The Modern Design of Experiments for Configuration Aerodynamics: A Case Study

Richard DeLoach*
NASA Langley Research Center, Hampton, VA 23681

The effects of slowly varying and persisting covariate effects on the accuracy and precision of experimental result is reviewed, as is the rationale for run-order randomization as a quality assurance tactic employed in the Modern Design of Experiments (MDOE) to defend against such effects. Considerable analytical complexity is introduced by restrictions on randomization in configuration aerodynamics tests because they involve hard-to-change configuration variables that cannot be randomized conveniently. Tradeoffs are examined between quality and productivity associated with varying degrees of rigor in accounting for such randomization restrictions. Certain characteristics of a configuration aerodynamics test are considered that may justify a relaxed accounting for randomization restrictions to achieve a significant reduction in analytical complexity with a comparably negligible adverse impact on the validity of the experimental results.

Nomenclature

ANOVA	=	Analysis of Variance
AoA	=	Angle of Attack
CBN	=	Critical Binomial Number; minimum number of successes expected with a specified confidence level if there are a given number of Bernoulli trials in which there is a specified probability of success in any one trial
CCD	=	Central Composite Design
CL_{max}	=	maximum lift coefficient
CRD	=	Completely Randomized Design
df	=	degree(s) of freedom
MDOE	=	Modern Design of Experiments
MS	=	Mean Square
PSP	=	pressure sensitive paint
SPD	=	Split Plot Design
SS	=	Sum of Squares
Alternative hypothesis	=	an assertion that two levels of a variable are different
Covariate	=	an uncontrolled ("nuisance") variable such as temperature that influences system response
Dispersion	=	a measure of the variance in a sample of random variable levels
Factor	=	an independent variable (e.g., angle of attack)
F statistic	=	ratio of a component of explained variance to the unexplained variance
Inference Space	=	a Cartesian coordinate system with one axis associated with every independent variable in an experiment; points correspond to combinations of independent variable levels set in the experiment
Level	=	the setting of a particular variable (e.g., AoA = 2°)
Location	=	an estimate (typically, a sample average that may or may not be weighted in some way) of the value of random variable
Null hypothesis	=	an assertion that no difference exists between two levels of a variable
Parameter	=	a population estimate of dispersion, location, or other characteristic

* Senior Research Scientist, Aeronautical Systems Engineering Branch, NASA Langley Research Center, MS 238, 4 Langley Blvd, Hampton, VA 23681, Senior Member.

Population	=	the (generally, only theoretically achievable) totality of all relevant measurements; a sample with an "infinite" number of data points
P statistic	=	probability of observing a given *F* statistic by chance under the null hypothesis
Random variable	=	a quantity for which the specific level is not deterministic, but is determined according to a probability distribution
Residual	=	difference between a measured value and some reference (e.g., a convenient constant, a sample mean, or a prediction)
Response surface	=	a mathematical relationship describing system response as a function of independent variables
Statistic	=	a measure of dispersion, location, or other characteristic of a data sample
Sample	=	a set of data points limited in volume by resource constraints or other considerations

I. Introduction

In 1997, NASA Langley Research Center began examining the costs and benefits of applying formal experiment design methods to wind tunnel testing. These techniques, described collectively at Langley as the Modern Design of Experiments (MDOE), differ in fundamental ways from classical wind tunnel test methods referred to in the literature of experiment design as One Factor at a Time (OFAT) testing.[1] The OFAT method places a high premium on data volume and the quality of individual data points, stressing quality assurance measures that rely upon efforts to improve the measurement environment by reducing unexplained variance in the data.[2,3] MDOE methods assume that real-world measurement environments are inherently imperfect, and rely instead upon tactical measures in the design of test matrices to achieve high quality results even in the presence of systematic and random variation in the measurement environment, which are recognized as inevitable.[4,5,6]

While high data volume is a measure of productivity in conventional wind tunnel testing, MDOE practitioners view data volume as a cost metric, since increases in data volume are accompanied by increases in cycle time, direct operating expenses, and direct and indirect labor costs. In an MDOE test, data volume requirements are defined from an inference error risk management perspective, in terms of precision requirements and inference error risk tolerance levels.[7] Ample data are specified to meet precision and inference error risk requirements defined in the experiment design process, but resources that would otherwise be expended by acquiring additional data beyond this are preserved.

Conventional OFAT testing methods have been adequate in testing environments in which 1% precision levels were considered acceptable. However, ground-testing precision requirements are now approaching the 0.1% level, with fractional drag-count error budgets increasingly the norm in precision performance testing. Furthermore, the trend is toward ever greater emphasis on the need to stretch relatively scarce research resources. This order-of-magnitude reduction in acceptable error budgets, coupled with an increasing focus on full-cost accounting and other incentives to increase efficiency[8], provides a compelling motivation to fundamentally reexamine approaches to wind tunnel testing and other expensive elements of aerospace research. MDOE methods were introduced to the aeronautical ground testing community at NASA Langley Research Center in 1997 as an element of such a reexamination.

This paper presents a case study in which MDOE methods were applied to a configuration aerodynamics wind tunnel test in the 16-foot transonic tunnel at Langley Research Center. Configuration aerodynamics can be distinguished from other forms of experimental aeronautics by a particular attribute of configuration variables that is especially relevant from an experiment design perspective. The levels of some independent variables in a wind tunnel test can be changed relatively conveniently, such as Mach number and the angles of attack and sideslip. They are generally changed by commands to a control system that are issued while the facility is running. Because they require no interruption of wind tunnel operations, they are relatively inexpensive and convenient to change.

By contrast, configuration variables tend to be rather less convenient to change, and relatively more labor intensive and time consuming. They often require physical entry into the tunnel test section, and a relatively extended interruption of tunnel operations to make changes to the model's mechanical configuration. Examples of such variables include control surfaces (flaps, ailerons, elevons, etc.), and various combinations of such elements as landing gears, strakes, speed-brakes, or other components that change the overall mechanical configuration of the test article. The time and effort required to effect configuration changes makes configuration aerodynamics a prime candidate for the application of MDOE methods, in which the volume of data is explicitly specified at minimum levels that are still ample to achieve particular technical objectives.[9]

MDOE practitioners seek to improve quality through operational tactics employed during the execution of the experiment that involve optimizing the run order of the test matrix. One such tactic is to randomize the run order to defend against systematic components of unexplained variance that can otherwise adversely impact the

reproducibility of research results. The hard-to-change nature of configuration variables makes them less amenable to randomization than typical model attitude or flow state variables, which can generally be changed relatively easily. A special type of designed experiment known as split plot design (SPD) is often prescribed when such restrictions on randomization are in play, but the SPD requires a significantly more complex analysis than is required for a completely randomized design (CRD) that features no restrictions on randomization.[10,11]

This paper examines the effect that restrictions on randomization had on a configuration aerodynamics test recently conducted at Langley Research Center as a split plot design. An analysis that respects the restrictions on randomization inherent in this test is compared with an analysis that ignores such restrictions, in order to quantify the penalty associated with this significant simplification in the analysis. Some unanticipated insights were achieved regarding the general nature of conditions under which randomization restrictions must be rigorously taken into account. Practical circumstances are described under which ignoring such restrictions may have no significant effect on the analysis of experimental results.

The sections that follow will review systematic variations and their effects on sample means and variances, which compromise their utility as reliable estimators of corresponding population parameters. The effectiveness of randomizing set-point order as a defense against the effects of such systematic error is discussed. The quality implications of practical restrictions on randomizing hard-to-change variables are reviewed and the special complexity of an analysis that rigorously respects restrictions on randomization is outlined. A specific configuration aerodynamics case study is used to examine costs and benefits of practical options for coping with restrictions on randomization. The discussion section reflects on general features of an experiment that may impact how randomization restrictions can be treated. Brief concluding remarks summarize the key findings.

II. Systematic Components of Unexplained Variance

Unexplained systematic variation is caused by persisting effects such as temperature changes, which vary over time periods that are not short compared to the dwell time associated with a typical data structure such as a pitch polar. Because of the relative time scales, the researcher cannot rely upon simple replication to cancel out these systematic, non-random errors. They behave as slowly varying bias errors, which are often large compared to normal chance variations in the data, and which are also much harder to detect and to quantify than random error.

When systematic variation is present, the experimental errors in individual measurements are not independent of each other. For example, if a trend of increasing temperature causes the most recent measurement to be too high so that its experimental error is characterized as "positive," it is likely that the next measurement will be positive as well. That is, there exists some correlation between measurement errors, and they cannot be said to be independent. We say in such circumstances that the "random sampling hypothesis" is not valid.

Researchers often believe that their measurements are independent, under the good-faith assumption that as long as standard measurement procedures are followed, no additional effort is required to ensure that the random sampling hypothesis can be reliably invoked. As Box, Hunter, and Hunter[12] put it, *"[Researchers] frequently make the assumption of independence at the beginning of their writings and rest heavily on it thereafter, making no attempt to justify the assumption, even though it might have been thought that 'a decent respect to the opinions of mankind requires that they should declare the causes which impel them' to do so. The mere declaration of independence, of course, does not guarantee its existence."*

Unfortunately, covariate effects can induce correlation in experimental data without the researcher's knowledge. Covariates are factors such as frictional heating in the test section of a wind tunnel that can influence the bias and sensitivity calibration constants of force balances and inertial angle of attack sensors, or subtle variations that may occur in flow angularity over time, or any of a myriad other factors that may influence forces, moments, and other response variables in a test, but that are not under the control of the researcher. They are "nuisance variables," the effects of which are not practical to eliminate entirely even when their presence is known.

Unknown covariates are a threat to the reproducibility of research results because their influence can vary from test to test. A polar acquired at the end of two shifts of near-continuous mid-summer running is likely to have been influenced by a completely different temperature environment than an ostensibly identical polar acquired at the start of the first shift of a mid-winter test, for example.

Relatively mild, slowly varying covariate effects can induce a significant bias in the sample mean, and the correlated residuals that result from time-varying covariate effects can significantly bias the variance in the distribution of sample means, as will be demonstrated below. Under such circumstances, these sample statistics are not reliable estimators of the population parameters they are intended to represent. This is a very serious situation, in that the raison d' etre for experimental research is to make reliable estimates of the population parameters that resource limitations prevent us from measuring directly. We must rely instead upon sample statistics from a finite

volume of data that is limited always by a variety of practical considerations. Furthermore, because correlated residuals induce a bias error and not a random error, the acquisition of data in high volume cannot be relied upon to ameliorate the effect, as the errors do not cancel. In fact, when each data point carries a systematic component of unexplained variance, additional data points can simply exacerbate important aspects of the problem, as will be seen below.

A. Effect of Systematic Variation on Sample Means

If there is a time-dependent bias error in the i[th] measurement of an n-point sample so that $y_i = y_{0i} + b_i(t)$, where y_{0i} is the measured value in the absence of bias error and $b_i(t)$ is the bias error at time t, then the sample mean is

$$\bar{y} = \sum_{i=1}^{n} \frac{y_{0i} + b_i(t)}{n} \tag{1}$$

or

$$\bar{y} = \bar{y}_0 + B_{\bar{y}} \tag{2}$$

where $\bar{y}_0$ represents what the sample mean would have been in the absence of systematic covariate effects, and

$$B_{\bar{y}} = \sum_{i=1}^{n} \frac{b_i(t)}{n} \tag{3}$$

is a bias error in the estimate of the sample mean that depends on how the bias errors in the individual measurements depend on time.

Consider a pre-stall lift polar acquired in the presence of a persisting systematic effect that causes earlier lift measurements to be biased somewhat lower – and later lift measurements to be biased somewhat higher – than they otherwise would be. (We may consider individual measurements made at a given angle of attack as a special case of Eq. (1) in which $i=1$; i.e. a single-point "sample.")

If the AoA set-points are scheduled in sequential order, starting with the lowest AoA level and monotonically increasing each subsequent AoA level in defined increments as is the usual custom, the acquired polar will be rotated counterclockwise with respect to the polar that would have been acquired otherwise, in the absence of a systematically changing error. Figure 1 shows this effect schematically.

Rotation in the biased polar of Fig. 1 is only apparent by comparison with the true polar presented in the same figure. The true polar is inconveniently absent in most practical circumstances, and since there is nothing about the biased polar itself to suggest that it is deficient, the effects of systematic error tend to go undetected unless the polar is replicated. But because time-varying bias errors are a function of local conditions that are not likely to reproduce identically from test to test, the researcher who replicates a polar may be left with two unmatched polars with no way to determine which (if either) accurately depicts the lift dependence on AoA.

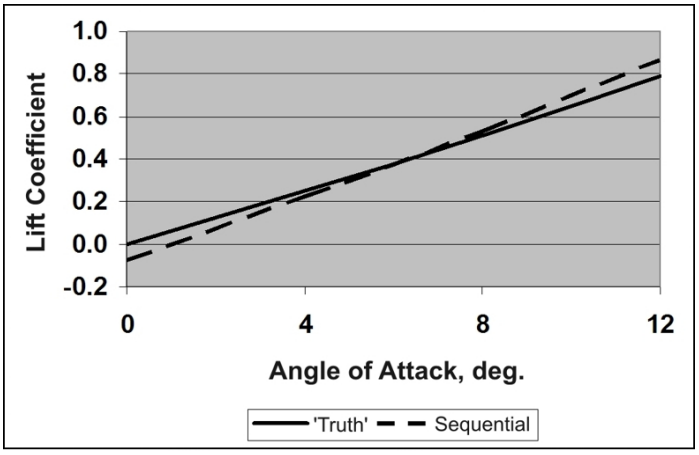

Figure 1. Sequential lift polar rotated by effect of time-varying bias error.

