# Comparison of Resource Requirements for a Wind Tunnel Test Designed with Conventional vs. Modern Design of Experiments Methods

Richard DeLoach* and John R. Micol†
*NASA Langley Research Center, Hampton, Virginia, 23681*

**The factors that determine data volume requirements in a typical wind tunnel test are identified. It is suggested that productivity in wind tunnel testing can be enhanced by managing the inference error risk associated with evaluating residuals in a response surface modeling experiment. The relationship between minimum data volume requirements and the factors upon which they depend is described and certain simplifications to this relationship are realized when specific model adequacy criteria are adopted. The question of response model residual evaluation is treated and certain practical aspects of response surface modeling are considered, including inference subspace truncation. A wind tunnel test plan developed by using the Modern Design of Experiments illustrates the advantages of an early estimate of data volume requirements. Comparisons are made with a representative One Factor At a Time (OFAT) wind tunnel test matrix developed to evaluate a surface to air missile.**

## Nomenclature

| | |
|---|---|
| *Adequate* | Property of a response model for which predicted responses do not differ from physical measurements by more than what the investigator considers acceptable |
| *ALPT* | Model total angle of attack, deg |
| *ANOVA* | Analysis of variance |
| *AoA, ALP* | Model angle of attack, deg |
| *ALPTUN* | Model angle of attack, normal force in vertical plane, deg |
| *BETTUN* | Model angle of sideslip, normal force in vertical plane, deg |
| *CAF* | Forebody axial force coefficient (Body *and* Missile Axis) |
| *Candidate list* | Fixed set of achievable independent variable combinations available to fit a response model |
| *CLMNR* | Non-rolled pitching moment coefficient, tunnel fixed (Missile Axis) |
| *CLL* | Rolling moment coefficient (Body *and* Missile Axis) |
| *CLNNR* | Non-rolled yawing moment coefficient, tunnel fixed (Missile Axis) |
| *CNNR* | Non-rolled normal force coefficient, tunnel fixed (Missile Axis) |
| *CYNR* | Non-rolled side force coefficient, tunnel fixed (Missile Axis) |
| *DEL1* | Canard #1 deflection angle, deg |
| *DEL2* | Canard #2 deflection angle, deg |
| *DEL3* | Canard #3 deflection angle, deg |
| *DEL4* | Canard #4 deflection angle, deg |
| *df* | Degrees of freedom |
| *Design space* | A coordinate system with each axis corresponding to one independent variable. Each “site” in this space represents a unique combination of factor levels. Also called the “inference space.” |
| *F* | Ration of effects variance to error variance |
| *Fcrit* | Critical F: Criterion for significant effect |
| *Factor* | A variable for which levels changes are planned in the course of an experiment |

* Senior Research Scientist, NASA Langley Research Center, MS 238, 4 Langley Blvd, Hampton, VA 23681, Associate Fellow.
† Ground Facilities Testing Technical Lead for Business Partnership, NASA Langley Research Center, MS 225, Bldg 1236, Rm 208, Senior Member.

| | |
|---|---|
| $H_0$ | Null hypothesis |
| $H_A$ | Alternative hypothesis |
| *Inference space* | A coordinate system with each axis corresponding to one independent variable. Each “site” in this space represents a unique combination of factor levels. Also called the “design space.” |
| *Level* | A specific value of a factor or independent variable (2° is a level of the factor angle of attack) |
| *LSD* | Least Significant Difference |
| *Mach* | Mach number |
| *MDOE* | Modern Design of Experiments |
| *MS* | Mean Square, also Variance |
| *OFAT* | One Factor At a Time |
| *P-value* | Probability that an effect is due to chance |
| *PHIS* | Balance roll angle, deg |
| *Prediction error* | Difference between measured and modeled response. Attributable to modeling imperfections and ordinary experimental error |
| *PS* | Static pressure, psia |
| *PT* | Total pressure, psia |
| *Q* | Dynamic pressure, psia |
| *QA* | Quality Assurance |
| *Regression* | A process of modeling responses as a function of factors, typically in such a way as to minimize prediction error |
| *Residual* | Predicted response minus measured response |
| *Response* | A variable that depends on the values of various factor levels |
| *RN* | Reynolds number |
| *RSM* | Response Surface Model, Response Surface Methods |
| *Sample* | A collection of N individual data points, where N can be any integer, including 1 |
| *Site* | A point in the design space representing a unique combination of independent variable levels |
| *SS* | Sum of Squares |
| *TS* | Tunnel static temperature, deg F |
| *TT* | Tunnel total temperature, deg F |
| *UPWT* | Unitary Plan Wind Tunnel |

## I. Introduction

A surface-to-air missile model (Fig. 1) tested in the Unitary Plan Wind Tunnel (UPWT) at NASA’s Langley Research Center using a conventional One Factor At a Time (OFAT) experiment design resulted in a database that the authors were recently asked to re-examine. Specifically, we were asked to determine if the Modern Design of Experiments (MDOE) could generate improvements in quality and productivity in a missile test similar to those that have been achieved in fixed-wing wind tunnel tests.[1–9]

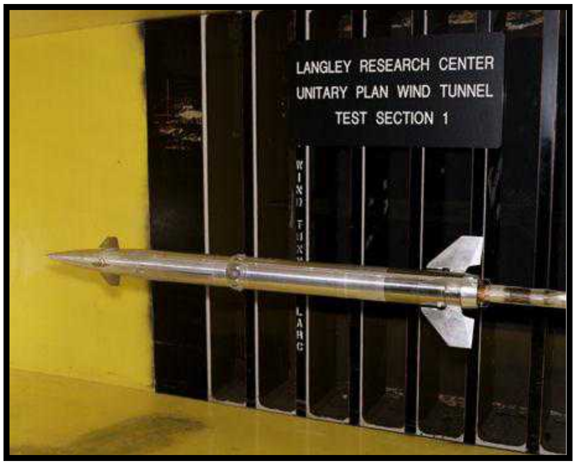


**Figure 1. Missile test article used in wind tunnel test.**

The initial phase of this OFAT/MDOE comparison focused on an objective assessment of the UPWT measurement environment. Both the random and the systematic components of unexplained variance in the T1878 data set were quantified. This information is useful in the planning stage of any MDOE experiment design, as MDOE data volume requirements depend on the magnitude and nature of the unexplained variance that can be anticipated in a worst-case scenario, and minimizing the volume of data required for a test is a key cost control tactic. These unexplained variance estimates were reported[10] at the 27th AIAA Aerodynamic Measurement Technology and Ground Testing Conference in Chicago.

This paper focuses on the resource savings that MDOE testing can achieve compared to conventional OFAT testing, using T1878 as specific but representative case study. Flight system development projects that rely upon wind tunnel testing are generally very expensive, and the cost of capital required to fund such projects can run into the hundreds of thousands of dollars per day for a development project as expensive as a new commercial jet transport, for example. One effective tactic to reduce such total testing costs is to limit the volume of data acquired to the minimum needed to achieve specific technical objectives. The question of data volume is therefore addressed early in an MDOE experiment design.

The data volume question is also addressed in conventional OFAT testing, of course, but from a rather different perspective. Data volume is considered a metric of productivity in OFAT testing, and such tests are therefore designed to *maximize* data volume rather than to minimize it, the adverse impact of this approach on cost and cycle time notwithstanding.

Both the OFAT and the MDOE practitioner understand that they can afford to stop consuming test resources as soon as the objectives of the test are achieved, but OFAT testing is an exhaustive enumeration methodology in which the objective is to make a measurement at each of the largest possible number of factor combinations that may be of interest. In practice, the number of factor combinations of interest in a wind tunnel test generally far exceeds the number that can be set with available resources. As a consequence, it is rare in an OFAT test to acquire all the data that is desired, and the authors' experience suggests that it would be unusual to terminate an OFAT wind tunnel test while resources are still available to acquire additional data. For practical purposes, then, the classic OFAT answer to the question of how much data to acquire is, "all the data we can take before the available resources (time and money) are exhausted." A testing methodology featuring an exit strategy that is based on the exhaustion of all available resources can never be "low cost," whatever else it may have to recommend it.

Another motive for minimizing data volume, besides the reduction in associated direct operating costs and cycle time, is that it preserves limited test resources so that they can be used for various quality assurance and quality assessment tactics that must be ignored with a high data volume acquisition strategy. We therefore begin the paper with a brief review of the necessity of such quality assurance and quality assessment in Section II. A means of estimating data volume requirements sufficient to satisfy quality assurance specifications is presented in Section III. In Section IV, the original OFAT experiment design for test T1878 is described, and an MDOE alternative is proposed. The two designs are compared in Section V on the basis of quality and productivity. Certain key points are expanded upon in the Discussion section, Section VI, which is followed by a summary of the principal points of the paper in Section VII.

## II. Quality Assurance and Quality Assessment as Motivation for Minimizing Data Volume

OFAT experiment designs are generally predicated on the erroneous assumption that chance variations in the data ("random error") occur about sample means that are stable with time. That is, the general assumption is that if an $n$-point sample is reproduced after an interval of arbitrary duration, the sample means will agree within an error that is proportional to the square root of the random error variance. For large $n$, this anticipated error can be quite small, and it is therefore not uncommon for rather ambitious quality assertions to accompany OFAT wind tunnel results. Unfortunately, it is equally uncommon for the mean values of reproduced samples to actually agree with such high precision.

There are many candidate reasons to explain the practical difficulties encountered by experienced aerodynamicists in obtaining consistently reproducible empirical results. The results of a test reproduced in another facility may depend on the effects of between-facility differences, for example. These include differences in instrumentation, wall effects and other scaling factors, flow characteristics, and even personnel skills and experience. Some subset of these factors may be in play even if the test is reproduced at a later date in the same facility. However, one of the most likely explanations, and one that is commonly overlooked in OFAT testing, is that the basic premise of time-invariant sample means is flawed. The fact is that chance variations in the response measurements occur about sample means that often change gradually over time. That is, in addition to the well-recognized random component of experimental error, and the time-invariant bias errors that are also well-known (which may be colloquially described as "11-inch ruler" errors), there is a time-dependent bias error resulting from a systematic (not random) component of the unexplained variance.

This systematic variation is attributable to "covariate effects," comprised of natural factors that influence the response variables measured in an experiment, but that are not under the control of the experimentalist. Perhaps the archetypical example is temperature, systematic changes in which can influence the calibration constants of various instruments, can induce thermal expansions and contractions that influence wall effects, and can change the Young's modulus of material comprising the sting, affecting its stiffness and thus the sting bending corrections. Other covariate effects include systematically varying flow angularity as described by Oberkampf et al.,[11–14] and the desiccation of force balance strain gages as moisture leaches out of them during the course of a wind tunnel test, altering the gage sensitivity over time, per Steinle et al.[15–17] Even personnel performance can change systematically, either diminishing through fatigue, or improving through what are called "learning effects" that result in improved performance as a certain procedural routine settles in over the long term of an extended test. These are but a few of what is an uncountable number of potential covariate effects, including many that are unknown and unknowable.

Individual covariate effects can be small in an absolute sense, but their combined effects can still induce errors that are a significant fraction, and often a significant multiple, of the small error budget that is typical of a high-precision wind tunnel test. Furthermore, the effect of this variable bias error is to cause the errors in a series of sequential measurements to be correlated; if the previous measurement is biased slightly higher than the true value due to some covariate effect that is persisting in one direction (is not random), the next measurement is also more likely to be high than to be low. The effect of correlated (non-independent) experimental errors is to bias both the location statistics (sample means) and the dispersion statistics (variance estimates) in any sample of data afflicted with such a lack of statistical independence.[3] Since every sample of experimental data is comprised of random variables characterized entirely by their location and dispersion statistics, such biases in location and dispersion virtually guarantee an erroneous experimental result whenever covariate effects are in play. Furthermore, the magnitude, direction, and rate of change of covariate effects reflect localized conditions that are not replicated precisely from test to test. This no doubt explains much of the historical difficulty in achieving reliably reproducible results from one OFAT wind tunnel test to the next when statistical independence is simply declared and not explicitly assured.

The MDOE method employs quality assurance tactics intended to ensure statistical independence by overtly disrupting the correlation of experimental errors in a sequence of factor level changes when covariate effects are in play. This ensures that sample statistics are unbiased estimators of the population parameters they are intended to represent. However, to be able to afford the resources (mainly time) needed to execute various quality assurance tactics in a wind tunnel test, it is necessary to develop a test plan that minimizes the volume of data necessary to adequately predict test article performance over the range of independent variables that is of interest.

## III. Data Volume Requirements

Specific data volume requirements depend on the nature of the experiment, but a typical force/moment/pressure test is often designed as a Response Surface Modeling (RSM) experiment. The authors would have cast the missile test under evaluation in this exercise as an RSM experiment had it been developed from the beginning using MDOE methods.

An RSM experiment seeks to develop a mathematical relationship between response variables such as forces and moments, and independent variables such as angle of attack, Mach number, and angle of sideslip. If there is any basis for selecting a specific functional form of such a model, perhaps based on physical first principles or the researcher's subject matter expertise regarding the test article under study, then it is possible to simply fit a set of experimental data to such a function.

The more common case is that there is no a priori knowledge of a preferred function, and the researcher relies upon a general Taylor series expansion about some specified design-space site to represent the response variable over some limited range of variables comprising a near neighborhood of that site. A linear transformation of independent variables can convert such a Taylor series into a low-order polynomial in those conveniently transformed variables.[18]

Experience suggests that polynomials of no higher order than four are often adequate to represent aerodynamic forces, moments, and pressures over useful ranges of the independent variables. The RSM process then consists initially of estimating from experience and subject-matter expertise the order of polynomial adequate to represent a given response over a specified range of independent variables. This uniquely determines the number of terms in such a polynomial. The smallest volume of data necessary to fit such a polynomial is one point per term in the polynomial, so this represents a lower limit on the volume of data that must be specified in the experiment design.

The following formula can be used to quantify the number of parameters, *p* (including the intercept term) in a $d^{th}$-order polynomial function of *k* variables:

$$p = \frac{(d+k)!}{d!k!} \tag{1}$$

Consider, for example, a fourth-order polynomial in three variables. In such a case, $d = 4$, $k = 3$, and *p* therefore is 35. Such a polynomial would have 35 terms and it would therefore require at least 35 data points to fit it, using ordinary regression methods for example.

It is important to note that a relatively low-order polynomial cannot always adequately represent a given response over the entire design space of interest. One of the initial steps in the RSM process might then be to estimate the range of independent variables over which a polynomial of specified low order (fourth-order, say) is adequate to represent the data. The entire design space can then be represented in a piecewise continuous way, using response models representing multiple adjacent "neighborhoods" to span it. Some related issues are considered in the Discussion section below.

Even when a polynomial with *p* terms is adequate to span a given range of the design space, more than *p* terms might still be needed within that limited range in order to assure that the response model prediction uncertainty is small enough to satisfy precision requirements. The number of additional data points that are needed beyond the *p* required per Eq. (1) to fit a $d^{th}$-order polynomial function of *k* variables depends on three factors.

The first factor influencing how many more than *p* points are needed is the variability of the measurement environment. It is intuitively clear that a greater level of intrinsic variation in the data due to experimental error would call for a greater volume of data to generate a mathematical response model with a given level of precision. Those data acquired beyond the minimum *p* points needed to generate any model at all combine through a kind of "internal replication" in the fitting process, to cause some of the positive and negative errors to cancel. The concept of *unexplained variance* is key to understanding the relationship between experimental error and response model data volume requirements.

All samples of experimental data not comprised of identical replicates feature the quantifiable "differentness" property known as variance; and since perfectly identical replicates are an abstraction not observed in nature, it is safe to assume that any sample of data used to fit a response surface model will be characterized by variance. Most of this variance is intentionally induced in the course of the experiment, and can be explained in terms of changes that were made to various independent variables. Absent such explained variance, the data sample would contain little useful information to reveal how responses such as the forces and moments depend on the independent variables.

