The Impact of Truth Surrogate Variance on Quality Assessment/Assurance in Wind Tunnel Testing

Richard DeLoach[1]
NASA Langley Research Center, Hampton, VA, 23681

Minimum data volume requirements for wind tunnel testing are reviewed and shown to depend on error tolerance, response model complexity, random error variance in the measurement environment, and maximum acceptable levels of inference error risk. Distinctions are made between such related concepts as quality assurance and quality assessment in response surface modeling, as well as between precision and accuracy. Earlier research on the scaling of wind tunnel tests is extended to account for variance in the truth surrogates used at confirmation sites in the design space to validate proposed response models. A model adequacy metric is presented that represents the fraction of the design space within which model predictions can be expected to satisfy prescribed quality specifications. The impact of inference error on the assessment of response model residuals is reviewed. The number of sites where reasonably well-fitted response models actually predict inadequately is shown to be considerably less than the number of sites where residuals are out of tolerance. The significance of such inference error effects on common response model assessment strategies is examined.

Nomenclature

Accuracy	Measure of how well a response estimate approximates the true response
Alternative hypothesis	Assertion that there is a significant bias between a response estimate and the true response
CBN	Critical Binomial Number
Confidence Interval	Response range within which the true response is expected to lie with a prescribed probability
Degrees of freedom, df	The number of independent observations in a sample
Design space	A coordinate system with each axis associated with an independent variable
Factor	An independent variable, such as angle of attack or Mach number
Inference error	Erroneous rejection of null hypothesis (Type I) or alternative hypothesis (Type II)
Lack of Fit	A systematic bias error due to inadequate response model formulation
Level	Setting of a given factor
MDOE	Modern Design of Experiments
Null hypothesis	Assertion that there is no significant difference between a response estimate and the true response
OFAT	One Factor At a Time
PDF	Probability Density Function
Precision	Measure of how well response estimates approximate the mean of a sample of genuine replicates
Precision interval	Response range within which the mean of a distribution of response estimates is expected to lie with a prescribed probability
Prediction interval	Response range within which an individual response estimate is expected to lie with a prescribed probability
Replication	Repetition of an experimental condition so that the variability associated with the phenomenon can be properly estimated
Residual df	Residual degrees of freedom. The total number of independent observations in a sample of data minus the number of parameters that must be estimated from the sample.
RSM	Response Surface Model(ing)

[1] Senior Research Scientist, Engineering Directorate, MS 238, AIAA Associate Fellow

Significance	The probability of committing a Type I inference error by erroneously rejecting the null hypothesis. Denoted by α
site	A point in a design space. Represents a specific combination of factor levels
Truth surrogate	Unbiased estimate of true response
G_A	Accuracy Gain Factor. Ratio of λ_0 to λ
G_p	Precision Gain Factor. Ratio of γ_0 to γ
N	Number of fitted points in a response surface modeling experiment
N_0	Minimum number of fitted points required to assure quality specifications
m	Number of truth surrogate replicates at a given RSM validation site
m_o	Optimum number of truth surrogate replicates. Corresponds to smallest volume of data required to meet quality specifications.
out of tolerance	Differing by more than a prescribed amount from a specified reference such as an independent estimate of the true response
p	Number of terms in a response surface model, including the intercept
p_s	Per-trial probability of success in a series of Bernoulli trials
p_f	Per-trial probability of failure in a series of Bernoulli trials: $1 - p_s$
P_f	Percentage of validation residuals that are out of a specified tolerance
P_s	Percentage of validation residuals that are within a specified tolerance
S	Number of validation sites in a response surface modeling experiment
T_0	Smallest total number of fitted and validation points required to assure quality specifications
α	Maximum acceptable Type I inference error probability
β	Maximum acceptable Type II inference error probability
γ	Specified precision interval half-width
γ_0	Smallest precision interval half-width that can be achieved without residual degrees of freedom
λ	Smallest acceptable response model lack of fit bias error
λ_0	Smallest lack of fit bias error for which the Type II inference error probability would be no greater than β without residual degrees of freedom
σ	Standard deviation of a normal distribution
σ_0	Standard random error of the measurement environment

I. Introduction

CONVENTIONAL wind tunnel tests are generally conducted as high-volume data acquisition activities in which the objective is to acquire good data and lots of it. The question of how much data to acquire is seldom asked, as the answer is implicitly understood to be, "as much as possible." Unfortunately, resources—money and time—are required to obtain each new data point, so any strategy for maximizing the volume of acquired data is also a strategy for maximizing test costs.

An alternative approach to wind tunnel testing was introduced at NASA Langley Research Center in the late 1990s under the name of the Modern Design of Experiments[1-11] (MDOE). This method seeks to obtain the same kind of information as in a conventional test, but by minimizing test costs by acquiring the smallest volume of data necessary to do so. Key to the MDOE method is to fit a relatively small number of direct physical measurements to a mathematical "Response Surface Model" (RSM) that can be used to adequately estimate other responses that were not measured directly.

