Force calibrations using errors-in-variables regression and Monte Carlo uncertainty evaluations

Thomas Bartel[1], Sara Stoudt[2] and Antonio Possolo[1]
[1] National Institute of Standards and Technology, US Department of Commerce, Gaithersburg,MD, USA
[2]Department of Statistics, University of California, Berkeley, CA, USA

Content submitted to and published by:
Metrologia **53**, pp. 965-980 (2016)

U.S. Department of Commerce
Penny Pritzker, Secretary

National Institute of Standards and Technology
Willie E. May, Director

National Institute of Standards and Technology • U.S. Department of Commerce

DISCLAIMERS

Certain commercial entities, equipment, or materials may be identified in this document in order to describe an experimental procedure or concept adequately. Such identification is not intended to imply recommendation or endorsement by the National Institute of Standards and Technology, nor is it intended to imply that the entities, materials, or equipment are necessarily the best available for the purpose.

Any link(s) to website(s) in this document have been provided because they may have information of interest to our readers. NIST does not necessarily endorse the views expressed or the facts presented on these sites. Further, NIST does not endorse any commercial products that may be advertised or available on these sites.

Force Calibrations using Errors-in-Variables Regression and Monte Carlo Uncertainty Evaluations

Thomas Bartel

National Institute of Standards and Technology, U.S. Department of Commerce, Gaithersburg, MD, USA

Sara Stoudt

Department of Statistics, University of California, Berkeley, CA, USA

Antonio Possolo

National Institute of Standards and Technology, U.S. Department of Commerce, Gaithersburg, MD, USA

E-mail: antonio.possolo@nist.gov

Abstract. An errors-in-variables regression method is presented as an alternative to the ordinary least-squares regression computation currently employed for determining the calibration function for force measuring instruments from data acquired during calibration. A Monte Carlo uncertainty evaluation for the errors-in-variables regression is also presented. The corresponding function (which we call *measurement function*, which in gas metrology is often called *analysis function*) necessary for the subsequent use of the calibrated device to measure force, and the associated uncertainty evaluation, are also derived from the calibration results. Comparisons are made, using real force calibration data, between the results from the errors-in-variables and ordinary least-squares analyses, as well as between the Monte Carlo uncertainty assessment and the conventional uncertainty propagation employed at the National Institute of Standards and Technology (NIST). The results show that the errors-in-variables analysis properly accounts for the uncertainty in the applied calibrated forces, and that the Monte Carlo method, owing to its intrinsic ability to model uncertainty contributions accurately, yields a better representation of the calibration uncertainty throughout the transducer's force range than the methods currently in use. These improvements notwithstanding, the differences between the results produced by the current and by the proposed new methods generally are small because the relative uncertainties of the inputs are small and most contemporary load cells respond approximately linearly to such inputs. For this reason, there will be no compelling need to revise any of the force calibration reports previously issued by NIST.

1. Introduction

Many national measurement institutes maintain force laboratories for the dissemination of force measurement standards through the calibration of force transducers that serve as transfer standards. These calibrations make use of test facilities, measurement procedures, and data analyses consistent with documentary standards including those issued by ASTM International (American Society for Testing and Materials, ASTM) ASTM E74-13a [1] and by the International Organization for Standardization (ISO) ISO 376:2011(E) [15]. This paper introduces several statistical methods that overcome limitations of procedures currently in use at the National Institute of Standards and Technology (NIST), which are consistent with those standards, thus increasing the reliability of calibration results and uncertainty evaluations.

NIST conducts calibrations of force transducers through the use of primary force standards consisting of six machines for applying discrete forces generated by stainless steel deadweights, spanning a range of 44 N to 4.448 MN [16]. The objective of each calibration is to characterize the transducer with a calibration function C that relates the transducer's response R to the applied force F, as $R = C(F)$.

When the transducer is employed to measure force in practice, the inverse relationship is used, that provides an estimate of force as a function of the instrumental response, $F = M(R)$. Being concerned with chemical composition of gas mixtures, ISO [14] calls M the *analysis function*. Here we will call M the *measurement function*, therefore in a very different sense in which this expression is used in the *International Vocabulary of*

Metrology (VIM) [20, 2.49], but very appropriately and expressively in the context that we are concerned with.

The measurement function may be computed directly from the experimental data collected in a calibration experiment, or it may be obtained as the mathematical inverse of the calibration function. In general, these alternative approaches need not produce the same function. M being the inverse of C means that $M(C(F)) = F$ for all values of the force F for which C is defined, which is often abbreviated $M = C^{-1}$. (In this expression, the superscript -1 denotes the compositional inverse rather than the multiplicative inverse, in the same sense that the arcsine function is the compositional inverse of the sine function, while the cosecant function is its multiplicative inverse, or reciprocal).

Since NIST is required to characterize each calibrated transducer and to document such characterization in a calibration report, we must compute the calibration function C every time, and also evaluate the associated uncertainty. To ensure consistency between the calibration and measurement functions, we define the latter as the mathematical inverse of the former, and evaluate uncertainty for M accordingly. This presumes that C is invertible over the relevant range of its argument: which has been the case in all calibrations of force transducers that we have performed at NIST, because for all of them C has turned out to be a monotonically increasing function of the applied force.

Section 2 reviews the force calibration procedures currently in use at NIST, and examines the assumptions that call for a reconsideration of the statistical methods employed. Section 3 details a proposed errors-in-variables regression. Section 4 presents a comparison of calibration results using the current and proposed methods, and Section 6 summarizes and discusses some conclusions and lessons learned.

Section 5 describes a Monte Carlo uncertainty analysis that may be more appropriate for the current state of transducer technology than the conventional uncertainty evaluations described by Bartel [2], which employ the methods described in NIST Technical Note 1297 [32] and in the *Guide to the expression of uncertainty in measurement* (GUM) [17].

The Monte Carlo method that we employ for uncertainty evaluation is consistent with the guidance in the Supplement 1 to the GUM [18], and provides great freedom and flexibility to model all uncertainty contributions accurately. In particular, it allows us to model and propagate properly the persistent (or, systematic) effects attributable to the orientation of the force transducer being calibrated relative to the deadweight machine used for calibration. Since many transducers are sensitive to the loading geometry, these effects may contribute significantly to the calibration uncertainty. For this reason, our calibration protocol is designed to express these effects deliberately by changing the orientation of the transducer between different runs of application of the sequence of forces. We recognize the corresponding uncertainty contribution by modeling the transducer as effectively changing every time that it is rotated, with these changes persisting throughout each run, as detailed in §3.2.3 and in §3.3.

Furthermore, and in general, the Monte Carlo uncertainty propagation is immune to non-

linearities in the function that produces estimates of the measurand for given values of the inputs [18]. NIST Technical Note 1900 describes several acceptable alternative methods for uncertainty evaluation, including Monte Carlo methods, and provides many realistic illustrations of their use [27]. It should be noted, however, that the Monte Carlo method would also be an alternative for the method currently in use at NIST to evaluate measurement uncertainty in the calibration of load cells [2].

In Section 3 we point out that the developments we are proposing here parallel very closely the developments that Guenther and Possolo [12] have proposed, and that have already been adopted for routine characterization of the composition of NIST gas mixture certified reference materials.

2. Ordinary Least Squares (OLS) Analysis

For the great majority of the transducers calibrated at NIST, which are designed to be as linear as possible, the calibration function may be taken as a polynomial,

$$R = A_0 + \sum_{i=1}^{p} A_i F^i, \tag{1}$$

that expresses the transducer response R (typically a voltage-ratio) as a polynomial of degree p in the applied force F. The $\{A_i\}$ are coefficients characterizing the transducer-specific relationship between transducer response and applied force. The degree p of the polynomial is usually chosen to be 2 or 3, and it should not exceed 5 as stipulated by ASTM E74-13a, the documentary standard requested for most transducers submitted to NIST for calibration.

It has historically been regarded as the customer's obligation to determine the measurement function, or inverse of the calibration function, to use the calibrated device to measure force. In Section 3 we describe how we may, in the future, also provide the inverse of the calibration function in our reports of calibration, for the customer's convenience, and also the means to evaluate the uncertainty of the measured forces (*cf.* Section 5).

The uncertainty associated with NIST's current force calibration measurement procedure has been described previously [2]. This uncertainty comprises contributions from multiple sources: the applied forces, the measurement of the transducer response, deviations from the assumed transducer response model, and transducer effects such as hysteresis, creep, and sensitivity to loading alignment.

Even though the measurement uncertainty typically varies with the applied force, the measurement uncertainty stated in calibration reports issued by NIST, which is determined in accordance with established guidelines [17; 32], has been generally expressed as a single, conservative value intended to apply over the full range of forces that the transducer is intended to measure.

The coefficients $\{A_i\}$ in Equation (1) are derived from a dataset whereby the responses $\{R_j\}$ of a customer's transducer are measured for each of several applied reference forces $\{F_j\}$, with the forces applied in a sequence in accordance with the appropriate test method. The sequence may be repeated for several transducer orientations relative to the loading platens of the deadweight force standard machine, in order to compensate for transducer sensitivity to loading geometry. Spacing of the forces within a sequence is as uniform as possible within the confines of a discrete set of deadweights. The transducer responses, which are also termed *deflections*, incorporate corrections to the transducer's indication readings, to account for the calibration of the indicating instrument, and for the readings at zero force. For current NIST calibrations the coefficients $\{A_i\}$ are estimated by ordinary least-squares (OLS) using the calibration dataset $\{(F_j, R_j)\}$.

Reliance on an OLS fit assumes that all significant components of uncertainty are associated with the responses $\{R_j\}$, and that the applied forces $\{F_j\}$ have no significant uncertainty. For the majority of force calibrations performed at NIST, the sources of uncertainty associated with the applied force have contributed less than 25 % of the total uncertainty reported for the calibration.

Recent years have seen refinements in transducer technology that challenge this assumption. It is no longer unusual to complete a calibration with a reported relative expanded uncertainty $U_{95\,\%}$ approaching 20×10^{-6}, thus finding a relative standard uncertainty u_f in the NIST applied force (5×10^{-6}) approaching 50 % of the combined standard uncertainty u_c of the calibration.

NIST has refined its force calibration analysis to increase the reliability of the estimates of the reference forces being applied, by accounting for the vertical gradient of the acceleration due to gravity over a deadweight set, and by calculating the air density for each applied weight, using real-time measurements of atmospheric pressure, temperature, and humidity. While this has the potential of reducing the relative standard uncertainty in most NIST applied forces to less than 3×10^{-6}, it is no longer appropriate to ignore the uncertainty in the applied forces during the calculation of the calibration function.

