Stepwise Regression Analysis of MDOE Balance Calibration Data Acquired at DNW

Richard DeLoach*
NASA Langley Research Center, Hampton, Virginia 23661

and

Iwan Philipsen†
Instrumentation and Controls Department, German-Dutch Wind Tunnels, Emmeloord, The Netherlands

This paper reports a comparison of two experiment design methods applied in the calibration of a strain-gage balance. One features a 734-point test matrix in which loads are varied systematically according to a method commonly applied in aerospace research and known in the literature of experiment design as One Factor At a Time (OFAT) testing. Two variations of an alternative experiment design were also executed on the same balance, each with different features of an MDOE experiment design. The Modern Design of Experiments (MDOE) is an integrated process of experiment design, execution, and analysis applied at NASA's Langley Research Center to achieve significant reductions in cycle time, direct operating cost, and experimental uncertainty in aerospace research generally and in balance calibration experiments specifically. Personnel in the Instrumentation and Controls Department of the German Dutch Wind Tunnels (DNW) have applied MDOE methods to evaluate them in the calibration of a balance using an automated calibration machine. The data have been sent to Langley Research Center for analysis and comparison. This paper reports key findings from this analysis. The chief result is that a 100-point calibration exploiting MDOE principles delivered quality comparable to a 700+ point OFAT calibration with significantly reduced cycle time and attendant savings in direct and indirect costs. While the DNW test matrices implemented key MDOE principles and produced excellent results, additional MDOE concepts implemented in balance calibrations at Langley Research Center are also identified and described.

Nomenclature

DNW	=	German-Dutch Wind Tunnels (Duits-Nederlandse Windtunnels/Deutsch-Niederländische Windkanäle)
SVS	=	Single Vector System
BCM	=	Balance Calibration Machine
MDOE	=	Modern Design of Experiments
OFAT	=	One Factor At a Time
AoA	=	angle of attack
95% CIHW	=	95% confidence interval half width
a_i	=	accuracy coefficient of the i^{th} component
F_i	=	maximum load of i^{th} component
F_n	=	actual load acting on n^{th} component
$F_{n,\,max}$	=	maximum load of n^{th} component
δ_I	=	tolerated root mean square averaged error of i^{th} component
LLF	=	Large Low-Speed Facility, a DNW wind tunnel

* Senior Research Scientist, Aeronautical Systems Engineering Branch, NASA Langley Research Center, MS 238, 4 Langley Blvd, Hampton, VA 23681, Senior Member.
† Head, Instrumentation and Controls Department, German-Dutch Wind Tunnels, Emmeloord, The Netherland.

ANOVA = analysis of variance
LSD = least significant difference

I. Introduction

Personnel in the Instrumentation and Controls Department of the German Dutch Wind Tunnels (DNW) have been involved in the efforts to improve the calibration methods for internal strain-gauge balances. They have recently investigated the application of MDOE methods for this purpose. The Modern Design of Experiments (MDOE) is an integrated process of experiment design, execution, and analysis proposed at NASA's Langley Research Center to improve quality and productivity in aerospace research.[1] MDOE methods have been applied successfully in numerous disciplines at Langley and elsewhere, including balance calibration. The innovation of a practical single-vector system[2] (SVS) for applying compound calibration loads was an essential enabling technology to permit MDOE test matrices to be executed with dead-weight loading, which features simplicity and low cost per load point. Relative to traditional dead-weight calibration methods that utilize a One Factor At a Time (OFAT) systematic loading schedule, MDOE loading schedules implemented via the Langley Single Vector System have lead to a substantial reduction in cycle time (from three-to-four weeks to two-to-three days) with attendant savings in direct operating costs, as well as improvements in calibration quality.

DNW personnel have now applied MDOE principles to the design and execution of a balance calibration experiment utilizing a different loading system—an automated balance calibration machine. Just as the Langley Single Vector System is able to do, this machine, designated the QinetiQ Balance Calibration Machine (BCM), is capable of applying the combinations of force and moment calibration loads that are required to implement an MDOE balance calibration. It is obviously also capable of executing conventional, non-MDOE loading schedules. As noted above, such conventional test matrices are generally the product of a class of experiment designs known in the testing technology literature as One Factor at a Time (OFAT) designs, because they tend to hold all variables constant except one and vary that one systematically, usually in convenient increments from smallest to largest value in a structured pattern.

DNW personnel recently executed a 734-point conventional OFAT balance calibration experiment using the QinetiQ BCM. They repeated the calibration on the same balance using a loading schedule that featured some of the key aspects of an MDOE calibration design. As this was an initial foray by DNW into MDOE calibration methods and essentially a learning experience, not all of the features of a full MDOE calibration were included in this first effort. The DNW MDOE design focused primarily on the MDOE imperative to minimize data volume as a means of reducing cycle time and direct operating costs. The selection of independent variable combinations that minimize uncertainty in the balance calibration math models was not attempted in this initial experiment, although a key quality assurance tactic in MDOE operations was investigated; namely, the act of randomizing the loading order of the test matrix.

MDOE practitioners randomize the set-point order of their test matrices to defend against slow, systematic variations that may be in play during an experiment.[3,4] Such variations can be due to any number of factors, but among the most common are instrument drift, warm-up effects, thermal expansion, and even operator fatigue (or operator learning effects, by which the operator's performance improves after an initial acclimation period). These variations have in common the fact that they are not random, so that ordinary replication will do little to cancel them out. They are troublesome because if the independent variables are also changed systematically over time, there is no way to separate the effects of independent variable changes from the effect of the systematic error source. We say in such circumstances that the independent variable effects are "confounded" by systematic error. Furthermore, they are "stealthy," in that they are difficult to detect because of their gradual nature.

Experience at Langley Research Center suggests that no matter how carefully a test is conducted, systematic long-term variations are not unlikely to occur in a given experiment. The systematic component they generate in the unexplained variance of an ensemble of test data invariably dominates the random component for any test that takes place over a sufficiently long period of time for these effects to come into play. Because systematic errors are so difficult to detect, and because they can have such a serious impact on the reproducibility of experimental results when they are in play, MDOE practitioners proactively defend against them by randomizing the set-point order of the test matrix. This makes each independent variable level is equally likely to be set early as late. When systematic changes are in play, the error at each such point is then equally likely to be positive as negative relative to the mean of all the data. Randomizing the set-point order therefore has the effect of converting systematic variations into simply another component of random error, which is easy to detect and easy to minimize by replication.

There is generally some extra work associated with randomizing a test matrix, in that the operator must pay closer attention if the next increment and direction of change for a variable can be either up or down, and can be of

differing increments. However, MDOE practitioners regard randomization as necessary insurance against systematic effects that may be in play, and which can be quite serious if they are. The peace of mind that is achieved by the knowledge that one is defended against such effects is generally regarded as worth the small additional effort of executing the test matrix in a non-systematic way.

There are practical restrictions on randomizing certain variables. For example, Reynolds number in a cryogenic wind tunnel is not practical to randomize because of the long time required to transition from warm to cold operating temperatures, not to mention the high cost of the liquid nitrogen used to achieve these transitions. When there are such practical restrictions on randomization, a special class of experiment designs is usually invoked to defend against systematic unexplained variance. This class of designs includes "Split Plot Designs" and other design types. The subject of tactical quality assurance under restrictions on randomization is beyond the scope of this paper, however this is not an issue for balance calibrations with a BCM, for which one loading sequence is generally as easy to implement as another.

On the other hand, it can be argued that over a sufficiently short period of time, systematic effects may be negligible simply because there is insufficient time for substantive changes to develop. Indeed, this is the basis for another MDOE quality assurance tactic called "blocking," in which the system is assumed to be stable over short "blocks" of time. An MDOE analysis can remove "block effects"—changes in mean system response from one block of time to another—which significantly reduces the unexplained variance and therefore improves resolution. Blocking is beyond the scope of the current paper as it was not implemented in this MDOE design, but it is discussed elsewhere in the context of aerospace ground testing.[5,6]

From a practical implementation perspective, the question then becomes, "How short a time interval is short enough to safely abstain from randomization?" It is considered good experimental practice to circumvent this question by simply randomizing whenever it is practical to do so. This insures the experiment against potentially serious systematic effects that may or may not be present, and which may or may not be large enough to be of concern if they are. However, for the purpose of examining MDOE methods, the DNW instrumentation engineers executed the load schedule of their initial MDOE experiment both in random order and in sequential or systematic order.

Thus, for this study there were three calibrations performed on the same balance using the same loading hardware. The balance calibrated in this study is a six-component moment type balance pictured in Fig. (1). The first of the three calibrations was a 734-point conventional OFAT calibration, which was followed by a series of simulated polars and other loadings in a 114-point series of confirmation points that could be used to test the results of the balance calibration. (The balance calibration math models should be able to predict these confirmation points within acceptable error limits.) This was followed by a loading schedule with a significantly smaller data volume—101 points—to test the MDOE hypothesis that conventional load schedules contain substantially more data than are necessary to achieve acceptable results. To assess the effects of randomization in this experiment, his smaller load schedule was executed twice, once in a systematic, sequential order and once in a randomized order. Certain other differences in the two smaller load schedules will be discussed.

The four sets of data—the OFAT data, the confirmation point data, and the two MDOE data sets—were all transmitted to Langley Research Center for analysis using MDOE response surface modeling techniques to define optimum calibration math models for each data set and to compare the results of the three experiment designs on the basis of accuracy and precision. A total of 18 calibration models were developed, one for each of the six balance outputs, with each of the three sets of calibration data. Each data set was used as confirmation points to evaluate the other two, in addition to the 114 confirmation points. A formal analysis of variance was performed to test for significant differences in accuracy and precision from one experiment design to the next, and from one balance component to the next.

The next section provides some tutorial background on the mechanisms by which MDOE experiment designs achieve improvements in quality and productivity. The balance used to acquire all three data sets is then discussed as is an overview of the balance calibration machine used to apply the loads. The test schedules are further discussed, as is the data analysis method employed. A brief summary and concluding remarks follows a discussion of the results.

II. Role of the Design Matrix in the Modern Design of Experiments

MDOE methods achieve productivity and quality improvements by optimizing an experiment's design matrix. The design matrix plays such a critical role in the MDOE method that this section will describe it in some detail, and will discuss its relationship to quality and productivity.