B. Effect of Systematic Variation on Sample Variance

The previous section demonstrated that persisting covariate effects can induce time-dependent bias errors in the individual measurements of a sample of data that result in a bias in the mean of that sample. Such effects also introduce correlation among the measurement errors, which results in a bias in the sample *variance* as well as in the sample mean. To see this, recall the following general error propagation formula[13,14] by which the variance in a function of multiple variables can be estimated from the variance in each of those variables:

If

$$y = f(x_1, x_2, \cdots, x_n) \tag{4}$$

is a known (or assumed) function of n independent variables, x_i, then

$$\sigma_y^2 = \left(\frac{\partial y}{\partial x_1}\right)^2 \sigma_{x_1}^2 + \left(\frac{\partial y}{\partial x_2}\right)^2 \sigma_{x_2}^2 + \cdots + \left(\frac{\partial y}{\partial x_n}\right)^2 \sigma_{x_n}^2$$

$$+ 2\left(\frac{\partial y}{\partial x_1}\right)\left(\frac{\partial y}{\partial x_2}\right)\rho_{x_1 x_2}\sigma_{x_1}\sigma_{x_2} + 2\left(\frac{\partial y}{\partial x_1}\right)\left(\frac{\partial y}{\partial x_3}\right)\rho_{x_1 x_3}\sigma_{x_1}\sigma_{x_3} + \cdots + +2\left(\frac{\partial y}{\partial x_{n-1}}\right)\left(\frac{\partial y}{\partial x_n}\right)\rho_{x_{n-1} x_n}\sigma_{x_{n-1}}\sigma_{x_n} \tag{5}$$

where σ_y is the standard deviation in the response variable y (square root of the variance in y), σ_x is the standard deviation in the independent variable x, and ρ_{uv} is the correlation coefficient between variables u and v.

Let us apply Eq. (5) to the following simple function of the sample mean:

$$n\bar{y} = y_1 + y_2 + \cdots + y_n \tag{6}$$

Note that a great simplification results from the fact that in this instance all the partial derivatives of Eq. (5) are just 1. We further simplify the application of Eq. (5) to Eq. (6) by dropping all the subscripts on the σ values under the assumption that the standard deviation in each measurement is the same, and we introduce the notation ρ_m to represent the correlation coefficient for two measurements separated by m time intervals. For example, ρ_1 represents the correlation coefficient for two measurements taken one measurement apart in a time series (that is, in immediate succession), ρ_2 represents the correlation coefficient for two measurements taken two measurements apart (that is, with one intervening measurement), and so on. For a given m, all such pairs are further assumed to have the same correlation coefficient. With these simplifications, applying Eq. (5) to Eq. (6) results in the following:

$$Var(n\bar{y}) = n\sigma^2 + \sum_{m=1}^{n-1} 2(n-m)\rho_m \sigma^2 \tag{7}$$

We are interested in the variance in the sample mean itself, rather than the function of the sample mean represented by Eq. (6). If we apply Eq. (5) to the following self-evident relationship:

$$\bar{y} = \left(\frac{1}{n}\right) n\bar{y} \tag{8}$$

we get

$$Var(\bar{y}) = \left(\frac{1}{n}\right)^2 Var(n\bar{y}) \tag{9}$$

Inserting Eq. (7) into Eq. (9):

$$Var(\bar{y}) = \left(\frac{1}{n}\right)^2 \left[n\sigma^2 + \sum_{m=1}^{n-1} 2(n-m)\rho_m \sigma^2 \right] \tag{10}$$

or, after rearranging terms:

$$Var(\bar{y}) = \frac{\sigma^2}{n} \left[1 + 2\sum_{m=1}^{n-1} \left(\frac{n-m}{n} \right) \rho_m \right] \tag{11}$$

Note that in the absence or correlated errors, the ρ_m are zero for all m, and Eq. (11) reduces to the familiar expression outside the bracket, representing the variance in the sample mean when all measurements are independent. We see that Eq. (11) has this simple form:

$$Var(\bar{y}) = Var(\bar{y}_0) + B_{Var(\bar{y}_0)} \tag{12}$$

where $Var(\bar{y}_0)$ represents what the variance in the sample mean would have been in the absence of systematic covariate effects and

$$B_{Var(\bar{y}_0)} = \frac{\sigma^2}{n} \left[2\sum_{m=1}^{n-1} \left(\frac{n-m}{n} \right) \rho_m \right] \tag{13}$$

is a bias error in the estimate of variance in the sample mean that depends on how errors in the individual measurements of a time series are correlated. The weighting represented by the *(n-m)/n* term within the summation of Eq. (13) is greatest for small *m*, corresponding to measurements separated by the smallest time intervals. Larger values of *m* correspond to measurement pairs separated by intervening measurements.

Random errors tend to cancel more and more completely as more and more data are acquired. However, Eq. (13) reveals that the acquisition of additional data does nothing to ameliorate the effect that correlated residuals have on variance bias error. On the contrary, since the sign of the correlation coefficient in Eq. (13), ρ_m, tends to be the same for all *m* (i.e., the bias errors in successive data points tend in the same direction in the presence of persisting covariate effects), when a degree of correlation exists among all point pairs, the errors do not cancel and the additional data points simply exacerbate the problem by providing more non-canceling terms in the summation of Eq. (13).

C. Impact of Biased Sample Statistics

Unbiased sample means are obviously desirable in any experiment because the sample mean is assumed to be an unbiased estimate of the population mean we seek to discover, and a bias in the sample mean per Eq. (2) will represent an error in estimating that population mean. The variance of the distribution of sample means is also a crucial statistic because it is key to assessing uncertainty. A bias in this quantity per Eq. (11) will result in an improper quality assessment.

Such unbiased sample statistics are important in designed experiments for additional reasons. The analysis of a designed experiment generally involves objective decisions about whether the magnitude of some effect is sufficiently large to distinguish it from zero. A null hypothesis is formed which asserts that the effect is in fact zero. This hypothesis is only rejected if the effect is located sufficiently far from zero that to do so entails an acceptably low risk of error, given the degree of dispersion (variance) in the estimate of the effect's location (mean). Significant bias errors in estimates of either the location or the dispersion of an effect can complicate the decision of whether or not to reject its null hypothesis, and lead to inference errors. An example will clarify this point.

Consider a response surface experiment design common in wind tunnel testing, in which the intent is to mathematically express system responses such as forces and moments in terms of the independent variables that influence them, such as angle of attack and Mach number. A generalized mathematical relationship is hypothesized

(called a response surface function) – typically a low-order polynomial comprising a Taylor-series representation of the response over some restricted range of independent variable levels. (Often in practical applications the restricted range over which a Taylor series of modest order can adequately approximate the system response is insufficient to span the whole independent variable range of interest. In such cases it is customary to represent the response over the entire range of independent variables as a piecewise-continuous response surface, consisting of multiple adjacent response surfaces.[15,16] The response surface modeling process entails an objective procedure by which decisions are made about which terms to include in the response model and which terms to drop.

Coefficients for the terms in the proposed response model are determined from experimental data using regression or some other means. A null hypothesis is formulated for each term in the model, asserting that the true value of its coefficient is zero and that the term does not belong in the model. The uncertainty in estimating each coefficient is estimated as well as its value. If the regression coefficient for a given term is located sufficiently far from zero given the dispersion in its estimate, the null hypothesis for that coefficient is rejected and the term is retained in the model. Otherwise the term is dropped on the basis that it is located too close to zero to be resolved as a real effect with a sufficiently high degree of confidence, given the dispersion in estimating it.

Equation 14 represents an initial reduced cubic response model for lift coefficient as a function of angle of attach, α, and Mach number, M, as formulated in a recent wind tunnel test at Langley Research Center.

$$C_L = b_o + b_1\alpha + b_2 M + b_3\alpha M + b_4\alpha^2 + b_5 M^2 + b_6\alpha M^2 + b_7\alpha^2 M \tag{14}$$

This order of model was initially conjectured from experience with similar aircraft tested over the same range of independent variables; α from -5º to +5º and Mach number from 0.35 to 0.80 in this case. We expected the lift coefficient to be dominated by a first-order dependence upon angle of attack in this pre-stall range, with some slight curvature at the higher α end reflecting the start of an approach to CL_{max}. We did not expect the lift coefficient to exhibit any higher order dependence on Mach number than for angle of attack. For this reason, the initial mixed cubic model was expected to be of more than adequate order to represent the data, although the fit of any such response model to the data is always subject to verification. If there is a need for a more complex model to adequately represent the data, this is generally revealed through an analysis of residuals and possibly confirmation points – data acquired not to fit the model but to test it. That is, a certain number of Mach-alpha points might be held in reserve, to compare with predictions based on the fitted response model.

The b_i in Eq. (14) are numerical coefficients that were determined empirically by fitting this equation to a set of lift data acquired over the prescribed range of angle of attack and Mach number. Values that were estimated for the coefficients in Eq. (14), as well as values of the standard error ("one-sigma") in estimating each coefficient, are given here.

$$\begin{aligned} 10,000\times C_L = {} & (43.6\pm 8.5) + (1881.8\pm 10.4)\alpha + (1.4\pm 10.4)M + (75.0\pm 10.5)\alpha M \\ & -(37.4\pm 14.9)\alpha^2 - (11.6\pm 7.7)M^2 - (8.8\pm 7.7)\alpha M^2 + (8.0\pm 14.8)\alpha^2 M \end{aligned} \tag{15}$$

The "10,000" multiplier facilitates a clearer comparison of the relative magnitudes of the coefficients, and converts the otherwise dimensionless lift coefficient units to "counts", where one count = 0.0001. The α and M independent variables have undergone a centering and normalizing transformation to convert them to numbers in the range of -1 to +1 to facilitate the regression computations.

Terms with insignificant (near-zero) coefficients could remain in this response model without having a significantly adverse effect on the value of response predictions, but it is very desirable to identify them and drop them from the model for two reasons. A clearer insight into the underlying process can be achieved if the response model is not cluttered by irrelevant terms, but even more importantly, the uncertainty in response predictions made by models such as Eq. (14) depends on the number of terms in the model. The average prediction variance is directly proportional to the number of terms, independent of the order of the model[17]:

$$\overline{Var(\hat{y})} = \left(p \big/ n \right) \sigma^2 \tag{16}$$

where p is the number of parameters in the model (including the b_0 intercept term), n is the number of data points used in the regression, and the term on the left is the prediction variance averaged over all the points in the

regression. There is thus some considerable pressure to eliminate as many unnecessary terms as possible, to clarify the model and improve the precision of its predictions. Specifically, we seek to eliminate small-coefficient terms that contribute more to the prediction uncertainty than to the prediction itself.

An objective decision to either retain or reject each term in the response model is made by appealing to a normal "reference distribution" centered on zero with a standard deviation reflecting the standard error in estimating the coefficient. Figure 2 illustrates the reference distribution for Eq. (14). Each vertical line in this figure represents a coefficient from Eq. (15), expressed in multiples of its standard deviation. The dispersion in the reference distribution reflects the resolution of the experiment. Regression coefficients outside the rectangle centered on zero are located sufficiently far from zero to reject the null hypothesis that their coefficients are insignificant, and to do so with no more than a 5% probability of an error in such an inference due to chance variations in the data. That is, we claim with at least 95% confidence that the coefficients outside the rectangle are non-zero and should be retained in the response model. The true value for coefficients located within the rectangle may be zero, with a non-zero numerical value attributable simply to random experimental error. Or they may be non-zero and simply small in magnitude. In either case, we are unable (with at least 95% confidence) to reject the null hypothesis for such coefficients given the dispersion in their estimates as represented by the normal reference distribution of Fig. 2.

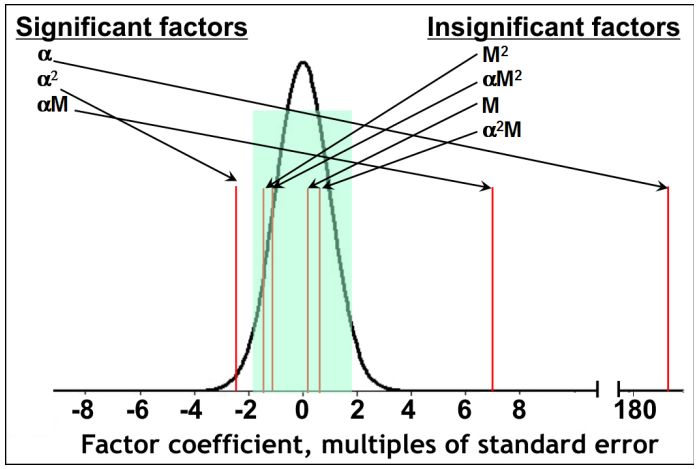

Figure 2. Reference distribution for detecting significant factors in a response surface model for pre-stall lift coefficient as a function of angle of attack (α) and Mach number (*M*).

For this particular response function, we note that three of the seven terms in the model (beyond the constant intercept term) have coefficients located sufficiently far from zero to reject the null hypothesis, and the remaining four terms are too close to zero to be adequately resolved, where again "adequately" was defined for this test to mean "with at least a 95% probability of a valid inference." In particular, the regression coefficient for the linear α term is seen to be over 180 standard deviations to the right of zero, leaving little doubt that this term is statistically significant. The quadratic α term is much smaller, suggesting only modest curvature, but it is still located far enough away from zero to satisfy our 5% maximum inference error risk criterion. The coefficient for an interaction term involving α and Mach is over seven standard deviations to the right of zero, comfortably justifying a rejection of the null hypothesis for this term. The presence of this term in the response model suggests that the slope of the (essentially) linear dependence of C_L upon α is greater for higher Mach numbers than for lower Mach numbers. Rewriting Eq. (14) to reflect only the significant terms we have:

$$C_L = b_o + b_1\alpha + b_3\alpha M + b_4\alpha^2 \tag{17}$$

This gives a much less cluttered view than Eq. (14) of the dependence of lift on angle of attack and Mach number over the α and Mach ranges for which this relationship was developed. Also, per Eq. (16), since this response model features half the number of terms of Eq. (14) – four instead of eight – the average prediction variance is cut in half.