The following section gives a description of a statistical analysis procedure, known as "Errors-in-Variables Regression" [6; 11], that accounts for the uncertainties in both the forces and the transducer responses as it performs a regression over the dataset $\{(F_j, R_j)\}$. The result of the regression is still a polynomial equation; however, its coefficients are computed to minimize the combined deviations along both axes (F and R) between the data points and the calibration function, rather than just along the R axis as is done with OLS.

3. Errors-in-Variables (EIV) Analysis

This section describes an alternative procedure for data reduction that takes into account the uncertainties associated both with the forces and with the deflections, and presents

the results of a simulation study showing that this alternative procedure estimates forces more accurately than the current calibration procedure.

This alternative procedure is not new: it has been in use for quite some time to build calibration functions used to assign values to gas mixtures, and is described in international standard ISO 6143:2001(E) [14]. A variant of this procedure developed by Guenther and Possolo [12] has been adopted by the NIST Gas Sensing Metrology Group for the production of gas mixture Standard Reference Materials (SRMs).

The alternative procedure for the calibration of force transducers accommodates the typical case where different instances of the application of the same deadweight produce different actual forces and associated uncertainties, owing to varying buoyancy corrections called for at different times.

In the following subsections we will first describe the problem that motivates the EIV calibration procedure (§3.1). Then we will discuss the elements of the proposed solution to this problem (§3.2), and describe the solution in general (in §3.2.3). We use a concrete, simple illustrative example (§3.2.4), to describe the conventional, ordinary least squares procedure currently in use (§3.2.5) and the proposed EIV procedure (§3.2.6). Finally (§3.2.7), we compare the two procedures and evaluate their performance (§3.2.8) in the context of that concrete, simple example. Comparisons involving real calibration data will be described in Section 4.

3.1. Problem

The force transducer calibration procedure that has been in use at NIST involves the application of forces that produce deflections (net corrected readings of the indicating instrument, usually expressed as voltage ratios), and then building a calibration function that maps values of force into values of the deflection, up to measurement errors. The calibration function typically is a polynomial of low degree, and its parameters are estimated by the method of least-squares.

In reality, however, the forces have associated (relative) uncertainties that are not negligible when compared with the (relative) uncertainties associated with the deflections. In these circumstances, it is well-known that ignoring the uncertainties surrounding the forces, and estimating the calibration function by ordinary least-squares, produces a biased estimate of this function [6].

3.2. Solution

3.2.1. Forces, Deflections, and Associated Uncertainties Suppose that the calibration experiment consists of applying n forces $F_1, \ldots, F_n$ to the transducer, and observing the corresponding deflections $R_1, \ldots, R_n$. For the modeling undertaken here, it is immaterial whether those forces correspond to multiple applications (*runs*) of the same sets of deadweights or not.

Suppose also that uncertainties associated with each of those quantities are available, as standard measurement uncertainties $u(F_1),\ldots,u(F_n)$ for the forces and $u(R_1),\ldots,u(R_n)$ for the deflections, and that all of them are based on very large numbers of degrees of freedom. For the purposes of this illustration, there is no need to discuss the nature and magnitudes of the components of uncertainty that each of these subsume, but we will review them in Section 5.

3.2.2. Reference Forces and Deflections Since the values of both forces and deflections are clouded by uncertainties, we consider the corresponding, unknown expected values $\varphi_1,\ldots,\varphi_n$ for the forces, and $\rho_1,\ldots,\rho_n$ for the deflections, and assume that the calibration function C relates them as $\rho_j = C(\varphi_j)$ for $j = 1,\ldots,n$. And each of these expected values is related to its observed counterpart according to the following measurement error models: $F_j = \varphi_j + \delta_j$ and $R_j = \rho_j + \epsilon_j$, for $j = 1,\ldots,n$, where the $\{\delta_j\}$ and the $\{\epsilon_j\}$ are independent Gaussian random variables with mean 0 and standard deviations $\{u(F_j)\}$ and $\{u(R_j)\}$. Since these standard deviations typically are roughly proportional to the forces and to the deflections, very often one focuses on their relative values, which are the ratios $\{u(F_j)/F_j\}$ and $\{u(R_j)/R_j\}$.

3.2.3. EIV Calibration Procedure The proposed calibration procedure involves the following five steps.

(EIV-1) Design and execute a calibration experiment that produces a set of n pairs of applied (observed) forces and transducer responses, (F_1,R_1), …, (F_n,R_n). The values of the measured forces include corrections for the buoyancy of the deadweights, based on real-time measurements of air temperature, pressure, and humidity, and also corrections for the vertical gradient of the gravitational acceleration as it impacts the actual force corresponding to the deadweights applied at each test point.

(EIV-2) Evaluate the standard uncertainties of the relevant components of uncertainty (which are described in detail in subsection 5.3): $\{u(F_j)\}$ for the forces and $\{u(R_j)\}$ for the deflections. Each $u(R_j)$ comprises a contribution $u_{\mathrm{v},j}$ from the uncertainty associated with the electrical response of the transducer (including calibration uncertainty for the voltage measurements), and a contribution $u_{\mathrm{b},j}$ reflecting the dispersion of values of transducer responses at the same nominal forces in different calibration runs.

> The dispersion of values captured in $u_{\mathrm{b},j}$ corresponds to differences in orientation of the transducer relative to the loading platens of the deadweight machine. This orientation is altered deliberately between consecutive runs by rotating the transducer by a specified angle (typically 0°, 120° and 240°, in the case of 3 runs). The values of all of these standard uncertainties typically vary between force set-points.

The between-run standard uncertainty $u_{b,j}$ is evaluated statistically (Type A method, in the sense of the GUM 3.3.5) as the standard deviation of the deflections observed in different runs for the same nominal value of the force. This leads to a slightly conservative evaluation of this uncertainty component because the forces actually applied in different runs that correspond to the same nominal force are not exactly equal, but differ owing to changes in buoyancy from run to run.

Since $u_{b,j}$ in fact pertains to the nominal force corresponding to the jth test point, and this nominal force is applied in multiple runs, all test points with identical nominal force will have the same value of $u_{b,j}$. Still, the perturbations applied to the deflections in different runs, that correspond to the same nominal force, are neither identical nor correlated: only the perturbations attributable to this effect that are applied to test points in the same run are correlated, as explained in step (UC-3) under §5.3, and these within-run correlations are duly propagated using the Monte Carlo method.

In sub-section 5.3 we call δ_j the error, with standard deviation $u(F_j)$, that represents the difference between the measured and reference force at the jth test point, and ν_j the error, with standard deviation $u_{v,j}$, that represents the difference between the measured and true voltage ratio. We also introduce an error ω_j to "explain" the difference between the deflection observed at the jth test point, and the deflection that the transducer would yield if it were geometrically perfect and insensitive to orientation relative to the the loading platens of the deadweight machine.

(EIV-3) For each candidate value $p = 1, \ldots, p_{MAX}$ (where usually $p_{MAX} = 5$) of the degree of the polynomial to be used as calibration function C, estimate the values of the coefficients $\boldsymbol{A} = (A_0, A_1, \ldots, A_p)$ of the calibration function $C_{\boldsymbol{A}}$, and the reference values $\boldsymbol{\varphi} = (\varphi_1, \ldots, \varphi_n)$ of the forces, that minimize the criterion $S(\boldsymbol{A}, \boldsymbol{\varphi})$ defined in Equation (3), and compute the values of the model selection criteria AIC (Akaike's Information Criterion), AICc (AIC corrected for the number of data points used to fit the model), and BIC (Bayesian Information Criterion), treating $-S(\boldsymbol{A}, \boldsymbol{\varphi})$ as the logarithm of the log-likelihood function involved in the definition of these criteria [3; 4].

(EIV-4) Determine the degree p of the polynomial that will be used as calibration function based both on the results of (EIV-3), and on graphical examination of residuals $\{R_j - \widehat{\rho}_j\}$ and $\{F_j - \widehat{\varphi}_j\}$ for each candidate value of p, where $\widehat{\rho}_j = C_{\widehat{A}}(\widehat{\varphi}_j)$ for $j = 1, \ldots, n$, and the $\{\widehat{\varphi}_j\}$ are the estimates of the $\{\varphi_j\}$. The smaller the value of the model selection criteria the better the model. Even though we examine both AIC and AICc, we tend to rely mostly on BIC. And the less structured the residuals, also the better the model. In practice, the smallest degree p is chosen for which there has been a marked decrease in the value of BIC, and for which the more preeminent features of the residuals will have clearly abated. Figure 5 supports a model selection exercise where BIC and graphical examination of the residuals are used in tandem to select the most appropriate degree of polynomial to adopt for the calibration function.

(EIV-5) Evaluate the uncertainty associated with the values produced by the calibration function, in a manner that facilitates its use in subsequent uncertainty propagation exercises, in particular to derive uncertainty evaluations for measurements of force made using the calibrated transducer (§5).

In step (EIV-3), the criterion $S(\boldsymbol{A}, \boldsymbol{\varphi})$ is minimized numerically using the global optimizer `genoud` implemented in package `rgenoud` [22; 30] for the R environment for statistical computing and graphics [28]. We have made this choice, which is computationally expensive, because the optimization problem is challenging, involving more adjustable parameters ($n+p+1$) than there are test points (n). The global optimizer mitigates both the sensitivity of the solution to the starting values, and the potential for entrapment in a local minimum of the optimization criterion defined in Equation (3). The starting values for the optimization are the coefficients of the OLS polynomial of the same degree, and the measured forces.

The optimizations that are done repeatedly in the course of the application of the Monte Carlo method for uncertainty evaluation described in §5.2 use R function `nloptr`, defined in the package of the same name [34], using the the "Subplex" algorithm [29] with constraints on how far from the observed forces F_j the estimates of the reference forces φ_j may be (typically within one or two standard uncertainties). For these optimizations, the starting values are the coefficients of the EIV polynomial fitted using `genoud`, and the corresponding estimates of the reference forces. In addition, we have used the following stopping criteria: relative tolerance 1×10^{-7}, maximum number of evaluations of the criterion 5×10^5, and maximum compute time 60 s.

A more drastic solution to convergence difficulties involves replacing F_j by the average of the measured forces over the different runs that correspond to the same nominal force, and inflating $u(F_j)$ to take into account the dispersion of the measured forces that are averaged, for $j = 1, \ldots, n$. In cases with three runs where the same nominal forces are applied in each run, this solution reduces the number of $\{\varphi_j\}$ threefold. This more drastic solution may also make practicable the modeling alternative discussed in 3.3.

3.2.4. Illustration To compare the performance of the conventional (OLS) procedure and of the proposed alternative (EIV) procedure, we will assume that the true calibration function that relates the expected values of the deflections to the reference values of the forces is known, and that it is this particular polynomial: $\rho = C(\varphi) = 0.1 + 3\varphi - 4\varphi^2 + 2\varphi^3$, where the force φ takes values between 0 and 1 (in arbitrary units).