The design matrix is an extension of the familiar test matrix that adds columns corresponding to terms in a math model used to predict responses as a function of independent variable levels. For example, consider a simple experiment in which some response, y, is a function of two independent variables, ξ_1 and ξ_2. This might be a wind tunnel experiment, for example, for which y is lift, say, and ξ_1 and ξ_2 are angle of attack (AoA) and Mach number. Under mild conditions that hold under a wide range of practical circumstances, the unknown lift function can be represented by a Taylor series expansion about some reference Mach number and angle of attack; cruise Mach and AoA, say. Let ξ_{01} and ξ_{02} represent these reference levels. Theoretically, all the terms in an infinite Taylor series would be required to represent the lift function exactly, but if the expansion is carried out within a sufficiently small neighborhood of ξ_{01} and ξ_{02}, the lift function might be adequately represented by only the first few terms in such a series. In Eq. (1), all terms of order 3 and higher in such a Taylor series are represented by R_3, which we assume are small enough to drop.

$$\begin{aligned} y(\xi_1,\xi_2) = y(\xi_{01},\xi_{02}) &+ \frac{\partial y}{\partial \xi_1}(\xi_1-\xi_{01}) + \frac{\partial y}{\partial \xi_2}(\xi_2-\xi_{02}) \\ &+ \frac{\partial y}{\partial \xi_1 \partial \xi_2}(\xi_1-\xi_{01})(\xi_2-\xi_{02}) + \frac{\partial^2 y}{\partial \xi_1^2}(\xi_1-\xi_{01})^2 + \frac{\partial^2 y}{\partial \xi_2^2}(\xi_2-\xi_{02})^2 + R_3 \end{aligned} \tag{1}$$

All derivatives in Eq. (1) are evaluated at ξ_{01} and ξ_{02}. Consider now a simplifying change of variables:

$$x_i = k_i(\xi_i - \xi_{0i}) \tag{2}$$

where the k_i are scaling constants introduced to circumvent certain practical problems in floating-point electronic computations and for other reasons. Neglecting third-order terms and higher, Eq. (1) can then be expressed as a simple polynomial function:

$$y = b_0 + b_1x_1 + b_2x_2 + b_{12}x_1x_2 + b_{11}x_1^2 + b_{22}x_2^2 \tag{3}$$

where b_0 is the value of y when all $x_i = 0$ and the other b values are constants that depend on the derivatives of y[‡].

We typically estimate the b values by measuring y at a number of combinations of ξ_1 and ξ_2. After performing the transformations indicated in Eq. (2), we then use regression methods to fit the math model in Eq. (3) to the data. The ξ_1 and ξ_2 combinations comprise the *test matrix* familiar to all experimentalists. The *design matrix* is a straightforward extension of the test matrix in which additional columns are added so that there is one column corresponding to each term in the math model that is proposed to represent the response function of interest, y.

To illustrate, assume that we have the simple nine-row by two-column test matrix that consists of the shaded entries in Table 1, where the scaling transformation of Eq. (2) has been applied to produce the array of values indicated. A column of 1s is added to the left of the test matrix under the Intercept column header and additional columns in Table 1 are derived from the x_1 and x_2 columns of the test matrix by performing the operations indicated in the column headers. The entries in Table 1 comprise a nine-row by six-column matrix we will designate by **X**, and call the *design matrix*. The design matrix depends on the schedule of independent variables selected for the test matrix, and it also depends on the math model chosen to represent the response as a function of the independent variables. While Table 1 is based on a nine-point test matrix used to fit a second-order polynomial in two variables, the design matrix can be generalized in a straightforward way to accommodate any number of data points, any order of math model, and any number of independent variables, so quite a wide array of practical experimental circumstances can be represented in this way. For example, consider the design matrix for a conventional 734-point OFAT balance calibration experiment intended to generate second-order calibration models in six load variables. It will have 734 rows and 28 columns, one row for each data point and one column for each term in the proposed math model.

[‡] Note that in balance calibration experiments, this polynomial would be a 28-term function of the six loading variables and b_0 would require certain additional iterative procedures to account for the fact that an unloaded balance still experiences applied forces due to the weight of the loading hardware and the balance itself. These procedures are beyond the scope of the current paper, but are addressed in detail in the current AIAA Recommended Practice on balance calibration.[7]

Table 1. Design matrix for second order function of two variables, generated from the test matrix.

Intercept	x_1	x_2	x_1x_2	x_1^2	x_2^2
1	0	0	0	0	0
1	-1	1	-1	1	1
1	-1	-1	1	1	1
1	1	-1	-1	1	1
1	1	1	1	1	1
1	0	√2	0	0	2
1	√2	0	0	2	0
1	0	-√2	0	0	2
1	-√2	0	0	2	0

When it is expressed matrix-vector form, the polynomial response function math model is a function of the design matrix, **X**, and has the same form regardless of the order of the model or the number of independent variables.

$$\hat{\mathbf{y}} = \mathbf{Xb} \tag{4}$$

where "y-hat" is a vector of predicted response values, **b** is a vector of model coefficients, and **X** is the design matrix as in Table 1.

For a given vector of measured response measurements, **y**, the vector of regression coefficients, **b**, depends only on the design matrix:

$$\mathbf{b} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y} \tag{5}$$

From Eqs. (4) and (5), we have

$$\hat{\mathbf{y}} = \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y} \tag{6}$$

Predictions made by a model developed from experimental data will have uncertainty that reflects the experimental error in the data. For a model developed from a set of data with an intrinsic variance, σ^2, the resulting variance in model predictions depends, again, entirely on the design matrix:

$$\mathrm{Var}\left(\hat{\mathbf{y}}\right) = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\sigma^2 \tag{7}$$

Equations (6) and (7) reveal the central role of the design matrix in determining the quality of experimental results obtained from a given sample of data. Equation (6) shows that the accuracy of response model predictions depends entirely on the design matrix, while Eq. (7) indicates that the precision of those predictions is likewise determined by the design matrix.

The details of the design matrix can also impact the quality of research results in other ways. For example, there is theoretically an infinite degree of uncertainty in estimating the regression coefficients when any two or more columns in the design matrix are linearly dependent.[8] In such a case, the regression coefficients quite literally can each be "anything," and the regression model is meaningless. While perfect linear dependencies are rare, in practical circumstances there can be some degree of dependence that degrades the quality of model predictions.

The uncertainty in estimating model coefficients is minimized when each column in the design matrix is orthogonal to all others. Perfect orthogonality between two columns is achieved when the products of corresponding

terms in each column sum to zero.[§] Note that the design matrix in Table 1 is perfectly orthogonal—the sum of products of corresponding terms from every pair of columns is zero. This is by design. (Table 1 represents a special class of experiment designs called Central Composite Designs, which for two independent variables and a second-order model feature this perfect orthogonality.) The uncertainties in estimates of the regression coefficients of a fitted model are as small as possible under these conditions. That is, any departure from orthogonality among the columns would result in an inflation of the variance in coefficient estimates relative to the orthogonal case. While perfect column orthogonality is often elusive in design matrices that must satisfy multiple practical constraints, it is regarded as a highly desirable trait of a good experiment design.

The design matrix also influences the productivity of experimental research, as well as the quality. The number of rows defines the volume of data to be acquired, which influences direct operating costs and what is often more important, cycle time.

The Modern Design of Experiments focuses on optimizing the design matrix for two important reasons: 1) the design matrix the central determinant of both quality and productivity in a response modeling experiment as noted above; and 2) the details of the design matrix are entirely within the researcher's control. For calibration data acquired in a given measurement environment, the productivity and quality of a strain gage balance calibration depends on four factors, all of which are reflected in the design matrix and all of which are under the control of the researcher. These factors are as follows:

1) **The number of calibration points.** A minimum of one data point is required for every term in the calibration math model of a given balance component, but the more points there are beyond this minimum, the higher the quality of the calibration. That is, the precision of a calibration model prediction tends to increase with the volume of data fitted to produce the model. There is a tradeoff between cost and benefit, however, as each additional point requires additional time to acquire and analyze, and entails additional direct operating costs as well. Furthermore, it can be shown that the incremental value of each additional data point is a monotonically decreasing function of the volume of data already in hand—a data point is much more valuable when little is known about a system than when much is already known. While the value of each new point declines as more data are acquired, the cost of acquiring each point remains about the same (discounting the amortization of set-up costs, which are often negligible compared to cycle-time costs). This means that there is a point of diminishing returns beyond which the value added by the next new point falls below the cost of acquiring it, so that there is an upper limit on the volume of data that optimizes the cost/benefit ratio of experimental research. We wish to optimize the design matrix by specifying an ample number of rows to meet precision requirements, but no more than that in order to preserve resources.
2) **The selection of calibration points.** There can be constraints on the combination of loads that it is possible to set simultaneously. For example, there may some maximum total load that cannot be exceeded due to a limitation in the loading system. In the case of a single vector loading system using dead weights,[2] the total force vector and total moment vector are orthogonal[9] so that the dot product of these two vectors is zero. This results in a constraint on three force-moment products consisting of a force and the moment about the force axis. All three such pairs must sum to zero, meaning that only five loads can be set independently with such a system. One degree of freedom is lost to the constraint so that the sixth load is determined when any five of the other loads are set. Within such constraints, however, the calibration engineer has considerable latitude in selecting load combinations once the total number of points has been determined. The uncertainty in estimating math model coefficients depends on the selection of load combinations in the calibration data set. This means that for a given calibration model, the elements of the design matrix can be specified to minimize prediction uncertainty.
3) **The order in which the data are acquired.** Randomizing the set-point order of a test matrix has been discussed as a quality assurance tactic when systematic (non-random) variations in response measurements may be in play. Slowly varying, persisting effects—instrument drift, thermal changes, and the like—cause such systematic errors, which cannot be distinguished from legitimate independent variable effects. Randomizing the set point order ensures that all measurements in a time series are statistically independent, a condition for obtaining meaningful results by regression analysis. This implies that the order of the rows in the design matrix will affect the quality of the experimental result.

[§] Recall from vector algebra that the dot product of two vectors is proportional to the cosine of the angle between them, and equal to the sum of products of corresponding elements. When this sum is zero, the vectors are at right angles and are therefore orthogonal.

4) **The math model to which the data are fitted.** There is a tradeoff in accuracy vs. precision that is driven by the number of terms retained in the math model, and therefore the number of columns in the design matrix. As noted earlier, the response of a given system can be represented by an infinite Taylor series. For balance calibrations, this is usually truncated to second order. Retaining higher order terms may increase the accuracy of the math model, but since each term carries with it some uncertainty, the total uncertainty degrades with the number of terms in the model. In fact, it can be shown that the prediction variance described in Eq. (7), averaged over all points used to fit the math model, is directly proportional to the number of terms in the model.[10] Therefore, the fewer terms in the math model the better, as long as there are an adequate number of terms of sufficient order to predict responses within specified requirements.

We have seen that an experienced MDOE practitioner can adjust the rows, columns, and cells of the design matrix, as well as the order of the rows, to maximize quality and productivity in a response modeling experiment such as a balance calibration. By contrast, the number of columns in the design matrix of a typical OFAT balance calibration experiment is often fixed by historical convention at 28—the maximum number necessary to fit a full second order model in six load variables. The possibility of reducing the number of columns in the design matrix of a balance calibration experiment has only recently been introduced at Langley Research Center through the application of MDOE methods, and independently at Ames Research Center with the introduction of new methods for automatically selecting math models.[11-13]

Likewise, the number of rows and the cell contents of a typical OFAT balance calibration design matrix tend to be dictated by other considerations besides productivity and prediction quality. For example, the OFAT design developed at Langley Research Center in the mid-1950s and used since then with very little change, was originally created to facilitate graphical estimates of the response model coefficients. This technique simply reflects the computational state of the art of that era.[14] The design matrix for this calibration features 729 rows to accommodate 81 load sequences executed systematically over time. Each sequence consists of a tare point, four increments, three decrements, and a return tare point, for a total of nine data points per sequence. No thought is given in this design matrix to the possibility of systematic, slowly varying effects of the kind that MDOE randomization defends against.