Note also that there is as much information about the underlying physical process to be gleaned from the *insignificant* terms as from the significant ones. In this instance, it is interesting to note that while Mach number influences the rate at which lift changes with angle of attack (via the interaction term), there is no net average dependence of lift on Mach number (this may be attributable to the symmetric ±5° α range centered on $\alpha=0$) and there is no curvature in Mach as there is in α. The fact that the $\alpha^2 M$ term is insignificant reveals that there is no interaction between Mach number and the curvature term for α. This suggests that the same nonlinearity in α that is observed at lower levels of the range of Mach numbers tested is also observed at the higher end of that range. A

subject-matter specialist can achieve many such interesting insights into the underlying physics by examining how "the chaff is separated from the wheat" when such response surface model coefficients are objectively determined to be significant or insignificant.

Note how this process of discovery depends on unbiased estimators of the sample means (required for reliable coefficient estimates) and on unbiased estimators of the variance in the distribution of sample means (required for a reliable reference distribution that is key to distinguishing between significant and insignificant effects). Persisting covariate effects that were demonstrated above to bias the estimates of both means and variances can play havoc with this process, and can result in response models that are incapable of reliably predicting responses for independent variable combinations of interest.

Consider the quadratic Mach term from Eq. (15), seen in Fig. 2 not to be located sufficiently far from zero for its null hypothesis to be rejected. A relatively subtle bias in the estimate of this coefficient – enough to increase its small magnitude by about a third – would have been sufficient to reject its null hypothesis and claim some quadratic Mach dependence for lift. Very little bias is needed for such an occurrence – enough to correspond to a shift in lift coefficient of less than 0.001. The pure error standard deviation for this test was 0.002, so this is a bias shift of only "half a sigma," and a small sigma at that.

Likewise, consider the coefficient of the quadratic angle of attack term, seen in Fig. 2 to be statistically significant (located far enough from zero to reject the null hypothesis with 95% confidence), but just barely so. A small bias in the variance of the reference distribution per Eq. (11) would be sufficient to erroneously infer that the null hypothesis for the quadratic α term should not be rejected. For example, an increase of about 30% in the "$\pm 2\sigma$" acceptance criterion interval half-width of the reference distribution would be sufficient to reverse the inference for the quadratic α term. This translates into a numerical value of about 0.35 for the dimensionless summation term in Eq. (11), which can be achieved under correlation conditions so mild that a given measurement is influenced only by the preceding measurement and none earlier ($\rho_m=0$ for $m>1$, a so-called "lag-1 autocorrelation"), and that $\rho_1=0.4$, a very mild degree of correlation indeed. Such a small level of correlation would have been sufficient to miss the subtle curvature in the α dependence on lift coefficient and to introduce errors in the prediction of lift that would not have included contributions from the quadratic α term of Eq. (17).

In this example, it is unlikely that any level of realistic systematic variation could have masked the statistical significance of the first-order α term in the response function for lift (180+ standard deviations away from zero!) and the alpha-Mach interaction at 7+ standard deviations away from zero would also have been hard to obscure. But somewhat more subtle effects such as the curvature of lift with α could easily have been masked by bias in the sample statistics induced by a significant systematic component of unexplained variance. Likewise, totally spurious effects could have been introduced into the description of lift's dependence on Mach number and angle of attack.

III. Quality Assurance Through Randomization of Set-Point Order

The previous sections have established that proper inferences about system responses can be compromised as a result of long-period response variations induced by systematically changing covariates that bias the estimates of sample means and variances. Furthermore, the level of systematic unexplained variance and the degree of correlated residuals required to induce erroneous inferences is very subtle. This suggests a certain underappreciated urgency for the need to eliminate systematic components of unexplained variance from experimental data by proactively engaging in quality assurance tactics designed to ensure the random sampling hypothesis.

A number of such efforts have been focused in the past on attempts to improve the measurement environment in certain large-scale facilities such as wind tunnels, by identifying systematic error sources ("assigning causes" to systematic error) and physically wringing those sources of systematic error out of the system. Of course this should be done to the full extent that it is practical to do so. However, one can only address systematic error sources that are known, and unfortunately it is not unlikely that a number of such sources are in play at any given time without our knowledge of them. In any case, as it is impossible to prove a negative, it is simply not possible to guarantee a state in which there are no sources of systematic variation in a wind tunnel, no matter how sincerely we strive to identify and eliminate them.

Fortunately, a conceptually straightforward quality assurance tactic exists by which the random sampling hypothesis can be induced, even in the presence of slowly varying covariate effects. This tactic, introduced by Ronald Fisher and his peers early in the 20th century[18], consists of randomizing the set-point order of data acquired in a time series to effectively convert unknown and undetected systematic components of unexplained variance into simple random error that is easy to detect and which can be addressed by the simple tactic of replication.

To illustrate how randomization can help in a wind tunnel test, let us revisit Fig. 1. This figure illustrates a lift polar that has been rotated counterclockwise under the influence of transient systematic phenomenon that had the

effect of biasing earlier measurements low and later measurements high. A slowly-varying change in test-section temperature could have this effect, through the influence in might have on calibration constants for onboard force and model attitude instrumentation, for example.

Now consider the effect of simply randomizing the set-point order in which the same AoA levels are set. The earliest setting might be a relatively large angle or a relatively small one, determined completely at random, and likewise for all subsequent measurements. AoA levels that are set earlier will be biased somewhat low in this example, and those set later will be biased somewhat high. The errors would display a systematic time dependence if plotted in run order, but they would be randomly distributed if presented in order of angle of attack. This is because each of the lower angles will have been just as likely to have been set earlier (when the systematic error was negative) as later (when the systematic error was positive), and likewise for the higher angles. The errors will be randomly distributed about a mean value that more closely represents the true functional dependence of lift upon AoA than if the AoA levels had been set sequentially, effectively confounding the AoA effect on lift with whatever effect is causing the systematic error.

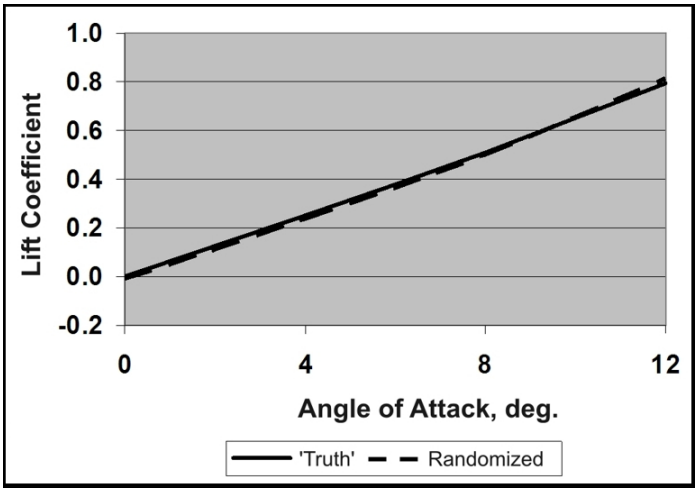

Figure 3. Effect of systematic variation on a randomized lift polar.

Figure 1 was actually generated from a simulation that represented lift coefficient as a simple first-order function of one variable – angle of attack. Random error was added via a Monte Carlo simulation that featured random selections from a normal distribution with mean of zero and standard deviation of 0.005, with all the experimental errors then correlated with a lag-1 autocorrelation coefficient of magnitude 0.4. The polar labeled as "truth" consisted of the original first-order function with no error components added, and the rotated polar resulted from a linear regression on a set of data consisting of the base function with random and systematic error components added, acquired in a monotonically increasing sequence of angle of attack set-points. This procedure ensured that AoA effects were completely confounded with the covariate effects responsible for the correlation, and resulted in the rotation of the polar relative to the "true" polar.

Figure 3 is identical to Fig. 1, except for a curve fit to data acquired with the angle of attack set-point order randomized. The randomization of set-point order has ensured that some of the lower-alpha set-points were acquired early, when the bias error was negative, and some was acquired late, when the bias error was positive, and likewise for the higher-alpha set-points. The result was a randomized distribution of former bias errors about the true AoA dependence.

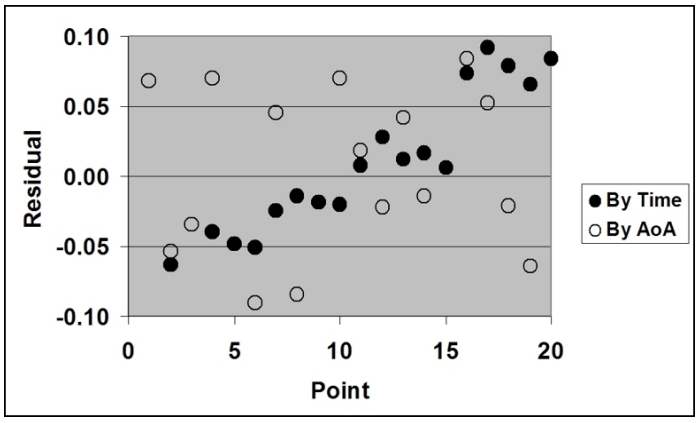

Figure 4. Residuals from randomized polar in order of alpha and in run order (by time).

Figure 4 presents the residuals from Fig. 3 in two ways. The open circles represent the residuals plotted in order of angle of attack. These exhibit a featureless swath centered on zero, indicating pure random error. The filled circles show the residuals plotted in the order the randomized set-point data were acquired. That is, the filled circles are residuals from points acquired in time-order. Note the pronounced lower-left-to-upper-right trend in the time-series of residuals, indicating clearly that data acquired early and data acquired late were biased in opposite directions due to the time-varying trend that was in fact simulated n this data set. But notwithstanding the pronounced trend in the residuals plotted against time, there is only a random

distribution of errors in the residuals plotted against angle of attack, clearly illustrating how randomizing set-point order effectively converts systematic variation to random error.

We note in passing that there are certain practical advantages to monotonic AoA schedules that have to do with the fact that flow attaches to the wing in different ways when the angle of attack is increased than when it is decreased. This so-called "hysteresis effect" can cause the forces and moments on a wind tunnel model positioned at a given angle of attack to be different if that AoA level was approached from below than if it was approached from above. Hysteresis can be avoided in a randomized schedule of AoA settings by a policy that requires negative AoA transitions to be preceded by a "home state setting" consisting of an AoA level lower than the smallest one to be set for data acquisition. In this way, all AoA data settings are acquired after a positive transition in AoA.

We also note in passing that randomization of test matrices can be viewed as an impediment to productivity by those who equate wind tunnel productivity with data volume. A randomized polar takes longer to execute than a sequential polar, typical by a factor of 1.5 to 2.5. However, as alluded to in the introduction, designed experiments compensate for reduced data acquisition rates by minimizing the volume of data necessary to achieve technical results. Significant improvements in both quality and productivity have in fact been achieved when designed experiments have been employed in practical ground testing problems. A more detailed general discussion is beyond the scope of the current paper, but the interested reader can consult other references for more information.[19,20]

IV. Restrictions on Randomization and the Split Plot Design

A significant portion of the unexplained variance in a high-precision data sample can often be attributed to slowly varying covariate effects that are subtle yet persistent, and which introduce what is in effect a time-dependent bias error than can cause sample means to vary systematically over relatively long time periods. There is an imperative to randomize the run order of a test matrix in order to ensure that experimental errors in points acquired in such an environment are all independent of each other. This is a necessary condition for sample statistics, such as the means and variances of random variables estimated in an experiment, to be unbiased estimators of the population parameters they are intended to represent. Absent such randomization, even the most subtle of systematic error effects can overwhelm the tiny error budgets that are typically associated with today's high-precision experimental aeronautics investigations. This can result in improper inferences and experimental results that are difficult to reproduce with required precision.

The need to randomize set-point order as a quality assurance tactic conflicts with certain attributes of configuration variables that were mentioned above. Specifically, configuration variables can be described as "hard to change" variables, the levels of which are generally not practical to randomize from point to point. Once the configuration of a wind tunnel model has been established by making the necessary mechanical changes to it, the tunnel is typically secured for running and only "easy to change" variables such as angle of attack and Mach number are altered until tunnel operations are again interrupted to permit the next configuration to be established. It is possible to randomize the set-point order of easy-to-change variables, and in fact the practical capability to randomize set-point order serves as a working distinction between variables that are classified as easy-to-change and those that are classified as hard-to-change.

The presence of hard-to-change variables necessitates certain modifications in the design and analysis of an experiment to accommodate them. These modifications are often implemented through a special designed experiment commonly used to cope with restrictions on randomization, called the Split Plot Design (SPD). This section will briefly introduce key features of split plot designs that highlight certain undesirable elements such as the complexity they introduce into the analysis of experimental data.

Split plot designs are so-named because they were developed originally in agricultural applications involving plots of land that featured randomization restrictions. For example, in an experiment to examine potato yield for three varieties of potato plant and two types of fertilizer, it would generally be more practical to distribute the three plant varieties among randomly selected subplots in two fields that are each treated with one type of fertilizer, than to try to fertilize each individual plant with one fertilizer or the other, selected at random. (Envision the fertilizer being applied by crop duster!)

Field A might be surveyed into 36 equal subplots for example (perhaps a six-by-six array of subplots), with potato variety #1 assigned to a third of them (12 of the 36 subplots selected at random) and the other two varieties likewise assigned at random to a third of the subplots. This random assignment of plants to subplots within a field defends against the possibility that yield differences might be due to large subsections of the field having different soil types, or moisture levels, or degrees of insect infestation, say. When the plants are assigned to subplots distributed randomly within the field, each plant type is just as likely to experience rich soil as poor, moist soil as dry, and so on, with any net yield differences across plant types attributable only to the plant-type differences

themselves. Field B would be planted similarly, and because it is convenient to do so, all of Field A would be treated with one type of fertilizer and all of Field B would be treated with the other.

