  $(y_0, \hat{\beta})$ .

## 5. Simulation study

### 5.1 Design

Data are generated from the model of Section 2 with known  $\beta$ , known  $\sigma_{\delta,i}$  (heterogeneous, drawn to resemble realistic certificates) and a known variance function. The following factors are varied.

The number of levels is  $m \in \{4, 6, 8\}$  and the number of replicates per level  $n \in \{2, 3, 5\}$ . The strength of heteroscedasticity, expressed as the ratio of the response variances at the ends of the range, is set to 10 and 100. In one misspecified case the data are generated from a power-of-the-mean or a quadratic variance function and fitted with the two-component form. In a model-selection scenario the data are generated from a straight line and the polynomial order is selected with the criteria of Section 4.6; in a coefficient-test scenario the true intercept is zero (type I error of the intercept test) or small but non-zero (power). The measurement order is either block-wise or interleaved, with a slow linear or random-walk drift added to the response. The number of Monte Carlo runs per scenario, the random seed and the software will be reported together with the results.

## 5.2 Estimators compared

Four estimators are compared:

1. Unweighted EIV ( $\sigma_\varepsilon$  constant, estimated from pooled residuals)
2. EIV with per-level replicate variances  $s_i^2$  — current practice
3. EIV with variance-function weights, replicate route
4. EIV with variance-function weights, residual route / joint ML

## 5.3 Metrics

Performance is assessed for  $\hat{\beta}$  and, more importantly, for the predicted value  $\hat{\xi}_0$  at three levels (low, middle and high in the range, including a response between calibration levels). The metrics are the bias; the ratio of the Monte Carlo standard deviation to the mean reported standard uncertainty, as a measure of the realism of the uncertainty statement; the coverage probability of the 95 % interval; the distribution of the goodness-of-fit statistic; the fraction of runs in which the correct polynomial order is selected; the size (rejection rate at a true  $\beta_0 = 0$ , nominal 5 %) and the power of the intercept t-test; and the number of convergence failures.

## 5.4 Results

The statements below are the expected findings, based on the theory of Sections 3 and 4; they remain to be confirmed by running the simulation study. With per-level weighting the estimates show little bias, but the variance of  $\hat{\beta}$  is strongly inflated and the coverage is erratic, most of all at low  $n$  and at the low end of the range; the goodness-of-fit statistic is unreliable. With variance-function weighting the efficiency comes close to that of a fit with the true weights for  $n \geq 3$ . At  $n = 2$  the method still outperforms per-level weighting, but the functional form of the variance can no longer be verified. When the variance function is misspecified, performance degrades gracefully. In the model-selection scenario, per-level weighting over-selects higher orders at  $n = 3$ , whereas variance-function weighting restores the selection rule. For the intercept test, per-level weighting gives a type I error far from the nominal 5 %, whereas variance-function weighting gives the nominal size and usable power. Under drift with block-wise order, per-level weighting gives over-confident uncertainties and a biased  $\hat{\beta}$ ; variance-function weighting limits but does not remove the damage, and interleaved order removes the bias for all estimators. The design of the

measurement sequence thus still matters; the estimator determines whether a poor design gives a mediocre or a bad result.

## 6. Application to a real calibration dataset

This section outlines the intended application to a real calibration dataset. Such an application requires a description of the instrument, the component(s), the range, the number of levels and replicates, the measurement order actually used, and the origin and stated uncertainties of the calibration materials. In the planned analysis, the replicate standard deviations are plotted against the level means on log-log scales together with the fitted variance function, and the per-level weights are compared with the variance-function weights. The calibration fit, the residuals and the goodness-of-fit statistic are then compared under both weightings, as are the predicted value and its uncertainty for a set of check samples.

Because the true values of a real dataset are unknown, such an application can show that the method is plausible on real data; it cannot prove that it is correct.

## 7. Discussion

The gain of the method comes from pooling: a variance function estimated from all replicates jointly replaces  $m$  imprecise per-level variances by two well-determined parameters. The asymmetry argument of Section 3.1 is what justifies applying this pooling to the response axis only, while the reference uncertainties are kept as given.

Whether the response variance may be modelled at all is the central assumption of this work (Section 3.1). In gas chromatography the main threat is chromatographic: a peak that is not baseline-resolved from a neighbouring component, tailing of a large preceding peak or a perturbed baseline makes the integration, and with it the variance and possibly the mean of the response, depend on the composition of each calibration material. Good chromatographic resolution, careful integration settings and calibration materials with a similar matrix across levels limit this risk; where an interference cannot be avoided, a material-specific term  $\tau_i^2$  is needed (Section 3.1).

Model selection and coefficient tests are formally available in current practice but are not informative, because they rest on noisy weights; the method makes them informative. Selection nevertheless remains a combination of statistical testing and prior knowledge, in particular for the intercept. The form of the variance function is imposed rather than selected; extensions such as a power of the mean or a counting-statistics term fit the same framework.

The method is more robust to block-wise measurement order than per-level weighting, and the comparison of the residual and replicate routes (Section 4.3) provides a diagnostic that current practice lacks. Interleaved measurement nevertheless remains the correct remedy. Nor is the variance function a licence to reduce the number of replicates further: below  $n = 3$  its form cannot be checked, and the additive term  $a$  is determined only when calibration levels lie near zero.

The main limitations of the present work are the assumed independence of the reference uncertainties, which fails for materials with a shared parent mixture or shared normalisation inputs; the use of one variance function per component and instrument; and shared systematic components in the response (Section 3.3). Procedures that take per-point response uncertainties as given inputs, such as ISO 6143, can accept variance-function values without any change to the algorithm; the change lies in how the inputs are produced. A side benefit is that the variance function gives limits of detection and quantification directly.