Central to the design of an MDOE wind tunnel test is the question of how many data points are needed to fit a response model adequately; that is, to ensure that there is an acceptably high probability that the model will estimate responses within some prescribed tolerance for independent variable combinations of interest. There are other issues in the design of an MDOE wind tunnel test that are beyond the scope of this paper, such as *which* data points to acquire, which points to replicate, the sequence by which the points should be acquired to ensure independence, and so forth, but this paper focusses on what is known as "scaling" in an MDOE test. Scaling simply means determining the minimum number of points that must be fitted to the model to ensure adequate response estimates.

A few specialized vocabulary words and phrases facilitate the discussion of formal experiment design because they describe concepts that are easy to understand, but are somewhat clumsy to express without them. We begin by introducing a few of these terms here and will add a few more as the context warrants.

Responses such as forces, moments, and pressures are estimated in a wind tunnel test for combinations of independent variables such as Mach number and angle of attack that can be represented graphically as a *design space*. A design space is simply a coordinate system with one independent variable associated with each axis. The design space is a plane for a test with only two independent variables as in Fig 1. More independent variables result in multidimensional design spaces for which the extension from two variables is conceptually straightforward, if more difficult to represent graphically.

Each point within the design space is called a *site*, and represents a unique combination of independent variable values. The term, "*factor*," is commonly used to designate an independent variable and the term, "*level*," is used to designate a specific factor setting. One speaks of factors such as angle of attack and levels of that factor such as 1°, 2°, 3°, etc. Each site in the design space represents some unique factor/level combination.

By the end of the test we wish to be able to estimate the responses at all sites of interest within the design space, and to do so within prescribed tolerance levels and with specified levels of confidence. The test is considered to be successfully concluded as soon as we are able to do this, no matter how many or how few data points have been acquired.

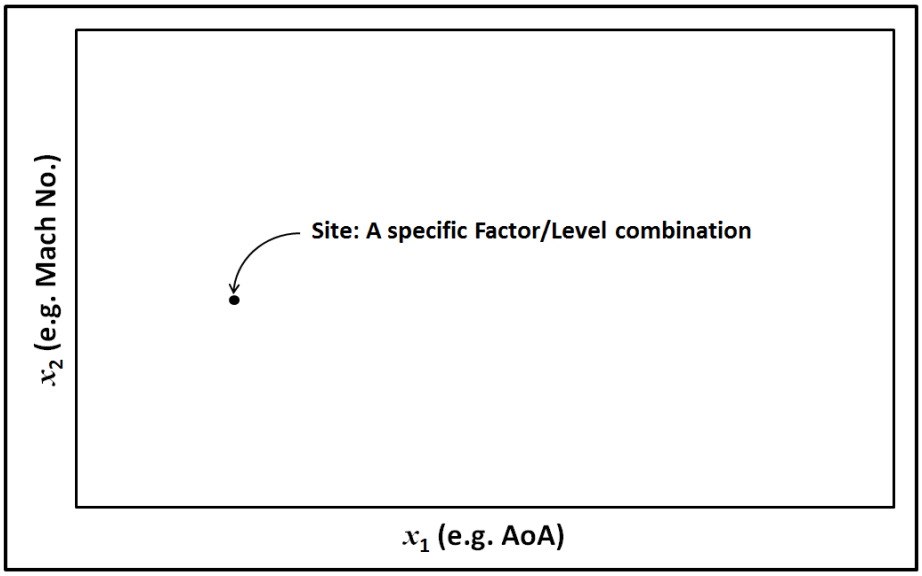

Figure 1. A two-factor design space.

Wind tunnel responses are estimated under the uncertainty that experimental error introduces. Whether responses are estimated by direct physical measurement or by a response model, this uncertainty can have two components. One is due to ordinary random experimental error and the other is due to a systematic bias error. The effects of random error can be reduced by replication and therefore vary inversely with the volume of acquired data; in general, more data can be made to translate into less uncertainty due to random error.

Replication in a response model is said to be "hidden." Hidden replication is the effective cancelation of random error effects that occurs when regression models pass near the means of measured data. Data with positive and negative random errors tend to align more or less symmetrically above and below the fitted model so that response predictions are relatively insensitive to them.

For direct physical measurements, the bias error is assumed to have been eliminated by calibration of the measurement systems so that all experimental error is random. The random error is expected to be normally distributed about a mean of zero, with a variance that characterizes the specific measurement environment of the test.

Responses that are estimated with a response model also have a random error component that results in some prediction variance, as well as a bias component due to systematic lack of fit (LOF) errors caused by fitting the data to a model that does not perfectly approximate the true underlying factor dependence. Details will be presented in subsequent sections but the general scaling strategy for coping with lack of fit is to ensure that enough data are acquired that LOF errors can be detected unambiguously, so that a better-fitting model can be substituted. In either case—whether responses are estimated by direct physical measurement or by using a response model—it is

desirable to eliminate bias errors so that all experimental error is random. This ensures that the uncertainty can be controlled by the volume of data that is acquired.