As noted earlier, this was an initial exploratory examination of MDOE design principles by DNW and a comprehensive optimization of the design matrix was not attempted. DNW compared a 734-point OFAT design with an MDOE design that differed primarily in the number of rows. Also, an MDOE-related analysis was performed at Langley Research Center, in which insignificant terms in the math models for the six balance outputs were deleted. This resulted in an ex post facto optimization of the number of columns. While both the OFAT and the MDOE results met quality standards quantified a priori, even better results might be achievable in future calibrations if the response models developed in this exercise are used from the beginning to design future calibrations of this balance. The DNW MDOE design did incorporate randomization as a quality assurance tactic, but not blocking. The design was also executed in standard (non-randomized) order for comparison.

III. Accuracy and Precision Requirements

There were quantitative accuracy and precision requirements established a priori for the calibration of this balance. These will be described here, as well as the criteria used for assessing the quality of the calibration.

A. Accuracy

The accuracy requirement of this balance is defined by the following formula[15,16]

$$\delta_i \le \frac{1}{1000} \times \left|F_i\right| \times \left[a_i + \sum_{\substack{n=1 \\ n \ne i}}^{6} \left| \frac{F_n}{F_{n,\max}} \right| \right] \tag{8}$$

Where

δ_I = tolerated root mean square averaged error of i^{th} component
F_i = maximum load of i^{th} component
a_i = accuracy coefficient of the i^{th} component (1 for all i in this study)
F_n = actual load acting on n^{th} component
$F_{n,\,max}$ = maximum load of n^{th} component

This same accuracy requirement was adopted for the response models in this study. That is, Eq. (8) is assumed to describe the largest permitted departure of predicted response measurements from measured responses. An independent estimate of the absolute accuracy of individual measured responses is beyond the scope of this paper, which focuses on the ability of math models developed from physical measurements acquired with different experiment designs to adequately predict those measurements. That is, we assess accuracy in terms of our ability to reproduce the *data*, and assume for the purposes of this study that there is no difference between the measured and true responses. We therefore neglect any bias errors that may exist in the calibration machine.

B. Precision

There is a repeatability requirement for the balance defined as "about 1/3 of its accuracy." In this study we assess repeatability, or precision, by quantifying the widths of the 95% confidence intervals associated with each prediction, based on Eq. (7). The 95% confidence interval half-width (95% CIHW) at points used to fit the model is the product of the standard error in the prediction (the square root of the prediction variance in Eq. (7)) and a t-statistic reflecting the number of degrees of freedom used to assess the unexplained variance:

$$95\%\ \text{CIHW} = t_{n-p,0.025}\sqrt{\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'}\sigma \tag{9}$$

where n is the number of data points used to fit a given model, p is the number of terms in the math model, σ is the standard deviation of the fitted data, and $\mathbf{X}$ is the design matrix. Note that for a "statistically significant" number of degrees of freedom (say, $n - p > \sim 20$), the t-statistic has a value of about 2, so that the 95% CIHW is interpreted in the familiar "2σ" sense, with "σ" modified for the case of response predictions by a factor of the square root term in Eq. (9). The influence of this term is illustrated in Fig. 2. Here, a response surface plot shows how standard prediction errors are distributed when $\sigma = 1$. For any combination of the two independent variables plotted, the actual confidence interval half width can be obtained by multiplying the z-axis by the product of the standard deviation in the response data and the t-statistic from Eq. (9). (Figure 2 shows the unit standard errors plotted as a function of F_x and F_y, but by symmetry the error distribution would be similar for any other combination of load variables). This figure illustrates how the square root term in Eq. (9) is responsible for a multidimensional "bow tie" effect, in which the standard errors are smaller near the center of the design and larger near the boundaries. Outside of the boundaries is an extrapolation region where the prediction uncertainty continues to grow rapidly. Only the area within the original range of fitted variables is the model considered valid.

Note also that Eq. (9) applies to the points used to fit the model. We can generalize that to other load combinations as follows:

$$95\%\ \text{CIHW} = t_{n-p,0.025}\sqrt{\mathbf{x_0}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{x_0}'}\sigma \tag{10}$$

where $\mathbf{x_0}$ is a vector corresponding to any row in the design matrix but with a specified combination of loading levels that may be different than any of the points used to fit the model.

IV. Calibration Experiment Designs

Three calibrations were performed in succession. The first one can best be described as a One Factor A Time (OFAT) load table with additional independent data to check the validity of the calibration. The second and third are designated as MDOE sequential (or structured) and MDOE randomized.

A. The OFAT Load Table

The OFAT load table begins with a first order calibration of pure loads, followed by a second-order calibration using the following procedure: Combined loads were applied for all 15 combinations of two loads. First, one load component was held constant at +75% of its load range and the other load component was varied from 0 to +75% and from 0 to –75% of its load range. Then the one load component was held constant at –75% of its load range and the other load component was varied from 0 to +75% and from 0 to –75% of its load range. After this the second load component was held constant at +75% and –75% of its load range and the first load component was varied from 0 to +75% and from 0 to –75% of its load range. The first order calibration required 253 load points (about a third of the total) and the second order calibration took 481 load points (about two-thirds).

Since the BCM allows simultaneous application of six load components, it was used to simulate a wind tunnel measurement of a typical DNW-LLF model. The simulated conditions represented a weight polar, a pitch polar, a yaw polar, the maximum load condition, and a repeat of the yaw polar. The load table was finally ended with a single component first-order calibration repeat of the axial component. The maximum load point is the maximum simultaneous load condition, 90% of the FS load range on all six components.

B. The MDOE Sequential (Structured) Load Table

The MDOE sequential (structured) load table consists of 101 points. The selection of loading combinations was not designed to exploit any of the MDOE techniques for minimizing uncertainty, but focused instead on reducing data volume significantly relative to the OFAT load table, with attendant reductions in cycle time and direct operating costs. The load table was derived from the OFAT load table out of convenience, and simply exploited the fact that to fit a pure second order model it should be sufficient to use only three load points for each component out the OFAT load table.[16] The maximum positive and negative loads on each component were chosen, as well as the zero load for that component, in order to minimize the influence of load errors on the slope fits. In addition, extra zero load points were measured to quantify repeatability.

Only one pair of loadings is necessary to fit each two-way interaction term in a full second-order model, but there are several reasons to acquire more data. The behavior of a moment type of balance is not always consistent in all four quadrants, for example, so it is prudent to acquire compound loads in each quadrant. If this compound loading consisted of the maximum of the two component loads, the resulting total load might exceed the load limit of the balance. Selecting component loads equal to $\frac{1}{2}\sqrt{2}$ of the single load would ensure that the total compound load vector had the same magnitude as the pure single loads that were applied, but using load points that are part of the OFAT calibration load makes it easier to compare results of the calibration. Therefore, the same 75% load levels used in the OFAT compound loadings were also used in the structured MDOE load sequence. The total volume of data was the smallest that could fit pure quadratic loads and all two-way interactions, subject to constraints imposed by the need to measure compound loads in every quadrant and also to provide sufficient replicates to assess the combined effects of random error in the data and any hysteresis there might be.

C. MDOE Randomized

The MDOE randomized load table is essentially the same as the MDOE structured load table except that the loading sequence is randomized. Unfortunately, during the randomization process a number of load points were omitted and replaced with a zero load point. The -100% load points of the single components were replaced by a zero load point for all but the Fz component. An unambiguous assessment of the effect of randomization was therefore not possible, as the comparison of the two MDOE calibrations reflect the combined effects of load-point sequencing differences and the asymmetry induced in the randomized load schedule by the omission of five of the six maximum negative pure loads.

V. Data Reduction and Analysis

Some preprocessing of the data performed prior to analysis will be described here, as will the general plan for analyzing the data. The results of this analysis are presented in the next section.

A. Data Reduction

Prior to analysis, loads applied to the balance were corrected for tare-forces due to the weight of the loading hardware. Two inclinometers measured the roll and pitch angles of the metric side of the balance. The weight tares are rotated over roll and then pitch to the metric axis system.

The zero signals of the balance are recorded before the balance is mounted to the BCM. This zero represents a situation where the balance mounted to a sting and no loads are applied to the balance (only the weighted part of the balance, the adapters, the BCM mounting block and the electronic offset is measured by the gauges). The balance is mounted in the BCM in such away that the signals of the balance correspond to the previously recorded zero. This zero is then subtracted from all further recorded signals.

The supply voltage is measured for each data point. The balance output signals are corrected for deviations to the nominal supply voltage of 10V. This is done for each output signal individually. The output signals are made “non-dimensional” with the nominal supply voltage of 10V.

The objective of a balance calibration experiment is to produce a math model by which loads applied to the balance and resulting electrical signals can be properly mapped. This experiment differs from a typical calibration, however, in that the development of the math model is not the ultimate objective, but is, rather, an element of the

approach to achieving the objective. The *objective* of this study is to compare the calibration results obtained from a 734-point OFAT data set with results obtained from another data set based on the MDOE imperative to acquire the smallest volume of data necessary to achieve all technical objectives subject to known constraints. A secondary objective was to assess the effects of randomization by comparing calibration models based on a randomized and sequential version of the same MDOE model. Unfortunately, this objective could not be entirely achieved for reasons outlined in the description of experiment designs presented in the previous section.

The development of the calibration math models, which represents the chief analysis thrust of a calibration experiment, is for this experiment simply a preliminary (albeit crucial) step. We therefore regard the estimation of the calibration coefficients as an element of data *reduction* in the current experiment, with the central analysis to consist of subsequent comparisons of the math models.

A reduced third-order model was fit to each of the six balance responses using standard stepwise regression methods[17] implemented in a commercial data analysis software package.[18] The stepwise regression method identifies a subset of significant regression coefficients from those in a relatively large initial model believed to be of sufficiently high order to capture the most important behavior of a response of interest.

A full 84-term cubic model in six independent variables was chosen as the initial model, since few terms if any that are greater than second order are expected to be significant in a modern strain gage balance calibration, and no significant terms of order four or higher are anticipated. The stepwise regression process identified a subset of these 84 terms with the property that adding the strongest term from the remaining pool of candidates would reduce the unexplained variance by an amount too small to resolve with a specified level of confidence (99% in this study). Likewise, removing the weakest term in the currently selected subset would increase the unexplained variance by an amount that can be detected with— in this case— 99% confidence. While results varied from response to response and depended on the experiment design, the stepwise regression process produced response models with as few as nine and as many as 16 terms, meaning that 68–75 candidate terms from the full 84-term cubic model (80–89%) were rejected as statistically insignificant. A total of 18 response models were developed, one for each of the six balance outputs and three experiment designs. Figure 3 shows which terms were retained as significant in each of these 18 models.