We call the two fields "whole plots" and we describe "fertilizer type" as a whole plot variable. "Plant type" is a subplot variable. Note that it is necessary to replicate this design to avoid confounding the whole plot variable effect with what are called "block effects." "Field A" might simply be superior to "Field B" for growing potatoes, no matter which fertilizer is applied. So we would want to plant multiple pairs of fields, deciding at random which of the two fields in each replicate to treat with one fertilizer type and which to treat with the other. That is, we would want to randomize the assignment of the whole plot variable (fertilizer) to whole plots (fields) just as we randomized the assignment of the subplot variable (potato variety) to subplots within each field.

This need for two types of randomization – whole plot and subplot – makes the analysis of split plot designs more complicated than the analysis of completely randomized designs that have no restrictions on randomization. A CRD features a single reference distribution by which the significance of all independent variable effects are judged. Because a SPD features two separate whole plot and subplot randomizations, two different reference distributions are required to assess separately the significance of whole plot effects and subplot effects. Interaction effects involving whole plot variables and subplot variables present a further complication.

The complication introduced by the separate randomization schemes that hard-to-change variables impose can be illustrated with an experiment that was recently executed at Langley Research Center in to assess the effect of pressure sensitive paint (PSP) on measurements of forces and moments. Pressure sensitive paint undergoes certain changes proportional to pressure that can be recorded optically to reveal global pressure distributions on the surfaces of wings and other aircraft components during a wind tunnel test. However, there is some potential that the application of the paint to a wind tunnel model could bias sensitive aerodynamic force and moment measurements, especially drag.

The actual PSP experiment was somewhat complicated, but to reduce its description to the simplest elements needed to explain split plot designs, imagine that the experiment involved only two independent variables, P and M. P is a binary state variable that assumes one level (+1, say) when the state of the model is "painted" and the other level (-1, say) when the state of the model is "clean" (no PSP). M is Mach number, which for the purpose of this discussion will also assume only two levels, a relatively low Mach number and a relatively high Mach number. The objective of this very simple test is therefore to assess the change in drag due to pressure sensitive paint for low and high Mach numbers at fixed levels of all other variables such as angle of attack and angle of sideslip.

In the simplest of execution plans, drag measurements would be made at the two Mach numbers when the model is clean and repeated when the model is painted. A total of four data points would be acquired in this simplified experiment, each contributing one degree of freedom to the analysis. One degree of freedom (df) would be consumed by the mean of the data (the intercept term of any predictive model to be developed from the data), with a total of three degrees of freedom remaining to assess all independent variable (paint and Mach) effects and the uncertainties in their estimation. Table 1 reveals the degrees of freedom available for this simple experiment.

One df is consumed by the paint effect, which quantifies the change in drag in going from a clean model to a painted model, averaged over Mach. One df is consumed by the Mach effect, quantifying the change in drag in going from low Mach to high Mach, averaged over paint states. One df is associated with the Mach x Paint interaction effect, which quantifies any change in paint effect that might occur in going from low Mach to high Mach. There are no degrees of freedom remaining to assess the uncertainty of any of these effects estimates, and thus there is insufficient information to construct a reference distribution such as the one used in Fig. 2 to objectively infer the significance of various experimental results.

Source	Degrees of Freedom
Paint	1
Mach	1
Mach x Paint	1
Error	0
Total	3

Table 1. Degrees of freedom for PSP experiment: No replication and no account of restriction on randomization.

It would be necessary to replicate this experiment to ensure that some df are available to establish a reference distribution after all the independent variable effects are estimated. This is important because the paint effect would be unlikely to be estimated as exactly zero even if there really was no effect, simply due to random experimental error. Absent a reference distribution reflecting the intrinsic precision of the experiment, we cannot objectively decide – with a prescribed level of confidence – whether the null hypothesis for the paint effect should be rejected or not. (The null hypothesis for the paint effect states that there is no difference in drag between the painted and unpainted cases.) This absence of a reference distribution is especially problematic if the paint effect is subtle, as it was expected to be (and as it in fact turned out to be).

Source	Degrees of Freedom
Paint	1
Mach	1
Mach x Paint	1
Error	4(r-1)
Total	4r-1

Table 2. Degrees of freedom for PSP experiment replicated r times: Restriction on randomization not taken into account.

Replicating the PSP experiment r times would provide a total of 4r df, or 4r-1 df beyond the one consumed in estimating the mean. The df budget for the case of r replications of the basic PSP experiment is shown in Table 2.

The replicated experiment features 4(r-1) error df that can be used to estimate the variance of a reference distribution. We would use this distribution to objectively assess the magnitude of the three potential effects – Paint, Mach, and Paint x Mach interaction – by determining if they can be reliably distinguished from zero given the intrinsic variability of the data. The intrinsic variability of the data is reflected by the unexplained variance of in the 4r data points acquired in this test (i.e., that portion of the total variance that cannot be explained by the two main effects – paint and Mach – and the interaction between them). We would conduct an analysis similar to the one represented graphically in Fig. 2, by which we objectively decided whether or not each regression coefficient in Eq. (14) was significant. In the PSP experiment, we would determine which of the three effects are located far enough away from zero to distinguish them from noise with an acceptable level of confidence – 95%, say.

Unfortunately, the df budget presented in Table 2 ignores an important restriction on randomization imposed by the fact that the paint state of the model cannot be changed conveniently. Because paint state is such a hard-to-change variable, it is much easier to set all Mach levels of interest for one paint state first and then the other, than to set 4r Mach/Paint combinations in random order as we would with a completely randomized design (CRD) of this experiment. We would still wish to randomize set point order as much as possible within this restriction on randomization, by determining at random (by coin toss, say) the order that the high and low Mach numbers are set for a given paint state, and by also randomizing the order that the paint states are run. For example, if r = 5, we would replicate the 4 combinations of low/high Mach with clean/painted model a total of 5 times, yielding 20 runs. Randomizing both Mach levels within a single paint-state level and then randomizing the order that the paint states are executed within a given replicate would make this a split plot design.

The df budget for a split plot design of the PSP experiment is more complicated than the one in Table 2 that neglects restrictions on randomization and assumes a CRD. The hard-to-change nature of the paint state variable in the PSP experiment results in two types of randomization and therefore two categories of error df, as indicated in Table 3.

As a specific implementation of the PSP experiment as a split plot design, let us assume that we replicate the PSP experiment on each of five consecutive days so that r = 5 in Table 3. That is, on Monday we flip a coin to decide whether to test the painted model first or the clean model first, and then flip a coin again to decide whether to run the low-Mach case first or the high-Mach case. We measure the drag for the Paint/Mach combination so determined, and then measure the drag at the other Mach number. We then prepare the model for the other paint state for that day (by removing the paint if the first paint state was "painted" or by painting the model if the first paint state was "clean"). We again run the two Mach number settings in a run order determined by coin toss, and repeat this same process on Tuesday, Wednesday, Thursday, and Friday, using coin tosses on each day to determine the order that the paint states will be run, and the order that the Mach numbers will be set for each paint state.

We will now have drag measurements for two Mach numbers at each of two paint states on each of five days, for a total of 2 x 2 x 5 = 20 runs. As before we lose one df to an estimate of the mean of this data sample, leaving a total of 19 df to assess the independent variable effects and the uncertainties (reference distributions) needed to test their significance.

We test the whole plot effect (Paint) against the r-1 df paint/replication whole plot error term in Table 3 as follows: For each of the five days, subtract the average of the drag measurements

Source	Degrees of Freedom
Replications	r-1
Paint	1
Paint x Replication	r-1 (whole plot error)
Mach	1
Mach x Paint	1
Mach x Replication	r-1 (subplot error)
Mach x Paint x Replication	r-1 (subplot error)
Total	4r-1

Table 3. Degrees of freedom budget for PSP experiment replicated r times with restriction on randomization taken into account.

made at both Mach numbers with the clean model from the average of the drag measurements made at both Mach numbers with the painted model. This provides estimates of the average paint effect for each of the five days. The mean of these five numbers is the average paint effect we seek.

To determine if the magnitude is significant, compute the standard deviation for these five paint-effect estimates and construct a *t*-statistic by dividing the mean by this standard deviation. This will express the paint effect in multiples of its standard deviation over the five replicates. Earlier we performed a similar analysis graphically to illustrate the concept of a reference distribution (see Fig. 2), but it is not actually necessary to do this graphically. The computed *t*-statistic can simply be compared with tabulated critical two-tailed *t*-values found in standard statistical tables that list them as a function of df and significance level – the complement of confidence level. For this case, standard tables list the critical *t*-statistic for 4 df corresponding to a significance level of 0.05 (95% confidence) as 2.776. If the magnitude of the average paint effect is at least 2.776 times greater than the 4 df estimate of the standard deviation associated with this average, then we are entitled to reject the null hypothesis of no significant paint effect and conclude that PSP actually does affect drag measurements, with no more than a 5% probability that this inference will be in error. Otherwise we are unable to reject the null hypothesis with at least 95% confidence. (The actual PSP experiment involved seven replicates, and included a more complex array of subplot variable combinations, but resulted in the conclusion that no significant paint effect could be detected at the 0.01 significance level, corresponding to a confidence level for this conclusion of at least 99%.)

There are two schools of thought for how to assess the significance of the subplot effect and the interaction between the subplot and whole plot variables in a split plot design. One approach is to proceed in an analogous fashion to the whole-plot case. Compute the subplot effect for each day (in this case by subtracting the average of the painted and clean drag measurements made at the low Mach number from the high-Mach estimate of the same quantity), and then average across all five days. Take the ratio of the average of these five numbers to their standard deviation and compare with a critical *t*-statistic as before. For the subplot/whole plot interaction (Mach/Paint), compute the low-Mach paint effect on each day and subtract it from the high-Mach paint effect for that day. This will be the interaction effect for that day. As before, average across all five days, using the standard deviation as a reference to compute a *t*-statistic that is compared with the appropriate critical *t*-value obtained from standard statistical tables.

The problem with this approach is that there is theoretically no real variation in subplot effects from one whole plot block to another. In the presence of slowly varying covariate effects we might expect the absolute level of drag measured at low Mach on Monday afternoon to be somewhat higher than it was on Monday morning, and similarly for the absolute level of drag measured at high Mach in the afternoon compared to the morning. But changing the Mach number from low to high should produce the same *change* in drag in the afternoon as it did in the morning, neglecting set-point errors in Mach and ordinary random experimental error.

If we permit the possibility of a legitimate block/Mach interaction, then the whole aerodynamic concept of Mach number is called into question. No longer can we regard a given differential Mach number as the change in Mach that produces a given drag change. We can only regard it as the change that produces this effect "on Monday afternoon," say. There are similar conceptual difficulties with the concept of three-way interactions among whole plot variables, subplot variables, and blocks. (Having questioned the validity of block/Mach interactions, it makes little sense to consider the interaction between this effect and paint state.)

An alternative solution to the conceptually difficult proposition of admitting block/factor interactions is to pool these effects with the rest of the unexplained variance to generate a single error term upon which to base a reference distribution for assessing the significance both of subplot main effects and of subplot/whole plot interaction effects. An explained sum of squares is computed by adding the sum of squares for whole plot effects, block effects (replicates), subplot main effects, whole plot interactions with replicates, and subplot/whole-plot interactions. This explained sum of squares is then subtracted from the total sum of squares to generate an error sum of squares. The error sum of squares is divided by its associated df, which would be $2 \times (r-1)$ for this PSP experiment, or 8 for $r = 5$ replicates. This results in an error variance estimate, the square root of which is the standard error that can be used to construct *t*-statistics for the subplot main effects and subplot/whole-plot interaction effects.

V. Split Plot Designs for Response Surface Modeling

The description of the PSP experiment described in the previous section reveals that restrictions on randomization can introduce substantial complexity into the analysis of even the most uncomplicated of experiments, including very simple two-level factorial experiments such as this one, with only two variables. The only objective of the PSP experiment was to assess the significance of two main effects and the interaction between them. Typical configuration aerodynamics experiments dominated by randomization restrictions are substantially

more complex than this demonstration experiment. The PSP experiment featured only one whole plot variable and one subplot variable, for example, while most experiments in configuration aerodynamics will feature multiple whole plot (hard-to-change) variables and typically more than one subplot variable.

Furthermore, the objectives of a typical configuration aerodynamics test are substantially more complicated than the PSP experiment. Configuration aerodynamicists may be interested in assessing the significance of potentially subtle effects as in the PSP test, but they are also generally interested in developing response models capable of making reliable, high-precision response predictions for complex combinations of many whole plot and subplot variables. In that process it is desirable to use objective significances tests to identify model coefficients that are too small to justify rejecting a null hypothesis that their effects are negligible, thereby improving the precision of model predictions and gaining important insights into the underlying processes. The rigorously correct reference distribution to use in such tests is complicated by the dual error terms of a split plot design, associated with separate randomization schemes for both easy-to-change and hard-to-change variables.

There are further complications in applying response surface methods to a split plot design. The vector of regression coefficients obtained in a completely randomized design is computed as follows:

$$\mathbf{b} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y} \tag{18}$$

where $\mathbf{y}$ is a vector of measured responses and $\mathbf{X}$ is the design matrix. The design matrix is an extension of the standard test matrix, with rows for each data point but with columns not only for each independent variable, but for every term in the model being fitted to the experimental data. Standard references provide more detail.[17,21]

Equation 18 must be extended to estimate regression coefficients in a split plot design, as follows[11]:

$$\mathbf{b} = \left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y} \tag{19}$$

where $\mathbf{V}$ is a variance covariance matrix that accounts for the fact that measurement pairs acquired within the same whole plot are correlated because they share a common error component. Unfortunately, $\mathbf{V}$ depends on the variance across whole plots and also the covariance between measurement pairs acquired within a whole plot. It is difficult to estimate either of these quantities without a level of replication that is not often practical in configuration aerodynamics testing.