Unfortunately, not all of the variance in a sample of experimental data can be explained in terms of known factor changes. There is always a residual component of the total variance that remains *unexplained*, which has the effect of inducing uncertainty in experimental results. The quality of the measurement environment, upon which data volume requirements depend in an RSM experiment, can be characterized by how much *unexplained* variance there is in the sample of data used to fit a response model. The unexplained variance is attributable both to random experimental error and to systematic (not random) error. The systematic component of the unexplained variance often dominates the random component, and is caused by persisting and slowly varying covariate effects such as temperature, flow angularity changes, instrument drift, and the like. The paper reporting the initial phase of this investigation focused on quality issues,[10] and contains more detail about the unexplained variance in the missile test that serves as an illustration of scaling principles in this paper. As will be shown below, an adequate response model requires that we augment the minimum *p* points by a factor that is directly proportional to the unexplained variance of the data (square of the standard deviation).

The second factor dictating how much more data is necessary than the minimum *p* points needed to fit a *p*-parameter model is the tolerance requirement imposed during the planning phase of the experiment. The specification of such a tolerance is necessary in order to quantify resource requirements in advance. It represents the smallest unacceptable difference between a measured response and one predicted by the model for the same combination of independent variable levels. That is, we require the model to represent the data well enough that the difference between measurement and prediction be *within* tolerance (*less* than the tolerance, not "less than or equal to"). We will show that the factor by which the minimum *p* points must be augmented to generate an adequate response model is inversely proportional to the square of the tolerance. This means that wider tolerances require fewer points than tighter tolerances, as one would expect. After deriving below a general formula for estimating data volume requirements in terms of arbitrary tolerance limits, we will consider a specific case in which the tolerance is identified as the smallest resolvable difference between two replicated response measurements.

### A. Inference error risk management

The third factor defining how many more points must be acquired beyond the bare minimum of *p* necessary to fit a *p*-parameter polynomial response model is the inference error risk one is willing to accept in evaluating the residuals. The residuals are simply the differences between measured responses and predictions generated by the response model at the same point. Absent information from sources external to the experiment, residuals provide the only available information about the adequacy of the response model. Their careful evaluation is therefore crucial to assessing the adequacy of a candidate response model.

The physical measurements themselves always feature some experimental error; and because the response model is fitted from an ensemble of such imperfect data points, even a model that fits the data well will generate response predictions that also have some error. Our task is to infer from the measured and predicted responses whether the model adequately represents the data, which we take as a surrogate for the unknown true response. But because the measurement error and prediction error generally differ at each point, we cannot demand perfect agreement as a condition for inferring model adequacy. That is, the residuals will be non-zero for adequate and inadequate models alike.

For every fitted point, we must therefore make an inference as to whether the corresponding residual, consistently non-zero, is small enough to safely attribute it to ordinary experimental error, or whether it is sufficiently large to suggest some systematic lack of fit in the response model. Whether we infer that the residual validates the adequacy of the model or not, our inference might be right or it might be wrong. The smaller the risk of an improper inference that we are willing to accept, the more data we have to acquire.

It is not possible to drive the probability of making either inference error to zero with any finite volume of data, but we can acquire a volume of data that is large enough to drive the probability of making either error to some acceptably low level, without acquiring significantly more data than is necessary to do so. We can therefore manage inference error risk through the size of the data sample. By understanding the relationship between inference error risk and data volume (a surrogate for cycle time and direct operating costs), we can reduce costs to the minimum level consistent with our risk tolerance.

To derive a useful mathematical relationship between cost (data volume) and risk, we must postulate some criterion by which to distinguish between residuals that are too large and those that are "small enough." If a specific residual is no larger than this criterion, we will infer that the response model adequately represents the data at that point, and if the magnitude of the residual is larger than this criterion, we will infer that the model does not adequately represent the data there. The strict binary nature of this outcome is by design, resulting in an inference with one of four possible outcomes for each data point used to fit the response model:

1) We might *correctly* infer that the model *does* adequately represent that data point.
2) We might *incorrectly* infer that the model *does* adequately represent that data point.
3) We might *correctly* infer that the model *does not* adequately represent that data point.
4) We might *incorrectly* infer that the model *does not* adequately represent that data point.

It is convenient to cast this problem in terms of formal hypotheses. We construct a null hypothesis, denoted $H_0$, asserting that there is no significant difference between measurement and prediction for a given data point and that the response model is therefore adequate to represent the measurement at that point. The corresponding alternative hypothesis, denoted $H_A$, asserts that the difference between measurement and prediction at a given point is too large to attribute to experimental error, and must therefore reflect an inadequate prediction model.

The null and alternative hypotheses are mutually exclusive, and so the formal process of making an inference simply consists of deciding which hypothesis to reject. If one erroneously rejects the null hypothesis, $H_0$, the resulting error is known as a Type I or alpha inference error. Erroneously rejecting $H_A$ results in a Type II or beta error. Thus, a Type I inference error will occur if we reject an adequate model, and a Type II inference error will occur if we validate an inadequate model. In considering how much data to acquire, we are primarily interested in situations that result in one of these two improper inferences, representing cases 2) and 4) above.

Figure 2a illustrates a reference distribution associated with the null hypothesis. The distribution is centered on a residual of zero, since there is no difference between measurement and model prediction under $H_0$. A normal distribution is assumed from the Central Limit Theorem, with a variance that reflects the model prediction uncertainty. The area under the reference distribution and to the right of the dashed line at $x^*$, denoted "$\alpha/2$," represents the probability of committing a Type I (alpha) error by erroneously rejecting the null hypothesis due to a prediction that in the case illustrated in the figure, overstates the response by $x^*$. By symmetry, there is an equal probability of *understating* the response by the same amount. We say in such a case that the null hypothesis is "two-sided." Therefore, given the prediction variance represented by the dispersion of the reference distribution in Fig. 2a, there is a total probability of α that the model will predict a response that differs from the measured value in either the positive or negative direction by *x** or more when the true difference is zero.

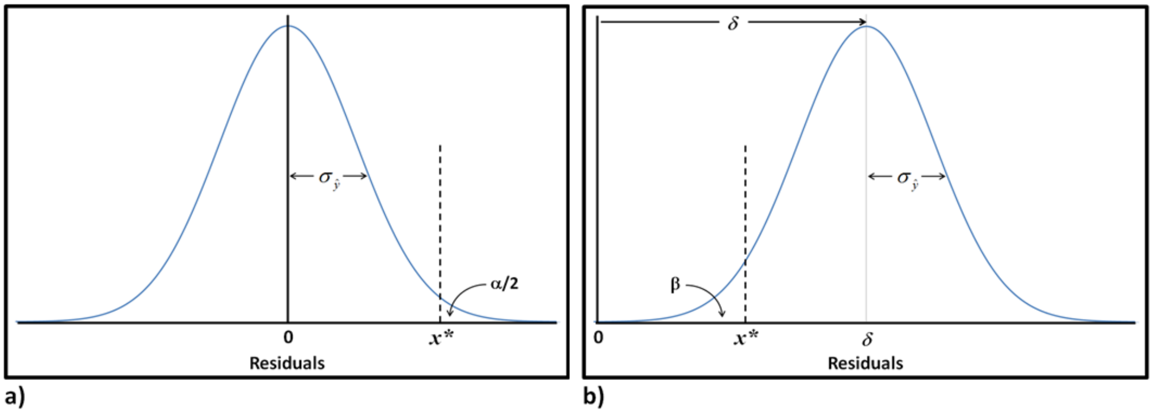


**Figure 2. Normal reference distribution for model predictions under a) the null hypothesis, and b) the alternative hypothesis.**

Note that for a specified α, *x** simply represents the prediction precision of the response model; it is unrelated to any adequacy requirement that may be specified for the test. For a polynomial response model, *x** depends on the volume of data used to fit the model and the number of parameters in the model, as well as the intrinsic variability of the measurement environment. It is also a function of the specific combination of independent variables for which the prediction is made (the specific site in the design space). It has nothing to do with any requirements for adequacy; *x** simply represents how good the model *is*, not how good it is *supposed* to be.

For a given model, we expect the predicted response to differ from the true response (assumed to be represented by the measured data) by an amount that is dictated by the reference probability distribution. In the case of Fig. 2a, there is a probability of 1 – α that the difference between the predicted response and the true response lies within ±*x**. This may or may not be satisfactory, depending on the adequacy specification. The adequacy specification, which is the difference between measured and predicted responses that is just too large to be satisfactory, is an externally generated number that depends only on the requirements of the principal investigator. It is completely arbitrary, and its specification is independent of *x**. (Note also that by this definition, we require the difference between each prediction and the corresponding measurement to be *less than* the adequacy specification, not "less than or equal to" it.)

Figure 2b illustrates a limiting case of the alternative hypothesis. In this figure, the RSM model predicts a response that differs from the corresponding measured response by just enough to be deemed inadequate. That is, the model misrepresents the measured response by an amount $\delta$, which is the adequacy level specified by the principal investigator. The reference distribution is centered on $\delta$ because under the alternative hypothesis, the model predicts the wrong value, one that is biased from the true value (again assumed to be represented by the measurement) by $\delta$ and is therefore just out of tolerance.

The quantity *x** in Fig. 2b is the same criterion level as in Fig. 2a. Here, the area under the reference distribution that is to the left of x*, denoted as β, represents the probability of erroneously rejecting the alternative hypothesis when the prediction is just out of tolerance. That is, any prediction resulting in a residual that is *smaller* than *x** is assumed to be drawn from the probability distribution in Fig. 2a that corresponds to the null hypothesis (model is adequate), while any prediction resulting in a residual that is *larger* than *x** is assumed to be drawn from the probability distribution in Fig. 2b that corresponds to the alternative hypothesis (model is not adequate). If an *adequate* model produced a residual with a magnitude larger than the *x** criterion because of an unusually large experimental error, we would erroneously reject the null hypothesis for that specific combination of factor levels, committing a Type I inference error. This is expected to occur with a probability of α. Likewise, if an *inadequate* model produced a residual smaller than the *x** criterion just by chance because of an unusually large experimental error, we would erroneously reject the alternative hypothesis for that specific combination of factor levels, committing a Type II inference error. This is expected to occur with a probability of β.

It is convenient to represent the $x^*$ criterion as a multiple of the standard deviation in the reference distribution. In Fig. 2a, if the variance of the reference distribution is well known, the number of standard deviations that $x^*$ is away from zero depends only on the value of $\alpha$, and is denoted by $z_\alpha$. For example, it is well known that for $\alpha = 0.05$, $z_\alpha$ is approximately two standard deviations (1.960, to be more precise). Such values, called standard normal deviates, are tabulated in typical statistics references for various values of $\alpha$ and they can also be computed using widely available computer software packages. If the variance of the reference distribution is not known precisely but is only estimated from a finite subsample of the data, say (especially a small subsample), then the number of standard deviations that $x^*$ is away from 0 in Fig. 2a depends not only on $\alpha$, but also on the number of degrees of freedom, $\nu$, that were available to estimate the standard deviation. In that case, instead of using the standard normal deviate, $z_\alpha$, one would use a t-statistic, $t_{\alpha,\nu}$, which can also be obtained from standard statistical tables for various combinations of $\alpha$ and $\nu$, or computed with readily available software.

### B. Scaling the Experiment—The General Case

By "scaling" we mean simply estimating the minimum volume of data (and associated cycle time and direct operating costs) necessary to construct a model that is consistent with our inference error risk tolerance, given the measurement environment and a specified requirement for adequacy in estimating the measured responses with the model. With obvious notational changes for the alternative hypothesis, we have from Fig. 2 the following two expressions for $x^*$:

$$x^* = z_\alpha \sigma_{\hat{y}} \tag{2}$$

and

$$x^* = \delta - z_\beta \sigma_{\hat{y}} \tag{3}$$

where the subscript on $\sigma$ implies that this is the standard deviation of the predicted response. From Eqs. (2) and (3), it is clear that

$$x^* = \delta - z_\beta \sigma_{\hat{y}} = z_\alpha \sigma_{\hat{y}} \rightarrow \delta = \left(z_\alpha + z_\beta\right)\sigma_{\hat{y}} \tag{4}$$

The standard deviation of the prediction varies with location in the design space, tending to be smaller where data points are clustered in close proximity and generally tending to be larger near the design space boundaries than at interior points. However, Box and Draper[18] show that for any order polynomial, in any number of independent variables, while the standard error of the prediction varies from site to site within the design space, the *average* prediction variance across all points used to fit the response model is always the same, namely:

$$\sigma_{\hat{y}} = \sigma\sqrt{\frac{p}{n}} \tag{5}$$

where $p$ is computed from Eq. (1), $n$ is the number of points used to fit the model, and $\sigma$ is the ordinary standard deviation in the data. Because there must be at least one data point for every parameter in the fitted model, $n$ is always at least as large as $p$ and is generally larger. The average standard error for a response model is therefore never greater than for conventional single-point measurements, and is generally smaller.

Inserting Eq. (5) into Eq. (4):

$$\delta = \left(z_\alpha + z_\beta\right)\sigma_{\hat{y}} = \left(z_\alpha + z_\beta\right)\sigma\sqrt{\frac{p}{n}} \tag{6}$$

Solving Eq. (6) for *n*:

$$n = p\left(z_{\alpha} + z_{\beta}\right)^{2} \frac{\sigma^{2}}{\delta^{2}} \tag{7}$$

We can insert Eq. (7) into Eq. (5) as follows:

$$\sigma_{\hat{y}} = \frac{\delta}{\left(z_{\alpha} + z_{\beta}\right)} \tag{8}$$

Equation (8) shows that reducing the tolerance for inference errors (making $\alpha$ and $\beta$ smaller, which increases the standard normal deviates $z_{\alpha}$ and $z_{\beta}$), results in smaller prediction errors and thus a more precise response model, as would be expected. Increasing the tolerance for error, $\delta$, increases the prediction error because a greater tolerance for error results in the need for fewer data points per Eq. (7). We pay for this savings in data acquisition resources through slightly more uncertainty in the average model predictions.

This result gives the volume of data that is necessary in a measurement environment characterized by a standard deviation of $\sigma$ in order to fit a *p*-term polynomial response model with sufficient precision that one can infer with probability $1 - \alpha$ that the model is adequate to represent the data within $\delta$, while identifying points that are marginally poor fits with a probability $1 - \beta$. (Residuals that are larger than $\delta$ can obviously be detected with even higher confidence.) As indicated above, the volume of data is greater than the minimum *p* points required to fit a *p*-term polynomial. The excess is proportional to the variance of the data, $\sigma^2$, and inversely proportional to the square of the tolerance, $\delta$. It is also a function of $\alpha$ and $\beta$, the probabilities specified as acceptable for committing Type I or Type II inference errors, respectively.

### C. Establishing accuracy criteria

Equation (7) is sufficient to estimate the volume of data required for a general RSM experiment, but it requires a consensus as to what the tolerance, $\delta$, should be, and it also requires that the experiment designer know the standard deviation for data to be acquired in the tunnel for which the test is being designed. Neither piece of information ever seems to be readily available in practical situations. Many facilities do not document their intrinsic variability in very useful ways, and researchers are often reticent to state in specific, quantitative terms just how good a response model has to be for it to serve as a surrogate for an exhaustive set of individual measurements.