B. Data Analysis

With the 18 models in hand, it is possible to perform the analyses necessary to objectively compare the OFAT and MDOE experiment designs on a cost/benefit basis. The cost comparisons are trivial, as the data volume serves as a surrogate for cycle time and direct operating costs. Clearly the conventional OFAT experiment design is much more time-consuming, requiring over seven times as many load-point settings as the MDOE experiment designs, which extends cycle time and direct operating costs correspondingly. The key question then becomes, "What benefits do we realize from this seven-fold increase in resource expenditure?" Benefits were assessed by quantifying model accuracy and precision, and comparing to accuracy and precision requirements described earlier.

The accuracy requirement is load dependent so Eq. (8) was used to compute an accuracy requirement individually for each of the 734 OFAT points, 103 MDOE points, and 114 independent confirmation points. For this comparative study, the balance models were not inverted to produce models of forces and moments as a function of electrical output signals. Rather, for the purposes of this study, calibration models were developed that describe the electrical responses in microvolts per volt as a function of applied load. The accuracy requirement is expressed in newtons and newton-meters when F_i, the maximum load component of the i^{th} component, is expressed in physical units. But because the summation term in Eq. (8) involves dimensionless ratios, the accuracy requirement was translated into electrical terms by replacing maximum loads with electrical signals corresponding to those loads.

The residual for each data point was used as a measure of accuracy. This is simply the difference between measured and predicted responses for a given load combination. A ratio was then computed by dividing the absolute value of the residual by the accuracy requirement for that specific loading combination as computed by Eq. (8). This ratio formed an error index with a value between 0 and 1 for all points that met the accuracy specification. The index represents the fraction of the accuracy error budget that is consumed. Any point that failed to meet the accuracy specification would have an error index value greater than 1.

The precision of the models developed from each experiment design was assessed by using Eq. (10) to compute 95% confidence interval half-widths for model predictions at each point. As noted earlier, the precision requirement was specified as "about 1/3 of its accuracy." An error index for precision was therefore created analogous to the accuracy error index by dividing the prediction 95% confidence interval half-width for each point by one third of the accuracy specification computed for that point using Eq. (8). Similar to the accuracy error index, the precision error index represents the fraction of the precision error budget that is consumed, with any point that failed to meet the precision specification having an error index value greater than 1.

For this study, average accuracy and precision error index values were computed for each experiment design and each balance component. Using the models developed for each balance output and from each experiment design, these error index values were calculated for two types of calibration loads, the original loads used to generate the data from which the models were developed, and a set of 114 independent loads that were not used in the model development. It is useful to compare model performance at the original calibration points and at independent points not used to create the models because the regression algorithms used to estimate model coefficients do so in such a way as to minimize residuals (in a least squares sense) at the calibration load points. We therefore expect reasonably good accuracy at the points used to fit the model. The confirmation points help determine how transferable the model predictions are to general loading combinations.

There were 36 average accuracy error index values and 36 average precision error index values computed, one for each combination of six balance components by three experiment designs by two point types—model calibration points and independent confirmation points. Table 2 lists them.

Table 2. Average Error Index Values for Calibration Model Points and Confirmation Points. Index values less than 1 imply accuracy/precision is within specifications.

Experiment Design (Data Volume)	Output	Average Accuracy Error Index		Average Precision Error Index	
		Cal Model Pts	Confirmation Pts	Cal Model Pts	Confirmation Pts
OFAT (734 Pts)	R1	0.254	0.419	0.264	0.155
	R2	0.131	0.145	0.206	0.077
	R3	0.115	0.163	0.174	0.055
	R4	0.079	0.078	0.123	0.042
	R5	0.142	0.160	0.204	0.081
	R6	0.136	0.110	0.152	0.047
Sequential MDOE (101 Pts)	R1	0.349	0.598	0.292	0.345
	R2	0.171	0.116	0.214	0.121
	R3	0.094	0.143	0.184	0.102
	R4	0.079	0.073	0.133	0.107
	R5	0.278	0.223	0.214	0.188
	R6	0.144	0.180	0.158	0.121
Randomized MDOE (103 Pts)	R1	0.124	0.406	0.285	0.201
	R2	0.122	0.142	0.208	0.156
	R3	0.094	0.166	0.179	0.131
	R4	0.066	0.076	0.130	0.090
	R5	0.127	0.144	0.209	0.153
	R6	0.087	0.136	0.154	0.117

It is difficult to see patterns in a table of numbers such as in Table 2. Furthermore, there is obviously some uncertainty in each of these numbers since they are all rooted in experimental data. It is therefore difficult to answer with high confidence such fundamental questions in this study as these:

- "Which experiment design yields higher precision math models, OFAT or MDOE?" "Which yields the higher accuracy models?"
- "Do the models predict responses at confirmation points as well as they do at model calibration points?"
- "Do all the balance outputs feature the same levels of precision and accuracy?"
- "Are there interactions; that is, are differences from one balance output to another dependent on experiment design or on point type, for example?"

An analysis of variance (ANOVA) was performed to reconcile the 72 numbers in Table 2 in such a way as to provide objective answers to these and other questions. The ANOVA was organized in terms of three factors:

Output (six levels corresponding to the outputs of the balance), Point Type (two levels—Calibration Model Point and Confirmation Point), and Experiment Design (three levels—OFAT, MDOE Sequential, and MDOE Randomized). It is convenient to assign Latin letters to these factors, as in Table 3.

Table 3. Experimental factors.

Factor	Name	Levels
A	Output	6
B	Point Type	2
C	Experiment Design	3

Table 3 allows us to use rather compact notation to describe the effects that interest us. We would say there is no significant "C" effect for precision, for example, if all three experiment designs generated models with comparable precision index values. If there is a difference in the precision of at least one of the experiment designs from the other two, or if all three differ from each other, we would say that there is a significant C effect and we would conclude that experiment design matters for precision.

In formal terms, we wish to test a *null hypothesis*, H_0, which can be expressed with our condensed notation as follows:

$$H_0: C=0$$

We will reject this hypothesis if we are able to detect some difference in the prediction precision of models developed from one experiment design to another.

There is corresponding *alternative hypothesis*, H_1, expressed as follows:

$$H_1: C\neq 0,$$

which we will reject if we cannot detect any differences among the prediction precision index values for models developed from one experiment design to another. There are analogous pairs of null and alternative hypotheses for the "A" and "B" effects as well.

The three factors in this study can interact with each other, and we can use the same notation to describe these interactions. We say there is a significant "AB" interaction for accuracy, for example, when accuracy differences from one balance component to another are different when assessed at model calibration points than at independent confirmation points. Likewise, there are AC and BC interactions that may or may not be significant. Theoretically, there is also a three-way ABC interaction that is possible, but for technical reasons it would have been necessary to replicate this experiment to detect the three-way interaction as well as the three main effects and three two-way interactions. That is, it would have been necessary to have two or more independent estimates of each error index for every unique combination of balance output, experiment design, and point type. For practical purposes, very little is lost by not evaluating the three-way interaction, as the main effects and two-way interactions are generally the dominant effects in such experiments.

With the analysis cast in this form, it is possible to make inferences about six specific hypotheses by rejecting either the null hypothesis or the alternative hypothesis associated with the three main effects of this study and the three two-way interactions. These hypothesis pairs are numbered for convenience and summarized here:

$H_{01}: A=0$ No significant difference in a specified quality metric from one balance output to another.
$H_{11}: A\neq 0$ At least one balance output differs in quality from the rest.

$H_{02}: B=0$ No significant difference in a specified quality metric between calibration model points and independent confirmation points.
$H_{12}: B\neq 0$ The quality assessed at confirmation load points differs from the quality assessed at model calibration points.

$H_{03}: C=0$ No significant difference in a specified quality metric when models are developed drom data acquired in one experiment design or another.
$H_{13}: C\neq 0$ The quality of models developed from one experiment design differs from the quality of models developed from another experiment design.

H_{04}: AB=0 Quality differences from one balance component to another are the same for model calibration points as for independent confirmation points.

H_{14}: AB≠0 Quality differences from one balance component to another are different for model calibration points and confirmation points.

H_{05}: AC=0 Quality differences from one balance component to another are independent of which experiment design was executed to produce the model.

H_{15}: AC≠0 Quality differences from one balance component to another depend on which experiment design was executed to produce the model.

H_{06}: BC=0 Any quality differences from one experiment design to another are independent of whether the quality is assessed at a model calibration point or an independent confirmation point.

H_{16}: BC≠0 Quality differences from one experiment design to another depend on whether the quality is assessed at a model calibration point or an independent confirmation point.

For each model quality metric examined in this study, one hypothesis is rejected for each hypothesis pair. The analysis of variance to which we alluded at the start of this section provides an objective means for selecting which hypothesis to reject. Tables 4 and 5 present the ANOVA results for the accuracy and precision error index values:

Table 4. ANOVA Table for Accuracy Error Index Values.

Source	Sum of Squares	df	Mean Square	F Value	p-value Prob > F
Model	0.45	25	0.02	29.8	< 0.0001
A-Output	0.29	5	0.06	97.6	< 0.0001
B-Point Type	0.02	1	0.02	36.6	0.0001
C-Experiment Design	0.03	2	0.01	20.9	0.0003
AB	0.06	5	0.01	21.5	< 0.0001
AC	0.04	10	0.00	6.5	0.0033
BC	0.00	2	0.00	2.5	**0.1346**
Residual	0.01	10	0.00		
Cor Total	0.45	35			

Table 5. ANOVA Table for Precision Error Index Values.

Source	Sum of Squares	df	Mean Square	F Value	p-value Prob > F
Model	8.14	25	0.33	20.3	< 0.0001
A-Output	3.23	5	0.65	40.2	< 0.0001
B-Point Type	2.43	1	2.43	151.4	< 0.0001
C-Experiment Design	1.21	2	0.61	37.8	< 0.0001
AB	0.18	5	0.04	2.3	**0.1228**
AC	0.15	10	0.01	0.9	**0.5422**
BC	0.93	2	0.47	29.1	< 0.0001
Residual	0.16	10	0.02		
Cor Total	8.30	35			

A detailed explanation of analysis of variance is beyond the scope of this paper, but the reader can consult standard texts for the details.[19] In short, the ANOVA tables partition the total variance in a set of experimental data into explained and unexplained components. The explained components include the main effects and interactions described above. For each of these, the ANOVA table lists variance metrics (Sum of Squares, degrees of freedom, and Mean Square) that are the basis of computing the quantities in the last two columns. These columns, "F Value" and "p-value," contain equivalent information about the magnitude of the effects listed in the first column, and are the heart of the ANOVA table. The F statistic is a signal-to-noise term that represents the variance associated with a

given effect as a multiple of the unexplained variance. Larfe F values imply significant effects. In Table 4, for example, the "Output" factor has the largest F-Value, implying that balance component has a greater influence on accuracy than either point type or experiment design. Likewise, from Table 5 it appears that Point Type is more important in determining precision than the other two factors 9or any of their interactions.