Notwithstanding the fact that the objectives are generally more complicated in a configuration aerodynamics response surface modeling experiment than they were for the relatively simple PSP test described above, for example, the available resources are seldom sufficient to support the level of replication employed in the PSP test, in which every combination of whole-plot variables was replicated five times and every combination of subplot variables was completely replicated within each whole plot. In general, the validity of significance tests required in a rigorously correct analysis of a split plot experiment depends on a level of replication that is simply impractical in most large-scale configuration aerodynamics tests.

None of the obvious alternatives are attractive. Mindful of the fact that restriction on randomization requires separate error variance estimates for hard-to-change and easy-to-change variables, one could simply incur the expense of providing the additional replication necessary to estimate these variance terms. Or one could simply abstain from randomization restrictions, executing every experiment as a completely randomized design no matter what the cost or how much trouble or extra time is involved.

Another alternative is to recognize that a split plot design is comprised of a series of completely randomized designs in the subplot factors, each of which can be analyzed legitimately as a CRD without having to take into account restrictions on randomization. (The subplot factors are all easy-to-change, and can be completely randomized.) This results in separate response models for every whole plot variable combination and fails to quantify interactions between subplot and whole plot variables, to say nothing of the fact that whole plot main effects are not estimated. However, there are circumstances in which these limitations may not be very important. For example, it may be sufficient to separately predict forces and moments as a function of Mach number and angle of attack for two separate wings, without developing an integrated response model that includes a categorical wing variable.

In general, however, one would prefer to develop response models that reveal main and interaction effects involving all of the independent variables, whether they are hard-to-change or easy-to-change. To avoid all of the additional analytical complexity alluded to above, as well as the ambiguity noted in the PSP analysis about differing schools of thought for how to test the significance of subplot variable main effects and interactions among subplot

and whole plot variables, there can be a temptation to analyze the results of an experiment that has been executed with restrictions on randomization as if there had been no such restrictions. That is, it is always possible, whether rigorously correct or not, to honor the restrictions on randomization that practicality dictates during the execution of the experiment but then to ignore those restrictions in the analysis, treating it as a completely randomized design. Such an enormous reduction in analytical complexity would be expected to have associated with it some cost in the form of considerably reduced validity in the conclusions reached, however, the author is unaware of any prior explicit attempt to quantify such costs in a configuration aerodynamics experiment. The remaining sections of this paper describe an analysis intended to quantify such costs for the case of a specific configuration aerodynamics test conducted at Langley Research Center.

VI. Quantifying the Cost of Ignoring Restrictions on Randomization

A test conducted in the 16-Ft Transonic Tunnel at Langley Research Center featured six hard-to-change configuration variables and two easy-to-change variables – angle of attack and Mach number. The configuration variables consisted of experimental lift augmentation devices arrayed along the leading edge of the port-side wing of a delta-winged vehicle. Each could be set at a level of effectiveness between zero and 100%.

There were a total of 52 configurations, consisting of various combinations of the six hard-to-change variables that were set as a face-centered central composite design (CCD) in those variables, with a half-fractional factorial block and eight replicated center points. The order in which the configurations were set was completely randomized. An identical schedule of 28 AoA/Mach combinations was executed for each of the 52 configurations, with the run order completely randomized separately for each configuration. The range of interest for angle of attack was -5º to +15º and for Mach number it was from 0.35 to 0.80. The AoA/Mach points were arrayed as two overlapping orthogonally blocked central composite designs. Both CCDs covered the same Mach range of 0.35 to 0.80. One of them covered the AoA range from -5º to +5º while the other covered the AoA range from 5º to 15º. Figure 5 shows the AoA/Mach inference space for this test and the sites within this space that were set identically for each lift augmentation configuration. Two of these AoA/Mach set points were each replicated six times to generate model-independent estimates of pure error, and two other points were replicated twice, so there were in fact 16 unique AoA/Mach settings for each configuration.

The use of dual CCDs designs to span the full AoA range of interest is an example of a common tactic employed at Langley Research Center to improve the fit that low-order models can provide to complex, real-world aerodynamic force and moment data. However, for the purpose of this investigation into the cost of ignoring randomization restrictions, analyzing two identical (except for range) subplot designs would double the already considerable number of regression computations, but would not add any more insight than considering only one of the subplot designs, so only the low-alpha CCD was examined in this comparison. The high-alpha CCD was completely randomized in exactly the same way, although in general the quality of the fits for this CCD were not as good as for the lower-alpha CCDs examined here, due to sever buffeting and possible shock reflection effects at the higher combinations of angles of attack and Mach number.

A "confirmation polar" was acquired for each of the 52 configurations tested, consisting of identical AoA settings for each configuration but at a different Mach number, selected at random from the range of Mach 0.35 to Mach 0.80. The AoA set-points for each confirmation polar were acquired in random order. These confirmation points were not combined for analysis with the other points in the test plan but instead were held in reserve to test response predictions made with regression formulas developed from the rest of the data.

To examine the consequences of ignoring randomization restrictions in this real-world configuration aerodynamics test, eight-factor response models was developed for each of the six stability axis forces and moments that were recorded. These (suspect) models were

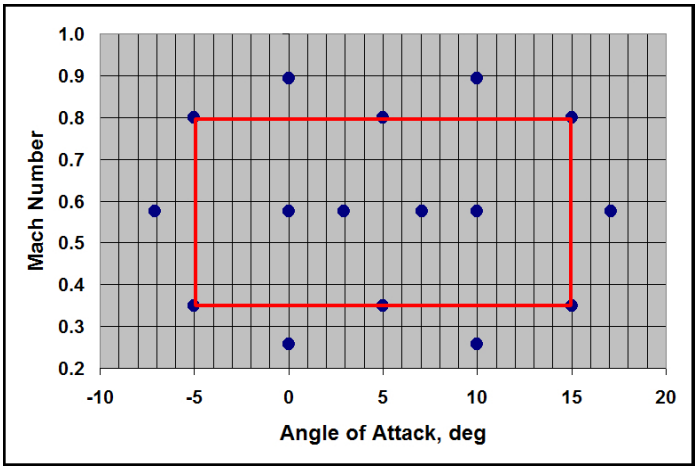

Figure 5. AoA/Mach points set for each of the 52 combinations of configuration variables. Two adjacent central composite designs.

constructed ignoring the restrictions on randomization that were actually in play during the execution of the experiment. That is, the data were analyzed as if the test had been executed as a completely randomized design. A single estimate of unexplained variance was used to test the significance of all regression coefficients rather than rigorously developing separate reference distributions for the hard-to-change and easy-to-change variables to reflect the differences in randomization schemes employed in the test. To the extent that such an analytical simplification is invalid, we would expect poor agreement between these models and the measured confirmation points.

AoA	Measured	2-Factor	8-Factor
0.0	0.0026	0.0043	0.0049
1.0	0.0362	0.0394	0.0397
-1.0	-0.0329	-0.0334	-0.0329
-3.0	-0.1025	-0.1079	-0.1079
3.0	0.1119	0.1147	0.1140
5.0	0.1864	0.1880	0.1856
-5.0	-0.1782	-0.1803	-0.1815

Table 4. Lift coefficient confirmation points for configuration 4501.

For each of the six forces and moments, individual two-factor models were also developed for each of the 52 lift augmentation configurations – a total of 6 x 52 = 312 individual regression models. Because the AoA/Mach run order was completely randomized within each configuration, there were no restrictions on randomization for these models so that a CRD analysis was rigorously justified. We would therefore expect adequate agreement between these models and confirmation points acquired at the configuration for which each model was valid, as they would not have been biased by any failure to properly account for randomization restrictions.

Forces and moments that were directly measured with seven confirmation points acquired with each configuration were compared with force/moment predictions based on the two types of regression models. One was a legitimate two-factor response model that did not feature restrictions on randomization and the other was a suspect eight-factor response model that ignored restrictions on randomization, developed from an analysis that erroneously assumed a completely randomized design.

The anticipated outcome of these comparisons was that the legitimate two-factor response model predictions would agree adequately with measured confirmation points because no faulty assumptions had been made about the randomization schedule, but that the suspect eight-factor models would predict forces and moments poorly. The task would then be to quantify the bias errors introduced by ignoring restrictions on randomization, and look for any patterns that might reveal the conditions under which such errors might be expected to be especially large or possibly small enough to be neglected within the tolerance specifications of the test.

Table 4 presents estimates of lift coefficient for the lift augmentation configuration designated 4501. This configuration consisted of a prescribed subset of six lift augmentation devices in the full "on" position with the remaining devices in the full "off" position. The first column of Table 4 lists the confirmation-point angles of attack in the randomized run order they were set. All of these points were acquired at a single Mach number. They were uniformly intermingled among the 16 unique AoA/Mach combinations (Fig. 5) with 12 replicates acquired in a completely randomized sequence for each configuration for the purpose of developing response models.

The second column of Table 4 lists values of lift coefficient that were measured at each confirmation point. The third lists predicted values based on the two-factor (AoA/Mach) lift model constructed for this configuration without any restrictions on randomization. The last column lists predicted values from the suspect eight-factor model (AoA, Mach, and the six configuration variables) that ignored restrictions on randomization.

The central question is this: Is there any significant difference among the three methods of determining the lift coefficient for these angles of attack? Specifically, we are asking if there are any differences in the data across columns that are too great to attribute to ordinary chance variations in the data.

If the seven confirmation points were simple replicates of the same AoA setting, we would use an ordinary one-way analysis of variance to compare the variation in the data across columns with the row-to-row variation within each column, assumed to represent the intrinsic random error of the data. However, the rows represent distinct AoA set points rather than simple replicates, so the row-wise variation reflects not only random error but also the substantial systematic variance attributable to AoA changes. (The row-wise variance in Table 4 is actually about four orders of magnitude greater than the column-wise variance.) For this reason a two-way ANOVA was performed on the data from Table 4, partitioning the variance into explained components that can be attributable to variations across rows and across columns, and a residual, unexplained variance. Table 5 displays the results of this analysis of variance.

The first column of Table 5 lists sources of variation. The 21 numbers comprising the data sample of Table 4 are not identical, implying that there is some total variance in this collection of numbers. We know that the numbers vary row-wise because lift coefficient is known to change with angle of attack. So "rows" is listed as a component of

Source of Variation	SS	df	MS	F	P-value
Rows	2.82E-01	6	4.70E-02	13211	3.37E-22
Columns	6.15E-07	2	3.08E-07	0.09	0.9176
Error	4.26E-05	12	3.55E-06		
Total	2.82E-01	20			

Table 5. Analysis of variance for lift coefficient confirmation data for configuration 4501.

the total variation. We are unsure if there is any real variation across columns, but this is a candidate source of explainable variation. There is also an entry in the first column called "Error," which represents the difference between the total variance in the data sample and that which can be explained by row-wise and column-wise variations in the data. That is, the "error" component of the total variance represents the "unexplained" variance in the data, presumed attributable to random measurement errors of unknown origin.

The second column lists sums of squares for each component of variation. The total SS is calculated in the usual way, by summing the squared values of all differences between each point and the average of these 21 numbers. The SS for rows is computed by adding the squared differences between each row average and the grand mean of all of the data, and then multiplying by the number of columns to normalize for the difference in the number of rows and columns. Likewise, the SS for columns represents the sum of squared differences between each of the three column means and the grand mean, multiplied by the number of rows. The error sum of squares can be computed by difference, by subtracting the SS for rows plus SS for columns from the total SS.

The third column in the ANOVA table lists degrees of freedom given the mean for each variance component. Since there are 21 data points and one is consumed in estimating the grand mean, there are $n - 1 = 20$ total df. Likewise, there are 6 df corresponding to the seven rows and 2 df corresponding to the 3 columns. The error df are computed by difference as before, subtracting the df for rows plus the df for columns from the total df.

The fourth column in the ANOVA table lists the mean square (MS) or variance for each source. It is simply the ratio of the SS to the df, obtained by dividing column 2 by column 3 in the ANOVA table.

The fifth column, labeled "F," is the ratio of the MS for each source of variation to the error MS. This *F*-statistic is an indicator of signal to noise, measuring the variance of each source relative to the intrinsic unexplained variance in the system. Note that the row-wise variation is 13,311 times greater than the variation that can be attributable to ordinary random error in the data. There is a vanishingly small probability that so much variation could exist from row to row just by chance if there really were no systematic changes occurring, and this probability is listed in the final column of the ANOVA table as the *P* value. The very high *F* value (or equivalently, the miniscule *P* value) reflects the fact that the lift coefficient is indeed very likely to change with angle of attack, an inference that will not exactly startle seasoned aerodynamicists. Of considerably greater interest because it is not known in advance, is whether or not there is significant variation from column to column. The ANOVA table reveals this in the same way, through the magnitude of the *F* and *P* values for columns.

For the data of Table 4, the column-wise variance is quite small – less than 10% of the variance attributed to ordinary chance variations in the data ($F = 0.09$). The *P* value is corresponding large – 0.9176 – implying greater than a 90% probability that this small amount of variation could occur from column to column just due to random fluctuations in the data, even if there was no true column to column difference in the data.