The left panel of Figure 1 shows the graph of C and the locations of $n = 15$ set points used for calibration (large red dots), and also the locations of $m = 14$ points (small blue dots) where the OLS and EIV predictions of force will be compared with reference values. These n set points track the marked curvature of the calibration function fairly well. In calibrations of real transducers, the calibration function is never far from being linear, and typically an even smaller number of different force set points suffices to characterize it.

[Figure 1 about here.]

Since C is strictly increasing for values of the force between 0 and 1, it has an inverse that we denote M, in the sense that $M(C(\varphi)) = \varphi$ for $0 < \varphi < 1$. In general, and also in this particular case (right panel of Figure 1), the inverse of a polynomial function is not a polynomial: refer to `http://wmueller.com/precalculus/newfunc/invpoly.html` for an example. Determining the value $M(\rho)$ when C is a polynomial can be done exactly by solving the polynomial equation $\rho = C(\varphi)$, or it can be done approximately.

Concerning the exact solution, we remind the reader that a polynomial of degree p has p roots (not necessarily all distinct), possibly complex numbers, which may be found using a numerical root-finding procedure (for example, as implemented in R function `polyroot` [28]).

For polynomials used to approximate the calibration functions of real, contemporary transducers, there will be only one real root within the relevant range of applied forces. Approximate solutions may be constructed by evaluating C over a fine grid of values of its argument φ, thus obtaining a set of pairs of values $(\varphi_1, C(\varphi_1))$, ..., $(\varphi_m, C(\varphi_m))$ and then, for example, computing an interpolating spline (smooth, piecewise polynomial function that passes through all of the given points without any unwieldy oscillations between them) [9; 31], and that expresses φ as a function of $\rho = C(\varphi)$ for any value of ρ between the smallest and largest of the $\{C(\varphi_i)\}$. In the case of this illustration, an interpolating spline fitted to the large blue dots (in the right panel of Figure 1) is essentially indistinguishable from M.

To facilitate appreciating the differences between the conventional and alternative calibration procedures, we will entertain relative errors that are much larger than those that are typically encountered in our force transducer calibrations. The calibration set points (large red dots in Figure 1) are at $n = 15$ equispaced nominal force values $\{\varphi_j\}$. The simulated observed forces and deflections are generated according to the measurement error models specified above, with the same (relative uncertainty) 0.075 for both forces and deflections: therefore, $u(F_j) = 0.075\varphi_j$ and $u(R_j) = 0.075\rho_j$ for $j = 1, \ldots, n$. The performance of the OLS and EIV calibration functions will be evaluated based on how accurately they predict deflections corresponding to $n-1$ nominal force values midway between consecutive force set points (small blue dots in Figure 1).

3.2.5. Ordinary Least-Squares Procedure The ordinary least-squares procedure used to determine the calibration function minimizes

$$S(\boldsymbol{A}) = \sum_{j=1}^{n} (R_j - (A_0 + A_1 F_j + A_2 F_j^2 + A_3 F_j^3))^2 \tag{2}$$

with respect to $\boldsymbol{A} = (A_0, A_1, A_2, \text{and } A_3)$, whose true values in this case are $A_0 = 0.1$, $A_1 = 3$, $A_2 = -4$, and $A_3 = 2$, as indicated in 3.2.4.

It is generally preferable to use orthogonal polynomials when fitting a polynomial regression model to data, instead of raw polynomials, because powers of the same variable (force in this case) tend to be highly correlated, and this may make the solution unstable. However, to use an orthogonal polynomials regression model to make predictions in practice requires that the values of the predictor be transformed using the same basis functions that were used when the regression model was built. Since many users will not have either the necessary information about the calibration, or the computational technology required to do this, we use raw polynomials both in this study and in our calibrations.

Irrespective of whether the calibration function is built using OLS or EIV, in practice it is also necessary to select the degree of the polynomial to use, which should be done as described in step (EIV-4) of the EIV procedure specified in §3.2.3. In this comparison of performance of OLS versus EIV, using an artificial example, we fit polynomials of the same degree as the polynomial known to be true, specified in the first paragraph under 3.2.4, that relates expected deflections to reference forces. In general and in practice there is no true polynomial, only an unknown function that one may choose to approximate using a polynomial.

3.2.6. Errors-in-Variables Procedure The errors-in-variables procedure to determine the calibration function minimizes

$$S^*(\boldsymbol{A}, \varphi_1, \ldots, \varphi_n) = \sum_{j=1}^{n} \left(\frac{R_j - (A_0 + A_1\varphi_j + A_2\varphi_j^2 + A_3\varphi_j^3)}{u(R_j)} \right)^2 + \left(\frac{F_j - \varphi_j}{u(F_j)} \right)^2 \quad (3)$$

with respect to $\boldsymbol{A} = (A_0, A_1, A_2, A_3)$, $\varphi_1, \ldots, \varphi_n$, under the assumption that the $\{u(R_j)\}$ and $\{u(F_j)\}$ are known and do not involve uncertainty components that need to be estimated from the data. This simplification suffices for this comparative evaluation centered on an artificial illustration. However, in real calibration experiments the between-run variability is evaluated using a Type A method, as described in 5.1.2.

3.2.7. Comparisons Figure 2 shows the data and the estimates of the calibration function derived from these data by the conventional, OLS procedure ($\widehat{C}_{\mathrm{OLS}}$), and by the errors-in-variables procedure ($\widehat{C}_{\mathrm{EIV}}$).

[Figure 2 about here.]

3.2.8. Performance To compare the performance of the conventional (OLS) and alternative (EIV) procedures, from the viewpoint of the reliability of the estimates of force corresponding to observed deflections, we repeated the following process $K = 10\,000$ times, by taking the following steps for $k = 1, \ldots, K$, always starting from the same set of reference forces and deflections $\{(\varphi_j, \rho_j)\}$ corresponding to the large red dots in Figure 1:

(1) Simulate observed forces as Gaussian random variables $F_{j,k}$ with means φ_j and standard deviations $0.075\varphi_j$, and observed deflections as Gaussian random variables $R_{j,k}$ with means ρ_j and standard deviations $0.075\rho_j$, for $j = 1, \ldots, n$;

(2) Determine the corresponding calibration functions $\widehat{C}_{\mathrm{OLS},k}$ and $\widehat{C}_{\mathrm{EIV},k}$;

(3) Simulate a value $R_{i,k}$ of the observed deflection (corresponding to one of the $m = 14$ small blue dots in Figure 1) for $i = 1, \ldots, m$;

(4) Compute the force corresponding to the least-squares estimate, as the permissible root $\widehat{\varphi}_{i,\mathrm{OLS},k}$ of the cubic equation $\widehat{C}_{\mathrm{OLS},k}(\varphi) = R_{i,k}$, for $i = 1, \ldots, m$;

(5) Do likewise for the errors-in-variables estimate, to obtain $\widehat{\varphi}_{i,\mathrm{EIV},k}$ as the permissible root of $\widehat{C}_{\mathrm{EIV},k}(\varphi) = R_i$, for $i = 1, \ldots, m$;

(6) Compute the root mean squared errors of the two sets of estimates: $\mathrm{RMSE}_{\mathrm{OLS},k} = (\sum_{i=1}^{m}(\widehat{\varphi}_{i,\mathrm{OLS},k} - \varphi_i)^2/m)^{1/2}$, and $\mathrm{RMSE}_{\mathrm{EIV},k} = (\sum_{i=1}^{m}(\widehat{\varphi}_{i,\mathrm{EIV},k} - \varphi_i)^2/m)^{1/2}$.

Figure 3 shows a smooth histogram of the ratio of the root mean squared errors for the two methods, revealing that the alternative, errors-in-variables procedure outperforms the conventional, least-squares procedure 65 % of the time (under the particular conditions considered in this simulation study).

[Figure 3 about here.]

3.3. Alternative EIV Model

In all of the above, we have focused on the EIV model specified in 3.2.2, which recognizes the difference between the observed and expected deflection at each test point j as an error ϵ_j. This error, however, comprises two very different effects: an error ν_j in the measurement of the voltage ratio, and an error ω_j that expresses the effect of the orientation of the transducer relative to the loading platens of the deadweight machine. This effect results from the sensitivity of the transducer to unavoidable minute variations in loading geometry.

Figure 4 shows the run-to-run variability for the two transducers examined in subsection 4.1, depicted as residuals from an OLS regression that expresses the mean value of the deflection as a polynomial of degree $p = 3$ in the nominal force.

[Figure 4 about here.]

Transducer T1 is obviously much less sensitive to orientation than transducer T2. Furthermore, not only are the differences between runs more pronounced for T2 than for T1, they are also more structured, expressing themselves as three perturbed versions of the same, underlying calibration function, and corresponding to the three different placements of T2 on the machine. The alternative approach described below would model this pattern explicitly.

However, we capture the variability apparent in this pattern during the Monte Carlo uncertainty evaluation, as explained in step (UC-3) under §5.3, in the process automatically recognizing the typically rather small number of degrees of freedom with which this source of uncertainty is evaluated. And, of course, the maximum likelihood procedure that we use to fit the EIV model to the calibration data does average these differences that are caused by spurious interactions between the transducer and the deadweight machine.

An alternative modeling approach, which we may implement in future versions of our calibration procedure, explicitly entertains a calibration function that can change from run to run, every time that the transducer is rotated, with these changes persisting throughout each run. In this conformity, we would have three calibration functions for the three runs. However, examination of the scale of the vertical axes in Figure 4 reveals that, although sensibly different, these calibration functions are close to one another, hence could be regarded as perturbations of a single, "average" calibration function that would be assigned to the transducer. The corresponding uncertainty would recognize the fact that generally it will be unknown how a transducer will be mounted for practical use.

In such alternative modeling framework, instead of introducing errors $\{\omega_j\}$ that account for deviations between observed deflections and putative, corresponding expected values, one entertains perturbations to the calibration function itself, that vary unpredictably from run to run, as the transducer is rotated.

To describe what is involved in this alternative approach, consider a calibration function that is a polynomial of the first degree, relating expected deflections to reference forces as $\rho = A_0 + A_1\varphi$, and suppose that the calibration experiment comprises three runs corresponding to three different orientations of the transducer relative to the deadweight machine.

The version of the calibration function corresponding to run $\ell = 1, 2, 3$ is $\rho = (A_0 + B_{0,\ell}) + (A_1 + B_{1,\ell})\varphi$. A_0 and A_1 define the "average" calibration function, while the $\{B_{i,\ell}\}$ define "perturbations" to this calibration function. Even though the values of these perturbations are unpredictable, they persist throughout the run. In the case of transducer T2 such persistence manifests itself in all the deflections changing in the same direction, either up or down, after each rotation.