The right-most column contains probability values that are interpreted as the probability that an F-Value as large as recorded could occur due to chance because of random variations in the data. A small p-value implies a low probability of chance effects and therefore a high probability that the effect is not due simply to chance variations in the data but is in fact real. We typically agree on a "significance level" prior to the test, which marks the p-value criterion for accepting an effect as real. In this study the significance criterion was selected to be 0.01, meaning that any effect with a p-value below 0.01 has a probability of less than 1% of being induced by random variations in the data, and therefore at least a 99% chance of being a real effect. We say that we can accept any effect as real by this criterion with 99% confidence. The effects that are accepted as real have black p-values in Tables 4 and 5, while the effects not believed to be real are listed in red. The results of the analysis of variance and other analyses performed in this study are discussed in the next section.

VI. Results and Discussion

The design matrix enables us to generate a distribution of unit standard errors for a given design before any data are acquired. The three parts of Fig. 2 reveal similarities in the standard error distribution for all three designs. In general, the standard error of the design is at a minimum near the center of the range of independent variables. These distributions exist in a six-dimensional space of which Fig. 2 represents only two dimensions, but the same general behavior applies when the unit standard errors are plotted against any two of the independent variables.

The OFAT error distribution achieves a deeper minimum near the center of the design than the MDOE error distributions because there are so many more residual degrees of freedom available to estimate the uncertainty. In either the MDOE or the OFAT designs, response estimates made in the extreme corners of the design space will feature significantly more uncertainty than points acquired in the middle.

This phenomenon is responsible for the interaction for precision that exists between point type (factor B) and experiment design (factor C), as the ANOVA results of Table 5 reveal. In this table, the BC interaction has a p-value of less than 0.0001, meaning that there is a probability in excess of 99.99% that the difference in precision from one experiment design to the next depends on point type. Figure 4 illustrates what is happening.

The "I-beam" marks on each data point in Fig. 4 represent 95% Least Significant Difference (LSD) bars. A difference can be said to exist with at least 95% confidence between points for which these bars do not overlap. In Fig. 4, we are unable to resolve a difference between the precision of confirmation points and model calibration points for the two MDOE designs, but we can say with at least 95% confidence that there is a difference for the OFAT design (since their LSD bars do not overlap). This difference can be attributed to the relative depth of the distribution of standard unit errors for the OFAT design compared to the two MDOE designs, as revealed in Fig. 2. The model calibration points are more or less uniformly distributed over the range of independent variables, but the confirmation points tend to be concentrated near the center of the design, with typically at least half of the six outputs at either zero or a relatively low load near the center of the relatively wide, symmetric load range. The precision is proportional to the magnitude of standard errors, with a larger difference existing between the standard errors in the center and the average standard error for the OFAT design than the MDOE designs.

Note that while this difference is large enough to resolve, it is still quite small in absolute terms. Figure 4 shows the difference in precision to be on the order of 0.1 of the precision error budget, which is one third of the accuracy error budget. The accuracy error budget is nominally 0.1% of full scale output by Eq. (8). Therefore, this difference amounts to roughly 0.003% of full scale. For practical purposes, then, there is no significant difference between the precision of model and confirmation points for any of the experiment designs.

Figure 5 illuminates another interaction that is seen to be significant in the ANOVA results of Table 4. The p-value is seen to be less than 0.0001 for the AB interaction, meaning that the difference in accuracy from model point to confirmation point is dependent on the output channel of the balance. Figure 5 shows that for all three experiment designs, outputs have generally similar accuracy at calibration points and confirmation points except for R1, where the accuracy in estimating responses at other load combinations besides those used to fit the model is degraded by an amount that can be resolved easily. While the accuracy for R1 still meets requirements even at the confirmation points (error index less than 1), clearly there is something about R1's behavior that distinguishes it from all five of the other balance outputs.

The one remaining significant interaction from the ANOVA tables is the AC interaction, which is significant for accuracy but not for precision according to the ANOVA results of Tables 4 and 5. In a sense, this is one of the most

interesting interactions. We are anxious to see if there is any difference in accuracy or precision from one experiment design to the next, and it is interesting to know if such differences depend on which balance output we are considering. Figures 6 and 7 illustrate the nature of this interaction for accuracy and precision.

The ANOVA results indicate no significant design-output interaction for precision. This means that the difference in the precision of the math model from one balance output to the next is independent of the experiment design used to develop that model. The significant AC interaction for accuracy can be seen in Figs. 6 and 7 to again be attributable to behavior in the R1 output that distinguishes it from all the other balance outputs, especially for confirmation points.

Absent the anomalous behavior of R1, there are no experiment design effects significant at the 0.01 level for accuracy. The only experiment design effects significant at the 0.01 level for precision are due to the selection of confirmation points near the center of the design, coupled with the seep minimum in the distribution of OFAT standard errors due to the large data volume of that design (seven times more loadings than the MDOE designs) Tables 6 and 7 present the ANOVA results neglecting R1.

Table 6. ANOVA Table for Accuracy Error Index Values, Neglecting R1. No significant experiment design effects at the 0.01 significance level.

Source	Sum of Squares	df	Mean Square	F Value	p-value Prob > F
Model	0.06	21	0.00	6.0	0.0107
A-Output	0.03	4	0.01	17.3	0.0010
B-Point Type	0.00	1	0.00	2.2	**0.1787**
C-Experiment Design	0.01	2	0.00	6.0	**0.0306**
AB	0.00	4	0.00	2.2	**0.1648**
AC	0.01	8	0.00	3.4	**0.0624**
BC	0.00	2	0.00	1.5	**0.2778**
Residual	0.00	7	0.00		
Cor Total	0.06	28			

Table 7. ANOVA Table for Precision Error Index Values, Neglecting R1. Only significant experiment design effects at the 0.01 significance level due to selection of confirmation points.

Source	Sum of Squares	df	Mean Square	F Value	p-value Prob > F
Model	0.07	21	0.00	25.7	0.0001
A-Output	0.02	4	0.00	35.3	< 0.0001
B-Point Type	0.04	1	0.04	272.7	< 0.0001
C-Experiment Design	0.01	2	0.00	36.2	0.0002
AB	0.00	4	0.00	3.7	**0.0627**
AC	0.00	8	0.00	1.3	**0.3861**
BC	0.01	2	0.00	20.8	0.0011
Residual	0.00	7	0.00		
Cor Total	0.07	28			

Figure 3 reveals another consequence of the relatively large volume of data acquired in the OFAT calibrations. This figure shows that the OFAT response models have on average about 50% more terms than the corresponding models developed from MDOE designs. The reason is that the substantial additional data results in higher precision levels for estimating model coefficients, and thus model terms that are too small to be detected with the resolution afforded by the volume of data in the MDOE designs are detected as significant in the OFAT models. Note that the biggest difference is in the higher-order terms, where the OFAT design detects more such terms than the MDOE designs. But as the ANOVA results indicate, there is virtually no difference between the accuracy or the precision of MDOE and OFAT models, notwithstanding the seven-fold increase in data volume for the OFAT experiment design. This implies that the higher order terms that the OFAT design is picking up are too small to be of practical

importance. The MDOE models are well within specifications, and the acquisition of more data after requirements are already met represents an unnecessary expenditure of resources. MDOE practitioners seek to scale the data volume in order to acquire an ample number of data points to satisfy precision and accuracy requirements, but no more. The preserves resources and reduces clutter in the response surface models, which helps reveal the underlying physics by identifying more clearly the most important factors.

It is also interesting to compare the models for the randomized and sequential MDOE models in Fig. 3. The MDOE design executed in sequential order will have been confounded with systematic effects if any were in play. If there were such effects, the outputs of the balance would be a function of seven variables, not six. They would depend on the six load variables and also a seventh variable—time. By fitting the data to only six variables, there is a tendency for curvature terms to become artificially inflated, as the model seeks to bend the responses to account for the variable that is not being explicitly fitted. While it is not possible to say that systematic effects were in play during the sequential MDOE experiment, the models developed from that data set have almost twice as many pure quadratic and higher order terms as the models developed from the randomized MDOE experiment.

We introduced a number of hypotheses in the previous section by saying that it was our intent to test these hypotheses using an analysis of variance. A slight inconsistency surfaced in the preliminary analysis involving the R1 output, as noted above. While the accuracy and precision of the R1 output are both well within tolerance, this balance component appears to behave sufficiently differently from the other five balance components that different conclusions are reached by analyses depending on whether R1 is included or not. The R1 phenomenon may be limited to the confirmation point data, and may reflect conditions during the acquisition of the few points rather than some general aspect of the balance. For these reasons we evaluate the hypotheses we have developed on the basis of ANOVA results presented in Tables 6 and 7.

A. Accuracy

1. Description and Computation

Average of "accuracy error index" values for models developed from a given experiment design for a given balance output. The accuracy error index is the ratio of a residual to a load-dependent tolerance level, representing the fraction of the accuracy error budget consumed on that point.

2. Observations

These observations are based on the ANOVA results in Table 6.

A1. The A (OUTPUT) effect is significant at the 0.01 level.
A2. The B (POINT TYPE) and C (EXPERIMENT DESIGN) effects are not significant at the 0.01 level.
A3. None of the two-way interaction effects is significant at the 0.01 level.

3. Conclusions

- We reject H_{01}: A=0, concluding that the accuracy of the response models is not constant over all six balance responses. This conclusion is supported by observation A1.
- We reject H_{12}: B≠0, concluding that the response models are just as accurate for predicting outputs at confirmation points as at model calibration points. That is, we conclude that the DNW calibration exhibits transferability with respect to accuracy, and the models are useful for predicting responses at general load combinations. This conclusion is supported by observation A2.
- We reject H_{13}: C≠0, concluding that the MDOE designs deliver the same accuracy as the OFAT designs with a reduction in loading combinations in excess of 85%. This conclusion is supported by observation A2.
- We reject H_{14}: AB≠0, concluding that the good transferability observed in the DNW balance calibration from model point predictions to predictions at independent load combinations holds individually for each balance component and not simply in an "average" sense. This conclusion is supported by observation A3.
- We reject H_{15}: AC≠0, concluding that the accuracy differences across balance components are not influenced by the design of the calibration experiment. This conclusion is supported by observation A3.
- We reject H_{16}: BC≠0, concluding that all three experiment designs generated the same degree of transferability from calibration model points to confirmation points. This conclusion is supported by observation A3.

4. Discussion

While each balance component satisfied accuracy requirements comfortably, there is some variation in accuracy from one component to another. The balance and its calibration models deliver good accuracy transferability in an average sense and also component by component. No accuracy decrement is detected with the MDOE designs relative to the OFAT design despite substantially fewer data points. There is no more difference in component-to-component performance with an MDOE design than with an OFAT design despite the substantial difference in data volume. The transferability in models developed from the MDOE calibrations was just as good as from the OFAT calibrations, notwithstanding so many fewer points in the MDOE calibration.