This finding is quite unanticipated. It says that for configuration 4501, there was no significant difference between estimating lift for these angles of attack by direct measurement, or by *either* of the two response models! The fact that the two-factor response model agreed well with the confirmation data is not surprising, since the two-factor model was

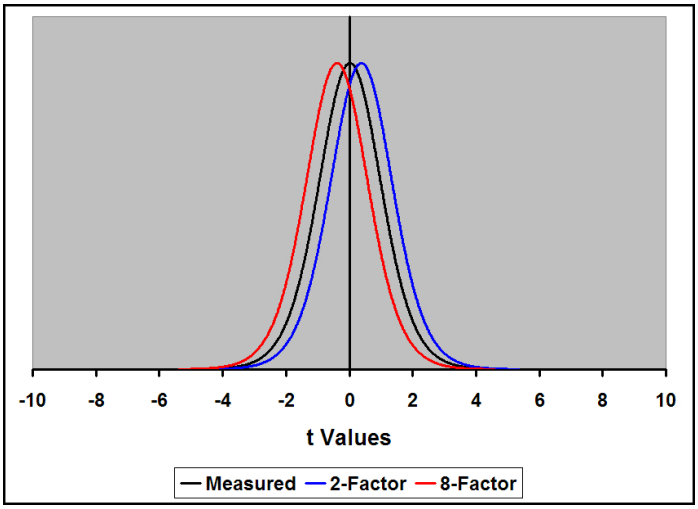

Figure 6. t-Distributions for column means, ANOVA for configuration 4501 lift coefficient comparisons.

based on a completely randomized schedule of AoA/Mach set-points acquired within a fixed configuration. No restrictions on randomization were in play so none were ignored. However, the eight-factor model was developed as if the experiment had been executed as a completely randomized design, completely ignoring the restrictions on randomization that were dictated by the hard-to-change nature of the configuration variables. A single reference distribution was used to test the significance of regression coefficients for both the hard-to-change variables and the easy-to-change variables, notwithstanding the fact that differences in randomization schemes should have necessitated different reference distributions for these two types of variables.

AoA	Measured	2-Factor	8-Factor
5.0	0.1801	0.1754	0.1766
1.0	0.0373	0.0352	0.0385
-1.0	-0.0320	-0.0360	-0.0320
-3.0	-0.1009	-0.1088	-0.1045
0.0	0.0040	0.0003	0.0039
3.0	0.1085	0.1052	0.1076
-5.0	-0.1806	-0.1821	-0.1779

Table 6. Lift coefficient confirmation points for configuration 4502.

Figure 6 displays 12 df *t*-distributions of sample means (reflecting the error term from ANOVA Table 5), centered on the normalized column means for the data in Table 4. This is a graphical representation of the analysis of variance which illustrates clearly why, given the dispersion in the unexplained variance, it is not possible to resolve with high confidence any real differences in the means of these distributions.

A similar analysis of variance was performed to compare how predicted lift coefficients compared with measured lift confirmation points for all of the 52 configurations. Table 6 displays the confirmation data for configuration 4502, for example, and Table 7 presents the corresponding analysis of variance.

The ANOVA for configuration 4502 differs from the configuration 4501 ANOVA in one important way. The significance of the row factor (AoA) is as unambiguous (and as uninteresting) for configuration 4502 as it was for configuration 4501 – astronomical *F* and miniscule *P* leave little doubt that changes in AoA cause changes in lift. But unlike configuration 4501, the column factor for configuration 4502 is also significant. Note that the *P*-value for columns is very small for configuration 4502. This means that the probability is very small that a column-wise variation as large as the one that was observed could be attributed to chance variations in the data if there was no real cross-column effect in play.

The inference we draw from the significance of the column-wise variation is that at least one of the three columns in Table 6 differs from the other two, and possibly all three differ from each other. The former case is in fact the result we have been anticipating. We expect similar results between the measured confirmation points and lift predictions made with a reliable two-factor model that does not ignore restrictions on randomization. We also expect the eight-factor model predictions to be biased due to model terms that were either erroneously retained or erroneously rejected during the model building process, because the analysis was based on an improper assumption of a completely randomized design. That is, we expect the eight-factor model to be "the odd man out."

Again the actual result was unanticipated. Figure 7 displays 12 df *t*-distributions of sample means (reflecting the error df from ANOVA Table 7), centered on the normalized column means for the configuration 4502 data in Table 6. This figure illustrates that the degree of dispersion in the data makes it difficult to resolve a significant difference between the eight-factor model predictions and the measured confirmation points, but even given that dispersion, it is easy to resolve a difference between the two-factor model predictions and either the measured results or the (suspect) eight-factor predictions.

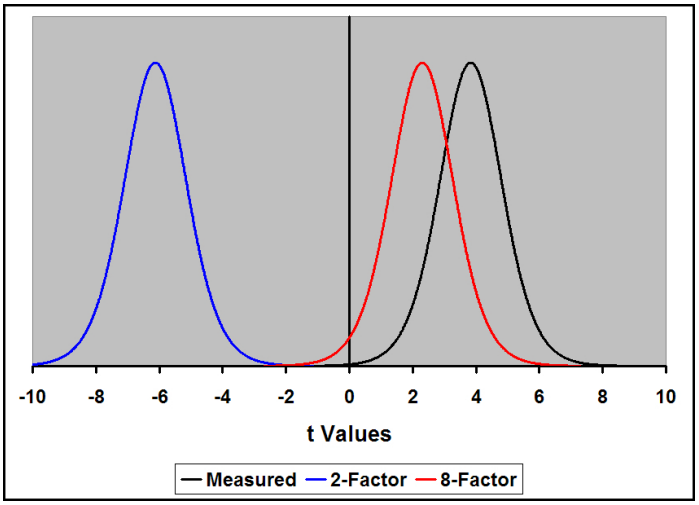

Figure 7. t-Distributions for column means, ANOVA for configuration 4502 lift coefficient comparisons.

The ANOVA for configuration 4501 implied that the suspect eight-factor model that failed to account for restrictions on randomization did just well at predicting lift as the rigorously valid two-factor model for which there were no restrictions on randomization. The ANOVA for configuration 4502 suggested that the suspect model did an even *better* job of predicting lift than the ostensibly reliable two-factor model, a surprising result indeed.

To investigate this further, we framed the confirmation analysis as a Bernoulli process in which each two-way ANOVA represented a "trial." ANOVA results with *P*-values for columns greater than 0.05 were defined as "successes." That is, we defined as "successes" those outcomes for which the column variation was so small that there was more than a 5% chance that it was simply due to chance variations in the data. In such cases it could not be said (with at least 95% confidence) that there was any true difference between estimating lift by a two-factor model, by an eight-factor model, or by direct measurement.

Source of Variation	SS	df	MS	F	P-value
Rows	2.66E-01	6	4.44E-02	24491	8.30E-24
Columns	6.08E-05	2	3.04E-05	16.8	3.35E-04
Error	2.18E-05	12	1.81E-06		
Total	2.67E-01	20			

Table 7. Analysis of variance for lift coefficient confirmation data for configuration 4502.

ANOVA results with *P*-values for columns less than 0.05 were defined as "failed" Bernoulli trials. For these failed trials, there was less than a 5% chance that column variation could simply be due chance, and thus a true column difference could be inferred with at least 95% confidence. That is, a failed trial was one in which either of the models (or both of them!) failed to predict the confirmation data within experimental error, or that a significant difference could be detected between the two-factor and the eight-factor model predictions.

There were a total of 52 trials, one for each of the confirmation polars acquired at each configuration. A crucial question that must be answered before this analysis can proceed is as follows: How many successes would we have to see in 52 trials to be convinced (at some prescribed level of confidence) that there is, say, a 95% probability of success in any one trial? Note that the answer is not anything so simple as "0.95 x 52," as this calculation would only the provide the *most likely* outcome under the prescribed assumption of a 95% success probability for each trial. This is rather analogous to asking how many times in a hundred trials that a tossed coin would have to come up "heads" to convince us that the coin is fair. The *most likely* outcome for a fair coin is 50 heads in 100 trials, but surely we would not assert that the coin is weighted if heads came up 49 times in an actual 100-toss examination. On the other hand, if heads came up only one time in 100 tosses, this would be interpreted as strong evidence that the coin was weighted. Somewhere between these extremes is a minimum number of heads that we would expect to see if the coin is fair. This is known as the Critical Binomial Number, CBN, which is available in standard statistical tables, or via the CRITBINOM worksheet function in the Excel spreadsheet, for example. (For 99% confidence, the CBN for the coin toss problem is 38 heads, incidentally, and by symmetry there is an upper limit of 62 heads, corresponding to 38 tails. That is, a fair coin would be expected to produce between 38 and 62 heads for 99% of the 100-toss tests that were administered to it.)

For our 52-trial test, the CBN for 52 trials with a per-trial success probability of 0.95 is 45 at the 0.01 significance level. That is, if the probability of success for any one trial is at least 95%, we would expect to see 45 or more successes in 52 trials at least 99% of the time. Note that we explicitly do not require a 100% success rate in a 52-trial test, because the assumed probability of success in any one trial is 95%, not 100%. The assumed 95% success rate actually implies an average of one failure in every 20 trials. We can accept up to $52 - 45 = 7$ failures in 52 trials before we declare that the per-trial success probability must have been less than 95%.

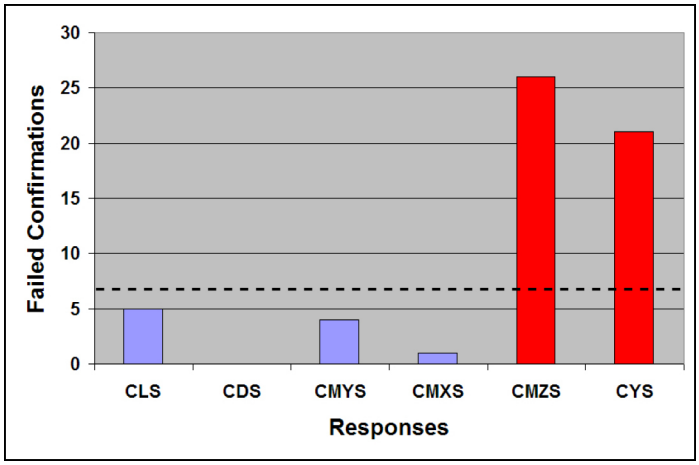

Figure 8. Critical Binomial Analysis applied to ANOVA tests of null hypothesis that both models agree with confirmation data. Seven or more failures in 52 trials (dashed line) are required to reject this hypothesis.

For each of the six stability axis forces and moments, 52 analyses of variance were performed as described above. Figure 8 compares the number of failures with the seven-failure criterion for rejecting the null hypothesis that both models produce the same result as a direct measurement. This figure reveals that on average, both models agree with direct measurement within experimental error 95% of the time for four of the six

stability axis responses. That is, for lift, drag, pitching moment, and rolling moment, the differences between the three column means in the ANOVA that corresponded to the two regression models (two-factor and eight-factor) and the direct measurement of confirmation points, were all small relative to the uncertainty in estimating the means. This is the situation illustrated by Fig. 6. However, there were significant differences for yawing moment and side force.

The yawing moment and side force discrepancies are attributed to poor signal-to-noise ratio, as indicated in Fig. 9. This figure displays the maximum dynamic range of the 52 confirmation data sets experienced for each force and moment, expressed as a multiple of the precision tolerance specified during the design of the experiment. Figure 10 displays the precision tolerances for each of the forces and moments. The precision tolerance specified for side force was ±50 counts, for example, but the maximum dynamic range encountered in the confirmation runs was only 37 counts! Likewise, the maximum yawing moment dynamic range was just over twice the minimum specified precision tolerance, which is not really adequate to develop a reliable regression model.

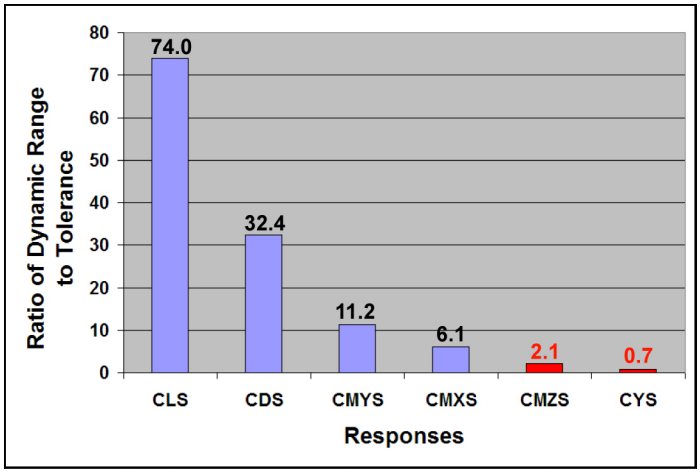

Figure 9. Dynamic range of confirmation points as multiple of error tolerance levels. Very little signal-to-noise available for yaw and side force.

The poor signal-to-noise ratio for the lateral/stability responses is attributed to the fact that for this portion of the test, there were no sideslip angles set. The configuration variables on the port-side wing generated some small differential lift that resulted a slight rolling moment and produced some second-order coupling into yaw and side force. But the primary effect of these lift augmentation devices at zero sideslip was on the axial responses.

The comparison of differences between column means that the ANOVA process facilitates, as illustrated in Figs. 6 and 7, can be quite instructive. There were a total of five confirmation data sets out of 52 that revealed significant differences among the lift estimates made by direct measurement or by the two response models. While this was fewer than the seven failures that would have been necessary to reject the hypothesis that agreement occurred 95% of the time, the fact that there were such cases provides an opportunity for insights into what may have caused them.

Figure 7 represents one of the five configurations where poor agreement was observed for lift. Figure 11 presents the other four cases. Considering Figs. 7 and 11 together, we see that in two of the five cases of poor agreement – configurations 4517 and 4521 – the models agreed with each other but not with the confirmation data. This suggests a fitting error of some kind that was common to both configurations, or possibly one or more bad data points in the two confirmation data sets.

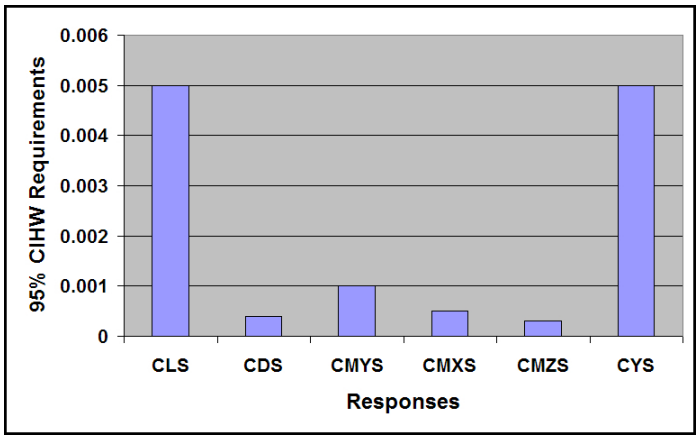

Figure 10. Tolerance requirements defining adequacy of response models and serving as a key test exit criteria.