If the response estimate is a single-point measurement, as is generally the case in a conventional wind tunnel test, the precision depends entirely on the intrinsic variability of the measurement environment. The quality of wind tunnel test results is therefore synonymous with *data* quality in conventional wind tunnel testing. However, if that single-point measurement were to be replicated, enough of the random error could be made to cancel to drive the precision interval and associated uncertainty to arbitrarily small levels, depending on the number of replicates and the intrinsic variability of the measurement environment. This is an example of a proactive quality assurance tactic that the experimentalist can use to insulate himself from a dependence on the intrinsic variability of the wind tunnel facility. The unexplained variance in the raw data then becomes more of a *cost* factor than a *quality* factor; arbitrarily high quality can be achieved by acquiring enough data, even if the measurement environment features considerable intrinsic random experimental error. This tradeoff between cost and quality is a central theme in the Modern Design of Experiments (MDOE), which strives to reduce operating costs and cycle time by acquiring no more data than is necessary to cover the design space while ensuring prescribed quality specifications.

Unfortunately, the One Factor At a Time (OFAT) method used in conventional wind tunnel testing makes such high demands of limited test resources that quality assurance tactics of this kind are not always practical. Figure 2 illustrates a typical OFAT test matrix for two independent variables. Each dot represents a site in the design space where response data are acquired, corresponding to a unique combination of Mach number and angle of attack in this two-factor example. The OFAT method would typically entail holding the Mach number constant (as well as all other factors in a more complex experiment) while one factor—angle of attack as a typical example—is changed over some prescribed range, often in fixed intervals. The Mach number might then be incremented and the angle of attack settings repeated. This process typically continues until all prescribed factor/level combinations are examined.

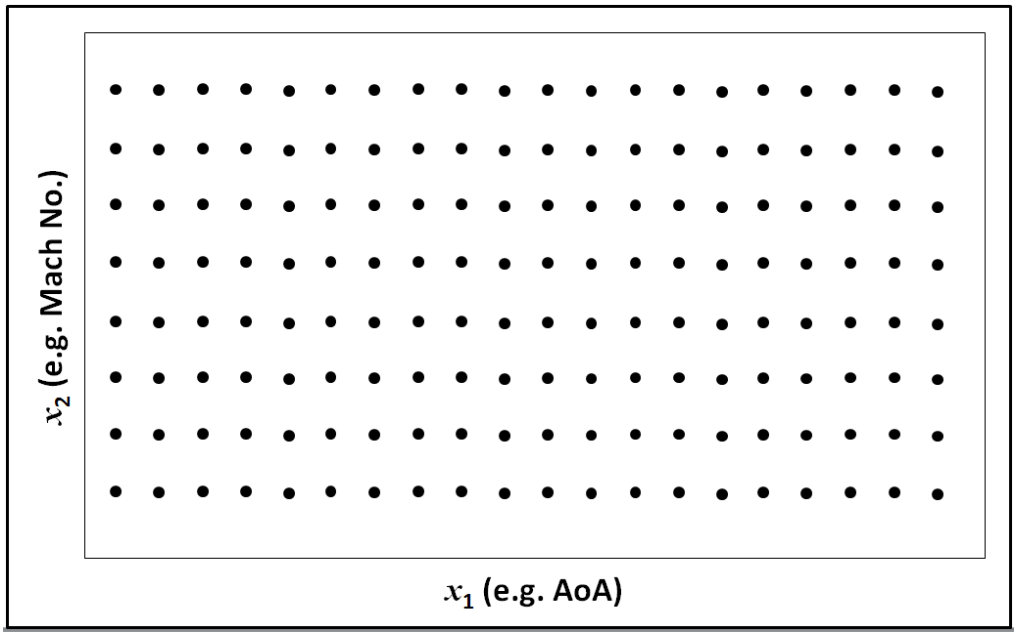

Figure 2. Two-factor design space with sites from a representative OFAT test matrix

The OFAT method covers the design space well in a two-factor test as illustrated in Fig. 2, but it is impractical to cover the entire design space in this way if there are more than a few independent variables, and even less practical if the data are replicated. For example, consider a relatively modest wind tunnel test in which only half a dozen factors are each set at ten levels. In such a test there are 10^6 or one million possible factor/level combinations, of which resource constraints will allow only a few thousand to be physically set in a typical two-to-four week tunnel entry. Five thousand data points would usually be regarded as a fairly productive wind tunnel test, yet this would represent only 0.5% of the possible factor/level combinations in this example, with 99.5% of them left unexplored.