B. Precision

1. Description and Computation

Average of "precision error index" values for models developed from a given experiment design for a given balance output. The precision error index is the ratio of a 95% confidence interval half-width for model prediction to a load-dependent tolerance level, representing the fraction of the precision error budget consumed on that point.

2. Observations

These observations are based on the ANOVA results in Table 7.

P1. The A (OUTPUT), B (POINT TYPE), and C (EXPERIMENT DESIGN) effects are all significant at the 0.01 level.
P2. The BC interaction is significant at the 0.01 level.
P3. Neither the AC nor the AB interactions are significant at the 0.01 level.

3. Conclusions

- We reject H_{01}: A=0, concluding that there are differences in precision from one balance output to another. This conclusion is supported by observation P1.
- We reject H_{02}: B=0, concluding that the precision is different for the calibration model points than the confirmation points. This conclusion is supported by observation P1.
- We reject H_{03}: C=0, concluding that not all of the experiment designs produced models with the same precision. This conclusion is supported by observation P1.
- We reject H_{14}: AB≠0, concluding that all components exhibit about the same difference in precision from OFAT to MDOE. This conclusion is supported by observation P3.
- We reject H_{15}: AC≠0, concluding that the precision differences across balance components are not influenced by the design of the calibration experiment. This conclusion is supported by observation P3.
- We reject H_{06}: BC=0, concluding that the MDOE designs and the OFAT design generate differing degrees of improvement in precision in going from a calibration model point to a confirmation point near the center of the design. This conclusion is supported by observations P2.

4. Discussion

- While each balance component satisfied precision requirements comfortably, there is some variation in precision from one component to another just as there was some variation in accuracy. While the confirmation points generated different levels of precision than the calibration points, the precision was in fact uniformly *better* for the confirmation points. This is attributed to the fact that the confirmation points were near the center of the design where prediction standard errors (and thus 95% confidence interval half-widths were at a minimum). In any case, there was no evidence of a negative transferability problem in which the model would have difficulty predicting responses anywhere besides the calibration points. These models performed at least as well with confirmation points as with calibration points. The precision of the OFAT model was generally greater than the precision of the models developed from the two MDOE designs, although both MDOE designs delivered precision levels well within specifications. The difference is attributed to the significantly greater number of data points acquired in the OFAT experiment. The precision enhancement resulting from so many OFAT data points applies to each balance component individually. Both MDOE and OFAT designs exhibited improvements in the precision near the center of the design, where unit standard errors of prediction were a minimum. We reject H_{06}: BC=0, concluding that the MDOE designs and the OFAT design generate differing degrees of improvement in precision in going from a calibration model point to a confirmation point near the center of the design. This conclusion is supported by observations P2.

VII. Summary and Conclusions

A strain gage balance has been calibrated using an automated balance calibration machine capable of setting the multiple compound loads required of a calibration using the Modern Design of Experiments (MDOE). The balance was calibrated using a conventional One Factor at a Time (OFAT) load schedule that required 734 loadings, followed by a load schedule that featured the kind of reduced data volume that MDOE practitioners are now using for balance calibration (103 points). The MDOE load schedule was executed both in random and in sequential order.

Precision and accuracy metrics were quantified for 18 combinations of balance output (six outputs) and experiment design (three designs). The response models were evaluated for both precision and accuracy at the same points that were fitted to estimate the model and also at 114 independent confirmation load combinations that were not used to fit the models. The average fraction of accuracy and precision error budget consumed were used as performance metrics for the calibration models developed from both OFAT and MDOE experiment designs.

Principal findings of this study are summarized as follows:

1. At 734 points, the OFAT experiment design required over seven times as many loadings as the MDOE designs.

2. A statistically significant difference was observed in the precision of OFAT vs. MDOE designs, with the OFAT design featuring the higher precision. The absolute difference in precision was on the order of 0.01% of full scale, and not deemed of practical importance, since both MDOE and OFAT designs satisfied precision requirements by comfortable margins. The slight OFAT advantage in precision is attributed to the seven-fold increase in data volume relative to the MDOE design.

3. While precision and accuracy differed from one balance output to another, all channels were within specifications. The R1 output channel displayed somewhat lower accuracy than the other five balance channels, although it was still within accuracy specifications except for a small number of extreme points. The precision of the R1 output was comparable to that of all five other channels.

4. The accuracy of the R1 output relative to the other five channels was better for calibration model points than for confirmation points, suggesting that some factor limited to the confirmation runs may have influenced the relative behavior of the R1 output.

5. The confirmation points were acquired near the middle of the design, where the model predictions are generally more precise than at points near the edges of the design. The estimated precision was therefore generally better for confirmation points than for calibration points.

6. The accuracies of calibration points and confirmation points were comparable. This suggests good transferability from points used to fit the model to general load combinations.

7. The MDOE designs delivered comparable accuracy and precision as the OFAT design despite requiring significantly fewer calibration loadings. Except for randomization, no MDOE quality assurance tactics were employed in this test. Specifically, the selection of loading combinations was not optimized for quality and the test matrix was not blocked, nor was the data volume rigorously scaled to the accuracy and precision requirements of the calibration. It is possible that some further improvement in quality and a further reduction in cost could be achieved by implementing these standard MDOE practices.

Acknowledgments

The authors acknowledge the support of the German Dutch Wind Tunnels and of the Wind Tunnel Enterprise at Langley Research Center. Dr. Norbert Ulbrich of Jacobs Technology at Ames Research Center was instrumental in initiating the collaboration of the authors on this project, and participated in many helpful discussions.

References

[1]DeLoach, R., "Applications of Modern Experiment Design to Wind Tunnel Testing at NASA Langley Research Center," *AIAA-98-0713,* 36th AIAA Aerospace Sciences Meeting and Exhibit, Reno, Nevada, January 1998.

[2]Parker, P.A., Morton, M., Draper, N., Line, W., "A Single-Vector Force Calibration Method Featuring the Modern Design of Experiments," *AIAA 2001-0170*, 39th Aerospace Sciences Meeting and Exhibit, Reno, Nevada, January 2001.

[3]DeLoach, R., "Tactical Defenses Against Systematic Variation in Wind Tunnel Testing," *AIAA 2002-0885*, 40th AIAA Aerospace Sciences Meeting and Exhibit, Reno, Nevada, January 2002.

[4]DeLoach, R., "Impact of Loading Selection and Sequencing on a Force Balance Calibration (Invited)," *AIAA 2006-3436*, 25th Aerodynamic Measurement Technology and Ground Testing Conference, San Francisco, California, June 2006.

[5]DeLoach, R., Hill, J. S., and Tomek, W. G., "Practical Applications of Blocking and Randomization in a Test in the National Transonic Facility (Invited)," *AIAA 2001-0167*, 39th AIAA Aerospace Sciences Meeting and Exhibit, Reno, Nevada, January 2001.

[6]DeLoach, R. "Blocking: A Defense Against Long-Period Unexplained Variance in Aerospace Ground Testing (Invited)," *AIAA 2003-0650*, 41st AIAA Aerospace Sciences Meeting and Exhibit,. Reno, Nevada, January 6-9, 2003.

[7]Recommended Practice: Calibration and Use of Internal Strain-Gage Balances with Application to Wind Tunnel Testing, AIAA R-091-2003.

[8]Myers, R. H., and Montgomery, D. C., *Response Surface Methodology: Process and Product Optimization Using Designed Experiments*, Wiley Series in Probability and Statistics, 2nd ed., John Wiley and Sons, New York, 2002.

[9]Beer, P. P., and Johnston, E. R. Jr., *Vector Mechanics for Engineers,* McGraw-Hill, 1962.

[10]Box, G. E. P., and Draper, N., *Empirical Model-Building and Response Surfaces,* John Wiley and Sons, New York, 1987.

[11]Ulbrich, N., and Volden, T., "Strain-Gage Balance Calibration Analysis Using Automatically Selected Math Models," *AIAA 2005-4084*, 41st AIAA/ASME/SAE/ASEE Joint Propulsion Conference and Exhibit, Tucson, Arizona, July 2005.

[12]Ulbrich, N., and Volden, T., "A New Approach to Strain-Gage Balance Calibration Analysis," 5th International Symposium on Strain-Gauge Balances, Aussois, France, May 2006.

[13]Ulbrich, N., and Volden, T., "Development of a New Software Tool for Balance Calibration Analysis," *AIAA 2006-3434*, 24th AIAA Aerodynamic Measurement Technology and Ground Testing Conference, San Francisco, California.

[14]Hansen, R. M., "Evaluation and Calibration of Wire-Strain-Gage Wind-Tunnel Balances Under Load," NACA Langley Aeronautical Laboratory, 1956.

[15]Eckert, D., et al., "Design and Construction of Internal Balances for the German/Netherlands Wind Tunnel (DNW)," 1st International Symposium on Strain Gauge Balance, 14 August 1997

[16]Philipsen, I., Hoeijmakers, H., and Alons, H. J., "A New Balance And Air-Return Line Bridges For DNW-LLF Models (B664 / RALD 2001)," 4th International Symposium on Strain-Gauge Balances, San Diego, California, 10–13 May 2004

[17]Draper, N. R., and Smith, H., *Applied Regression Analysis*, 3rd ed., John Wiley and Sons, New York, 1998.

[18]Design Expert®, Software Package, Ver. 7.03, StatEase, Inc., Minneapolis, Minnesota, 2006

[19]Scheffe, H., *The Analysis of Variance*, John Wiley and Sons, New York, 1959.

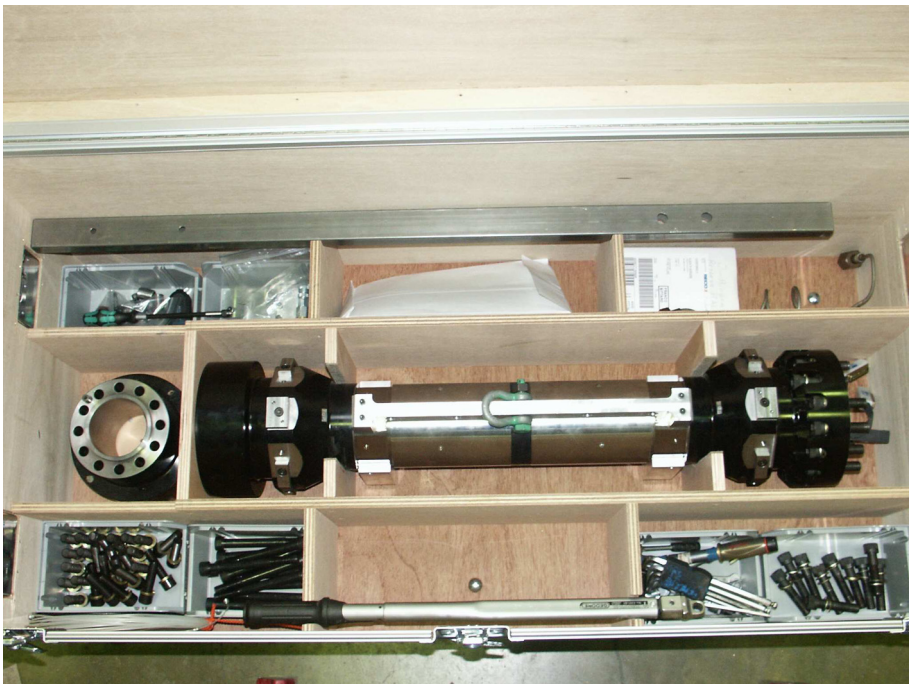

Figure 1. Balance calibrated in this study.