The definition of this alternative model is completed by characterizing the $\{B_{i,\ell}\}$ as random variables with mean 0, which allows us to think of an "average" calibration function [7; 21]. We would call a model like this a mixed effects model for errors-in-variables regression. The term "mixed" serves as a reminder that the model would comprise both fixed effects (the coefficients A_0 and A_1 of the "average" calibration function) and rotation effects (the $\{B_{i,\ell}\}$ that perturb the $\{A_i\}$). Neither of these sets of coefficients would be used in practice, and only the "average" calibration function would be used. In fact, we recognize (as has been pointed out by one of the anonymous referees), that they would pertain to spurious interactions between the transducer and the deadweight machine used during calibration, and merely provide the modeling backdrop for the uncertainty

evaluation.

Fitting this alternative model poses considerable challenges owing to the large number of parameters involved. For example, the calibration of transducer T2 comprises 3 runs with 11 test points each, hence it yielded 33 data points altogether. Fitting a cubic polynomial according to the EIV procedure described in 3.2.6 requires estimating 33 reference forces plus the 4 coefficients of each third-order polynomial. The alternative model just discussed would require estimating an additional $4 \times 3 = 12$ parameters because there would be 4 "perturbations" for the polynomial coefficients per run. For this reason, we will not use mixed effects models for errors-in-variables regression in this study, and will capture the between-run variability in the errors $\{\omega_j\}$, as described in 5.3. However, the simplification mentioned in the last paragraph of 3.2.3 may make this alternative modeling approach practicable.

4. EIV-OLS Comparison

This section presents comparisons, using data gathered in representative force calibration experiments carried out at NIST, of the existing analysis employing OLS regression and conventional uncertainty evaluation using the GUM technology [17], with the proposed analysis making use of errors-in-variables regression (EIV) and Monte Carlo uncertainty evaluation. Quantities intended for comparison are the calibration regression function polynomial coefficients and the associated uncertainty intervals, as well as the recommended degree of this polynomial.

4.1. Comparison Transducers

Calibration data from two load cell force transducers have been selected for comparison. The calibration measurements for both transducers were conducted in accordance with ASTM E74. In each case the strain gauge excitation and voltage-ratio indicating instruments were provided by NIST, with the electrical calibration of the indicating instruments performed at NIST [13]. Thus a relative standard uncertainty of 5×10^{-6} ascribed to this voltage-ratio instrument calibration [2] is incorporated into the analyses of these force transducers.

The first transducer, identified below as T1, yields deflections exhibiting small deviations from the regression curve, indicative of little sensitivity with respect to orientation about the loading axis of the force standard machine. Thus it is representative of devices for which the uncertainties in NIST's applied forces become significant, making especially relevant the application of techniques such as the proposed errors-in-variables analysis. Transducer T1, with a capacity of 266.89 kN, was calibrated in NIST's 498 kN deadweight machine.

The second transducer, identified here as T2, exhibits larger deviations of the deflections from the regression curve, with a relative standard deviation of the residuals of about

44×10^{-6} as opposed to about 9×10^{-6} for T1. Transducer T2, with a capacity of 111.21 kN, was calibrated in NIST's 113 kN deadweight machine.

4.2. Computation of Reference Forces

The force calibration datasets for both transducers T1 and T2 include readings of the air temperature, barometric pressure and relative humidity acquired during the calibration, from several atmospheric sensing instruments encompassing the deadweights. Thus the air densities at the locations of the weights are calculated for the time of application of each force point in the calibration, enabling real-time adjustments for buoyancy in the calculation of the applied forces. In addition to buoyancy, the individual mass values of the deadweights, as well as the value of the acceleration due to gravity computed at the location of each deadweight, are employed to obtain the best estimates of the reference forces applied to each transducer. The values of the acceleration due to gravity are derived from measurements made previously at several reference points in the laboratory, measured vertical gradient of acceleration due to gravity, and the elevations of the individual weights (taking into account the vertical movement during loading for sequenced weight stacks.)

The estimated reference forces, rather than the nominal force values to which the weight masses were originally adjusted, can now be incorporated into the determination of the calibration regression function for the force transducer. When atmospheric parameters change during the course of the calibration, slight differences in the reference forces can be incurred for subsequent repetitions of the nominal force sequence applied; however, both OLS and EIV can accommodate these reference forces in the computation of the calibration regression function.

The uncertainty intervals for the applied forces are now dominated by the mass uncertainties for the individual weights, with smaller contributions from the uncertainties in the measured values and gradient of the acceleration due to gravity at the reference points, weight elevation, air temperature, barometric pressure, and relative humidity. For the datasets for transducers T1 and T2, the relative standard uncertainties in the individual applied forces generally range from 2×10^{-6} to 3×10^{-6}. For the previous use of nominal forces in the analysis, the relative standard uncertainty in the force was conservatively estimated at 5×10^{-6} for all forces [2].

4.3. Comparison Results

The degree of the polynomial to use for the calibration function is selected based on two criteria used in conjunction: a numerical model selection criterion (we tend to rely mostly on the Bayesian Information Criterion, BIC,[4]), and the examination of plots of residuals (differences between observed and fitted transducer responses) against corresponding forces.

BIC serves to suggest a balanced compromise between model simplicity (the lower the degree of the polynomial the simpler the model) and goodness-of-fit of the model to the data (gauged by the size of the residuals): the smaller the values of BIC, the better the model. The plots of residuals serve to find the smallest degree of the polynomial that produces generally unstructured residuals (in particular, without perceptible linear or non-linear trends).

The procedure of ASTM E74 for determining the best fit suggests a polynomial of order of five for T1 and of order three for T2. The model selection criterion AICc and the residual plots, suggest $p = 3$ for T1; BIC and the residuals plots (Figure 5) suggest the same choice for T2. Therefore, for the comparisons outlined below, the EIV and OLS analyses for both transducers were conducted using cubic polynomials.

[Figure 5 about here.]

The corresponding sets of residuals are shown in Figure 6 for T1, and in Figure 7 for T2. Section 5 explains the meaning of the coverage bands, and how they were computed. The EIV residuals (differences between observed and fitted deflections) are generally smaller in absolute value than the OLS residuals: this is caused by the different weights assigned to different data points in the EIV procedure, according to the respective uncertainties, while in the OLS procedure all the data points are equally weighted.

[Figure 6 about here.]

[Figure 7 about here.]

Figure 8 compares the deflections predicted by calibration functions that are polynomials of the third degree fitted by OLS and EIV to the calibration data for transducer T2. The relative differences are largest in absolute value for the smallest forces but generally insensitive to the orientation of the transducer relative to the loading platens of the deadweight machine. The small sizes of the absolute values of these relative differences suggests that the new calibration procedure does not cast any shadow of doubt on the validity of our previous, high-precision calibrations, which used OLS.

[Figure 8 about here.]

5. Uncertainty Evaluation

The procedures that we use and describe below, to evaluate the uncertainty of calibration results obtained using the EIV approach, are consistent with the GUM and with the GUM Supplement 1, and with similar guidance documents that the NIST Quality Manual for our measurement services [26] requires that we comply with, in particular NIST Technical Notes 1297 [32] and 1900 [27].

We treat all recognized uncertainty components alike, and combine them on an equal footing, irrespective of their provenance or intrinsic nature, and of how they may have been evaluated (Type A or Type B). And we characterize them via fully specified probability distributions (or via samples from these distributions) because this facilitates such uniform treatment.

In particular, we rely on the Monte Carlo method of the GUM Supplement 1 for reasons of convenience and for these three principal orders of reasons: (i) for the conventional calibration procedure, based on ordinary least squares, it produces results identical to the approximation provided by the GUM; (ii) the measurand is a non-linear function of the inputs, and determining the relationship between inputs and outputs involves numerical, non-linear optimization; (iii) only the Monte Carlo method can model faithfully the persistent (that is, systematic), albeit unknown, effect of the typically important differences between calibration runs.

Some of the sources of uncertainty make volatile (often called "random") contributions, while other sources make persistent (often called "systematic") contributions. Provided their values are and remain unknown and cannot be estimated and corrected for, they are modeled as values of random variables, and treated accordingly. The term "random" as used here to qualify "variable" does not mean "chancy" as in a game of chance, but only that a choice has been made to model states of knowledge about unknown quantities using probability distributions.

In this conformity, and for example, the orientation of the transducer being calibrated relative to the loading platens of the deadweight machine is an example of a persistent effect because a change in such orientation may shift the transducer's response up or down at all set-points of the applied force, by unknown and possibly variable amounts, but all generally in the same direction. A contribution that varies sinusoidally with orientation (for example, also relating to differences between runs), but whose phase and amplitude are unknown, would be another example of a persistent effect that could be modeled using random variables. Buoyancy corrections, which depend on measurements of atmospheric temperature, pressure, and humidity, which are done in real-time, are examples of volatile effects.

5.1. Sources of Uncertainty

The uncertainties associated with the applied forces ($\{u_{\mathrm{f},j}\}$), which are diagrammed as variations along the horizontal axis in Figure 2, have been discussed in subsection 4.2. These uncertainties enter directly into the EIV regression computation, whereas in the OLS analysis they only appear as a separate contribution to the final calibration uncertainty, which is incorporated only after the regression has been computed [2].

The uncertainties associated with the deflections, depicted along the vertical axis in Figure 2, have components attributable to (i) the electrical calibration of the indicating instrument ($\{u_{\mathrm{v},j}\}$) and (ii) the lack of repeatability ($\{u_{\mathrm{b},j}\}$) in the different runs of the

calibration experiment. In our traditional analysis employing OLS, these two components have been kept separate, with the first being incorporated into the final uncertainty after the OLS regression is completed, and with the second being expressed in the standard deviation of the residuals about the fitted curve. These two components are discussed next.

5.1.1. Electrical Calibration Uncertainty The uncertainties attributable to the electrical calibration of the indicating instrument, $\{u_{v,j}\}$, represent the uncertainty in determining the instrument's corrections to be applied to its readings in order to yield measurements that are traceable to the NIST programmable Josephson Voltage Standard [13]. The indicating instrument has two channels, one scaled in millivolt and the other in volt, which are sampled together; the channel readings are divided internally by the instrument, with the result returned as a voltage ratio often denoted by the unit "mV/V".

Thus the $\{u_{v,j}\}$, whose relative size has been conservatively evaluated as 5×10^{-6} [2], and that is deemed applicable over the range of use, quantify the typical absolute value of an unknown bias that may be positive or negative, and that affects all of the deflections similarly. We do this in keeping with the choice we have made and explain in the introduction to §5, for how we use random variables and probability distributions to represent and describe effects whose contributions have unknown values, irrespective of whether they are persistent ("systematic") or volatile ("random"), and of whether they are evaluated using a Type A or a Type B method.