OFAT practitioners note that not all factor/level combinations are equally interesting, and claim sufficient expertise to rank order them so that the unexplored settings are of relatively little interest compared to those that are

examined. The author attributes the disproportion number of unexplored sites in the design space—typically >99%—to OFAT inefficiency rather than to any judgement about how interesting the insights would be there. The reader can decide for himself which motive is the more plausible but whatever the reason, relatively few resources are allocated to quality assurance in a typical OFAT wind tunnel test, even at design space sites with a priority high enough to be included in the data acquisition schedule.

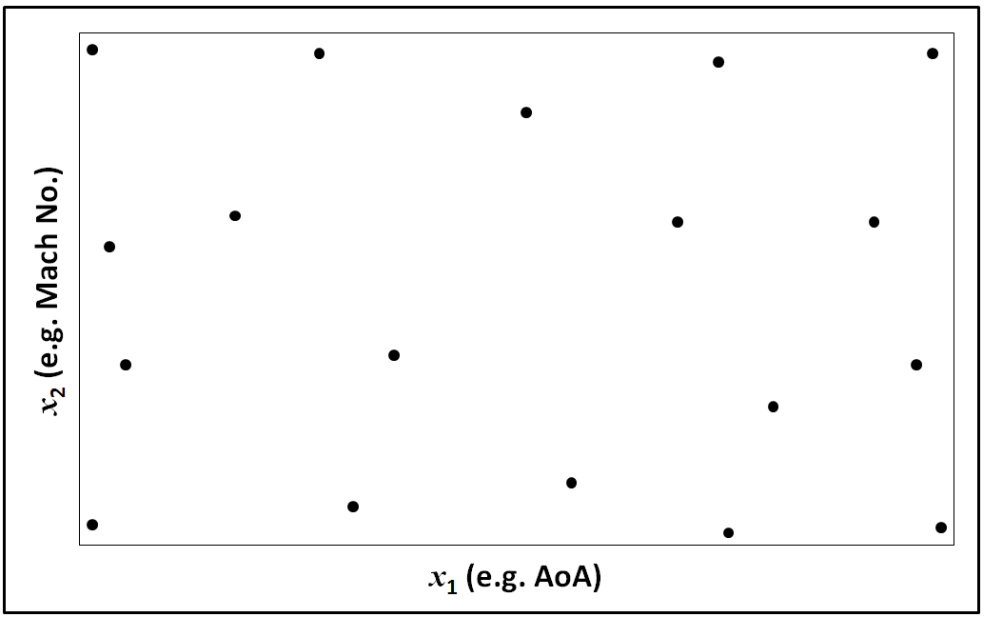

Figure 3. Two-factor design space with sites from a representative MDOE test matrix

The MDOE approach to wind tunnel testing, first advocated by the author in the late '90s[1-11], shares a similar objective with OFAT testing; that is, the OFAT practitioner and the MDOE practitioner each seeks to estimate certain responses such as forces, moments, and pressures, throughout the design space. The difference is in the *approach* to this objective. The OFAT practitioner makes direct physical measurements at every design space site of interest as illustrated in Fig. 2. As noted earlier, the MDOE practitioner acquires data at considerably fewer sites, distributing them strategically within the design space to enable adequate responses estimates throughout. Compare Fig. 3, a typical distribution of sites for a two-factor MDOE experiment, with its OFAT counterpart in Fig. 2.

The MDOE approach facilitates both quality *assurance* and quality *assessment* throughout the design space. Quality assessment will be addressed presently. Quality assurance is achieved by a number of tactics designed to impose statistical independence on the individual measurements of the data acquired in the test and to minimize the adverse effects of experimental error experienced at any given site. These tactics are beyond the scope of the present paper but are reviewed elsewhere in the literature[1-6]. This paper focusses on the relationship between uncertainty in an experimental result and the volume of data acquired in the test.

The practical objective of a wind tunnel test can never be to "minimize uncertainty" (by "maximizing data volume") for two reasons. First, resources—time and money—are consumed when each new data point is acquired. Cumulative resource expenditures can be quite large when arbitrarily large volumes of data are specified in the test plan. A strategy to acquire "as much data as possible" is tantamount to a strategy to "spend as much money as possible." In practice, the end of most conventional wind tunnel tests is dictated by the exhaustion of available resources (tunnel occupancy time, liquid nitrogen budget, etc.) rather than the achievement of specific technical objectives. This is not to say that such objectives are of secondary importance, but only to note that it is rare for a conventional wind tunnel test to end with test resources still available, simply because all technical objectives have been met. In fact, it is fair to say that the implicit objective of a conventional wind tunnel test is usually to acquire as much data as resources permit; to "make hay while the sun shines" as it were. The second reason that minimizing uncertainty by maximizing data volume is not recommended is that, while uncertainty can never be driven to zero for any finite volume of data, it can usually be reduced to levels small enough to be acceptable in most practical circumstances by acquiring a surprisingly small volume of data.