The basic structure of Eq. (7) suggests a practical means to solve both of these problems. Note that the volume of data needed for an RSM experiment depends on the ratio of the standard error, $\sigma$, to the tolerance, $\delta$. This suggests that instead of specifying individual numerical values for either quantity, one might express the tolerance as some function of $\sigma$. For example, one might specify as a tolerance the Least Significant Difference (LSD) between two measurements. With such a specification, the principal investigator agrees to the proposition that for any combination of independent variables, the response model prediction is to be regarded as adequate if it is no further from a physical measurement than the smallest difference in two physical measurements that can be resolved at that same point with, say, 95% confidence. That is, a predicted response will be considered adequate if it is not possible to resolve a significant difference between it and a physical measurement made at the same point.

To compute the least significant difference between two measurements, we begin with the general error propagation formula for independent factors. This well-known formula allows one to propagate errors in various factors into some other quantity that is a known function of those factors. Consider some quantity, *y*, that is a function of other factors, $y_i$, so that $y = f(y_1, y_2, \ldots y_n)$. Assume the uncertainty in $y_i$ is known for all *i*, and that we wish to know the uncertainty in *y*. we can calculate it using this formula, assuming all the $y_i$ values are independent:

$$\sigma_{y}^{2} = \left(\frac{dy}{dy_{1}}\right)^{2} \sigma_{y_1}^{2} + \left(\frac{dy}{dy_{2}}\right)^{2} \sigma_{y_2}^{2} + \cdots + + \left(\frac{dy}{dy_{n}}\right)^{2} \sigma_{y_n}^{2} \tag{9}$$

Apply this to the simple case in which *y* is the difference between two replicated measurements, each with the same standard deviation, σ. We then have:

$$y = y_2 - y_1 \tag{10}$$

Since the square of the derivative of *y* with respect to either $y_1$ or $y_2$ is 1, and since the standard deviation in both $y_1$ and $y_2$ is σ, Eq. (9) for this case reduces to:

$$\sigma_y^2 = 2\sigma^2 \tag{11}$$

The standard error corresponding to the variance in Eq. (11) is simply its square root, and if we define the Least Significant Difference as nominally two of that, and set the tolerance, *δ*, to be the Least Significant Difference, we have:

$$\delta = 2\sqrt{2}\sigma \tag{12}$$

It only remains to insert Eq. (12) into Eqs. (7) and (8):

$$n = \frac{1}{2}p\left(\frac{z_\alpha + z_\beta}{2}\right)^2 \tag{13}$$

$$\sigma_{\hat{y}} = \sigma\sqrt{\frac{p}{n}} = \left(\frac{2\sqrt{2}}{z_\alpha + z_\beta}\right)\sigma \tag{14}$$

Assume that a consensus is reached involving the inference error risk probabilities, α and β. Recall that α is the acceptable probability of erroneously rejecting a null hypothesis claiming no significant difference between predicted and measured responses. To make this inference error is to conclude that a given residual is large enough to imply an invalid response model when the model is in fact perfectly adequate. Likewise, recall that β is the acceptable probability of erroneously rejecting an alternative hypothesis claiming that there is a significant difference between predicted and measured responses. To make this inference error is to conclude that a given residual is small enough to imply an adequate response model when the model is in fact invalid. If in both cases we adopt the common convention in aerospace testing that “95% confidence” is sufficient to make an inference under uncertainty, this would establish both α and β as 0.05. The corresponding standard normal deviates, $z_\alpha$ and $z_\beta$, would then be 1.960 and 1.645, respectively. (They differ because the null hypothesis is two-sided in this instance, while the alternative hypothesis is always one-sided).

Inserting these values for the standard normal deviates into Eqs. (13) and (14):

$$n = 1.625p \tag{15}$$

$$\sigma_{\hat{y}} = 0.785\sigma \tag{16}$$

Equation (15) confirms the earlier assertion that in order to conform to specific tolerance levels and inference error risk requirements, it would be necessary to acquire more data than the minimum of one point per fitted model parameter. In this case, that bare minimum needed to be increased by 62.5%. Note, however, that it is not necessary to acquire *more* data than this to meet the response model adequacy requirements documented in this typical example, in which the model would be expected to generate response model predictions that on average cannot be distinguished with 95% confidence from physical measurements made at any given point in the design space. Note also that a data acquisition strategy based on acquiring data until all resources are exhausted is likely to produce a data sample that is excessive. Recall, however, that Eq. (15) describes only the volume of data necessary to fit a response model over one inference subspace, and if it is necessary in the experiment to fit multiple such subspaces, then this data volume would have to be adjusted accordingly.

Equation (16) simply describes for this specific example what Eq. (5) reveals for the general case; namely, that because we acquire more data than the minimum necessary to fit a polynomial with $p$ terms, the uncertainty in response model predictions will be less than if we acquired the absolute minimum volume of data necessary to fit the model. In such a case, in which $n = p$, the standard error in the response model and the standard error in an individual physical measurement would be the same, per Eq. (5). But the 62.5% increase in data volume beyond that minimum that was necessary to satisfy modeling tolerance requirements resulted in a reduction in uncertainty compared to single-point measurements. In this case, the uncertainty associated with a model prediction would be over 20% smaller than the uncertainty associated with an individual physical measurement.

If we erroneously reject the null hypothesis (acceptable probability of $\alpha$), we will invalidate an adequate model, the consequences of which would be relatively mild. At worst, such an error would result in time and effort lost in search of an improved model, when no substantive improvement is necessary. However, if we erroneously reject the alternative hypothesis (acceptable probability of $\beta$) by declaring a poor model to be adequate, that could have very serious consequences.

It is therefore the inference error probability, $\beta$, that is crucial, and Eq. (7) reveals how this probability is related to the volume of data that is acquired. From a formal experiment design perspective, it is not too strong a statement to say that the only rationale for incurring the expense and taking the time to acquire an additional data point is to drive the probability of generating an inadequate response model below the acceptable threshold of $\beta$. In other words, the benefit of incurring the cost of the next point is the incremental reduction it provides in the probability of proposing an inadequate model. That probability can never be driven to zero with a finite data sample, no matter how large, but it can be driven arbitrarily close to zero be acquiring enough data. The key to cost containment in an MDOE experiment design is to specify enough data to drive the probability of proposing an inadequate model sufficiently low to satisfy one's tolerance for risk, without specifying substantially more data than is necessary to do so.

For the example considered here of a 15-term model for which the adequacy criterion is defined as the Least Significant Difference between two replicated measurements, Fig. 3 shows the relationship between data volume and the benefit of acquiring that data when the probability of committing the less-serious Type-I inference error (rejecting a good model) is no greater than 0.05. In Fig. 3, Eq. (7) was used to plot $1 - \beta$, the probability of recognizing an inadequate model, vs $n$. The red arrow marks the volume of data for which this probability exceeds 95%, which is 25 points for the example we have been developing.

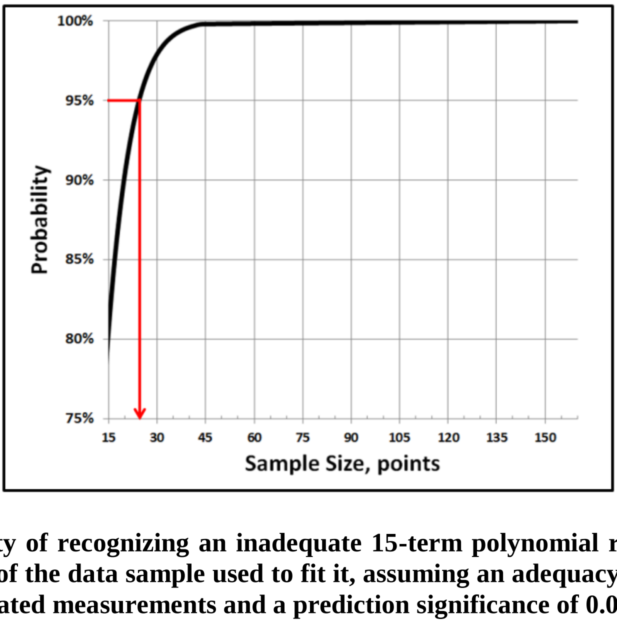


**Figure 3. Probability of recognizing an inadequate 15-term polynomial response model as a function of the size of the data sample used to fit it, assuming an adequacy requirement of the LSD between replicated measurements and a prediction significance of 0.05.**

Note the characteristic shape of the curve in Fig. 3. The benefit of acquiring additional data is a monotonically increasing function of data volume—the more data one acquires, the higher the probability of proposing an adequate model, approaching 100% asymptotically but never reaching it. However, the rate at which this probability increases is highly non-linear. When only a few data points are in hand, each additional point adds significantly to the probability of proposing an adequate model. But a volume of data is rapidly reached beyond which more data has relatively little effect. In other words, the value of the next data point is a monotonically decreasing function of the volume of data already in hand; the next data point is always worth more when relatively little data are available than when a huge volume of data has already been acquired.

Because the value of the next data point decreases with every new point that is acquired, it must eventually fall below the cost of acquiring it. If we continue to acquire data beyond that point, then each point will cost more than it is worth. Such a point of diminishing returns is often reached after a surprisingly small volume of data has been acquired. This contrasts with OFAT testing, in which there is virtually no limit to the value that is placed on more data, and thus no limit on the cost one is willing to incur in acquiring a volume of data that can be regarded as suitable.

## IV. OFAT/MDOE Designs

In this section, we outline an OFAT experiment design executed in the Langley Unitary Plan Wind Tunnel in the fall of 2000 using the missile of Fig. 1 as a test article. We then compare that OFAT design with what would have been the MDOE design of that same test. Differences between OFAT and MDOE approaches to assuring and assessing quality have been reported in a separate paper.[10] This paper focuses on differences in the test matrices, highlighting the resource savings (assumed to be a function of data volume requirements) that can be achieved with a formal experiment design.

### A. Test Article

The original test plan involved measurements of forces and moments on the test article displayed in Fig. 1, a surface to air missile model. The missile had four large tail fins displaced azimuthally at 90° intervals near the aft end of the model, and four smaller control surfaces also displaced azimuthally at 90° intervals, located near the nose of the model. The forward canards could be deflected. The tail fins were fixed in that they did not deflect in this test; however, the entire tail fin assembly was attached to a cylinder that was mounted on bearings, allowing it to rotate freely around the axial axis of the missile. Figure 4 displays the control surface deflection sign conventions.

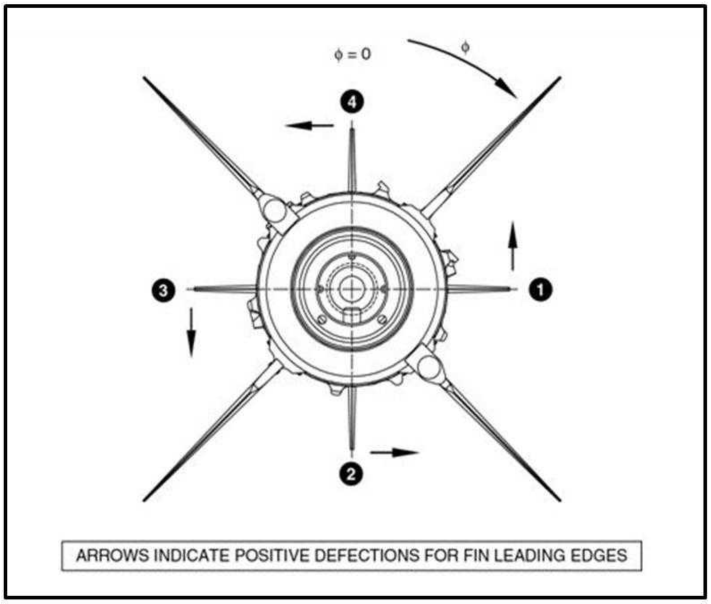


**Figure 4. Deflection sign convention for control surfaces.**

One of the reasons missiles use free-to-roll tails is to eliminate adverse roll effects. For certain missiles, when canards are deflected to roll the missile, the vortices from the canards will interact with the tail fins and cause a rolling moment in the opposite direction. So the rolling moment generated by the deflected canards is reduced, eliminated, or reversed by the tail fin rolling moment. By allowing the tails to rotate freely, the tails can no longer produce a rolling moment and the adverse roll effects of the tail are eliminated.

The tail fin assembly spun very rapidly at some conditions and it remained completely still at other conditions. Whether the fins rotated or not was dependent on the model configuration (primarily the deflection angles of the forward canards) and the model attitude. There was no control of the tail fin roll rate so it was not an independent variable. The tails were free to roll at any rate and in either direction, depending on the model attitude, model configuration, and Mach number. During this test, data were acquired at 30 frames/second for two seconds per data point. All 60 frames of data were averaged to produce a single data point. For the two seconds that data were acquired, the tail could have rotated 0, 10s, or 100s of times. So each data point represents an averaged effect of the spinning tail.

There were no stops or mechanisms to lock the tail in a particular position. If the tail did not spin, it was only because the aerodynamic forces on the fins were balanced. In addition, there was no instrumentation to measure the position or RPM of the tails.

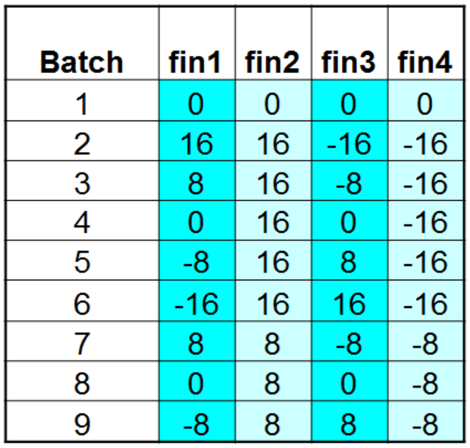
### B. OFAT Experiment Design

The original OFAT experiment design as executed in this test sought to quantify forces and moments as a function of roll angle, total angle of attack, and deflections in the four forward canards, for two discrete Mach numbers. Sideslip angle was recorded but it was not systematically varied as an independent variable. All data were acquired at a constant Reynolds number of $2 \times 10^6$, except for three runs acquired at $4 \times 10^6$ on the next-to-last day of the test.

The forward canards exhibited coordinated deflections in the OFAT test, with the trailing edges of opposite control surfaces moving always in the same direction. In the phi = 0 roll orientation, in which the four fins lie in a vertical plane and a horizontal plane, the two in the vertical plane (called the "yaw plane") were either both deflected left or both deflected right in this test, and always by the same amount. Likewise, the two horizontal ("pitch-plane") canards were either both deflected up or both deflected down, again by the same amount. This means that in terms of control surface deflections of interest, there are really only two independent variables ("planes") rather than four ("canards"), each set at one of five levels.

Each control surface could be set at one of five discrete deflection angles: 0°, ±8°, or ±16°. There were thus $5^4 = 625$ unique configurations possible from the four control surfaces that could each be set at five discrete levels. Because the coordinated nature of the deflections reduced the control surface variables effectively from four to two, there were actually only 25 possible coordinated control surface configurations, of which nine were examined.

**Table 1. Coordinated deflections of forward canards.**



| Batch | fin1 | fin2 | fin3 | fin4 |
|---|---|---|---|---|
| 1 | 0 | 0 | 0 | 0 |
| 2 | 16 | 16 | -16 | -16 |
| 3 | 8 | 16 | -8 | -16 |
| 4 | 0 | 16 | 0 | -16 |
| 5 | -8 | 16 | 8 | -16 |
| 6 | -16 | 16 | 16 | -16 |
| 7 | 8 | 8 | -8 | -8 |
| 8 | 0 | 8 | 0 | -8 |
| 9 | -8 | 8 | 8 | -8 |

The data acquired for each control surface configuration was described by a unique batch number, following a convention of this tunnel. Table 1 displays the pattern of deflections that were considered in the test. There were two additional batches of data for this test: a Batch 10 that featured three end-of-test “quality assurance polars” representing replicates of three similar polars that were acquired in Batch 1 near the start of the test, and a Batch 11 on the last day of the test that consisted two runs with some cable containment rails removed from the side of the missile. The first of these runs was a polar at a fixed roll angle of 0$^0$ spanning the range of ±3$^0$ at 0.5$^0$ increments, and the second was a replicate of that run.