Many transducers calibrated at NIST are combined with indicating systems which are not separated from the transducers. Examples are proving rings, which have their own mechanical indicator physically incorporated into the transducer, a proprietary indicator which may be matched to its connected transducer, and a commercial multimeter intended by the customer for use as a load cell's sole indicator. When such a provided indicator is calibrated with the transducer as a system, all the $\{u_{v,j}\}$ are set equal to zero. The calibration of the indicator is then embodied in, and not separable from, the calibration provided for the transducer itself.

It should be noted that the readings of the indicating instrument will contribute other variations to the deflections, which may be associated with the resolution of the indicator as well as with electrical "noise" seen in successive readings made while the applied force to the transducer remains constant. These variations, which are indistinguishable from the transducer related variations discussed in 5.1.2 immediately below, are accounted for in the traditional OLS analysis as contributing to the residuals of the least-squares fit.

5.1.2. Lack of Repeatability Uncertainty The uncertainty component attributable to the lack of repeatability during calibration, and manifest in between-run variability, usually is the largest source of variation in the deflections. In addition to effects of instrument noise and possible residual weight motion during sampling, these variations derive

from imperfect modeling of the transducer response represented by Equation (1), and also from the transducer characteristics of hysteresis, creep, and sensitivity to loading geometry (which is deliberately varied during calibration by reorientation about a vertical axis). These variations have traditionally been characterized in the OLS analysis by the standard deviation of the residuals about the least-squares fit, with corresponding standard uncertainties $\{u_{\mathrm{b},j}\}$ [2].

The relative values of the $\{u_{\mathrm{b},j}\}$ have historically been much greater than the relative values of either the $\{u_{\mathrm{f},j}\}$ or the $\{u_{\mathrm{v},j}\}$, but have been approaching their magnitudes as transducer technology improves. The evaluation of the $\{u_{\mathrm{b},j}\}$ (which typically depend on the nominal force), was explained in step (EIV-2) of 3.2.3.

While Figure 2 employs simulated data with exaggerated uncertainties for illustration purposes, the correct perspective, as seen in the plot in the upper right panel, can be visualized for actual calibration data. The vertical components of the gray line segments will have lengths characterized by a probability distribution comprising the uncertainties described in both 5.1.1 and 5.1.2 above, with a standard deviation that can be very large depending on the characteristics of the transducer. The horizontal components of these gray lines will have lengths limited by a distribution having a relative standard deviation of 5×10^{-6} or less, depending on whether the nominal forces or estimated forces (corrected for changes in buoyancy determined in real-time) are used in the regression.

5.2. Monte Carlo Method

We have developed a Monte Carlo method to evaluate the uncertainty associated with the calibration function, and also to enable the customer of our calibration service to evaluate the uncertainty associated with the force corresponding to an observed response produced by the calibrated transducer. These evaluations are consistent with NIST Technical Notes 1297 and 1900 [27; 32], and with the GUM Supplements 1 and 2 [18; 19]. In fact, the Monte Carlo method implements the parametric statistical bootstrap [10]. The same method could also be used to evaluate the uncertainty associated with the calibration functions traditionally determined using OLS.

The general idea of a Monte Carlo method for uncertainty evaluation [23] involves the application of stochastic perturbations to the measured forces and deflections, and their propagation to the calibration function. Since probability distributions are used to describe the uncertainty of the inputs $\{(F_j, R_j)\}$, the uncertainty of the outputs (calibration function and forces measured by the calibrated transducer) also is characterized by probability distributions. In most practical applications, however, it suffices to provide summary descriptions of the dispersion of values of these distributions, either in the form of coverage intervals, or of standard measurement uncertainties, as will be illustrated below.

Subsection 5.1 reviews the different sources of uncertainty that build up the uncertainty associated with the values of the calibration function. Some of them express persistent effects (sometimes also called “biases” or “systematic effects”), while others express

transient effects (also called "noise" or "random effects").

The orientation of the transducer being calibrated relative to the loading platens of the deadweight machine is an example of a persistent effect because a change in such orientation may shift the transducer's response up or down at all set-points of the applied force, by unknown and possibly variable amounts, but all generally in the same direction, as shown in Figure 4 and in Bartel [2, Figure 5]. This effect will persist during a whole calibration run when the transducer is subjected to a sequence of forces without being repositioned. The dispersion of the different readings of the indicating device corresponding to a constant applied force with the transducer in the same orientation relative to the machine, are an expression of transient effects.

Persistent effects are modeled and propagated differently from transient effects (*cf.* (UC-1) and (UC-3). The collective contribution made by transient effects tends to decrease in size as the number of force set-points, or the number of replicates at each force set-point increases, while the contributions made by persistent effects will not "average out" under increased replication, but will effectively act as a physically meaningful lower bound on the uncertainty associated with the values of the calibration function.

5.3. Calibration Function — Uncertainty Evaluation

We will simulate three sets of measurement errors, already introduced in sub-section 3.2.3, as raw materials needed to evaluate the uncertainty of the calibration function:

- The $\{\delta_j\}$ affect the reference forces and comprise contributions from multiple sources of uncertainty, including those affecting the determination of the mass of the deadweights, the real-time buoyancy corrections, and the estimation of the local acceleration due to gravity;
- The $\{\nu_j\}$ affect the transducer response (voltage-ratio) and represent the uncertainty in the calibration of the instrumentation used to measure voltage-ratios at NIST (these errors, as discussed in 5.1.1, are not in play when the transducer being calibrated incorporates an indicating instrument that is part of the device itself);
- The $\{\omega_j\}$ describe the differences between transducer responses at the same nominal forces applied in different calibration runs, and are attributable in large part to differences in the orientation of the transducer relative to the loading platens of the deadweight machine.

Suppose that, as a result of the data reduction done for the purpose of calibrating a transducer, as described in 3.2.3, based on n pairs of applied forces and transducer responses $(F_1, R_1), \ldots, (F_n, R_n)$, one will have obtained estimates of the corresponding reference values of the forces and of the transducer response, $(\widehat{\varphi}_1, \widehat{\rho}_1), \ldots, (\widehat{\varphi}_n, \widehat{\rho}_n)$.

The following algorithm is computation intensive, and in our implementation it is performed in parallel whenever it is done using a CPU that has multiple cores. It involves

repeating the following steps for $k = 1, \ldots, K$, where K is a large integer (in most cases, $K = 5000$ suffices to produce reliable uncertainty evaluations):

(UC-1) Draw a sample value $\delta_{j,k}$ from a Gaussian distribution with mean 0 and standard deviation $u(F_j)$, and form a replicate of the applied force $F^*_{j,k} = \widehat{\varphi}_j + \delta_{j,k}$, for $j = 1, \ldots, n$;

(UC-2) Draw a sample value z_k from a Gaussian distribution with mean 0 and standard deviation equal to the relative standard uncertainty associated with the voltage ratios, and then compute $u_{\mathrm{v},j,k} = z_k \widehat{\rho}_j$, for $j = 1, \ldots, n$ — these represent errors affecting the electrical calibration of the device used to indicate the transducer response;

(UC-3) The simulation of the $\{\omega_j\}$ is done separately for each run, as follows: First, partition them into groups according to the run they belong to. Suppose that there are L runs, and that run $\ell = 1, \ldots, L$ comprises m_ℓ force set-points. Then, do the following for each of the L runs:

(i) Simulate tossing a fair coin: if "heads" comes up, then all of these errors for this run will be positive; if "tails" then they will all be negative;

(ii) For each force set-point used in this run, draw a value from a Gaussian distribution with mean 0 and with standard deviation equal to the relevant between-run standard deviation $\{u_{\mathrm{b},j}\}$;

(iii) Compute the $\{\omega_{j,k}\}$ for this run as the products of the absolute values of the values drawn in the previous step and either $+1$ or -1 depending on the outcome of the coin toss.

(UC-4) Form a replicate of the transducer response $R^*_{j,k} = \widehat{\rho}_j + \nu_{j,k} + \omega_{j,k}$, for $j = 1, \ldots, n$;

(UC-5) Compute $u_{\mathrm{b},j,k}$ as the standard deviation of the deflections $\{R^*_{j,k}\}$ observed in different runs for the same nominal value of the force, as already described in step (EIV-2) of 3.2.3.

(UC-6) Compute $u^*(R_j) = (u^2_{\mathrm{v},j,k} + u^2_{\mathrm{b},j,k})^{1/2}$, for $j = 1, \ldots, n$

(UC-7) Compute the EIV estimate $\boldsymbol{A}^*_k$ of the coefficients of the calibration function polynomial, and an estimate of the reference forces $\{\varphi^*_k\}$, by minimizing the criterion defined in Equation 3, with the $\{F^*_{j,k}\}$ and the $\{R^*_{j,k}\}$ instead of the measured forces and transducer responses, and with the $\{u(F_j)\}$ and the $\{u^*(R_j)\}$ as their associated standard uncertainties.

This process produces a sample $\boldsymbol{A}^*_1$, …, $\boldsymbol{A}^*_K$ of size K of the vector of the coefficients of the polynomial that defines the calibration function. Next, take the following additional steps:

(a) Given a suitably small resolution Δ_F for the values of force, define a set of equispaced values of force between the minimum F_{MIN} and the maximum F_{MAX} forces used during the calibration experiment: $\widetilde{F}_i = F_{\mathrm{MIN}} + i\Delta_F$, for $i = 1, \ldots, m_F$, where m_F is the largest integer such that $\widetilde{F}_{m_F} \leqslant F_{\mathrm{MAX}}$.

(b) Define a matrix $\widetilde{\boldsymbol{R}}$ with K rows and m_F columns whose entry in row k and column i is $\widetilde{R}_{k,i} = f_{A_k^*}(\widetilde{F}_i)$. Each row of this matrix is a discrete version of the calibration function. Taken together, the rows portray the dispersion of values of the calibration function attributable to the uncertainty components affecting the inputs that determine the calibration function.

(c) Finally, compute a coverage band for the calibration function whose upper and lower boundaries include a specified proportion, typically 95 %, of these K versions of the calibration function in their entirety. For this computation we use the R function `envelope`, defined in package `boot` [5], which implements a technique described by Davison and Hinkley [8, §4.2.4].

Since the boundaries of the coverage band may be somewhat irregular, unless K be impracticably large, we fit regression B-splines to the boundaries, constrained to be increasing and not to crisscross each other or the calibration curve, using R function `cobs` defined in the package of the same name [24; 25].

5.4. Measurement Function — Uncertainty Evaluation

What here we call the *measurement function* M is what ISO [14] calls the *analysis function*: it is the (mathematical) inverse of the calibration function C. Given a value R of the transducer response, it produces an estimate of the force that induced that response, as $F = M(R)$. This is the function that a user of the transducer will require in order to use the transducer to measure forces in practice.