## 8. Conclusions

The response and reference axes of an analytical calibration differ in the origin of their uncertainties, and this asymmetry justifies modelling the response variance by a variance function while the reference uncertainties are kept as given. A two-component variance function, estimated from all replicates jointly, replaces per-level variances that are largely noise at the usual numbers of replicates. As a result, the goodness-of-fit test, model selection and the significance test of the intercept become informative, and the uncertainty of the predicted value can be evaluated at any response, including responses between calibration levels. The method does not remove the need for a good measurement design: interleaved measurement remains necessary to prevent drift from being absorbed into the level means, and interference between components limits the use of a variance function. Because the method changes only how the response uncertainties are produced, it can be used within existing EIV procedures without modification.

## Data and code availability

The simulation code and the calibration dataset will be made available for download alongside this note when they become available.

## Glossary

Short explanations of the main terms, abbreviations and symbols used in this note.

| Term or symbol                              | Meaning                                                                                                                                                       |
|---------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>Calibration level</b>                    | One calibration material, measured one or more times. $m$ is the number of levels.                                                                            |
| <b>Replicate</b>                            | A repeat measurement of the same calibration material. $n_i$ is the number of replicates at level $i$ .                                                       |
| <b>Reference value: <math>x, \xi</math></b> | The value of a calibration material as stated by its producer ( $x$ ) and its unknown true value ( $\xi$ ).                                                   |
| <b>Response: <math>y, \eta</math></b>       | The signal of the instrument ( $y$ ) and its unknown true value ( $\eta$ ).                                                                                   |
| <b><math>\delta, \varepsilon</math></b>     | The errors in the reference value and in the response.                                                                                                        |
| <b><math>\sigma_{\delta,i}</math></b>       | Standard uncertainty of the reference value of material $i$ , as stated by its producer.                                                                      |
| <b>EIV regression</b>                       | Errors-in-variables regression: fitting a calibration function while taking the uncertainties on both axes into account. Deming regression is a special case. |
| <b>Heteroscedasticity</b>                   | The variance of the response is not constant but changes with the signal. 'Structured' means that the change follows a smooth function.                       |

|                                                                 |                                                                                                                                 |
|-----------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------|
| <b>Variance function <math>\sigma^2_{\epsilon}(\eta)</math></b> | Gives the variance of the response as a function of the true response. In this note: $a^2 + (b \cdot \eta)^2$ .                 |
| <b>a, b</b>                                                     | Parameters of the variance function: the noise floor at zero signal (a) and the relative standard deviation at high signal (b). |
| <b>Weight</b>                                                   | The inverse of the variance of a point. Points with a smaller uncertainty count more heavily in the fit.                        |
| <b>Response function f, analysis function G</b>                 | The calibration function from reference value to response (f), and its inverse (G), which is used to analyse unknowns.          |
| <b><math>\beta</math></b>                                       | The coefficients of the calibration function, for example the intercept $\beta_0$ and the slope.                                |
| <b>Inverse prediction</b>                                       | Estimating the reference value of an unknown from its measured response $y_0$ with the analysis function G.                     |
| <b><math>\chi^2</math> (goodness of fit)</b>                    | A statistic that measures how well the fitted function agrees with the data, given the weights.                                 |
| <b>v</b>                                                        | Degrees of freedom.                                                                                                             |
| <b>AICc</b>                                                     | Akaike information criterion with a small-sample correction; used to compare calibration models.                                |
| <b>Block-wise and interleaved order</b>                         | Measuring each material several times in a row, or alternating between the materials (Figure 3).                                |
| <b>Drift</b>                                                    | A slow change of the instrument response over time.                                                                             |
| <b>Monte Carlo simulation</b>                                   | Repeatedly generating artificial data with known properties to test how well a method performs.                                 |

## References

- Carroll, R.J., Ruppert, D. *Transformation and Weighting in Regression*. Chapman & Hall, 1988.
- Davidian, M., Carroll, R.J. Variance function estimation. *J. Am. Stat. Assoc.* 82 (1987) 1079–1091.
- Fuller, W.A. *Measurement Error Models*. Wiley, 1987.
- Rocke, D.M., Lorenzato, S. A two-component model for measurement error in analytical chemistry. *Technometrics* 37 (1995) 176–184.
- Tellinghuisen, J. Impact of variance function estimation in regression and calibration. *Methods Enzymol.* (2004).
- Tellinghuisen, J. How not to do WLS fitting in calibration with heteroscedastic data. *Anal. Chem.* 98 (2026) 11641–11644.
- Hawkins, D.M., Kraker, J.J. Precision profile weighted Deming regression for methods comparison. *J. Appl. Lab. Med.* 11 (2026) 379–392. <https://doi.org/10.1093/jalm/jfaf183>
- Linnet, K. Estimation of the linear relationship between the measurements of two methods with proportional errors. *Stat. Med.* 9 (1990) 1463–1473.
- ISO 6143:2001. Gas analysis — Comparison methods for determining and checking the composition of calibration gas mixtures. ISO 11095:1996. Linear calibration using reference materials. ISO/TS 28037:2010. Determination and use of straight-line calibration functions.

## Appendix A — Derivations

This appendix is planned and will give the joint likelihood; covariance of  $(\hat{\beta}, \hat{\alpha}, \hat{\mathbf{b}})$ ; uncertainty of the inverse prediction; degrees of freedom  $\nu = N - p - q$ ; equivalence of Wald and LR tests for linear f.

## Appendix B — Simulation details

This appendix is planned and will describe the generation of heterogeneous  $\sigma_{\delta,i}$ ; drift models; convergence criteria; software versions.