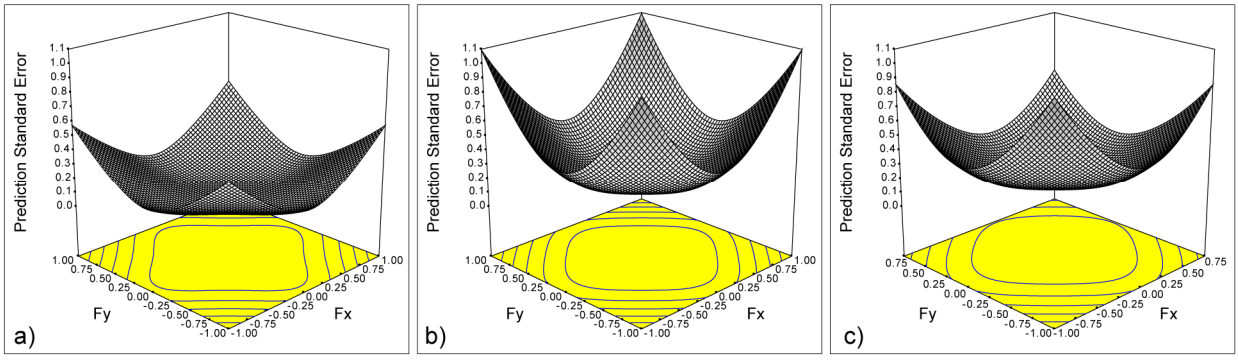

Figure 2. Distribution of Standard Errors. a) 734-Point OFAT experiment design; b) 101-Point MDOE experiment design, standard order; C) 103-Point MDOE experiment design, random order.

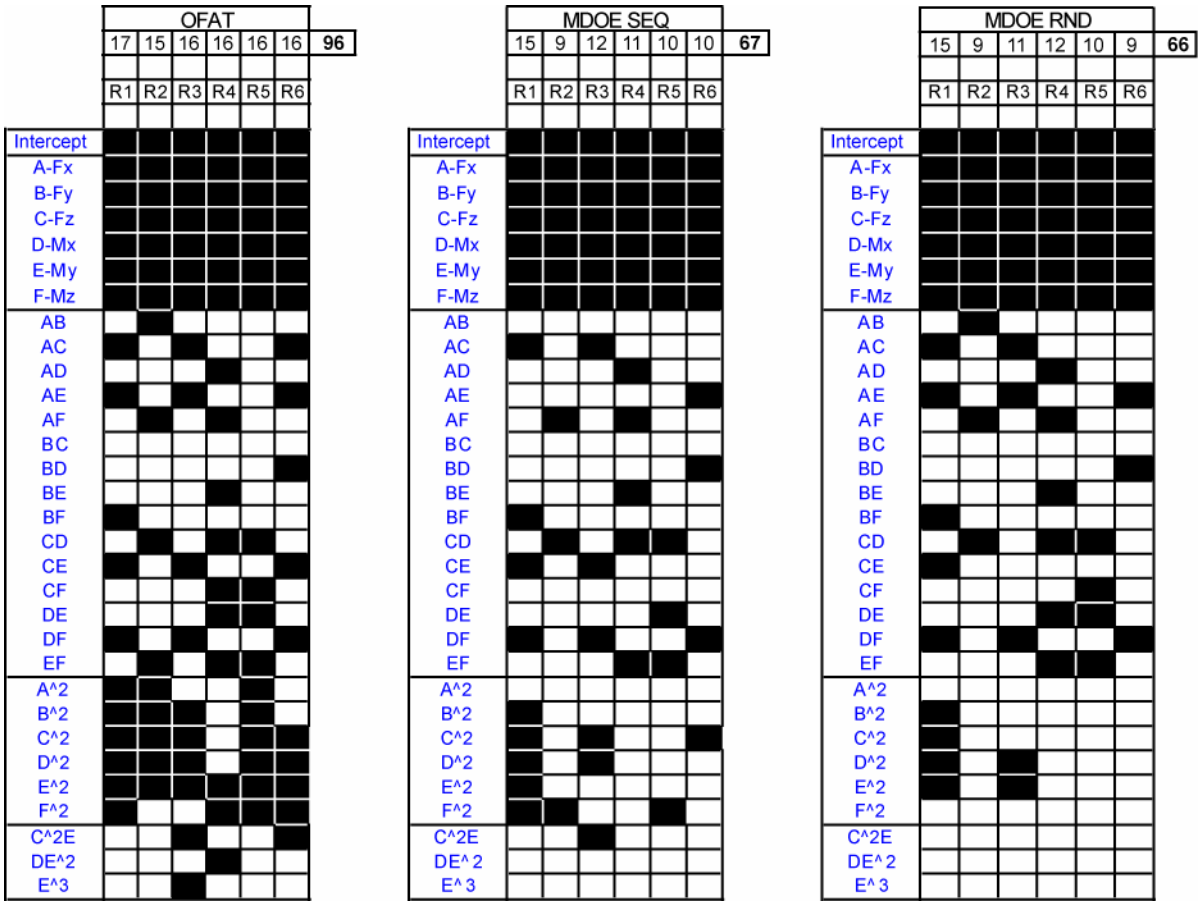

Figure 3. Significant model terms marked in black for six balance responses and three experiment designs. The number of significant terms is recorded above each model. Total significant terms in each experiment displayed to right of per-experiment term counts.

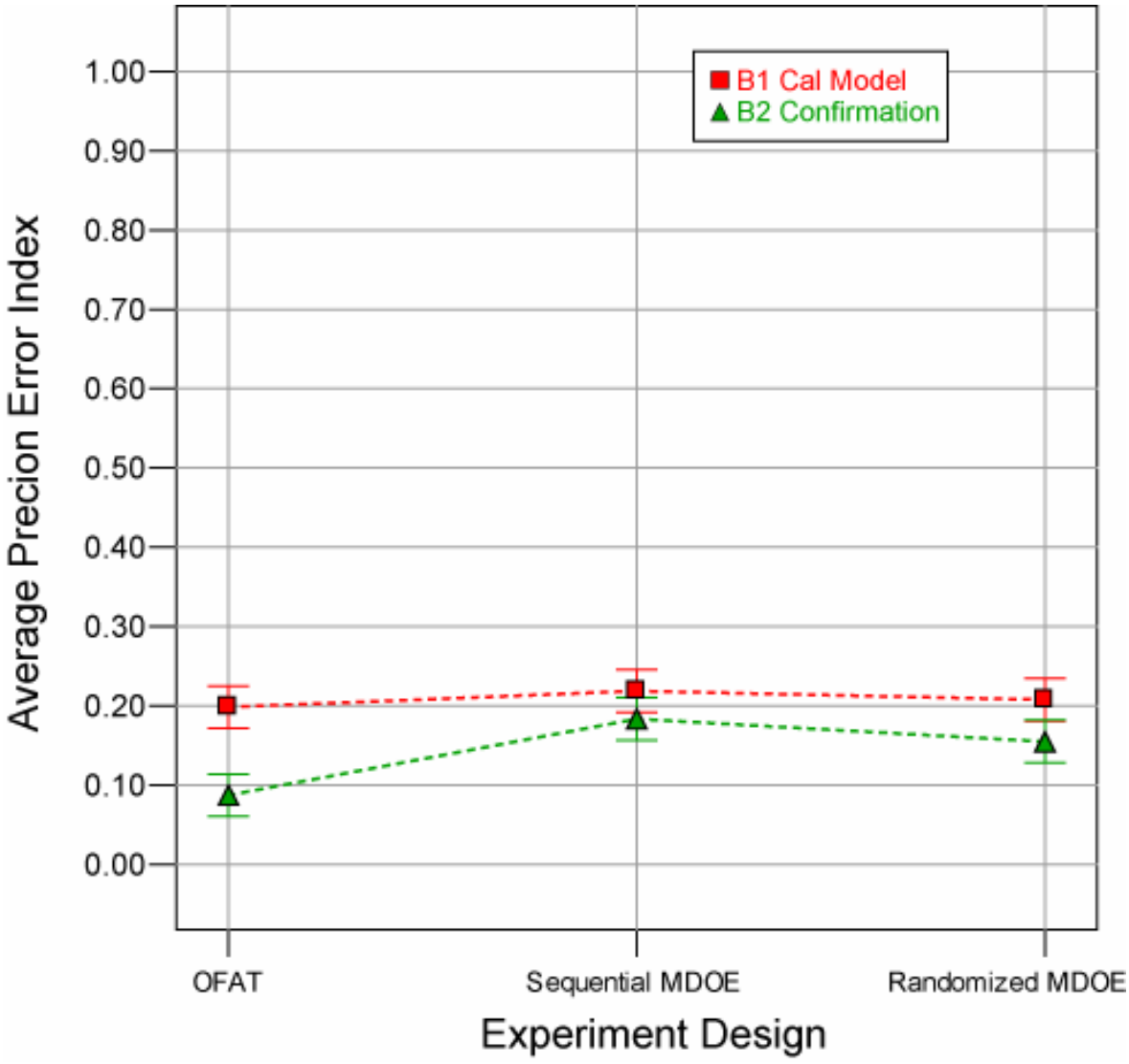

Figure 4. Significant interaction between experiment design and point type for precision at representative balance output (R5). Confirmation points were acquired near center of design space where OFAT distribution of standard errors achieves a lower minimum due to large number of residual degrees of freedom.