We characterize M by first building a spline s (smooth, piecewise polynomial function [9]) such that $\widehat{\varphi}_j \approx s(\widehat{\rho}_j)$, which approximates the $\{\widehat{\varphi}_j\}$ as a function of the $\{\widehat{\rho}_j\}$ (described above, in 3.2.3). The function s is built using R function `gam` defined in package `mgcv` [33].

Second, given a suitably small resolution Δ_R for the values of the transducer response, chosen so that it is suitable for the use that will be made of the transducer, define a set of equispaced values of this response between the minimum R_{MIN} and the maximum R_{MAX} values of the response observed during calibration: $\widetilde{R}_i = R_{\text{MIN}} + i\Delta_R$, for $i = 1, \ldots, m_R$, where m_R is the largest integer such that $\widetilde{R}_{m_R} \leqslant R_{\text{MAX}}$.

Finally, produce a look-up table of paired values $\{(R_i, F_i = s(R_i))\}$ intended to be used in practice, and then produce a corresponding coverage band similarly to what was done for the calibration function. Figures 9 and 10 depict the measurement functions for transducers T1 and T2, the values of the actual force and of the transducer response that the function has been fitted to, and 95 % coverage bands.

[Figure 9 about here.]

[Figure 10 about here.]

6. Conclusions

From calibration data sets for force transducers calibrated at NIST, the differences between the EIV and OLS calibration regression functions are not large. The differences between the deflections computed from EIV and OLS regressions have been seen to lie within 10×10^{-6} of the deflection at maximum applied force. Figure 8 corroborates this fact and suggests that the new calibration analysis procedure does not put at risk the validity of previous force calibration reporting at NIST which has employed OLS regression analysis.

The standard deviations of the residuals about the regression curves are larger for EIV than for OLS by about 0.6×10^{-6} of the output at maximum force for transducer T1, and 0.02×10^{-6} for transducer T2. For individual force points within the range of calibration, the computed deflections from the EIV and OLS regressions agree within 8×10^{-6} of the output at maximum force for T1, and 3×10^{-6} for T2. This can be seen visually in Figures 6 and 7. Thus while the errors-in-variables regression presents the more proper model of a transducer's calibration function, its adoption appears unlikely to impact a transducer's rating and other market related parameters.

It is noted that, while small, the relative differences between EIV and OLS regression are of the same order of magnitude as the standard uncertainties in the applied forces. For high precision force transducers of recent design, relative uncertainties associated with transducer characteristics are also approaching this low order of magnitude. The EIV regression and Monte Carlo uncertainty methods provide an analytical tool commensurate with the instrumentation now being employed for high precision force metrology.

The uncertainty intervals yielded by the Monte Carlo method provide a more realistic picture of the calibration uncertainty than the traditional method currently employed [2]. In addition, it yields coverage bands representing the uncertainties of the calibration and measurement functions.

The current method provides an uncertainty band for the calibration function, which has a constant value throughout the range of forces applied to the transducer. Implicit in the current method are the assumptions that (a) the residuals are uniformly distributed about the regression function without discernible structure, thus ignoring the systematic effects in the transducer response related to its orientation about the loading axis, and (b) the uncertainties in the applied force and the indicator calibration are conservatively represented by single numbers independent of position within the force range. These assumptions often lead to over-estimation of the uncertainties at the lower end of the range, and possible under-estimation at the upper end. The Monte Carlo method can readily incorporate the variation of these uncertainties over the force range, providing uncertainty bands that more adequately represent the dispersion in the data as seen in Figures 6–7.

Of particular value to the customer employing the calibrated device are the calculation of the measurement function and its respective coverage band, shown in Figures 9 and 10. While computational tools may be readily available to the user for finding the

roots of the calibration function, the estimation of the measurement function uncertainty, incorporating the NIST uncertainty components listed above, may be less tractable and thus especially suitable for inclusion in NIST force calibration reporting. We intend to deliver tabulated values of the measurement function and of the boundaries of the associated coverage band, for a sequence of values of the instrumental indication (typically a voltage ratio) pre-specified by the customer.

For the two transducers used in this comparison, the computations generating the results depicted in Figures 6 through 10 were performed with the R programming environment; conducted on "typical" office desktop personal computers, the computation for each transducer required about 10 minutes, most of which involved the Monte Carlo calculations. Continuing work at NIST is focusing on integration of the new algorithms with existing software employed at NIST for force calibration reporting, with practical implementation of a reporting structure for displaying the results of both the current OLS-based analysis and the EIV and Monte Carlo analyses. This new reporting structure may be utilized for ongoing force calibrations at NIST with a view toward facilitating the adoption of errors-in-variables regression within the relevant documentary standards.

References

[1] ASTM. *ASTM E74-13a, Practice of Calibration of Force-Measuring Instruments for Verifying the Force Indication of Testing Machines*. West Conshohocken, PA, 2013.

[2] T. Bartel. Uncertainty in NIST force measurements. *Journal of Research of the National Institute of Standards and Technology*, 110(6):589–603, 2005.

[3] K. P. Burnham and D. R. Anderson. Multimodel inference: Understanding AIC and BIC in model selection. *Sociological Methods & Research*, 33(2):261–304, November 2004. doi: 10.1177/0049124104268644.

[4] K.P. Burnham and D.R. Anderson. *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. Springer-Verlag, New York, NY, 2nd edition, 2002.

[5] A. Canty and B. Ripley. *boot: Bootstrap R (S-Plus) Functions*, 2013. URL `cran.r-project.org/web/packages/boot/`. R package version 1.3-15.

[6] R. J. Carroll, D. Ruppert, L. A. Stefanski, and C. M. Crainiceanu. *Measurement Error in Nonlinear Models — A Modern Perspective*. Chapman & Hall/CRC, Boca Raton, Florida, second edition, 2006.

[7] M. Davidian and D. M. Giltinan. *Nonlinear Models for Repeated Measurement Data*. Chapman and Hall/CRC, Boca Raton, Florida, 1995.

[8] A. C. Davison and D. Hinkley. *Bootstrap Methods and their Applications*. Cambridge University Press, New York, NY, 1997. URL `statwww.epfl.ch/davison/BMA/`.

[9] C. de Boor. *A Practical Guide to Splines*. Number 27 in Applied Mathematical Sciences. Springer-Verlag, New York, NY, 2001.

[10] B. Efron and R. J. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall, London, UK, 1993.

[11] W. A. Fuller. *Measurement Error Models*. John Wiley & Sons, New York, NY, 1987.

[12] F. R. Guenther and A. Possolo. Calibration and uncertainty assessment for certified reference gas mixtures. *Analytical and Bioanalytical Chemistry*, 399:489–500, 2011.

[13] Y. h. Tang, T. W. Bartel, and J. E. Sims. Ratio calibration of a digital voltmeter for force measurement using the programmable Josephson voltage standard. *NCLSI Measure*, 3(2):70–75, June 2008.

[14] ISO. *Gas analysis — Comparison methods for determining and checking the composition of calibration gas mixtures*. International Organization for Standardization (ISO), Geneva, Switzerland, 2001. International Standard ISO 6143:2001(E).

[15] ISO. *Metallic materials — Calibration of force-proving instruments used for the verification of uniaxial testing machines*. International Organization for Standardization (ISO), Geneva, Switzerland, 2011. International Standard ISO 376:2011(E).

[16] Z. L. Jabbour and S. L. Yaniv. The kilogram and measurements of mass and force. *Journal of Research of the National Institute of Standards and Technology*, 106(1): 25–45, January–February 2005.

[17] Joint Committee for Guides in Metrology. *Evaluation of measurement data — Guide to the expression of uncertainty in measurement*. International Bureau of Weights and Measures (BIPM), Sèvres, France, 2008. URL `www.bipm.org/en/publications/guides/gum.html`. BIPM, IEC, IFCC, ILAC, ISO, IUPAC, IUPAP and OIML, JCGM 100:2008, GUM 1995 with minor corrections.

[18] Joint Committee for Guides in Metrology. *Evaluation of measurement data — Supplement 1 to the "Guide to the expression of uncertainty in measurement" — Propagation of distributions using a Monte Carlo method*. International Bureau of Weights and Measures (BIPM), Sèvres, France, 2008. URL `www.bipm.org/en/publications/guides/gum.html`. BIPM, IEC, IFCC, ILAC, ISO, IUPAC, IUPAP and OIML, JCGM 101:2008.

[19] Joint Committee for Guides in Metrology. *Evaluation of measurement data — Supplement 2 to the "Guide to the expression of uncertainty in measurement" — Extension to any number of output quantities*. International Bureau of Weights and Measures (BIPM), Sèvres, France, 2011. URL `www.bipm.org/en/publications/guides/gum.html`. BIPM, IEC, IFCC, ILAC, ISO, IUPAC, IUPAP and OIML, JCGM 102:2011.

[20] Joint Committee for Guides in Metrology. *International vocabulary of metrology — Basic and general concepts and associated terms (VIM)*. International Bureau of Weights and Measures (BIPM), Sèvres, France, 3rd edition, 2012. URL `www.bipm.org/en/publications/guides/vim.html`. BIPM, IEC, IFCC, ILAC, ISO, IUPAC, IUPAP and OIML, JCGM 200:2012 (2008 version with minor corrections).

[21] N. T. Longford. *Random Coefficient Models*. Oxford University Press, Oxford, UK, 1993.

[22] W. R. Mebane, Jr. and J. S. Sekhon. Genetic optimization using derivatives: The rgenoud package for R. *Journal of Statistical Software*, 42(11):1–26, 2011. URL `www.jstatsoft.org/v42/i11/`.

[23] M. G. Morgan and M. Henrion. *Uncertainty — A Guide to Dealing with Uncertainty in Quantitative Risk and Policy Analysis*. Cambridge University Press, New York, NY, first paperback edition, 1992. 10th printing, 2007.

[24] P. Ng and M. Maechler. A fast and efficient implementation of qualitatively constrained quantile smoothing splines. *Statistical Modelling*, 7(4):315–328, 2007.

[25] P. T. Ng and M. Maechler. *COBS — Constrained B-splines (Sparse matrix based)*, 2015. URL `CRAN.R-project.org/package=cobs`. R package version 1.3-1.

[26] NIST. *NIST Quality Manual for Measurement Services — NIST QM-I*. National Institute of Standards and Technology, U.S. Department of Commerce, Gaithersburg, Maryland, November 2015. URL `www.nist.gov/qualitysystem/`. Version 9.

[27] A. Possolo. *Simple Guide for Evaluating and Expressing the Uncertainty of NIST Measurement Results*. National Institute of Standards and Technology, Gaithersburg, MD, 2015. doi: 10.6028/NIST.TN.1900. NIST Technical Note 1900.

[28] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2015. URL `www.R-project.org/`.