The tradeoff between data volume (therefore data acquisition cost) and uncertainty is addressed through "scaling" the experiment, the process by which a minimum volume of data is determined that is sufficient to ensure

prescribed response tolerances with no more than a specified level of inference error risk (e.g. ±0.5 drag count with 95% confidence, which is to say a probability of less than 0.05 that random error will cause a response estimate to differ from the true response by more than half a count.) It is fair to say that the proper strategic objective of most wind tunnel tests is the same; namely, to gather enough new information about the test article that its future behavior can be predicted at any site of interest in the design space, within a specified tolerance and with no more that prescribed levels of inference error risk. Scaling is the means by which this strategic objective is achieved.

This concept of inference error risk is crucial to quality assurance. In formal terms, there is always an implied null hypothesis stating that there is no significant difference between a reported response estimate and the truth. We define "no significant difference" using some prescribed tolerance specification, such as "0.5 drag count" in this example. The null hypothesis is rejected when there is cause to suspect that a response estimate is out of tolerance, but either inference about the null hypothesis—that it is valid or that it is not valid—is made under uncertainty and can be right or wrong. The probability of an inference error depends on data volume, dropping rapidly as more data are acquired. The essence of quality assurance in an MDOE wind tunnel test is to determine how much data is needed to drive inference error risks low enough to be acceptable.

Much of this paper focusses on scaling as a key mechanism of quality assurance, expanding on earlier work[8] to accommodate evolving assumptions by the experimental aeronautics community as to what constitutes an adequate response model. Early skeptics of formal experiment design regarded direct physical measurements as "truth" and viewed predictions by a fitted response model with suspicion. Under the null hypothesis, residuals consisting of differences between predicted and measured responses were assumed to be normally distributed by the Central Limit Theorem, with a mean of zero and a variance attributable entirely to response model prediction variance. Variance in the physical measurements serving as surrogates for "truth" in the residual calculations was ignored in the scaling process, consistent with the assumption that these measurements represented the known truth against which model predictions were to be compared. The current paper revises these scaling calculations to explicitly account for truth surrogate variance. This truth surrogate variance has a number of interesting consequences for quality assurance in formally designed wind tunnel tests, which are explored in the sections that follow.

Quality *assessment* differs from quality *assurance*. The result of a quality assessment exercise is some demonstration that quality specifications declared during the design of the experiment have been satisfied. This typically involves a series of tests by which the adequacy of a candidate response model is evaluated. Inference errors described earlier play an important role in response model quality assessment. It is shown, for example, that while the probability may be high that a response model residual will be out of tolerance if the model predicts inadequately, the reverse is not necessarily true. It is not necessarily true that the model predicts inadequately at a site in the design space where the residual is out of tolerance. It will also be shown that for well-fitted models, the probability is surprising high that a residual will be out of tolerance at a site for which the model actually predicts adequately.

Inference error risk limits are established through a scaling process that facilitates quality assurance in an MDOE wind tunnel test, as will be demonstrated in the sections that follow. These inference error risk limits are used to compute a proposed new response model adequacy metric that is related to a conventional "tolerance limit." This metric represents the fraction of the design space within which tolerance and inference error risk specifications can be expected to be met. For example, a drag coefficient model might be deemed adequate if in 95% or more of the design space, model predictions are expected with a prescribed level of confidence to be within a half count of the true drag coefficient.

II. Quality Assessment with a Constant Truth Surrogate

We begin with an elementary review of the relationship between data volume and the precision with which experimental results can be estimated, in order to introduce some nomenclature and to develop a foundation for more complex cases to follow. Consideration of the effects of truth surrogate variance will be postponed briefly, until this foundation is laid, but will be treated in some detail in Section III. Precision in direct physical measurements is reviewed, as a precursor to considerations of precision in a response surface modeling experiments. The topic of Lack of Fit bias errors is then introduced. This topic is especially important in response surface modeling experiments because it is a new error source that does not have to be considered in conventional wind tunnel testing. The relationship between Lack of Fit errors and response model accuracy is discussed, as well as scaling to generate the precision needed to ensure that LOF errors do not result in improper inferences about the adequacy of candidate response surface models.

A. Precision in Direct Physical Measurements

A response estimate is a random variable featuring dispersion that can therefore be represented by a probability density function (PDF) as in Fig. 4, which displays the PDF of an *N-point* sample mean. This distribution is assumed to be normal by the Central Limit Theorem, with a standard deviation of $\sigma = \sigma_0/\sqrt{N}$, where σ_0 is the standard deviation of the measurement environment's irreducible random error and N is the sample size. The quantity σ_0 is therefore a measure of the intrinsic variability of the measurement environment. It is key to scaling an experiment because the volume of data required to meet given quality specifications is proportional to variance of the measurement environment, which is the square of σ_0.