The other three cases of poor agreement – configurations 4502, 4507, and 4546 – all involved situations in which the two-factor model was "voted out" by the good agreement between the measured data and the results of the eight-factor predictions. Conspicuous by its absence is any case in which the suspect eight-factor model for lift was the "odd man out." The only time the eight-factor model predictions failed to agree with the data was when the data differed from both the assumed-reliable two-factor model as well as the suspect eight-factor model.

A failure to properly account for randomization occurred as a result of the eight-factor models being analyzed as if their data had been generated in a completely randomized design, when in fact there were significant

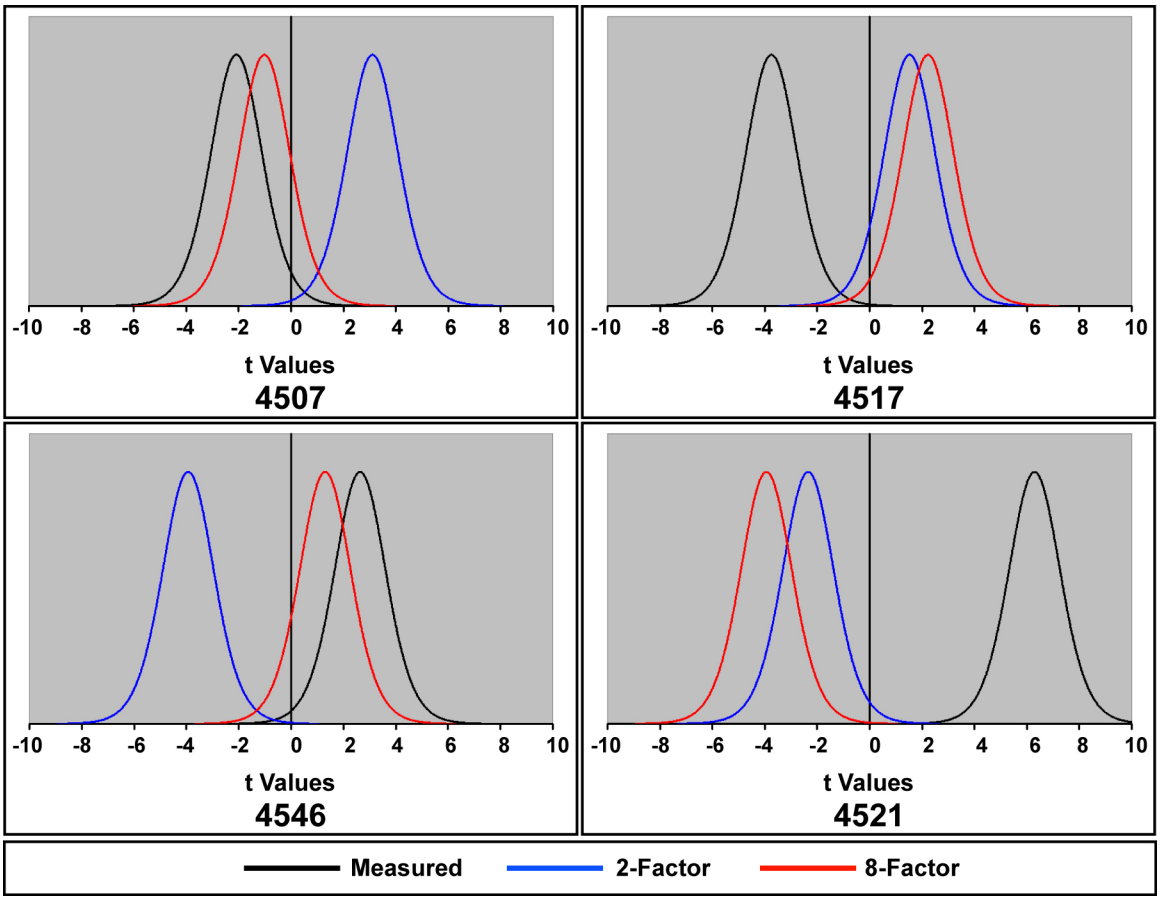

Figure 11. t-Distributions for column means, ANOVA for configurations with significant differences for lift coefficient.

restrictions on randomization. However, no evidence can be seen in Figs. 7 and 11 that this distinguished the eight-factor models from the two-factor models that were legitimately analyzed as CRDs, or from the direct measurements of lift. For the four forces and moments for which there was adequate dynamic range to construct reliable models, the critical binomial analysis of the ANOVA results suggest that ignoring restrictions on randomization in this experiment did not result in any detectable penalty in the predictive capability of the response models.

We established early in the paper that persisting systematic variations during the execution of an experiment generate correlated experimental errors, and that this correlation results in a bias in variance estimates that could either reduce or inflate the dispersion of the reference distributions used to test the significance of candidate regression coefficients in a proposed response surface model. (We also established that systematic errors introduce bias into the estimates of sample *means* as well as variances, so the reference distribution could be biased in location as well as dispersion.)

We described how randomizing the set point order restores the random sampling hypothesis by disrupting the correlation in experimental errors that is responsible for much of this mischief, but that hard-to-change variables introduce certain restrictions on randomization that would presumably result in a biased estimate of residual variance if not properly taken into account. This suggests that a failure to account for restrictions on randomization would result in erroneous decisions to retain or reject certain terms in the regression models. This should then bias those models in such a way that they would perform relatively poorly as predictors of independent confirmation data points, compared to models developed under circumstances in which no restrictions on randomization were in play. We can test this directly for each response variable by comparing the residual variance of the confirmation points for both the two-factor and eight-factor models. If our failure to account for restrictions on randomization biased the eight-factor models, their residuals should be different than the two-factor models which did not entail any randomization restrictions.

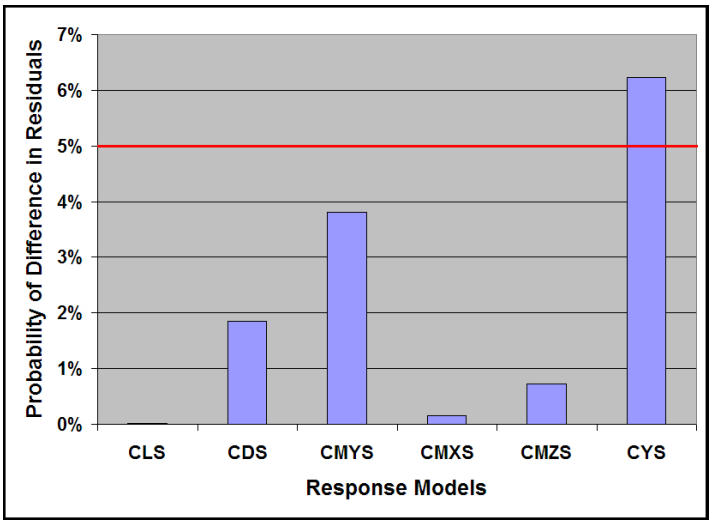

Figure 12. Result of F test comparing variance in confirmation-point residuals for two-factor and eight-factor response models. No significant difference (0.05 level) except for side force.

A standard *F* test was employed to test the difference in confirmation point residuals between the two-factor and eight-factor models. This test is largely redundant to the ANOVA tests previously described, but it has the advantage of generating a specific probability that the two models produced residuals of the same magnitude. The complement of this probability describes how likely it is that ignoring the restrictions on randomization in a test like this one would result in any significant bias in the model predictions. These latter probabilities are presented in Fig. 12, which show that except for side force (with its extraordinarily low signal-to-noise ratio), the probability that ignoring randomization restrictions would bias the model predictions is comfortably below the 0.05 level necessary for us to assert with at least 95% confidence that restrictions on randomization – at least for this test – can be safely ignored.

The analyses described above all indicate that the *accuracy* of the two models – defined in terms of their ability to estimate measured responses – is equivalent. Figure 13 reveals that the *precision* of the eight-factor models that ignored restrictions on randomization is actually superior to that of the two-factor models. Equation 16 reveals why this is so. Even though the eight-factor models featured more terms than the two-factor models by almost a factor five on average (Fig. 14), the number of data points available to assess the eight-factor regression coefficients was larger by a factor of 52 than for the two-factor models. The *p/n* ratio in Eq. 16 was therefore roughly an order of magnitude smaller for the eight-factor models that for the two-factor models. The widths of the precision intervals are proportional to the square root of the prediction variance, which explains the approximate factor-of-three improvement in precision revealed in Fig. 13.

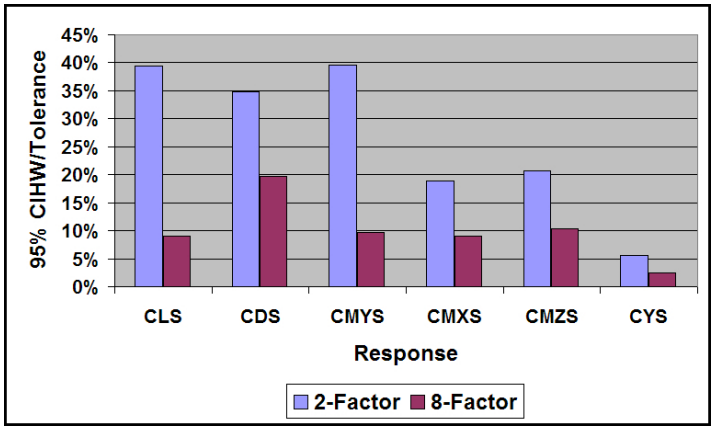

Figure 13. Comparisons of prediction precision for two-factor and eight-factor models. All models well within tolerance, but eight-factor models featured higher precision than two-factor models.

VII. Discussion

The importance of randomizing run order as a quality assurance tactic to secure the random sampling hypothesis is not in question. Nor has it been the intent of this paper to suggest that restrictions on randomization are generally irrelevant. The necessity for split plot designs and the attendant complexity in analyzing them has been recognized since the first applications of designed experiments to agricultural experiments almost a hundred years ago. However, the results of this paper do suggest that at least under certain circumstances, the benefits to be derived by a rigorous split plot analysis of an experiment executed with restrictions on randomization do not justify the considerable added complexity of such an analysis. It is interesting, therefore, to contemplate what conditions require such rigor, and under what other circumstances it might be reasonable to simplify the analysis by ignoring

the randomization restrictions and proceeding as if the experiment had been conducted as a completely randomized design.

Several features of this experiment may help to explain why the restrictions on randomization had such a negligible practical effect. One may have been the relatively large number of whole plot blocks.

Restrictions on randomization result in a confounding of whole plot variable effects with block effects unrelated to the independent variables. In the PSP experiment used to introduce the basics of split plot designs, for example, replication was necessary to introduce enough error df to properly assess the inherent variance in the experimental results. In an unreplicated experiment, the difference between the drag measured with a painted model and the drag measured with a clean model would have been attributable to two sources, the effect of the paint, and the effect of any other change to the measurement system, data system, apparatus, or facility that may have occurred between the time of the first measurement and the time of the second. This latter source is what is meant by the term, "block effect," and refers in this case to changes of unknown origin that result in a systematic difference in response measurements from one block of time to the next.

Replicating the PSP experiment (and randomizing the order that the pained and clean configurations were run) forced cancellation among much of the block effects. In general, the impact of restricted randomization is greatest when the opportunity for block effects to cancel is smallest. In the current experiment, there were 52 blocks in which whole plot variable levels were set in random order. The fact that configuration variables were randomly assigned to such a large number of whole plot blocks may have been responsible for diminishing the confounding of blocks effects with whole plot variable effects.

The general stability of the facility may have also contributed to the negligible effect of ignoring the split plot structure of this experiment. Absent significant temporal block effects, there is a reduced need to randomize altogether. Under such circumstances, restrictions on randomization become even less relevant.

Another possible reason that restrictions on randomization had such a negligible effect in this experiment may have been the fact that the response variables (forces and moments) were dominated by the subplot variable effects – particularly angle of attack – and were influenced only slightly by the whole plot (configuration) variables. The regression coefficient for the first order AoA term for lift was more than three orders of magnitude greater than the largest configuration variable coefficient, for example.

In the limiting case in which the whole plot variables could be imagined to have exerted no influence whatsoever, this experiment would reduce to a simple randomized complete block design (RCBD). In such a design, blocking on replicates would improve the precision of the result by removing block effects from the unexplained variance, but no other special analysis would be needed to account for restrictions on randomization. For practical purposes, because of the enormous difference in the relative influence of the whole plot and subplot factors in this experiment, it was much more in the nature of a randomized complete block design than a split plot design.

Related to the relatively subtle effect of the whole plot variables in this experiment is the fact that a failure to properly account for restrictions on randomization is expected to bias both the location and the dispersion of reference distributions used to test the significance of candidate regression coefficients. Coefficients that are located far enough away from zero to be resolved when the reference distributions are properly constructed may be rejected in the response model building process, and likewise insignificant coefficients may be erroneously retained. However, the null hypotheses for regression coefficients that are associated with dominant factors in a high-precision experiment are very likely be properly rejected even if their corresponding reference distributions are slightly broadened, narrowed, or shifted. For example, it would be hard to imagine any practical situation in which the coefficient of the first-order AoA term in a lift response function would be erroneously regarded as insignificant simply because a subtle restriction on randomization was not rigorously taken into account.