[29] T. Rowan. *Functional Stability Analysis of Numerical Algorithms*. PhD thesis, University of Texas at Austin, Austin, TX, 1990. Department of Computer Sciences.

[30] J. S. Sekhon and W. R. Mebane, Jr. Genetic optimization using derivatives: Theory and application to nonlinear models. *Political Analysis*, 7:189–213, 1998.

[31] Y. N. Subbotin. Spline interpolation. In *Encyclopedia of Mathematics*. Springer & European Mathematical Society, 2002. URL `www.encyclopediaofmath.org/index.php?title=Spline_interpolation&oldid=11892`. Last modified on 7 February 2011.

[32] B. N. Taylor and C. E. Kuyatt. *Guidelines for Evaluating and Expressing the Uncertainty of NIST Measurement Results*. National Institute of Standards and Technology, Gaithersburg, MD, 1994. URL `physics.nist.gov/Pubs/guidelines/TN1297/tn1297s.pdf`. NIST Technical Note 1297.

[33] S. N. Wood. *Generalized Additive Models: An Introduction with R*. Chapman & Hall/CRC, Boca Raton, FL, 2006.

[34] J. Ypma. Introduction to `nloptr`: an R interface to NLopt, August 2014. URL `cran.fhcrc.org/web/packages/nloptr/`. Vignette for R package `nloptr`.

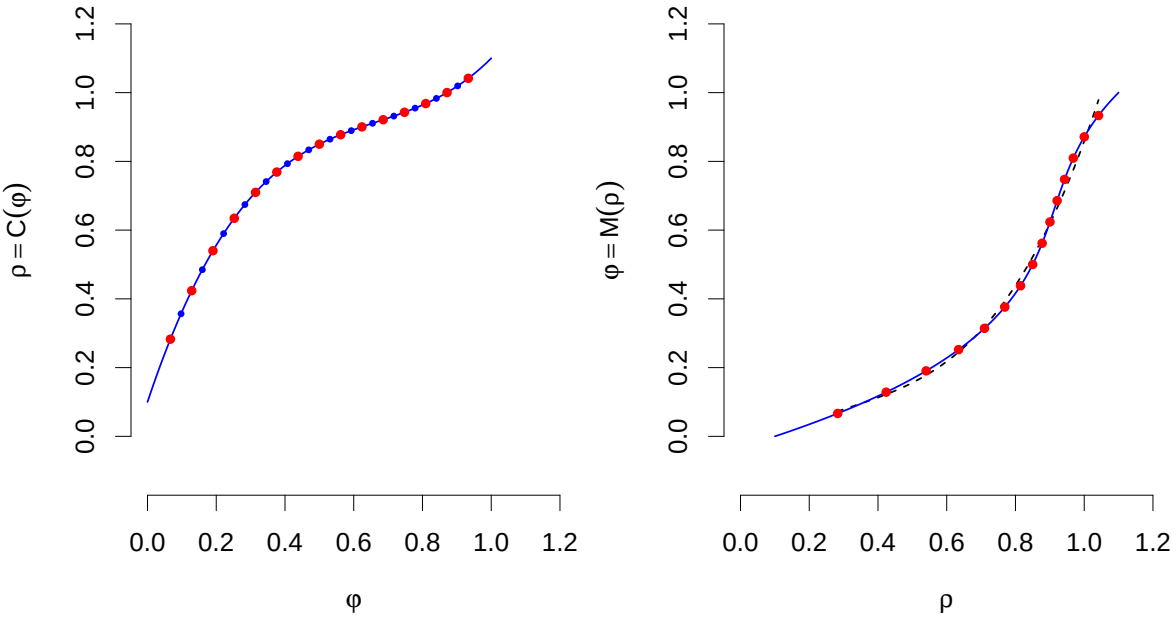

Figure 1. LEFT PANEL: Graph of the calibration function C (solid blue line) and calibration set points (large red dots). RIGHT PANEL: Graph of the measurement function M (solid blue curve), which is the inverse of the calibration function C, calibration set points (large red dots), and third-degree polynomial approximation (dashed black line) to M. The obvious difference between the third-degree polynomial approximation and M shows that the inverse of a polynomial of the third degree generally is not a polynomial of the third degree.

FIGURES 28

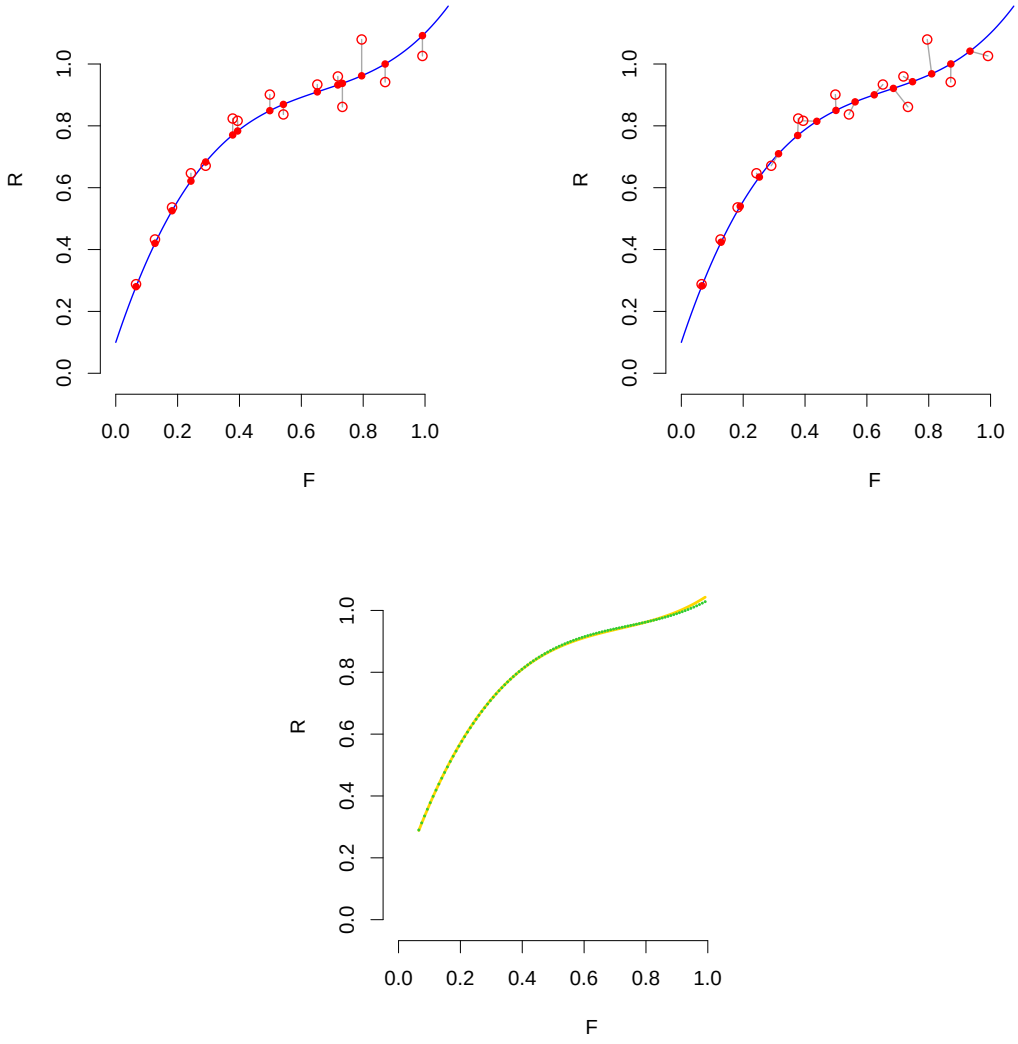

Figure 2. TOP LEFT PANEL: Graph of the true calibration function C (solid blue line), apparent (but incorrect because the corresponding measurement errors are assumed to be zero when in fact they are not) calibration set points (solid red dots), and observed forces and deflections (open red circles). The OLS procedure determines a curve that minimizes the sum of the squared lengths of the (gray) vertical line segments. TOP RIGHT PANEL: Graph of the true calibration function C (solid blue line), true calibration set points (solid red dots), and observed forces and deflections (open red circles), which are joined by (gray) line segments to the corresponding true set points. The errors-in-variables procedure seeks to minimize the sum of squared lengths of the horizontal and vertical components of these line segments, weighted by the reciprocals of the corresponding squared standard uncertainties. BOTTOM CENTER PANEL: Graphs of the OLS estimate of the calibration function (dotted green line), and of the errors-in-variables estimate (solid gold line).

FIGURES 29

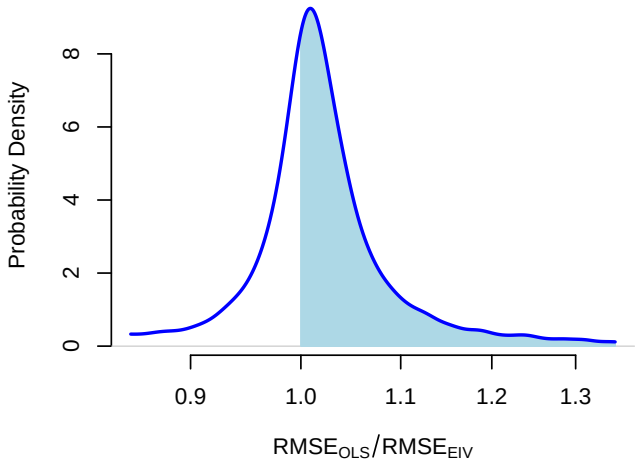

Figure 3. Smooth histogram of the ratio of root mean squared errors when measuring forces using the least-squares and the errors-in-variables calibrations. The shaded (light blue) area under the curve and to the right of 1 amounts to 65 % of the total area under the curve: it is the probability of the errors-in-variables procedure outperforming the least-squares procedure under the conditions considered in this comparative assessment.