For each control surface configuration, roll angle sweeps were executed that consisted of discreet roll angles set nominally at 11.25$^o$ increments over a range of 0$^o$ to 348.75$^o$. These 31 roll angles were each set at eight angles of attack in the range of -2$^o$ to +20$^o$. The roll sweeps were augmented by two angle of attack polar sweeps, each at a different fixed roll angle. The two angle of attack polars ranged from -2$^o$ to +20$^o$ in 2$^o$ increments. There were thus 31 × 8 = 248 points acquired per configuration in the roll sweeps and 2 × 12 points in the polar sweeps, for a total of 272 alpha-phi point combinations per Mach number, per control surface configuration. Figure 5 displays the distribution of alpha-phi combinations set for each of the configurations in Table 1.

There were two Mach numbers per configuration (1.6 and 2.16) and nine configurations per Table 1, or a total of 272 × 2 × 9 = 4896 individual measurements. There were also eight additional 13-point polars acquired in Batch 0 (no fin deflections) at fixed roll angles set nominally at 0 to 315 in 45 increments. Each of these polars spanned a range of ±3$^o$ in 0.5$^o$ increments, for a total of 104 additional points at each of the two Mach numbers, or 208 extra points for the no-deflection control surface configuration. This resulted in a grand total of 5104 individual measurements in the test matrix, not counting the three quality assurance polars noted above that comprised Batch 10, and the points acquired on the last day in the Batch 11 sample with a different rail configuration. There were actually 5575 total points acquired in this test including various miscellaneous points.

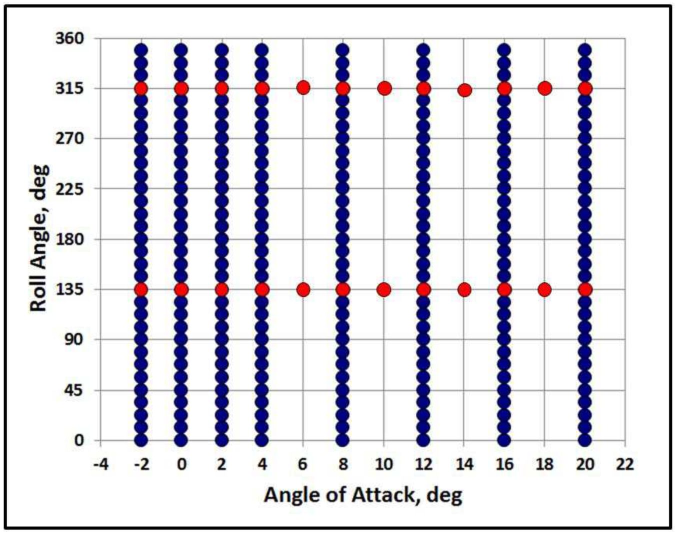


**Figure 5. Design space for OFAT configurations 2–9. Red points are angle of attack sweeps. Blue points are roll sweeps.**

## C. The MDOE Experiment Design

The first step in every MDOE wind tunnel test is always the same; namely, to state the objective of the test in quantitative terms that make its successful completion unambiguous when it occurs. Fortunately, the objective of a typical MDOE wind tunnel test is rather easy to state, and it is the same objective for a wide range of commonly occurring tests. The typical objective of an MDOE wind tunnel test is to gain enough knowledge about the test article to adequately predict its future behavior for all combinations of factor levels of interest within a specified range. “Future behavior” simply refers to the forces, moments, pressures, or other responses that will be measured if the test is reproduced sometime in the future. An “adequate” prediction is one that differs from the true response (for which direct measurements are assumed to serve as a surrogate) by no more than some prescribed tolerance. We limit this requirement to combinations of factor levels within the range actually tested; that is, we do not extrapolate our experimental results to predict the future behavior of the test article out of the range of variables over which it was tested.

Key components in the statement of objectives are then the response variables to be measured, some tolerance level for each of those variables, the independent variables to be changed during the test as a means of injecting explained variance into the resulting data set, and the upper and lower limits of the ranges over which those variables will be changed. For the existing test we might state the objective as follows: "We wish to understand the test article well enough to be able to adequately predict the coefficients of missile axis forces and moments at Mach numbers of 1.6 and 2.16 for any combination of roll angle and angle of attack from $0^o$ to $348.75^o$ in roll and from $-2^o$ to $+20^0$ in angle of attack, for nine discrete control surface deflections. By 'adequate' we mean that for the same combination of independent variable levels, predicted values shall differ from measurements by no more than the smallest difference between replicated measurements that can be resolved with 95% confidence."

We then consider the design space over which we will conduct the experiment, to decide whether to try to fit a single response model (almost never practical) or whether to partition the design space into subspaces. The design subspace boundaries must then be defined, through an appeal to subject matter expertise, an examination of prior data, or some combination. Figure 6 reveals how the normal force coefficient in the missile axis, CNNR, varies with angle of attack for a similar test article over the same angle of attack range we wish to study. We elect to establish an inference subspace boundary at $9^0$, as the dashed line indicates.

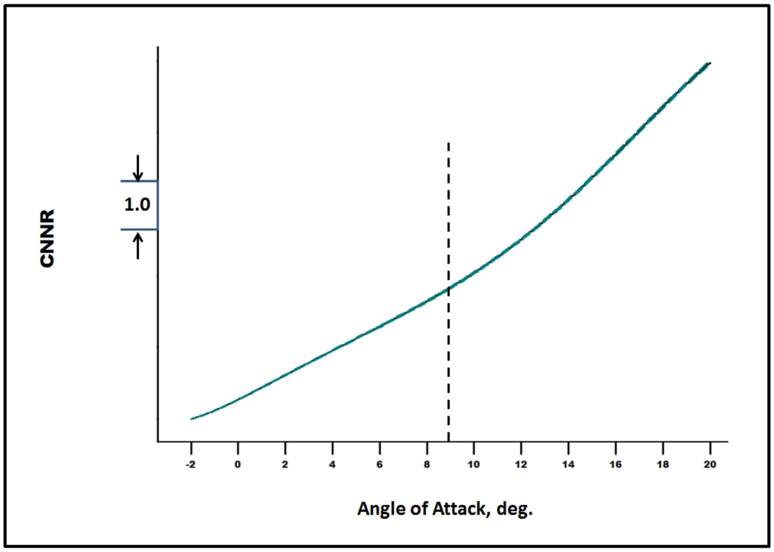


**Figure 6. Representative normal force polar, indicating inference subspace boundary decision.**

A description of the actual process by which the location of inference subspace boundaries is established is beyond the scope of this paper and would be difficult to describe generically in any case, as it depends on many factors associated with the details of the specific test. Numerous factors must be taken into account, including an examination of all response factors of interest and not just the normal force coefficient of Fig. 6. If no prior data exists upon which to base decisions about inference subspace boundaries, it may be prudent to acquire a small sample of what is called "survey data."

Survey data points are acquired rapidly with the sole intent of determining the "lay of the land" with respect to inference subspaces. If we use the colloquial terms "Kansas" and "Switzerland" to distinguish between inference subspaces characterized by relatively low-order or relatively high-order response functions, then survey data simply serves to distinguish the boundaries between Kansas and Switzerland. To press the analogy a little further, they help define boundaries that can subdivide some region of Switzerland into two or more regions that look more like Kansas.

We know by symmetry to expect a more or less sinusoidal roll angle dependence for the forces and moments, and for this experiment design we decide that we can adequately represent this with a fourth-order model. Thus our initial plan is that for each of the six forces/moments, we will fit four response models (low and high AoA, low and high Mach) for each of the nine configurations. We will evaluate these models during the execution of the test to see if the subspaces need to be further truncated to provide an adequate fit, or if additional data must be acquired.

Now that we know how many response models we intend to fit, we must decide on the order of model to fit. We have already decided that the model must be fourth order in phi, the roll angle, in order to fit the anticipated sinusoidal response. Experience suggests that responses often display a relatively complex function of angle of attack, even over relatively modest ranges. We therefore decide to fit response models of order 4 in both angle of attack and roll angle. That is, we will fit fourth-order models in two factors. Inserting $d = 4$ and $k = 2$ in Eq. (1), we calculate the number of parameters in the model, $p$, to be 15. We have adopted the tolerance criterion that lead to Eq. (15), which instructs us to multiply the number of model parameters by 1.625 to estimate the volume of data we will need for each inference subspace. We therefore estimate our data requirements to be $1.625 \times 15 = 24.4 = 25$. We will specify measurements at 25 AoA/phi combinations for each of the four combinations of low and high AoA and Mach number. This is then 100 points per configuration, or a minimum of 900 points for the test.

Having defined the scale of the experiment (the volume of data to be acquired), the last task in the design process is to decide how to distribute the 25 points in each of the subspaces. There are a number of options for doing this. One popular criterion is to distribute the points in such a way as to minimize the uncertainty in the coefficients of the response model. Such a distribution of sites within the inference subspace is called a D-optimal design. This provides the clearest indication of how each term in the model contributes to the response, and can lead to deep insights into the underlying physics of the test article. Another option is to distribute the points in what is called an IV-optimal way. Such a design seeks to minimize the integral of the prediction variance across the design space. It is a good choice when the primary interest is to make the best predictions possible, and is a good choice in many wind tunnel RSM applications. The MDOE design proposed for this study is IV-optimal.

We know at this point that we must acquire a minimum of 15 points to be able to fit a fourth-order polynomial response model in two factors, and by Eq. (15) we know that to satisfy our adequacy requirement we need 10 more points, for a total of 25. We have the option of specifying how many of these additional 10 points will be unique factor combinations, and how many will be replicates. Normally we would specify some number of replicates to quantify random error, and also to defend against leverage, which is the property of a so-called “influence point”—one that has a disproportionate impact on the shape of the response model. Influence points are harmless if the data acquired there are of high quality, but a moderate amount of experimental error at an influence point can have a significantly adverse impact on the response model. Replicates reduce the leverage of influence points, and are therefore highly desirable as a kind of “insurance” against an outlier acquired at an especially unlucky site within the design space.

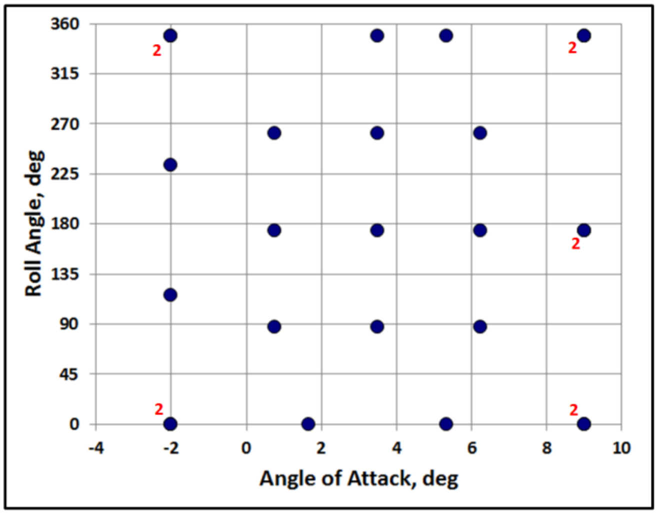


**Figure 7. AoA/Phi Design Space for Fourth-Order IV-Optimal Design Featuring Five Pure Error Degrees of Freedom. Numbers near points indicate replicates.**

The decision as to how many replicates to specify (pure error degrees of freedom) and how many additional unique data points (lack of fit degrees of freedom) there should be is a matter of taste and experience, but in this case a 5-5 split would not be a bad decision. The actual distribution of the sites in the design space is determined by specialized experiment design software, after specifying the number of factors, the order of the model, and the number of extra lack of fit and pure error degrees of freedom. Figure 7 displays the low-AoA alpha/phi design space of a IV-optimal design for a fourth-order model in two factors, with five replicated points. The numbers next to selected points indicate how many times they were replicated. Note that the replicates tend to occur at the design-space boundaries and especially in the corners, where points tend to have the greatest leverage.

## V. OFAT/MDOE Design Comparison

We have proposed an MDOE alternative to the original OFAT design for a wind tunnel test to study a surface to air missile model. In this section, we will compare the two approaches. We will compare the OFAT and MDOE test matrices themselves, as well as the information it is possible to glean from each of them and the uncertainties in response estimates that result when they are executed. We will also compare representative pitch-sweeps and roll-sweeps developed using each method.

We begin with a few remarks about a hypothetical MDOE/OFAT comparison strategy that is often suggested when a large OFAT database already exists, as is the case in the current study in which the T1878 OFAT wind tunnel test generated thousands of data points that are already in hand. This strategy appears as if it would be both cost effective and revealing, but we will comment on a few of its practical drawbacks.

### A. Commonly Proposed OFAT/MDOE Comparison Strategy That Does Not Always Work Well

Conventional OFAT practitioners with large data sets from prior tests often ask to see a demonstration of MDOE methods using the data they already have in hand. Rather than incurring the added expense of an MDOE/OFAT comparison in a new wind tunnel test, they suggest that an MDOE test matrix be developed from the existing OFAT database, and offer to estimate the response values that would have been measured in such an MDOE test (force/moments and/or pressures, say) using that database. If a prescribed MDOE data point fails to coincide with any of the acquired OFAT points, the assumption is that a serviceable response estimate can be obtained by interpolating between nearest-neighbors, using cubic B-splines or some other such interpolation method.

The idea is that response models can be developed by fitting the data from a simulated MDOE experiment (some subset of points interpolated from the original OFAT data), and then used to predict the OFAT data already in hand. The MDOE method can then be evaluated by how well these predictions compare with the measured responses that have already been recorded, using such metrics as the standard deviation of residuals, the number of measured responses that lie within prescribed RSM prediction intervals, and the ratio of RSM confidence intervals to customer-specified tolerance levels.

This comparison strategy seems eminently sensible, and we do use this method below. However, there are certain practical considerations that limit how realistic such a comparison can be. One of the most important is the difference in how much of the inference space is covered in an OFAT and an MDOE experiment design. An MDOE design is typically structured to span the entire inference space, but in an OFAT design, the fraction of the entire design space that can be covered one factor at a time is strictly limited by available resources.

For example, for even a moderate number of independent variables, each set at typical intervals over typical ranges, the number of factor/level combinations that can be supported by the usual budget of a representative tunnel entry is a fraction of the total number of factor combinations that could possibly be set. The current study featured four control surfaces with five deflections each, 31 levels of roll angle, 19 levels of angle of attack, and two Mach numbers. This represents $5 \times 5 \times 5 \times 5 \times 31 \times 19 \times 2 = 736{,}250$ possible factor combinations. With the conventional experiment design as actually executed, there was only enough time to examine 5,575 of these combinations one factor at a time, which is only 0.6% of the total.

It was therefore necessary to leave 99.4% of the possible factor combinations unexamined. This is not atypical of OFAT wind tunnel test designs generally, and does not even take into account the fact that there was literally an infinite number of intermediate settings that likewise were not set. This illustrates the classic conflict that OFAT testing imposes upon the experimentalist. On the one hand, he must profess so *little* knowledge of the test article that he can justify the high cost of an elaborate, time-consuming, labor-intensive, and expensive wind tunnel test. On the other hand, he must profess so *much* knowledge of the test article—obtained presumable through other experience—that he can adequately infer all that is necessary to know from such a small subset of the possible observations.

The inefficiency of OFAT testing does not reflect on the facility or the personnel conducting the test, but is instead attributable to the defining characteristic of conventional OFAT testing; namely, that each data point is only required to carry information about a single factor change. By contrast, the data points in an MDOE experiment each carry information about multiple factor changes, which enables much more of the design space to be navigated per unit of time. A much higher fraction of the total variance in an MDOE data sample can be explained by the intentional changes that are in the independent variables. MDOE testing therefore has a wider "knowledge bandwidth" than conventional OFAT testing. The additional information that MDOE methods pack into each data point is "unpacked" after a specified sample of data is acquired, using analytical methods that are not routinely employed in OFAT testing (regression, ANOVA, etc).