Two limits known as critical values for the distribution are established at symmetric distances of $\pm\gamma$ relative to the mean. They are each located a multiple of σ from the mean as indicated in Fig. 4. This multiple is designated z_α, and is defined by α, the total size of the area under the PDF and outside the critical values, which is shaded in the figure. Numerical values of z_α are tabulated in standard statistical tables for specified values of α and can also be determined using workbook functions in standard spreadsheet software and other statistical software packages.

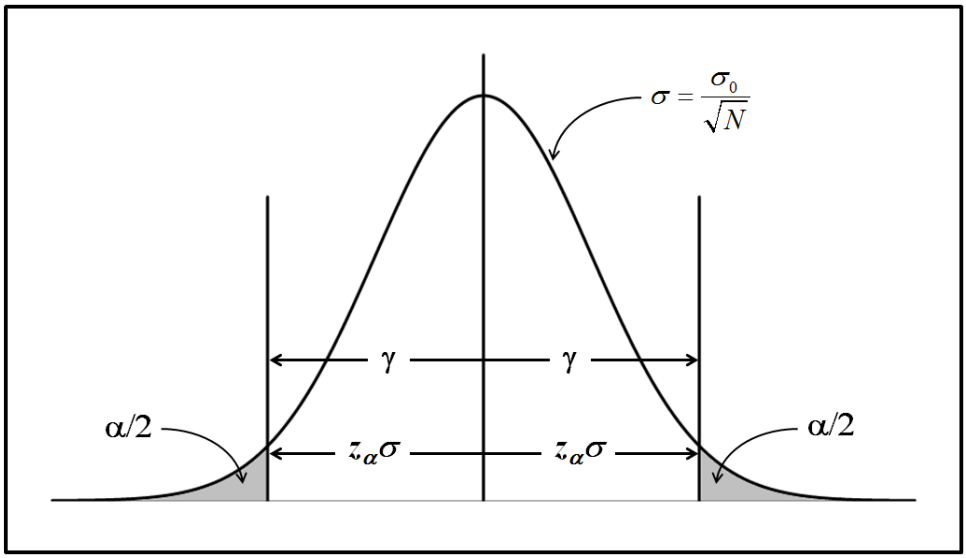

Figure 4. Probability Density Function for the mean of an *N-point* sample

The interval centered on the mean and spanning the range from $-\gamma$ to $+\gamma$ relative to the mean is a $(1 - \alpha) \times 100\%$ precision interval that indicates how precisely the sample mean has been estimated. If such intervals were to be generated for an infinite number of N-point samples drawn from the same population, $(1 - \alpha) \times 100\%$ of those intervals would be expected to contain the average of a future N-point sample drawn from the same population. Two cases of the precision interval are of special interest. When $N = 1$ the interval is known as a prediction interval and when $N = \infty$ it is known as a confidence interval. The former represents an interval within which the next single-point measurement is expected to fall with a probability of $(1 - \alpha)$ and the latter represents the interval within which the true response (the population mean) is expected to be found with a probability of $(1 - \alpha)$.

The quantity α represents the probability that a measured response estimate will be more than $\pm\gamma$ away from the true response simply because of random experimental error. If we construct a null hypothesis which states that there is no significant difference between the true response and the experimental response estimate, α is the probability of an inference error committed by erroneously rejecting that hypothesis. Erroneously rejecting the null hypothesis in this way comprises a "Type-I" or "alpha" inference error. We will also have to consider a "Type-II" or "beta" inference error that is often more serious, when we augment the current discussion of precision with remarks about accuracy. Distinctions between accuracy and precision will be addressed presently.

Consider now a response estimated for a given site in a wind tunnel test and assume it consists of the mean of an N-point sample of genuine replicates. For the special case of a single-point sample such as characterizes most conventional wind tunnel testing, $N = 1$ and the standard deviation of this PDF is just $\sigma = \sigma_0$, with a corresponding value of $\gamma = z_\alpha\sigma_0$ that is designated as γ_0 for this special case of $N = 1$. There is a probability of $1 - \alpha$ that the true response lies within $\pm\gamma_0 = \pm z_\alpha\sigma_0$ of the empirically estimated response, which for $\alpha = 0.05$ is a probability of 95%

that the true response lies within $\pm 2\sigma_0$ of the empirical response estimate. For example, if the response that is being estimated empirically is the drag coefficient and the measurement environment is characterized by a σ_0 of one drag count, a future unbiased single-point measurement is expected to be within ±2 counts of the present estimate of drag with a probability of 95%. If this precision is acceptable, then no replication of the measurement is required.