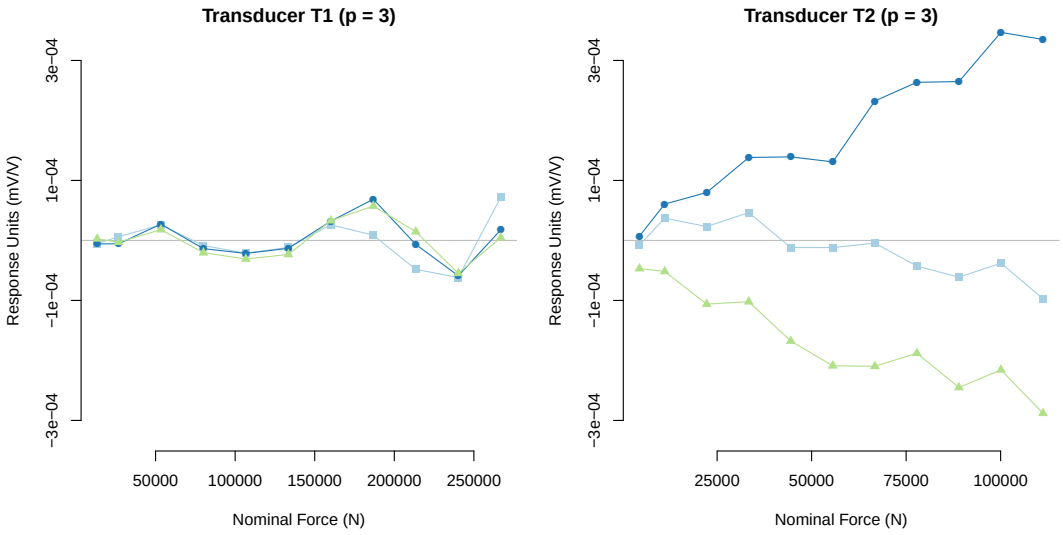

Figure 4. Residual deflections (mV/V) from an OLS regression that expresses the mean value of the deflection as a polynomial of degree $p = 3$ in the nominal force.

FIGURES 31

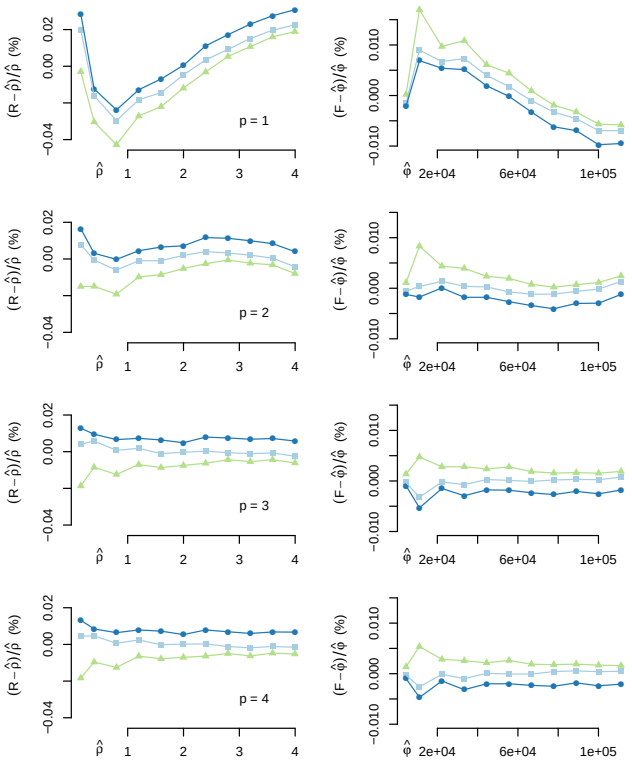

Figure 5. For transducer T2 and polynomials of degree $p = 1, 2, 3, 4$, plots of relative residual deflections $(R_j - \widehat{\rho}_j)/\widehat{\rho}_j$ versus estimated deflections $\{\widehat{\rho}_j\}$, and of relative residual forces $(F_j - \widehat{\varphi}_j)/\widehat{\varphi}_j$ versus estimated forces $\{\widehat{\varphi}_j\}$. For $p = 5$ the patterns of residuals (not shown) are essentially indistinguishable from the patterns for $p = 4$. The residual structure remaining for $p = 3$ reflects differences between runs that persist for all the force set points applied during the same run.

FIGURES 32

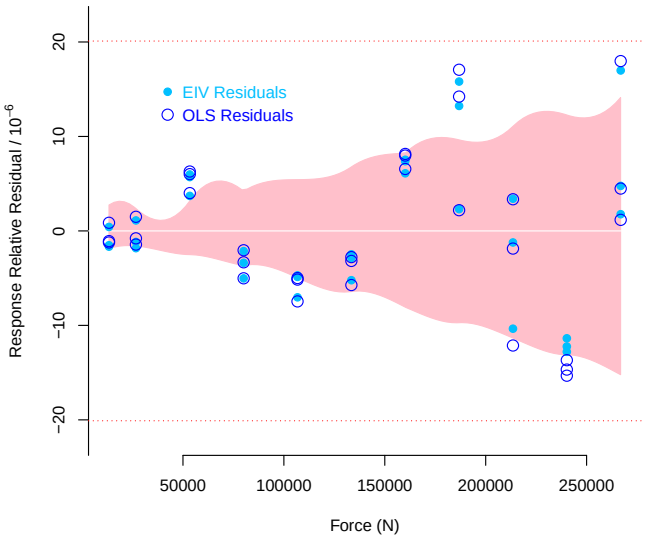

Figure 6. De-trended, simultaneous 95 % coverage band for the calibration function of transducer T1 computed using the Monte Carlo method for uncertainty evaluation, and EIV and OLS residuals relative to average deflection corresponding to maximum applied force. Horizontal dotted lines mark the boundaries of the analogous band corresponding to the present NIST OLS analysis.

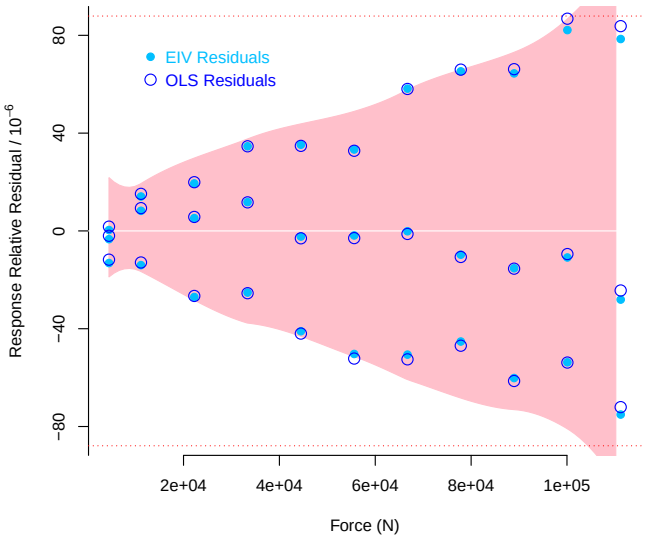

Figure 7. De-trended, simultaneous 95 % coverage band for the calibration function of transducer T2 computed using the Monte Carlo method for uncertainty evaluation, and EIV and OLS residuals relative to average deflection corresponding to maximum applied force. Horizontal dotted lines mark the boundaries of an analogous band corresponding to the present NIST OLS analysis.

FIGURES 34

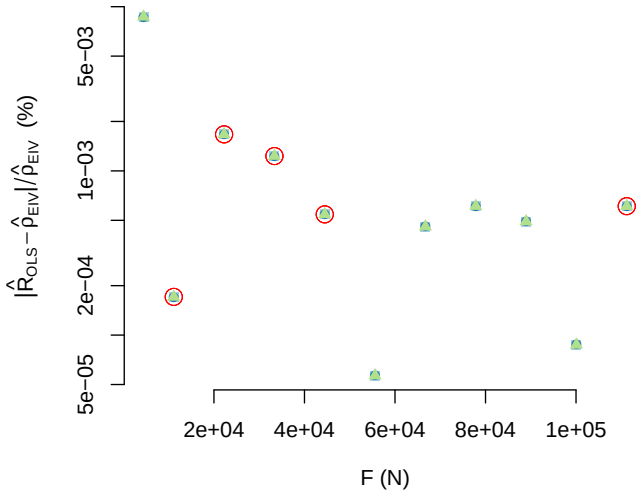

Figure 8. Absolute value of the relative difference between the deflections predicted by cubic polynomials fitted to the calibration data for transducer T2 by OLS and EIV, plotted against the observed values of the forces. The three different plotting symbols indicate the three different runs in the calibration experiment. Note that the scale of the vertical axis is logarithmic. The open (red) circles point out the negative differences

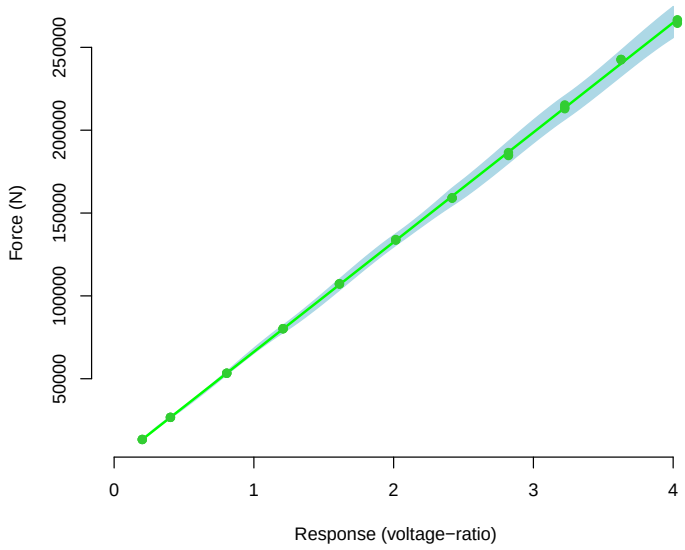

Figure 9. Measurement function for transducer T1 and 95 % simultaneous coverage band computed using the Monte Carlo method for uncertainty evaluation. The thickness (vertical cross-section) of the bands has been magnified 5000 times.

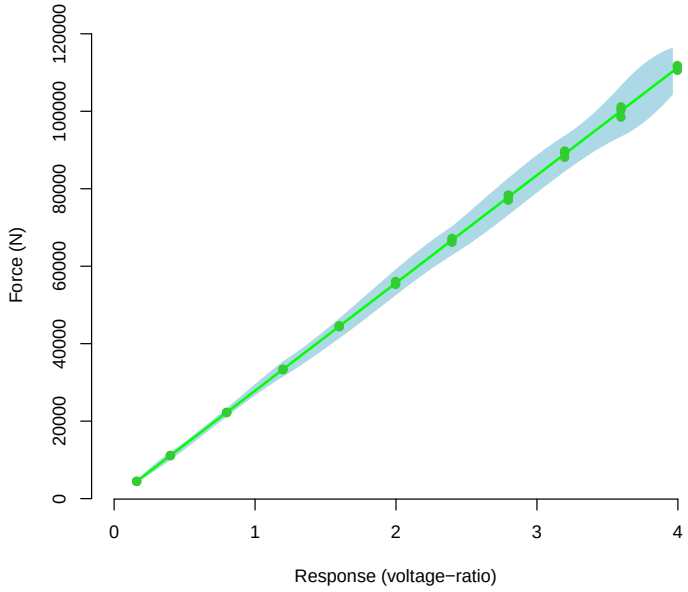

Figure 10. Measurement function for transducer T2 and 95 % simultaneous coverage band computed using the Monte Carlo method for uncertainty evaluation. The thickness (vertical cross-section) of the bands has been magnified 5000 times.