This difference in efficiency between OFAT and MDOE testing is responsible for what is often an unanticipated complication in MDOE evaluation efforts that are based on using OFAT data to simulate an MDOE experiment. Because an MDOE design can span the entire design space while the corresponding OFAT design spans a smaller fraction of it, the MDOE design invariably specifies factor combinations that were not only unexamined in the original OFAT test, but are often substantial distances from the nearest OFAT points. Thus, it often requires a significant degree of local *extrapolation* of the OFAT data to simulate a proposed MDOE data point, rather than simple interpolation over ranges small enough to generate reliable results. It is possible to incur considerable error by such long-distance extrapolation.

This situation is partially ameliorated when there are constraints among the factors that effectively limit the total number of factor/level combinations of interest. For example, certain combinations of control surface deflections may not occur in normal flight scenarios, and some may even be aerodynamically nonviable. However, an argument could be made for the utility of an experimental result that would permit system responses to be predicted even for pathological factor combinations, as it might be useful to understand other flight scenarios besides those that are intended or planned. Indeed, it can be even more important to understand unplanned scenarios than those that are commonly encountered.

Nonetheless, if we assume that no other factor combinations are of interest except those that are encountered in normal flight scenarios, the total number of factor combinations that are of interest can still be much greater than the total that can be set. For example, in the current study it was possible to reduce the $5^4 = 625$ possible control surface deflection combinations to only 25 by allowing only coordinated deflections. This reduced the number of factor combinations of interest to 29,450 from 736,250. Unfortunately, the 5,575 factor combinations that were set in the OFAT version of this test still only covered 18.9% of the possibilities, leaving 81.1% of the design space unexamined. This is still a lot of potentially useful knowledge to miss.

Another factor that complicates the comparison of OFAT and MDOE design results is the fact that OFAT testing almost never includes any quality assurance tactics of the kind that are needed to ensure statistical independence of the data. Unfortunately, ubiquitous covariate effects as outlined in Section II of this paper and described in some detail in the first phase of this study[10] virtually guarantee correlated measurement errors in a conventional OFAT wind tunnel test. This, in turn, assures bias errors in all estimates of population parameters such as the means and variances that are generated from data samples afflicted with such correlation. These bias errors contribute to the irreproducibility of wind tunnel test results generally, and can complicate the comparison of MDOE and OFAT results.

### B. Test Matrix Comparisons

Recall that the full angle of attack range from -2° to +20° was partitioned into two sub-ranges during the design of the MDOE experiment. Figure 8 compares the distribution of sites in the full alpha-phi design-space of the MDOE design with corresponding OFAT distribution. For each of the two Mach numbers, the points in Fig. 8 are acquired at each of the nine control surface configurations described in Table 1.

For this comparison we neglect the last two configurations of the original OFAT test, designated Batch 10 and Batch 11. Batch 10 consists of nothing more than a set of three replicated polars acquired during the first configuration and used only to quantify unexplained. These polars were examined in detail in the first phase of this study.[10] Batch 11 appears to be a simple target of opportunity, consisting of a single replicated polar with a different configuration of externally protruding rails.

The OFAT design represented graphically in Fig. 8a features 272 roll/pitch combinations organized as eight roll sweeps and two pitch sweeps. The corresponding MDOE design in Fig. 8b used a total of 50 points. Both the OFAT and the MDOE design called for the pitch/roll points of Fig. 8 to be executed at each of the two Mach numbers, Mach 1.6 and Mach 2.16, and for each configuration of canard deflections shown in Table 1.

Force and moment data acquired at each of the 50 pitch/roll points of Fig. 8b were fitted to fourth-order RSM models capable of predicting responses at all 272 OFAT points in Fig. 8a, plus an infinite number of intermediate point in the range of -2° to +20° in angle of attack and 0° to 360° in roll. (Reliable responses are actually only rigorously available within the range of 0° to 348.75° in roll over which data were acquired, but it requires an extrapolation of less that 3.5% of the full roll range to reach 360°, which appears from the analysis to be sustainable.)

Figure 9 is the low AoA (-2° to +9°) normal force pitch/roll response surface for the Batch 2 control surface configuration at Mach 2.16. See Table 1. At any given roll angle, it displays the customary near-first-order angle of attack dependence that most aerodynamicists would anticipate in the pre-stall lift regime. It also displays the sinusoidal dependence on roll angle that was anticipated, with normal force achieving a maximum at a roll angle of 315° and a minimum at 135°. Consulting Table 1 and Fig. 4, we see that the maximum lift occurs when the four forward control surfaces are oriented at 45° with respect to the horizontal and vertical axes of the figure (at the positions occupied in the figure by the larger tail fins), and with trailing edges all down. Likewise, the minimum lift occurs when the four forward control surfaces are at the same positions, but with trailing edges all up.

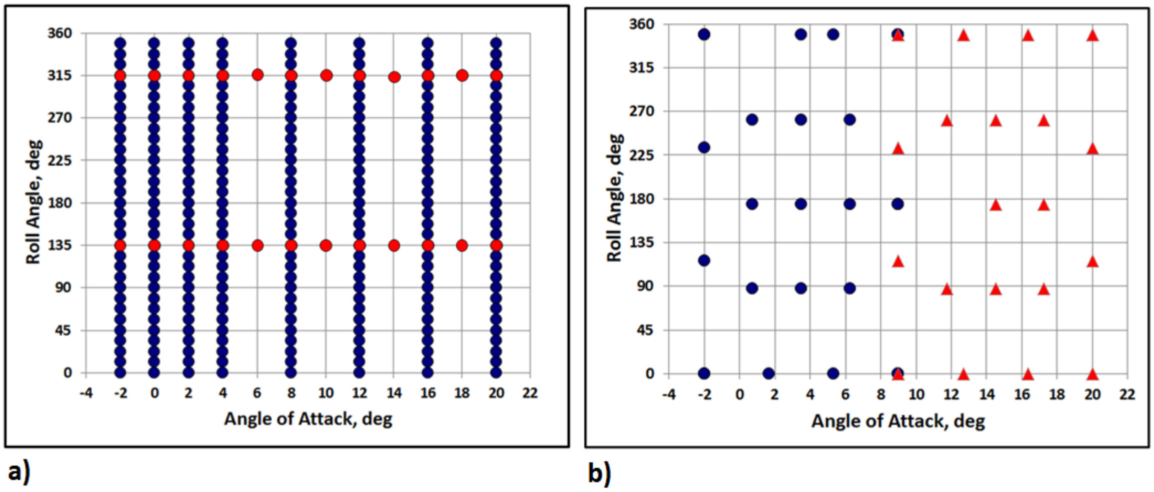


**Figure 8. Design space comparison, a) OFAT: Blue points are roll sweeps red points are pitch sweeps. b) MDOE: Low-AoA RSM design (filled circles) and high-AoA RSM design (triangles). Each design replicated at two Mach numbers for nine canard configurations.**

A close examination of Fig. 9 reveals that there is no twist in the response surface. That is, the slope in the AoA direction is the same for every roll angle, so that the left and right edges of the surface are parallel to each other and to any other slice through the surface parallel to the AoA axis. Likewise, every slice through the surface parallel to the roll axis reveals a sine wave with the same amplitude, frequency, and phase. This absence of twist implies that there is no interaction between roll and pitch. That is, a given change in angle of attack generates the same change in normal force at any roll angle, and likewise a given change in roll angle has the same effect at every angle of attack; roll and pitch are entirely independent, at least at Mach 2.16 and for the control surface configuration of Batch 1. Such independence among the independent variables of a wind tunnel test is exceedingly rare.

Recall that the tail fin assembly is mounted on bearings that enable it to rotate freely in either direction, and the orientation of the tail fins, while captured in each run on video, were not measured. The influence on normal force of the free-to-spin tail assembly is therefore unclear.

Response surface models such as the one displayed in Fig. 9 cover the entire design space rather than discrete slices through it at fixed angles of attack of fixed roll angles. It is therefore possible to use the results of the MDOE experiment design to generate pitch polars for any arbitrary roll angle and not just at the 135° and 315° roll angles for which pitch polars were physically executed in the OFAT design of Fig. 8a. Likewise, the MDOE results can generate roll sweeps for all intermediate angles of attack, and are not limited to the eight discrete AoA levels for which such sweeps were physically executed in the OFAT test.

Notwithstanding the fact that the MDOE design provides the researcher with more flexibility in studying force and moment response variables at arbitrary sites throughout the design space than the OFAT design, it required less than 20% of the data. Furthermore, genuine replicates were designed into the MDOE test matrix, permitting the same 50 data points to yield a 10 degree of freedom estimate of pure (i.e., random) error without the need to reconfigure the test article to acquire data in a dedicated quality assurance configuration at the end of the test. This further saves resources and provides a more reliable estimate of random error; one that is based on the actual data sample as acquired, and not on some alternative configuration for which the replicated data might be acquired days or weeks later.

We note that the three quality assurance polars acquired in Batch 10 near the end of the OFAT test, consisting of replicates of three nominally identical polars acquired near the start of the test, were not specified by the external customer and were never actually analyzed by test personnel. They were simply acquired because it was at that time a policy at Langley Research Center to do so. This illustrates the relatively low priority that historically has been placed on quality assessment in conventional OFAT wind tunnel testing, for which the low information density per point caused by changing only one factor at a time imposes the need for so much data that resources are seldom available to acquire replicates or to engage in any other quality assurance tactics.

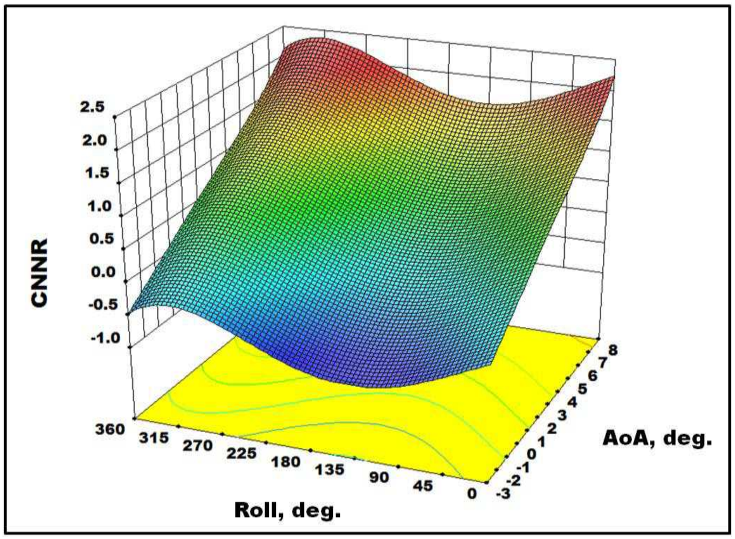


**Figure 9. Low AoA (-2$^o$ to +9$^o$) normal force coefficient pitch/roll response surface for the Batch 2 configuration at Mach 2.16.**

The savings of 272 – 50 = 222 points in Fig. 8 would have been realized at both Mach numbers at all nine canard configurations, for a total reduction in data volume of 222 × 2 × 9 = 3,996 points. If we also account for the three 13-point OFAT quality assurance polars that are no longer needed in Batch 10 because of the replicates included in the MDOE design, the total reduction in data volume exceeds 4,000 points, or roughly 80% of the OFAT database. The actual time and money that could be saved if objectives were achieved by acquiring only 20% of the data in a wind tunnel test will vary from facility to facility and from test to test within a given facility, but these resources could be used to conduct additional experiments on the same test article, if the investigator prepares in advance his “B-list” of alternative data points to be acquired if resources permit. This is one mechanism by which wind tunnel testing productivity can be enhanced through MDOE testing.

In a typical MDOE test, however, at least some of the saved resources are dedicated to quality assurance tactics intended to ensure statistical independence in the presence of covariate effects. This in turn assures that sample statistics serve as unbiased estimators of the population parameters they are intended to estimate. In non-statistical English, this means that there is less chance of getting the wrong answer because of slowly varying bias errors that depend only on local conditions, and do not reproduce with requisite precision from test to test (temperature variations, instrument drift, flow angularity changes, trip-dot ablation rates, etc).

Such quality assurance does not come without a price, however. MDOE quality assurance tactics reduce the rate at which data points can be acquired (hence the MDOE mantra, “go slow and get it right”). The reduction in data acquisition rate varies from test to test; but experience has shown that it takes something in the range of 1.5 to 2.5 times as long to acquire a given volume of data in an MDOE test as it would take to acquire the same volume of data in an OFAT test. However, because so many fewer data points are required in an MDOE test, it is still possible to save time, despite the reduced data acquisition rate. For example, even assuming a worst-case scenario in which the difference in data acquisition rate is a factor of 2.5, it would still take half as long to acquire the MDOE data as the OFAT data of the present study, since there is only 20% as much data to acquire in the MDOE test.

Equation (5) reveals that, averaged across all the points used to fit a response surface, the standard error in estimated response will be less for an MDOE response surface modeling experiment than for an OFAT experiment executed in the same environment, as long as the volume of data that is acquired in the MDOE experiment exceeds the number of parameters in the response model. In the present study, we used 25 points to fit fourth-order polynomial models in two factors. Each such model featured 15 possible parameters (including the intercept term). We describe these as “possible” parameters because not all of them were retained in the final model.

The coefficients of all 15 model terms were quantified, but so was the uncertainty in estimating each coefficient. Not all of the coefficients were large compared to the uncertainty in estimating them. Any coefficient with a magnitude that was too small to distinguish from zero with at least 90% confidence was defined as “statistically insignificant” and rejected from the model.

The standard error of model predictions increases with the number of parameters in the model per Eq. (5). This is because each term in the model contributes a degree of uncertainty to the response prediction, as each is computed from data that is necessarily afflicted with some level of experimental error. There is thus considerable incentive to eliminate as many insignificant terms from the model as possible, as they add more “noise” than “signal” to the response predictions.

It also happens that initial estimates of the order of the model can understate its complexity. The options then are to reduce the size of the inference subspace over which the response is fitted, or to increase the order of the fitted model by adding a few higher-order terms. In the present study, a quartic model in pitch and roll was sufficient to adequately fit rolling moment and the three longitudinal responses, but the yawing moment and side force models required a additional terms. Recall that in addition to five replicates, there were also five unique (so-called “lack of fit”) points added to the minimum 15 required to fit a fourth-order polynomial in two variables. Lack of fit points are included in a design for just this situation, to allow additional model degrees of freedom when it is necessary to fit a somewhat more complex model than originally anticipated. In this case it would have been possible to support a model with as many as 20 terms, had it been necessary to do so in order to achieve a good fit. The yawing moment model required 16 terms and the side-force model required 18. The remaining models required eight terms for rolling moment, nine for normal force and pitching moment, and 13 for axial force. In all cases, the total parameter count was less than the 25 points acquired in each design space, so it was possible to fit the models with residual degrees of freedom left over to reduce the standard error (“one-sigma value”) of the predictions, which by Eq. (5), ranged from 57% to 85% of the standard error for individual OFAT measurements.

A reduction in standard error relative to individual measurements is achieved through the internal replication that occurs when more data are fitted to a response model than the minimum needed to fit it. Had this experiment actually been executed as an MDOE test, however, it is likely that there would have been additional improvements in quality. As noted earlier, covariate effects induce correlations in the data that can bias the estimates of sample means and variances. MDOE quality assurance tactics break up these correlations, converting stealthy and potentially serious bias errors into additional components of random error that can be driven to acceptably low levels by replication. The response surface modeling approach therefore reduces the “fuzz band” around response estimates through internal replication, while MDOE quality assurance tactics reduce bias errors and ensure that the “fuzz” occurs about a more accurate estimate of the response.