If greater precision is required, it can be obtained by replicating the measurement. We can estimate the volume of data required to achieve greater precision by examining Fig. 4 and making use of the following relationship:

$$\gamma = z_\alpha \left(\frac{\sigma_0}{\sqrt{N}} \right) \rightarrow N = \left(\frac{z_\alpha \sigma_0}{\gamma} \right)^2 = \left(\frac{\gamma_0}{\gamma} \right)^2 \tag{1}$$

The quantity on the far right of Eq. 1 is the ratio of the intrinsic resolution of the measurement environment, γ_0, to γ, which is the tighter resolution that we require. We will call this ratio the "Precision Gain" required to satisfy our quality requirements, and designate it as G_P. We then have a very simple formula to compute how much data is required to satisfy our precision requirements:

$$N = G_P^2 \tag{2}$$

where

$$G_P = \frac{\gamma_0}{\gamma} \tag{3}$$

In the present example, $\sigma_0 = 1$ count. For $\alpha = 0.05$, $z_\alpha = 2$ so $\gamma_0 = z_\alpha \sigma_0 = 2$ counts. If we are satisfied with that resolution, that is, if $\gamma = \gamma_0$ suits us, the $Gp = 1$ and a single measurement per site is all that is required. However, if the greatest acceptable value of γ is, say, 0.5 drag count, then we have $G_P = 2/0.5 = 4$ and

$$N = G_P^2 = 16 \tag{4}$$

Thus 16 replicates are required to achieve a precision of ±0.5 drag count with 95% confidence in a measurement environment with an intrinsic standard error of $\sigma_0 = 1$ count. For single-point response estimates, ±0.5 drag count precision can only be achieved with 95% confidence if the measurement environment is characterized by an intrinsic standard error no greater than 0.25 drag count.

B. Precision in Response Surface Modeling

The previous section considered the relationship between precision and data volume when responses were estimated by direct physical measurements, as in an OFAT test. We now consider that same relationship for the case in which responses are estimated with a response surface model. All elements of Fig. 4 apply to this case, except that the dispersion of the PDF is different. This case is complicated by the fact that dispersion in the response model predictions is site specific. While the variation in dispersion over sites can be reduced by judicious site selection, there is typically somewhat more variance in the corners of the design space than elsewhere for typical site distributions. Fortunately, while the prediction variance differs from site to site, the *average* prediction variance is the same for all polynomial response models with a given number of terms in the model[12] that are fitted to the same volume of data with the same intrinsic random error variance:

$$\bar{\sigma}_{\hat{y}}^2 = \left(\frac{p}{N} \right) \sigma_0^2 \tag{5}$$

Here, p is the number of terms in the response model including the intercept, σ_0 is the standard error of the measurement environment as before, and N is the number of points fitted to the response model. A minimum of $N = p$ unique points are required to fit a p-term model so $N \geq p$ and therefore $p/N \leq 1$. This means that the average standard deviation associated with a response surface model prediction at a given site in the design space is never

greater than the standard deviation of a single-point physical measurement, and is generally less. This is due to "hidden replication," by which the random error in points fitted to a response model tend to cancel.

We proceed for the case in which responses are estimated with a response surface model along the same lines as when we considered responses estimated with direct physical measurements. With obvious substitutions, Eq. (1) can be wriiten for the RSM case as follows:

$$\gamma = z_\alpha \sigma_0 \left(\sqrt{\frac{p}{N}} \right) \rightarrow N = p \left(\frac{z_\alpha \sigma_0}{\gamma} \right)^2 = p \left(\frac{\gamma_0}{\gamma} \right)^2 \tag{6}$$

Inserting Eq. (3):

$$N = pG_P^2 \tag{7}$$

If one requires no more precision than is available "for free" from the measurement environment (that is, without incurring the expense of additional residual degrees of freedom), then $\gamma = \gamma_0$ as before, and $G_P = 1$. In that case, Eq. 7 reduces to $N = p$ and it is not necessary to acquire more than the minimum volume of data needed to fit a p-term polynomial. Also note that the single-point physical measurement case described in Eq. 2 is just a special case of the p-term RSM result in Eq. 7, with $p = 1$. This is because an N-point sample mean, even for $N = 1$, represents the intercept of an elementary polynomial response model in which all terms except the intercept have been neglected.

$$y = b_0 + \cancel{\sum_{i=1}^{p-1} b_i x_i}$$

Such a "zeroth-order response model in no factors" has exactly $p = 1$ term, so Eq. 7 is perfectly general.

The purpose of adding progressively higher order terms to the response model, starting with an intercept and then adding first-order terms, second-order, third-order, and higher order terms is to expand the factor range over which the model can be expected to predict adequately. As noted, zeroth-order (intercept-only) response models are effectively "fitted" in an OFAT test when the mean of an N-point sample of direct physical measurements is reported at each site of interest, where $N = 1$ is possible, and is in fact common. The weakness of low-order models generally, and zeroth-order models as a limiting case, is that the range over which they can be applied adequately is restricted. Single-point "response models" apply over an area of a normalized two-dimensional design space that can be represented by a circle with a radius of zero; that is, they describe the response at a single point only. Responses between points, strictly speaking, are entirely unknown, although methods such as linear and cubic-spline interpolation are often applied in post-test analyses to estimate responses between the sites where data are actually acquired. In effect, then, a kind of localized response surface modeling takes place even in an OFAT test. True response surface modeling with higher-order terms simply extends this between-point interpolation to cover a greater range of the design space.