Another potential contributor to the apparent irrelevance of restrictions on randomization in this experiment may have been the high potential parameter count resulting from the combination of model order and number of factors. A full third-order model in eight independent variables would feature 165 possible parameters, the coefficients of most of which would be located too close to zero to include in a reduced model. If a failure to properly account for restrictions on randomization resulted in slight errors in the variance or the mean of the reference distributions for these coefficients, some would be erroneously rejected while others would be erroneously retained. The small number of large coefficients driving the response will not be affected by such subtleties, but it is possible in such circumstances that if a large number of small coefficients were improperly identified as significant or insignificant, there could be a certain canceling of the effects of such inference errors. This may have occurred in this test, in which only about 10% to 25% of all the possible coefficients were judged to be significant, depending on the response variable being modeled. A model in fewer variables may have been more problematic. (A full third order

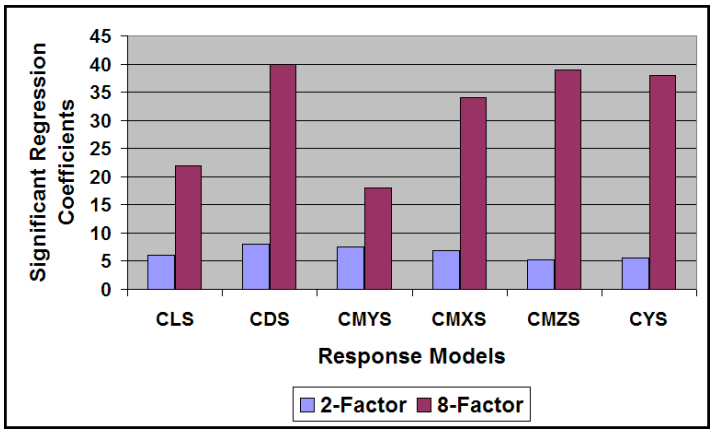

Figure 14. Complexity of two-factor and eight-factor response surface models revealed by number of significant terms in each model.

model in only two variables has a maximum of only 10 parameters, for example, compared to 165 if there were eight factors as in this experiment.)

Yet another possible explanation for the fact that restrictions on randomization seemed so unnecessary in this test may have been the high precision of the measurement environment. In an extremely precise measurement environment such as the one in which this test was conducted, the reference distributions will be very narrow. A given percentage error in characterizing their dispersion or location will translate into a relatively small absolute effect that will impact only a relative few coefficients, which themselves would have to be relatively small to be adversely impacted. Related to this is the possibility that split-plot analytical corrections may be less important when error tolerance levels are generous compared to the random and systematic components of unexplained variance, as was the case in this experiment.

Another aspect of this experiment is the fact that eight-factor models turned out to have so many residual lack of fit degrees of freedom compared to the two-factor models. While Fig. 14 indicates that there were on the order of five times as many significant terms in the eight-factor models as in the two-factor models, Fig. 15 shows that the two-factor models required a much greater percentage of the available terms than the eight-factor models, and in that sense had to "work harder" to achieve a good fit to the data. The eight-factor models may have performed enough better than the two-factor models on this account to compensate at least partially for small prediction biases that could have been introduced by erroneously rejecting or retaining small regression coefficients because of a failure to properly account for restrictions on randomization. Thus, it is possible that the failure to randomize actually had a somewhat great effect that these analyses revealed.

A number of factors have been discussed which may explain why a failure to rigorously account for restrictions on randomization apparently had such a small impact in this particular experiment. No representation is being made that randomization restrictions are generally unimportant, or that configuration aerodynamicists can be forever liberated from the responsibility for taking such restrictions into account. On the other hand, many of the features of this experiment which may have contributed to the reduced need to account for randomization restrictions are common in configuration aerodynamics, including high precision in the measurement environment, large numbers of independent variables, and the potential to randomize whole plot variables across many blocks. There is a tradeoff between quality and productivity inherent in split-plot designs, and ultimately the decision for how rigorously to account for restrictions on randomization must be based on a combination of experience and judgment.

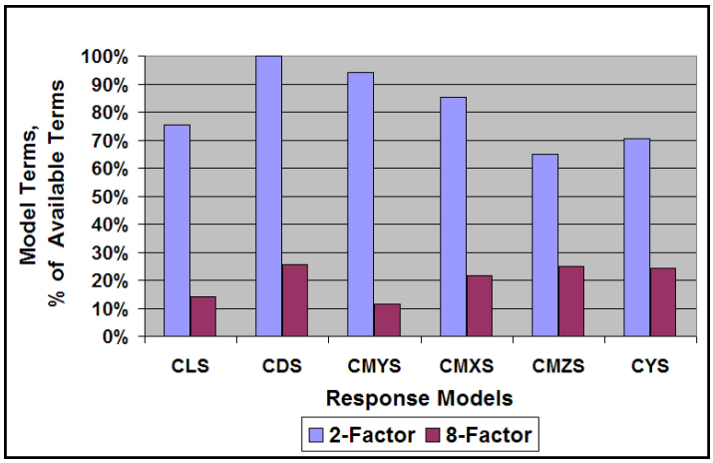

Figure 15. Percentage of total available terms (8 for two-factor models and 157 for eight-factor models) required for best fit of data.

VIII. Concluding Remarks

Bias errors are introduced into estimates of sample means and variances by slowly varying covariate effects that impact the intrinsic quality of experimental results as well as the ability to accurately assess quality. In an MDOE experiment, decisions to retain or reject individual terms in a regression response model depend on the magnitude of the corresponding coefficient relative to the variance in estimating it, both of which are adversely impacted by covariate effects.

Failure to properly identify significant terms in a response model adversely impacts the accuracy of system response predictions, and incorrectly portrays the underlying physics. Randomizing the set point order defends against these effects but restrictions on randomization caused by hard-to-change variables complicate such quality assurance tactics. Certain features of experimentation under restrictions on randomization have been considered in this paper. The principal conclusions are summarized as follows:

- Persisting systematic variations that occur during the execution of an experiment generate correlated experimental errors by upsetting the random sampling hypothesis stating that all measurement errors are independent.

- Correlated experimental errors prevent sample statistics such as means and variances from serving as unbiased estimators of the corresponding population parameters.

- Reference probability distributions centered on zero with a variance reflecting experimental uncertainty are used to objectively assess the significance of regression coefficients in the construction of regression equations used to fit experimental data to a mathematical response model.

- Biases in the location and dispersion of reference distributions can result in model terms that are either erroneously retained or erroneously rejected in the model building process, producing a bias in response predictions made with the resulting model.

- Randomizing the run order of a test matrix is an effective quality assurance tactic that restores the random sampling hypothesis by disrupting the correlation in experimental errors that occurs when there is a significant component of systematic unexplained variance is in the data.

- Hard-to-change variables commonly encountered in configuration aerodynamics introduce certain restrictions on randomization in that cause them to have to be randomized according to a different schedule that easy-to-change variables in a structure referred to as a split plot design.

- In a configuration aerodynamics test, blocks of time in which easy-to-change variables are completely randomized while hard-to-change configuration variables are held constant are called whole plots, and smaller blocks of time within each whole plot in which easy-to-change variables are set are called subplots. These terms reflect the agricultural heritage of the method.

- The two randomization schemes of a split plot design result in two types of unexplained variances and thus two reference distributions for testing the significance of whole plot variable coefficients on the one hand, and coefficients for subplot variables and interactions between subplot and whole plot variables on the other.

- The resulting analysis of a split plot design is very complicated and there is no unanimous agreement even among experts on how to proceed with some of the finer points of such an analysis, including how to quantify the variance associated with subplot main effects and subplot/whole plot interaction effects in the absence of block-factor interactions. Variance covariance matrices are also significantly complicated by split plot designs, making the estimation of prediction variances and associated precision intervals extremely complex.

- The complexity of a split plot analysis can be avoided by treating the results as if they had been acquired in a completely randomized design (CRD).

- The great simplicity that would be introduced by eliminating the complexity of a split plot analysis of data acquired under restrictions on randomization is presumed to have associated with it some significant cost in the form of reduced validity that this study sought to characterize.

- A specific configuration aerodynamics test executed at Langley Research Center as a split plot design with the usual restrictions on randomization was analyzed using methods that respected these restrictions, and methods that ignored them, yielding the following unanticipated results:

    - No significant differences were detected between the prediction accuracy of response models that respected restrictions on randomization and those that ignored them.

    - The precision of the response models that ignored restrictions on randomization was actually higher than the precision of the response models that respected the restrictions, due to the significantly greater volume of data that could be utilized in each analysis when randomization restrictions were ignored.

- The fact that no significant cost seems to have been attached in this test to the substantial simplicity in analysis afforded by ignoring the restrictions on randomization is attributed to a number of possible factors that were in play, including the following:

    - The relatively large number of whole plot blocks in this experiment that were executed in random set-point order, which provided ample opportunity for whole plot block effects to cancel.

    - The general stability of the facility that for this test may have minimized the general need for quality assurance tactics such as randomization and the further need to properly account for restrictions on randomization.

    - The dominant role of the subplot effects relative to the whole plot effects in this test, which may have resulted in a close approximation to a randomized complete block design in which no special analysis is needed to account for restrictions on randomization.

    - The existence of numerous dominant terms in the response models whose coefficients would be unambiguously resolved from zero even if a failure to account for restrictions on randomization introduced some bias in the dispersion and location of the reference distributions.

    - The high potential parameter counts in the response functions of this experiment that were due to large combinations of model order and number of factors, that may have resulted in the cancellation of potential bias errors attributable to erroneously rejecting or retaining large numbers of small candidate terms in the regression model because of a failure to account for restrictions on randomization.

    - The high precision of the measurement environment that would have had associated with it very narrow reference distributions for which a bias in either the location or dispersion induced by a failure to account for restrictions on randomization would have resulted in acceptance/rejection decision errors for a relatively small number of low-magnitude regression coefficients.

    - The relatively generous error tolerance levels compared to the random and systematic components of unexplained variance, which may have rendered split-plot analytical corrections less important than if extraordinarily tight error tolerances had been specified.

    - The relatively large number of available lack of fit degrees of freedom that may have resulted in improvements in the fit of the response models that partially compensated for some amount of bias error caused by ignoring restrictions on randomization.

- No representation is made that randomization restrictions are generally unimportant, but many of the features which may have contributed to the reduced need to account for randomization restrictions in this

experiment are common in configuration aerodynamics, including high precision in the measurement environment, large numbers of independent variables, and the potential to randomize whole plot variables across many blocks.

- There is a tradeoff between quality and productivity inherent in split-plot designs, and ultimately the decision for how rigorously to account for restrictions on randomization must be based on a combination of experience and judgment.

IX. Acknowledgements

The collaboration of Mr. Bobby Berrier of the Configuration Aerodynamics Branch at Langley Research Center is gratefully acknowledged, as well as the performance of the test engineers and technicians at Langley's 16 Ft Transonic Tunnel. This work was supported by the Langley Wind Tunnel Enterprise.

X. References

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

[2]Hemsch, M.J. "Development and Status of Data Quality Assurance Program at NASA Langley Research Center – Toward National Standards". AIAA 96-2214. 19th AIAA Advanced Measurement and Ground Testing Technology Conference, June 1996.

[3]Hemsch, M., et al. "Langley Wind Tunnel Data Quality Assurance – Check Standard Results (Invited)". AIAA 2000-2201. 21st AIAA Advanced Measurement and Ground Testing Technology Conference, Denver, CO. 19-22 June 2000.

[4]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.

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

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

[7]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.

[8]Kegelman, J.T. "Recent Cycle Time Reduction at Langley Research Center" AIAA 99-0178, 37th AIAA Aerospace Sciences Meeting and Exhibit. Reno, NV, Jan 1999.

[9]DeLoach, R., Cler, D., Graham, B. "Fractional Factorial Experiment Designs to Minimize Configuration Changes in Wind Tunnel Testing" AIAA 2002-0746. 40th AIAA Aerospace Sciences Meeting & Exhibit. Reno, NV. January 14-17, 2002

[10]Hicks, C.R. *Fundamental Concepts in the Design of Experiments,* Third Edition, CBC College Publishing. 1982.

[11]Myers, R.H. and Montgomery, D.C. (1995). *Response Surface Methodology: Process and Product Optimization Using Designed Experiments*. New York: John Wiley & Sons.

[12]Box, G. E. P., Hunter, W. G., and Hunter, J. S. (1978). *Statistics for Experimenters. An Introduction to Design, Data Analysis, and Model Building.* New York: Wiley.

[13]Coleman, H. W. and Steele, W. G. (1989). *Experimentation and Uncertainty Analysis for Engineers*. New York: Wiley.

[14]Bevington, P.R. and Robinson, D.K. (1992, 2nd Ed). *Data Reduction and Error Analysis for the Physical Sciences.* New York: McGraw-Hill.

[15]DeLoach, R. and Erickson, G.E. "Low-Order Response Surface Modeling of Wind Tunnel Data Over Truncated Inference Subspaces" AIAA AIAA-2003-456. 41st AIAA Aerospace Sciences Meeting & Exhibit. January 6-9, 2003

[16]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, Oregon, June 28-Jul 1, 2004

[17]Box, G.E.P. and Draper, N.R. (1987). *Empirical Model Building and Response Surfaces*. New York: John Wiley and Sons.

[18]Fisher, R. A. (1966). *The Design of Experiments, 8th ed.* Edinburgh: Oliver and Boyd.

[19]DeLoach, R. "MDOE Perspectives on Wind Tunnel Testing Objectives" AIAA 2002-2796 22nd AIAA Aerodynamic Measurement Technology and Ground Testing Conference. St. Louis, MO. Jun 24-26, 2002.

[20]DeLoach, R. "The Modern Design of Experiments: A Technical and Marketing Framework (invited)" AIAA 2000-2691. 21st AIAA Aerodynamic Measurement Technology and Ground Testing Conference. Denver, CO. Jun 19-22, 2000.

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