In the present case, it is unlikely that the original OFAT data is devoid of covariate effects, and there is thus no more guarantee that such a data set can be reproduced within tight specifications than there is for any other OFAT wind tunnel test in which no precautions are taken to insure statistical independence. However, the response models developed in this study may adequately represent the *data*, however well the data actually represents the true system responses. To determine how well the models represent the data, we can use some of the OFAT data as “confirmation points,” predicting these existing response estimates with the RSM models to see how well the models do. Specifically, we will use response models developed from data acquired with the test matrix represented graphically in Fig. 7, to reproduce representative OFAT data from Fig. 8a.

### C. Comparison of MDOE and OFAT Results

It is beyond the scope of this paper to perform a detailed comparison of OFAT experimental results for this test with MDOE results simulated from the same OFAT database. In any case, it has it has already been noted that such comparisons can be questionable. However, a few comparisons may be instructive.

One of the advantages of the response surface modeling approach is that various precision intervals can be computed at every data point for which a response is estimated. We define a precision interval to be the range within which the mean of an *n*-point sample of experimental data is expected to fall with a specified probability. Two cases are of special interest, when $n = 1$ and when $n$ is infinite. We refer to the former as a prediction interval—the range of responses within which an individual measurement can be expected to fall with a prescribed probability—and to the latter as a confidence interval—a range within which the true response can be expected to lie, neglecting bias errors.

It is usually more informative to display precision intervals than response model predictions, simply because the precision intervals indicate not only the *location* of the responses (within the intervals), but also the *dispersion* in response predictions, indicated by the widths of the intervals. To compare a response surface such as the normal force coefficient surface of Fig. 9 with a conventional normal force pitch sweep acquired at a given roll angle, we simply examine a slice of the response surface parallel to the angle of attack axis, at the prescribed roll angle.

For example, the red dashed lines in Fig. 10 represent the upper and lower limits of a 95% prediction interval for the normal force coefficient, displayed as a function of angle of attack at a fixed roll angle of 90°. Our claim is that at any angle of attack, there is a 95% probability that a normal force coefficient measured at that angle of attack will lie above the lower limit and below the upper limit. The individual points in this figure are physical measurements made during the OFAT test.

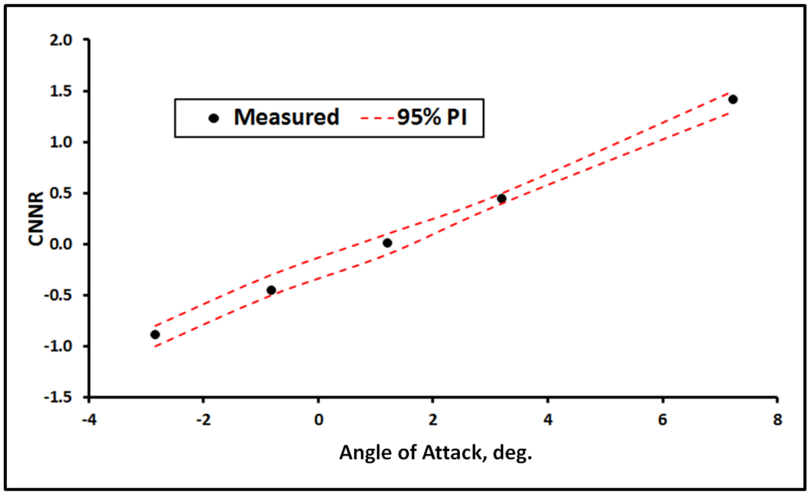


**Figure 10. Normal force coefficient polar generated from a Batch 2 MDOE response model at a roll angle of 90$^0$ and Mach 2.16.**

There are two important points about Fig. 10. The first is that the response model, and thus the upper and lower limits of the corresponding 95% prediction interval, was estimated from normal force values associated with the 25 points in Fig. 7. They were not the result of some linear regression on the specific points in the figure, unless one of them by coincidence happened to coincide with a point in Fig. 7. The second point, of course, is that the MDOE response model appears to represent the data well.

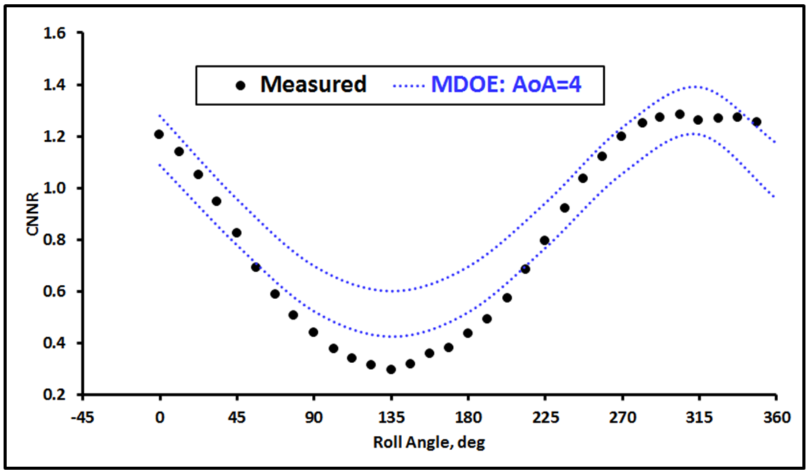


**Figure 11. Normal force coefficient roll sweep generated from MDOE response model at angle of attack of 4$^0$ and Mach 2.16 with control surfaces deflection set in Batch 2.**

One of the advantages of a response surface model is that it is as easy to display responses as a function of one variable as another. By "slicing" through the response surface parallel to the roll axis instead of the angle of attack axis, it is possible to generate the kind of result that would be acquired in a roll sweep executed at a fixed angle of attack in a conventional OFAT test. For example, the blue dotted lines of Fig. 11 define 95% prediction interval limits for the normal force coefficient modeled as a function of roll angle at an angle of attack of 4°, obtained from the response surface in Fig. 9. The Mach number is 2.16 and the control surfaces are set as described in Table 1 for Batch 2 for these model predictions.

The figure displays the kind of sinusoidal roll dependence that was anticipated, but this result is still troubling as it definitely displays a systematic departure of the response model from the roll sweep data reported in the OFAT test that it is supposed to represent. Specifically, for a range of roll angles centered on 135°, the measured normal force values are systematically lower than the levels that the response model predicts, by amounts that cannot be credibly attributed to ordinary random error. Likewise, while the agreement between measurement and prediction in the vicinity of 315° is better than it is around 135°, it does not seem to be as good as for other roll angles.

The apparent discrepancy in Fig. 11 can be explained by a comparison in Fig. 12 of the OFAT pitch/roll design space as planned, and as executed. Notwithstanding the OFAT imperative to change only one variable at a time, it is apparent that during roll sweeps when the intent was to hold the angle of attack constant, changes were actually made jointly to the angle of attack and the roll angle. This is believed to be attributable to sting knuckle alignment errors at preferred roll angles, which have since been corrected. However, the OFAT roll sweep data acquired in this test, including the data presented in Fig. 11, reflects the combined effects of changes in pitch and roll that are displayed on the right side of Fig. 12.

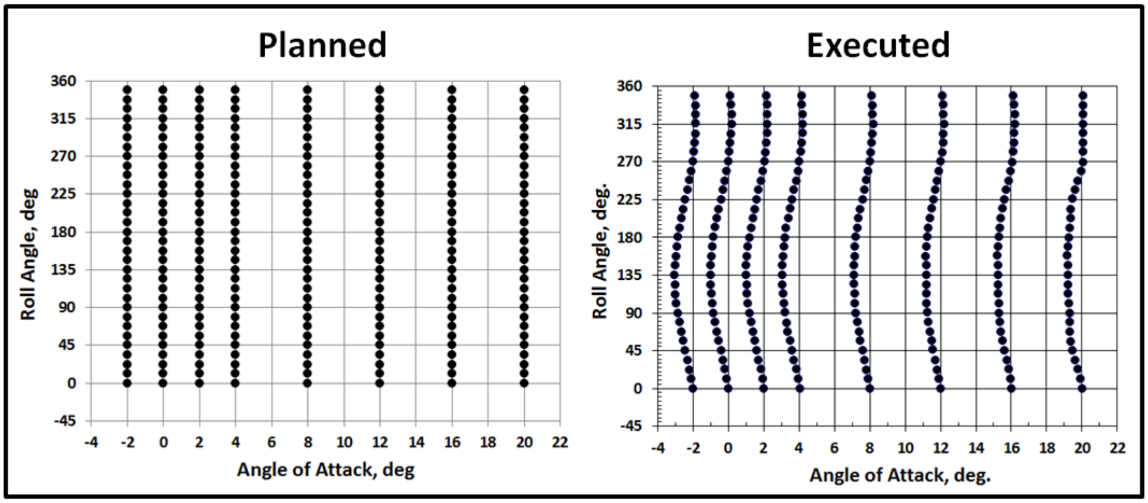


**Figure 12. OFAT design space as planned (left) and as executed (right).**

The OFAT roll sweep data displayed in Fig. 11 was thought to be acquired at a constant angle of attack of 4° but as Fig. 12 reveals, the angle of attack actually dropped systematically as the roll angle approached 135°, increasing with roll angle beyond that to return to levels relatively near nominal AoA design-points by the end of the roll sweep. This behavior is observed for every nominal angle of attack design point.

The systematic decrease in angle of attack centered on 135° of roll accounts for the discrepancy in the measured OFAT data relative to the MDOE response model as revealed in Fig. 11. The inadvertent coupling of pitch and roll angle changes that influenced the OFAT roll sweep so dramatically had no appreciable effect on MDOE response model predictions because, as Fig. 8b illustrates, the design of an MDOE test matrix intended to fit response surface models does not rely upon the systematic variation of one factor while others are held constant. Had the systematic AoA error of Fig. 12 occurred during the execution of an MDOE test, the effect would have been to displace some of the points in Fig. 8b slightly to the left and parallel to the AoA axis. This would have the effect of causing a minor increase in the integrated prediction variance across the design space, since the site selections displayed in Fig. 8b were chosen to minimize this. This may have caused a slight inflation in the 95% prediction interval widths, corresponding to a somewhat degraded estimate of uncertainty. While the dispersion estimates may have been slightly affected in this way, the location estimates that determine the center of the prediction interval would not have been seriously affected.

Figure 13 compares the same OFAT data of Fig. 11 with MDOE predictions revised to include the various AoA levels that were actually set. The red dashed lines represent the 95% prediction interval within which we expect to find measured normal force coefficients with a probability of 95%. This does not correspond to a conventional roll sweep at AoA = 4°, but is instead a sequential series of measurements in which AoA varied systematically with roll angle as in Fig. 12. It represents the normal force response surface of Fig. 9 that lies above the wavy series of systematically varying pitch/roll points located nominally at AoA = 4° on the right side of Fig. 12. Neglecting any covariate effects that may have been in play during the OFAT test, the 95% prediction interval of Fig. 11 represents an unbiased estimate of the true Mach 2.16 roll dependence at 4° angle of attack, when the control surfaces are set as described by the Table 1 entry for Batch 2.

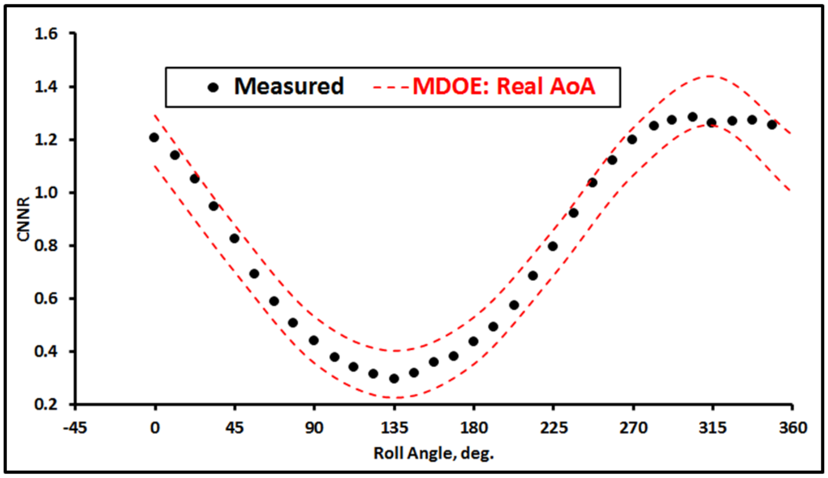


**Figure 13. Normal force coefficient roll sweep generated from MDOE response model at Mach 2.16 with control surfaces deflection set in Batch 2. Variable AoA as executed per nominal 4° set-point in Fig. 12.**

Figures 14–18 compare measured responses with model predictions for other responses besides normal force. The comparisons take the same form as those in Figs. 10, 11, and 13, in that the predictions represent pitch sweeps at a constant roll angle of 90°, and roll sweeps both at the intended AoA set-point of 4° and at various AoA levels that were actually set, as displayed for the nominal value of 4° on the right of Fig. 12. Because these responses were not as sensitive to AoA as the normal force coefficient, the difference between results obtained at AoA = 4° and at the actual AoA set-points was less. We emphasize that the prediction intervals in Figs. 14–18 result from response surface models fitted to the data acquired at MDOE design-point pitch/roll combinations as displayed in Fig. 7, and not from simply fitting the validation measurements in these figures directly.