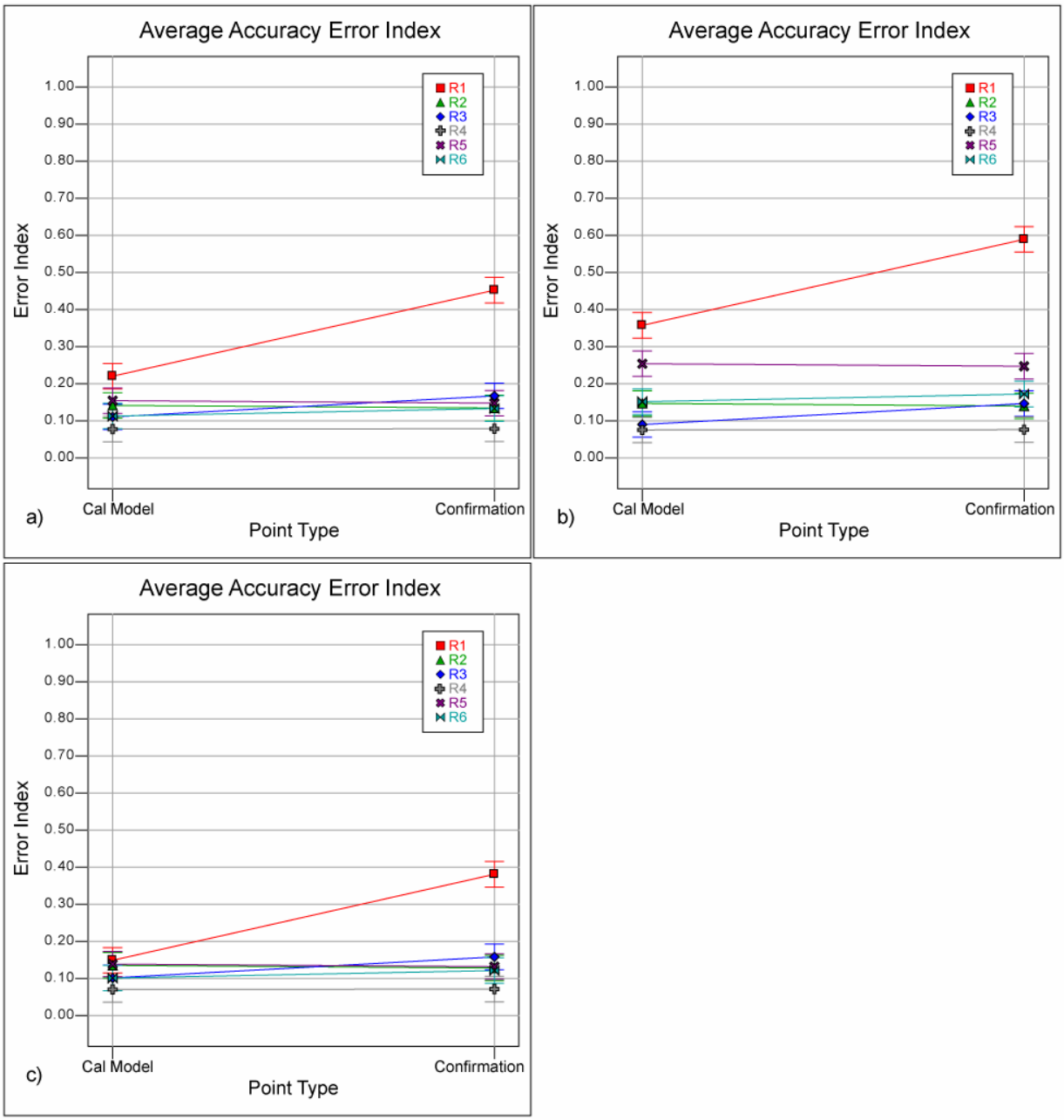

Figure 5. Significant interaction between point type and balance output for accuracy. a) OFAT; b) Sequential MDOE; c) Randomized MDOE. For all experiment designs, outputs have generally similar accuracy at calibration points and confirmation points except R1.

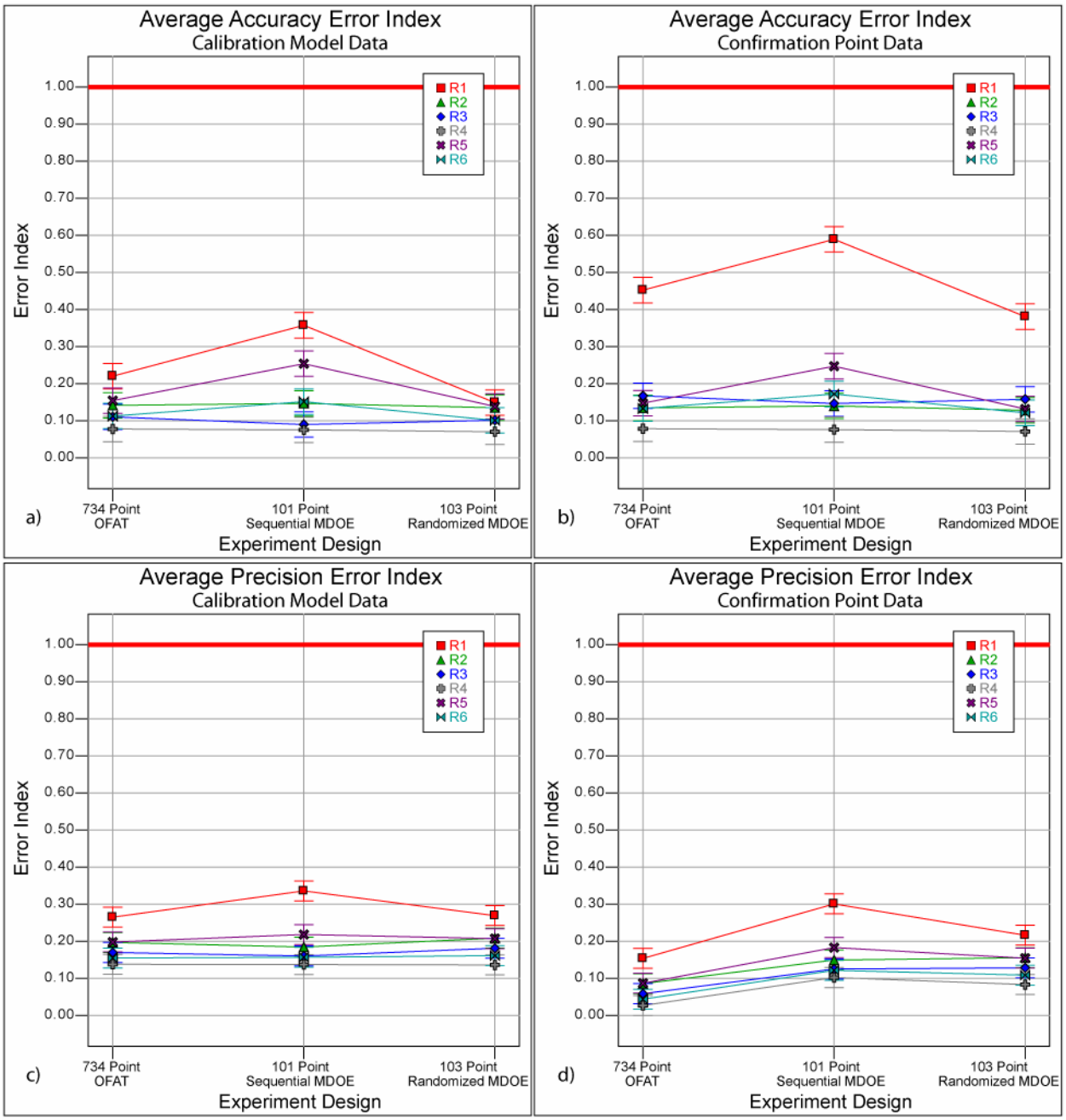

Figure 6. Interactions between experiment design and balance output. a) Accuracy for calibration model data; b) Accuracy for confirmation point data; c) Precision for calibration model data; d) Precision for confirmation point data. Insignificant interaction for precision, but significant interaction for accuracy due to output R1.

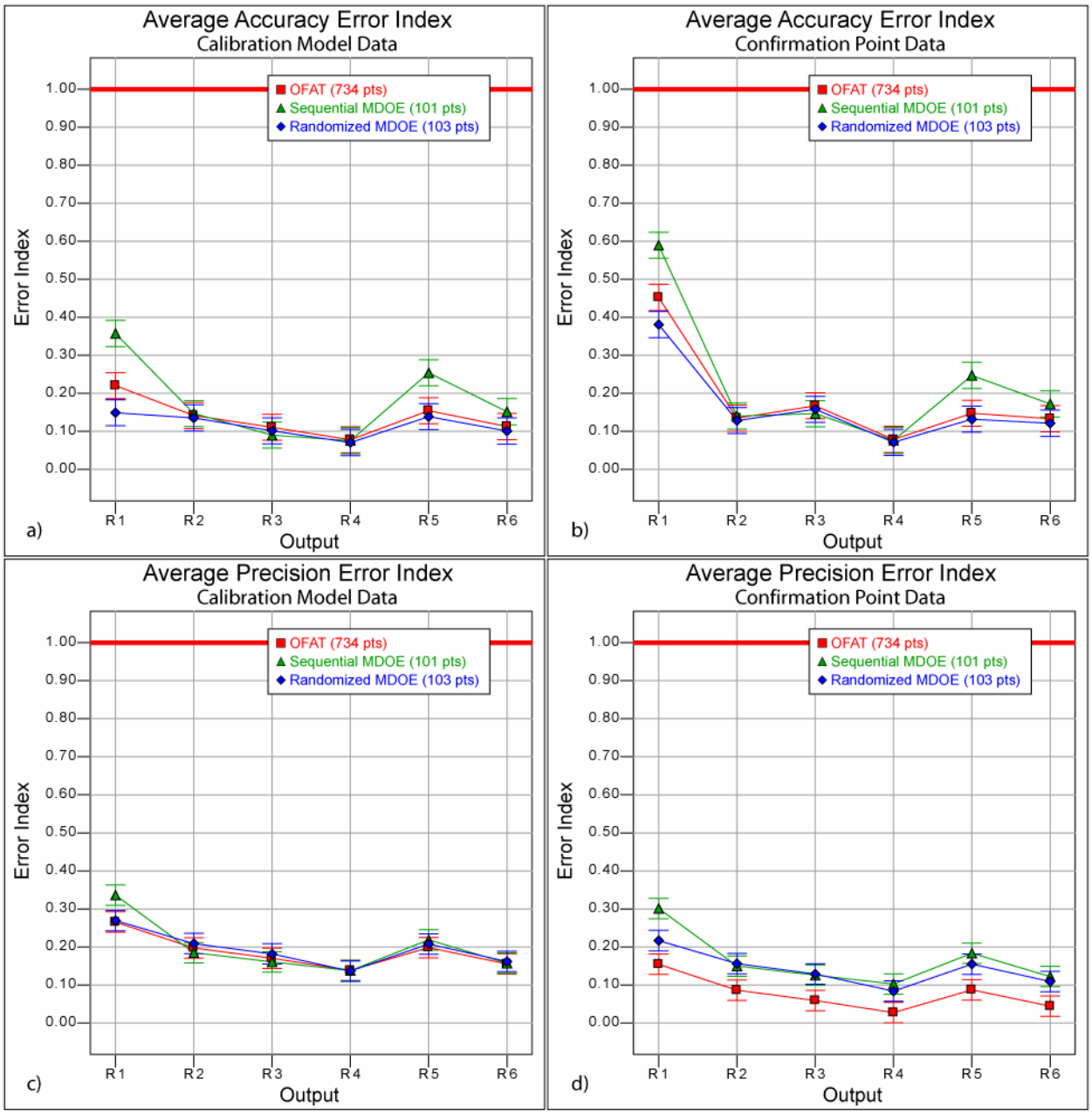

Figure 7. Interactions between experiment design and balance output. a) Accuracy for calibration model data; b) Accuracy for confirmation point data; c) Precision for calibration model data; d) Precision for confirmation point data. Insignificant interaction for precision, but significant interaction for accuracy due to output R1.