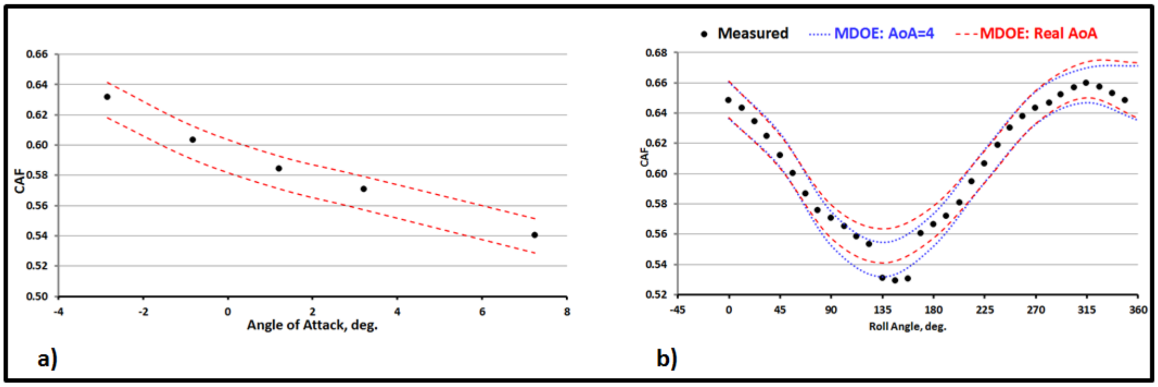


**Figure 14. Comparison of measured and predicted Axial Force Coefficient for Mach 2.16 when control surfaces are configured as in Batch 2. Lines are MDOE 95% prediction intervals: blue for AoA = 4°, red for variable AoA as measured. a) is pitch sweep, b) is roll sweep.**

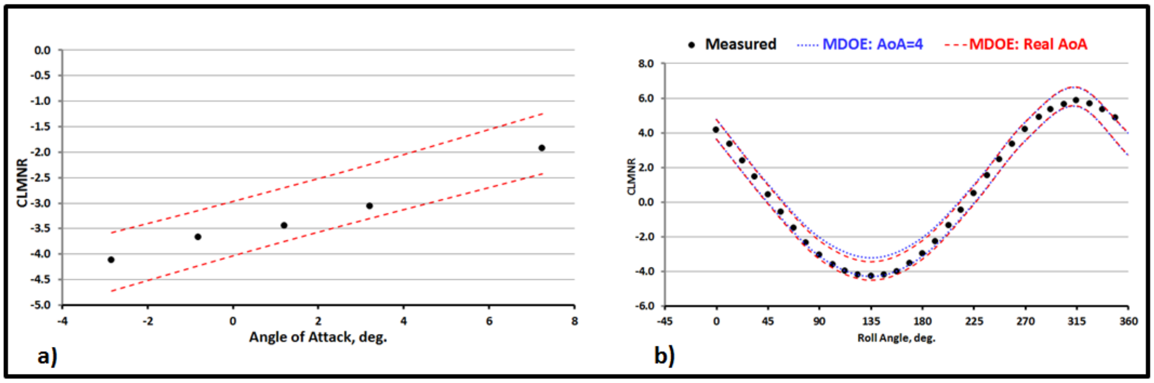


**Figure 15. Comparison of measured and predicted Pitching Moment Coefficient for Mach 2.16 when control surfaces are configured as in Batch 2. Lines are MDOE 95% prediction intervals: blue for AoA = 4$^o$, red for variable AoA as measured. a) is pitch sweep, b) is roll sweep.**

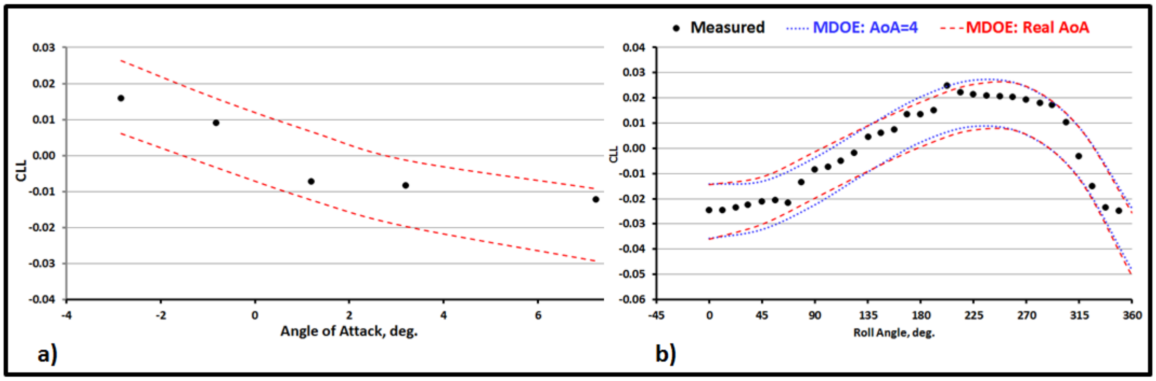


**Figure 16. Comparison of measured and predicted Rolling Moment Coefficient for Mach 2.16 when control surfaces are configured as in Batch 2. Lines are MDOE 95% prediction intervals: blue for AoA = 4$^o$, red for variable AoA as measured. a) is pitch sweep, b) is roll sweep.**

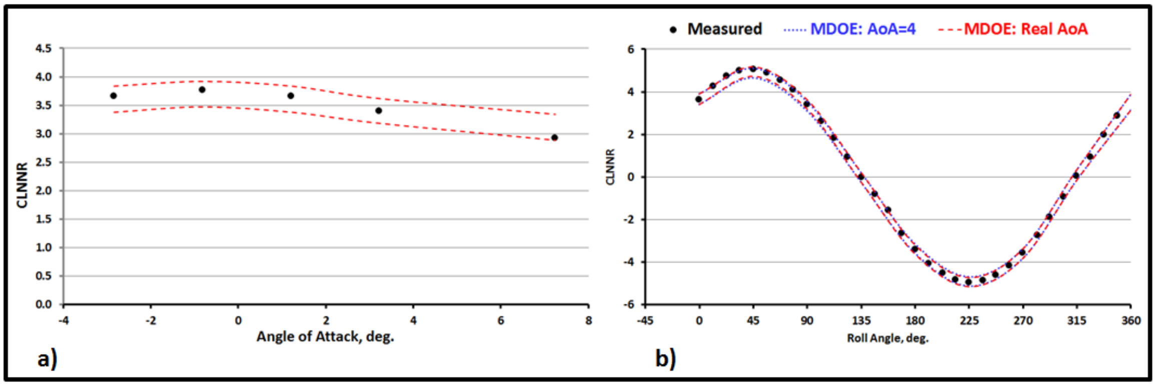


**Figure 17. Comparison of measured and predicted Yawing Moment Coefficient for Mach 2.16 when control surfaces are configured as in Batch 2. Lines are MDOE 95% prediction intervals: blue for AoA = 4$^o$, red for variable AoA as measured. a) is pitch sweep, b) is roll sweep.**

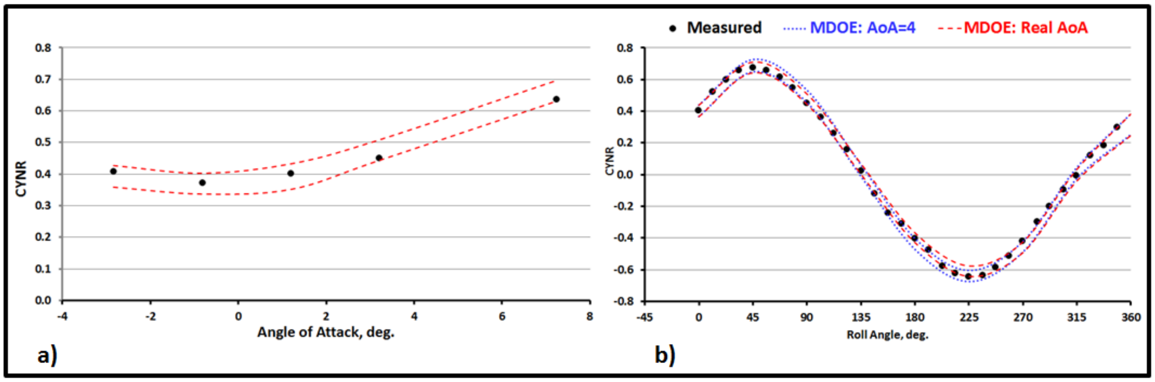


**Figure 18. Comparison of measured and predicted Side Force Coefficient for Mach 2.16 when control surfaces are configured as in Batch 2. Lines are MDOE 95% prediction intervals: blue for AoA = 4$^{o}$, red for variable AoA as measured. a) is pitch sweep, b) is roll sweep.**

## VI. Discussion

This section features further remarks on a few topics treated earlier in the paper. An alternative MDOE design is also introduced, which is capable of providing much more information than the existing OFAT design, with a modest increase in required resources (three additional shifts).

### A. Quality of the Fit

The comparisons between measured and predicted response estimates in Figs. 10, 13, and 14–18 suggest that the models captured the general independent variable dependence of the forces and moments, but a number of factors conspired in this test to degrade the quality of the fits. The first that should be mentioned is Lack of Fit error, which is always a factor in any response surface modeling experiment. This error is attributable to the fact that the low-order polynomial models used to represent the data are at best approximations. Our only claim is that we can make such approximations "good enough," that is, within some prescribed tolerance, by fitting the models over suitably modest ranges of the independent variables and/or fitting the data to an adequately high order polynomial. In practice, there will always be some residual contribution to the unexplained variance due to model imperfections. These consist of all the higher-order terms in an infinite Taylor series that would define the unknown response function perfectly, which we drop when we fit the data to a low-order polynomial approximation of such an infinite series. It is probably not an over-simplification to say that in the kind of high-precision measurement environment that characterizes most major wind tunnel facilities, the essence of model adequacy assurance is to drive lack of fit errors to suitably low levels by ensuring that the complexity of the response model matches that of the unknown underlying response function.

This study was not designed or executed using MDOE principles, so there was never a prescribed tolerance and therefore no guidance as to how small the lack of fit error would have to be in order to be "good enough." Figures 10, 13, and 14–18 represent what in an actual MDOE experiment would have been a first iteration, used to determine if more data or a more complex model were necessary to achieve precision and accuracy goals by reducing net random and lack of fit error, respectively.

The problem of managing lack of fit error is common to all response surface modeling experiments but there were additional factors unique to this study that probably also impacted the quality of the models. Two of these factors, standard covariates and the freely rotating tail assembly, are related. They are related because both phenomena are capable of potentially influencing response estimates, yet neither was modeled in this experiment.

The uncertainty reflected in the widths of the prediction intervals in Figs. 10, 13, and 14–18 is likely due in part to the fact that covariate effects as described in Section II were in play, which are not addressed in the execution of typical OFAT tests and were not addressed in this one. For example, the pitch polars in these figures were constructed by plotting responses acquired at a nominal roll angle of 90$^{o}$ from a sequence of roll sweeps acquired at different angles of attack as displayed in Fig. 12. These sweeps were acquired over an interval of more than a half an hour. Assume for a moment that during this time, slowly changing covariate effects were in play that could impact forces and moments (instrument drift, say, or effects induced by systematic changes in temperature, flow angularity, or other factors). Then, any model constructed as a function only of pitch and roll, and not of these covariate effects, would display some apparent lack of fit error.

Covariate effects are everywhere; systematic, prolonged variations generally dominate the random fluctuations that often comprise the only component of unexplained variance addressed in a typical OFAT quality assessment. Systematic unexplained variance, the tell-tale sign of covariate effects, was detected in the first phase of this test[10] in polar replicates acquired over intervals as short as a few minutes. It is therefore highly likely that the OFAT data used to generate the models displayed in Figs. 10, 13, and 14–18, acquired over a half hour interval, are confounded by covariate effects to some degree. This would contribute to lack of fit errors in any response model fitted from such data, which would widen the prediction intervals.

The ubiquitous nature of covariate effects, coupled with the fact that OFAT methods have no provisions for accounting for them, is one of the reasons, if not the principal reason, that OFAT testing methods are so universally criticized in textbooks on professional experiment design. Covariate effects impact the data but are essentially local phenomena that do not reproduce exactly from test to test. This no doubt contributes to the general difficulty of generating reliably reproducible high-precision wind tunnel results using OFAT testing methods.

The spinning tail represents a kind of covariate effect that is unique to this experiment. The direction and rotation rate of the tail were not controlled in the original OFAT test, but allowed to vary as a function of angle of attack, Mach number, roll angle, and canard deflection. Under some conditions, the tail assembly did not rotate at all; and in other cases, it might rotate up to a hundred times or more during the two-second interval over which individual frames of data, acquired at 30 measurements per second, were averaged. It is entirely possible that the tail rotation dynamics induced vibration in the test article that contributed to the unexplained variance in force and moment measurements, much as an automobile's wheel might cause excessive vibration when it is out of balance. Indeed, the precision of the axial force models relative to irreducible random error seems to be higher than the relative precision of other models, consistent with the fact that axial force is probably less susceptible to tail spin dynamics than the other forces and moments. In any case, response surface models such as the one represented in Fig. 9 quantify response changes as a function only of the independent variables that were modeled. If the response is actually a function of spin rate, say, as well as of pitch and roll, while the response model is constructed only as a function pitch and roll, the spin dynamics effect will not be accounted for and will therefore contribute to the lack of fit error. It is possible that a better experimental result could have been achieved by controlling various factors associated with the tail assembly, so that they might be better understood and less likely to contribute to the unexplained variance.

Another factor that adversely impacted the precision with which the OFAT data of this test could be represented with response surface models is site selection. A site is a point in the design space defined by some unique combination of independent variable levels. Fig. 8 displays the difference in site selection between an MDOE and an OFAT test design. Site selection decisions are generally taken in an OFAT experiment design for the purpose of maximizing data acquisition rate and therefore total data volume, or to maximize operator convenience. In an MDOE test, the sites are selected entirely on the basis of quality assurance. The prediction variance of a response surface model is directly affected by the distribution of sites at which the fitted data sample is acquired. The MDOE site distribution displayed in Fig. 8 reduces the prediction variance integrated over the entire design space, for example (a so-called "IV-optimal" design). Other MDOE site distributions might be selected to achieve other quality-related goals. For example, a so-called "D-optimal" site distribution would have the effect of generating response models for which the uncertainty in individual regression coefficients is minimized.

Unfortunately, the MDOE sites of Fig. 8, which would have been prescribed in an IV-optimal MDOE design, do not coincide with the OFAT sites where data were actually acquired. In order to simulate an MDOE experiment based on the available OFAT data, it was therefore necessary to approximate the IV-optimal MDOE site distribution by using a "candidate list."

A candidate list is necessary when arbitrary factor combinations cannot be set, and the factor combinations must be drawn from some finite list of available points. This situation often arises in configuration aerodynamics, for example, when there is only a finite set of discrete part sizes and shapes to examine, and in cases in which control surface deflections are represented with parts machined at discrete angles, with no intermediate deflections possible. In the present case, only the OFAT data points were available.

We did not use interpolation to approximate responses at the IV-optimal design-space sites, although that was an option. Such an approach would have been very labor intensive, and any advantages there might have been of this approach over the candidate list approach were unclear. That is, it is not clear if the quality improvement afforded by an optimized site distribution is enough to compensate for the errors inevitably introduced by using interpolated results rather than directly measured response values.

The practical effect of using a candidate list of OFAT measurements is that the site distribution is not optimal. We generated an MDOE design from points on the left of Fig. 8 that were as close as possible to the points on the right, but few points if any coincided exactly, and some of the candidate list sites were relatively distant from the nearest proposed MDOE site. This means that the prediction variance integrated over the design space was not minimized, so that model predictions made at individual sites in the design space have more uncertainty than they otherwise would. That is, the prediction intervals in Figs. 10, 13, and 14–18 are inflated by an unknown amount over what they would have been if the MDOE design had actually been executed.

Finally, a general weakness of response surface modeling with low-order polynomials must be mentioned, which is that absent a large number of higher-order terms for which data volume requirements could be prohibitive, it is not easy to model near-discontinuities in the response model with high fidelity. We see what might be such a situation in the roll sweep of Fig. 14, where the three points near 135° appear to form a step discontinuity in the axial force polar. It is possible that these points are outliers, resulting from some anomaly in play for the short interval of time in which they were acquired. However, it is also possible that this represents a real effect, in which case the smooth, low-order axial force response model is not adequate to represent the data over this interval exactly. In an attempt to accommodate this near-discontinuity, the response modeling software that was used to fit the data actually specified a rare fifth-order term, and higher-order terms would probably have been specified had there been a sufficient number of degrees of freedom to accommodate them and to maintain hierarchy, a property of well-formed polynomials that specifies lower-order sub-components of all higher-order terms (A and B if AB, etc).

The three points that are not well-fitted by the low-order model might comprise a “drag-bucket,” which is a phenomenon that can occur under conditions for which laminar flow is maintained over an extended chord-wise fraction of an airfoil, minimizing drag forces. Perhaps such laminar flow conditions formed when the missile rolled 135 degrees at a 4° angle of attack.

This situation highlights the need to validate response surface models with confirmation points, examining residuals not only at the design-space sites where data used to fit the models were acquired, but at other sites as well, as in this case. It should also be noted that an experienced aerodynamicist is likely to have been able to bring sufficient subject matter expertise to bear to have anticipated some non-linearity near drag-minimal conditions. In such a case, having initially determined from the response models that axial force was a minimum near 135° of roll, a further schedule of roll settings at finer intervals might have been specified in this region, if details of the drag bucket were indeed of interest. It is of course always possible that for this application, simply knowing the roll orientation that minimizes axial force may have been sufficient, without exploring the fine-structure of the axial force polar. As always, an unambiguous understanding of specific objectives helps inform the test procedures.

### B. Inference Subspace Truncation

One reason we prefer to approximate a complex underlying response function with low-order polynomial models over multiple, relatively limited sub-ranges of the independent variables is that high-order response models that are functions of many variables can feature scores if not hundreds of terms per Eq. (1). The effects of small modeling errors are magnified with such unwieldy and complex models, which can result in non-physical behavior especially at the design space boundaries that are relatively large distances from the “center of mass” of the data sample used to fit the model.

Furthermore, it is generally undesirable to use a single, high-order polynomial to try to represent situations in which the underlying physics changes from one subspace range to another, as is often the case in wind tunnel testing. One would not expect the same polynomial to provide an equally adequate approximation of the unknown true underlying response function over pre-stall and post-stall angles of attack, for example. Even over pre-stall angles of attack, it is often prudent to specify a two or three AoA ranges (possibly overlapping), over each of which a different polynomial might be fitted. When we partition the full AoA range into subspaces in this way for modeling purposes, Eq. (1) is used to estimate the volume of data required in each subspace.

Discontinuities at response surface subspace boundaries can at first glance seem to be a concern when these methods are applied, but such concern is not really justified. It is true that on the subspace boundaries, neither the surfaces nor their slopes will generally match, which are the two conditions for a continuous transition across boundaries. However, this simply reflects the experimental error in the underlying data used to generate the models, (plus any lack of fit error in the models themselves). Just as it would be unreasonable to expect two physical measurements made on an inference subspace boundary to replicate perfectly, so it is unlikely that two different model predictions, based on models developed from two different, finite samples of imperfect experimental data, would be identical in the same place. Our claim is not that the models in adjacent subspaces represent the response over those subspaces with infinite precision, but that the true response lies somewhere between the upper and lower limits of prediction intervals we can quantify for each subspace, reflecting prediction modeling uncertainty.

### C. A Comment on Explained Variance

MDOE methods are more productive than OFAT methods because in general, an MDOE sample of a given size will feature more variance than an OFAT sample of the same size. While both samples will have the same *unexplained* variance if they were both acquired under the same conditions, the MDOE sample will have more *explained* variance, since more than one factor level will have been changed per data point. A data sample with more explained variance possesses a higher density of information than one with the same volume of data but with less explained variance. This higher information density per point translates into a need for fewer points, which is a key factor in the cycle time and cost savings afforded by professional experiment design.

### D. Evaluation of Multiple Residuals

The quantity x* has been used in the development of a mathematical relationship between data volume and various quantities upon which it depends. See Eq. (7). It is useful for distinguishing between individual points that may be drawn from the distribution corresponding to the null hypothesis or from the distribution corresponding to the alternative hypothesis. To make such an inference, we would simply compare the observed residual to x* as computed by Eq. (2). A smaller residual would validate the model and a larger one would call it into question.

In practice, we do not evaluate the adequacy of a response model based on an inference from a single residual, or even on individual inferences from a number of residuals. The reason is that such a process would be characterized by cumulative errors that could lead to an erroneous conclusion about the adequacy of the model. Consider, for example, an individual inference for which the probability of an inference error is 0.05, so that there is a 95% probability of a proper inference. The volume of data that would actually be specified for a given response model would depend on a number of factors as illustrated above, but using Eq. (15) coupled with Eq. (1) as a guide, and allowing for multiple inference subspaces, even in the least ambitious of wind tunnel tests the number of model residuals would seldom be less than a hundred. If 100 residuals are evaluated and each evaluation has a 95% probability of being correct, the probability that every inference will be correct is only 0.006. There is thus an overwhelming probability that some number of residuals will be improperly assessed.

An alternative to individual inferences is to make a single inference based on the entire ensemble of residuals. The evaluation of residuals is treated as a Bernoulli process in which each residual is considered a trial that may be successful or may fail. A successful trial is defined as a residual that falls within some interval about zero for which there is a known probability of success for such a trial. The 95% prediction interval is convenient for such an interval because it is readily calculated and by definition, the success probability of individual trials is known for that interval assuming a model that fits the data well. For a given number of trials (residuals), there is some threshold number of successes that can be anticipated from the individual success probabilities. This threshold is known as the Critical Binomial Number, and it can be found in standard statistical tables or estimated with readily available software. For example, there is a 99% probability that out of 100 residuals, 89 or more will fall within the 95% prediction interval for a model with no significant lack of fit error.

### E. Split-Plot Alternative

Over 5000 points were acquired in the OFAT test as executed and the equivalent MDOE design required less than 20% of this. Even allowing for a factor of 2.5 reduction in data acquisition rates to accommodate MDOE quality assurance tactics, this translates into roughly a factor of two in the savings of all resources that are directly proportional to wind-on minutes. (Not all costs have this relationship, but clearly the fewer the number of data points that are acquired, the lower the costs of the test and the less time it will take.)

An MDOE experiment design is usually predicated on the assumption that response estimates are desired throughout the design space. It is not usually necessary to make the same kind of potentially painful decisions that are required in OFAT testing as to which regions of the test matrix to examine and which to leave for another time because of resource constraints. For example, the OFAT test was executed as a separate experiment for each of nine of the 25 possible combinations of two canard planes (pitch and yaw) set at five levels each. An MDOE experiment could have been designed to quantify forces and moments for the 16 canard plane combinations that were not examined, as well as the nine that were. Furthermore, such an experiment would have permitted forces and moments to be estimated at all combinations of intermediate pitch-plane and yaw-plane settings within the full range of ±16° that was tested, and for all intermediate pitch and roll angles within the full pitch/roll design subspace. In addition, forces and moments could also be estimated for all intermediate Mach numbers as well, instead of at only two discrete Mach numbers as in the present test.

There is a complication associated with an experiment design that would in this way cover the full pitch/roll/Mach/yaw-plane/pitch-plane design space. MDOE productivity tactics change multiple independent variables per point, and MDOE quality assurance tactics randomize the order that these points are taken in order to eliminate correlated errors due to systematic covariate effects, thereby assuring statistical independence of points acquired at sequential factor levels. This in turn assures experimental results that are not biased by locally changing covariate effects and are therefore capable of being reproduced in future experiments in which the same precautions are taken to defend against such covariate effects.

One practical consideration for MDOE quality assurance is that not all factors are as easy to change as others. Configuration variables such as canard deflection angle are designated “hard to change” variables, for example, and it would not be practical to change such variables on every point as a fully randomized design would prescribe. We must acquire all the necessary pitch, roll, and Mach number data for each canard configuration before changing to another configuration. This situation is called a “restriction on randomization,” which necessitates a special class of experiment designs known as “split plot” designs.

A split-plot design would organize the experiment in terms of what are called “whole plots” and “subplots.” The whole plots are comprised of the hard to change variables which, for this test, are the canard pitch plane and yaw plane deflection angles. Assuming that we can adequately describe canard effects with a fourth-order model in these two factors, by Eq. (1) we would require a minimum of 15 canard configurations, but by Eq. (15) we would increase this to 25 configurations, or “whole plots.” For each whole plot (unique canard configuration) we would design a subplot experiment suitable for fitting fourth-order polynomials in three factors for both a low-AoA and a high-AoA inference subspace. The three subplot factors are pitch, roll, and Mach. By Eq. (1), a minimum of 35 pitch/roll/Mach points would be needed to fit fourth-order models in each of the two AoA subspaces, for a total of 70 points minimum per configuration, but by Eq. (15) we would increase this to 114 sub-plot points per whole plot. There would thus be a total of 25 × 114 = 2850 points in this design, or roughly half of the 5575 points acquired in the OFAT version.

Pitch/roll data points were acquired at a rate of about three points per minute within each configuration of the OFAT test. Assuming this rate is reduced by a factor of roughly 2.5 when the points are randomized, it would take about an hour and a half to acquire all the subplot points in each whole plot. It also took an average of about an hour and a half to change configurations in the OFAT test, so approximately three hours would be required for each whole plot in an MDOE split plot design. It would therefore require an estimated 3 × 25 = 75 hours to complete this experiment, or roughly 10 shifts, compared to the seven shifts required by the OFAT test. However, for three additional shifts we would be able to achieve the following improvements over the OFAT design:

- Response estimates for all 25 canard plane configurations and any intermediate deflection within the ±16° range specified for this test, rather than at only the nine discrete configurations of the OFAT test
- Mach number dependence of all forces and moments rather than response values at only two discrete Mach numbers
- Forces and moments at all intermediate combinations of roll angle and angle of attack
- Randomization of set-point order respecting hard to change variables, which ensures statistical independence of data points and experimental results that are unbiased by time-varying covariate effects
- Uncertainty assessment for any factor combination of interest
- Insights into interactions among configuration and model attitude variables
- Results that are relatively insensitive to the level of inadvertent correlation among independent variable set points experienced in the OFAT test per Fig. 12

As is often said in MDOE/OFAT design comparisons, it is not a question of which design is “right” and which is “wrong,” but rather a question of costs and benefits for each approach. An MDOE approach would have been able to reproduce the OFAT results with roughly half the data volume, achieving about a 20% reduction in experimental error. An MDOE split plot design would have facilitated more information, as noted in the above list, but at an estimated additional cost of three more shifts. Test personnel would have to make the cost/benefit decisions associated with this choice.

## VII. Summary

The design of a conventional One Factor At a Time (OFAT) wind tunnel test has been revisited to quantify quality and productivity improvements that might be realized by applying the Modern Design of Experiments (MDOE). Formulae derived for scaling an MDOE experiment to satisfy prescribed accuracy and inference error probability criteria were applied to make data volume estimates, and the estimated volume of data was distributed within the design space so as to enable response surface models to be generated that minimize the integrated prediction error throughout the space. An alternative MDOE design was described with numerous advantages over the original OFAT design, but which cost an estimated three additional shifts to execute.

The following specific results are reported:

- The MDOE experiment design required less than 20% as many data points as the original MDOE design.
- Reduced data acquisition rates resulting from MDOE quality assurance tactics would have translated the 80% reduction in data volume into about a 50% reduction in wind-on minutes.
- The standard error in MDOE force and moment values would have less than for single-point OFAT response measurements, because MDOE error estimates exploit internal replication of multiple error degrees of freedom that are available whenever a response model is fitted to more points than there are terms in the model.
- The minimum number of data points required to fit a polynomial response model in a given region of the design space is a known function of the order of the model and the number of independent variables upon which it depends.
- To satisfy prescribed tolerances quantifying quality requirements for a response surface model, it is generally necessary to fit the model to more points than the absolute minimum needed to produce it.
- The additional volume of data needed to fit an adequate model is a known function of the intrinsic variability of the measurement environment, the severity of the adequacy requirement, and the researcher's tolerance for inference error risk in assessing the adequacy of the model from the residuals it generates.
- The value associated with each new data point can be assessed in terms of the amount of inference error risk it retires.
- Points acquired early in an experiment, when few data points are available, retire more inference error risk than points acquired later on, after some critical mass of data are in hand.
- The value of the next data point is a monotonically decreasing function of the volume of data already in hand. This implies that the value of acquiring the next data point must eventually drop below the cost of acquiring it. This in turn implies a point of diminishing returns beyond which the acquisition of additional data is not cost effective, which argues against traditional wind tunnel strategies that focus on maximizing data volume.
- MDOE methods can be applied to a wind tunnel test featuring a missile as the test article, reducing the data volume requirements and all cost factors related to data volume, while also reducing experimental uncertainty relative to a corresponding OFAT design.
- An MDOE split plot design would have been able to generate substantially more information than the OFAT wind tunnel test examined in this study, for the cost of three additional shifts of testing.

## Acknowledgments

This work was supported by the Aeronautics Test Program of the National Aeronautics and Space Administration. The authors gratefully acknowledge the helpful comments of Mr. Floyd Wilcox of Langley Research Center's Unitary Plan Wind Tunnel, who also provided the model hardware description reproduced herein.

## 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, NV, Jan. 1998.

[2] DeLoach, R., "Tailoring Wind Tunnel Data Volume Requirements Through the Formal Design Of Experiments," AIAA 98-2884, 20th Advanced Measurement and Ground Testing Conference, Albuquerque, NM, Jun. 1998.

[3] DeLoach, R., "Improved Quality in Aerospace Testing Through the Modern Design of Experiments (invited)," AIAA 2000-0825, 38th AIAA Aerospace Sciences Meeting and Exhibit, Reno, NV, Jan. 2000.

[4] Morelli, E. A., and DeLoach, R., "Ground Testing Results Using Modern Experiment Design and Multivariate Orthogonal Functions (Invited)," AIAA 2003-0653, 41st AIAA Aerospace Sciences Meeting & Exhibit, Reno, NV, Jan. 6–9, 2003.

[5] DeLoach, R., "Tactical Defenses Against Systematic Variation in Wind Tunnel Testing," AIAA 2002-0885, 40th AIAA Aerospace Sciences Meeting and Exhibit, Reno, NV, Jan. 14–17, 2002.

[6] Dowgwillo, R. M., and DeLoach, R., "Using Modern Design of Experiments to Create a Surface Pressure Database From a Low Speed Wind Tunnel Test," AIAA 2004-2200, 24th AIAA Aerodynamic Measurement Technology and Ground Testing Conference, Portland, OR, Jun. 28–30, 2004.

[7] 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, NV, Jan. 69, 2003.

[8] DeLoach, R., "Analysis of Variance in the Modern Design of Experiments," AIAA 2010-1111, 48th AIAA Aerospace Sciences Meeting and Exhibit, Orlando, FL, Jan. 4–7, 2010.

[9] Rhode, M.N., and DeLoach, R., "Hypersonic Wind Tunnel Calibration Using the Modern Design of Experiments," AIAA 2005-4274, 41st AIAA/ASME/SAE/ASEE Joint Propulsion Conference and Exhibit, Tucson, AZ, Jul. 10–13 2005.

[10] DeLoach, R. and Micol, J.R., "Analysis of Wind Tunnel Polar Replicates Using the Modern Design of Experiments (Invited)." AIAA 2010-4927. 27th AIAA Aerodynamic Measurement Technology and Ground Testing Conference. Chicago, IL, June 28–July 1, 2010.

[11] Oberkampf, W. L. and Aeschliman, D. P., "Joint Computational/Experimental Aerodynamics Research on a Hypersonic Vehicle, Part 1: Experimental Results," AIAA Journal 1992, 30(8): 2000–2009.

[12] Oberkampf, W. L., Aeschliman, D. P., Henfling, J.F., and Larson, D.E., "Surface Pressure Measurements for CFD Code Validation in Hypersonic Flow," AIAA 95-2273. 26th AIAA Fluid Dynamics Conference. San Diego, CA, June 19–22, 1995.

[13] Aeschliman, D. P. and Oberkampf, W. L., "Experimental Methodology for Computational Fluid Dynamics Code Validation", AIAA Journal 1998: 36(5): 733–741.

[14] Oberkampf, W. L. and Trugano, T. G., "Verification and Validation in Computational Fluid Dynamics". Progress in Aerospace Sciences 2002; 38: 209–272.

[15] Steinle, F. W., Jr., "Calibration of Strain-Gage Balance for Thermal Effects," AIAA 2002-0882, 40th AIAA Aerospace Sciences Meeting and Exhibit, Reno, NV, Jan 14–17, 2002.

[16] Steinle, F. W., Jr., and Richardson, S.G., "Transitioning Thermal and Anelastic Balance Calibration to Practice," AIAA 2006-3435, 25th AIAA Aerodynamic Measurement Technology and Ground Testing Conference, San Francisco, CA, Jun. 5–8, 2006.

[17] Steinle, F. W., Jr., and Peters, W. L., "Balance Calibration Procedures for Time-Dependent Phenomenon," AIAA 2007-352, 45th AIAA Aerospace Sciences Meeting and Exhibit, Reno, NV, Jan. 8–11, 2007.

[18] Box, G. E. P., and Draper, N. R., *Response Surfaces, Mixtures, and Ridge Analyses,* 2nd Ed., John Wiley and Sons, New York, 2007.