In both OFAT and MDOE testing, then, response models that are adequate over a limited range are employed to develop what is in effect a piecewise continuous response model that covers the entire design space. The advantage of a higher-order response model over the zeroth-order response models employed in OFAT testing is that by including higher-order terms the range of applicability of the model is increased significantly, and fewer subspace models are required to span the entire design space. Figure. 2, when extended to the practical case of a multidimensional design space, illustrates that thousands of zeroth-order models, each adequate over an infinitesimal range, are typically required to span the entire design space even approximately (that is, even neglecting responses between points). When higher-order terms are added to these intercept-only models, each fitted model spans a larger subspace and so fewer models are required altogether. In practice, a half-dozen or so response models that are no higher than fourth-order are sufficient to span even a multi-dimensional design space in a typical wind tunnel test. While each such model requires more points to fit than a single-point zeroth-order intercept-only model, so many fewer of them are required to span the whole design space that significant reductions in total data volume can usually be achieved, with attendant reductions in cost and cycle time.

C. Lack of Fit and Accuracy Considerations.

Response surface modeling entails an additional potential source of error that does not have to be considered in OFAT testing, which results in an understandable reticence by some OFAT practitioners to embrace the method. The adequacy of a response estimate made by direct measurement at a given site in the design space depends on the quality and quantity of the data acquired there (standard error of the measurement environment, σ_0, and number of replicates). The adequacy of an RSM prediction depends on those things also, plus something else: the specific functional form that is assumed for the fitted response model. A straight line will not adequately predict high-order responses, even if it was fitted to good data and lots of it, and even if it is the "best" straight line that could be fitted. Fitting the wrong model results in a systematic bias error in addition to the random prediction error inevitably associated with even a well-fitted model.

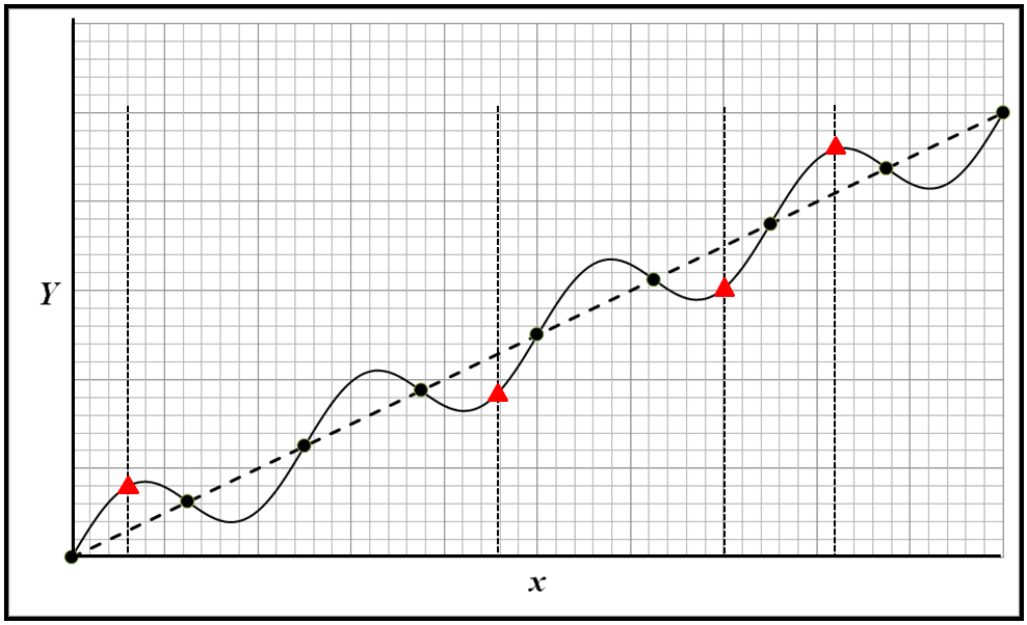

Figure 5. Lack-of-Fit Error. The heavy dashed line is a lower-order model fitted to discrete points from the higher-order solid black curve. Triangles indicate confirmation points acquired at random sites marked with the light vertical lines, which reveal inadequacy of the model.

Figure 5 illustrates the kind of systematic lack-of-fit (LOF) bias error that can be introduced when the wrong model is fitted. In this figure, the true response is the relatively high-order function represented by the undulating solid black line, and the heavy dashed line represents an inadequate model that is fitted to discrete response measurements represented by the black dots. The situation illustrated in Fig. 5 is somewhat exaggerated—it is not very likely that one would be so unfortunate as to obtain this clear an indication of a first order response from discrete samples of a higher-order function. Nonetheless, it is always possible that a relatively simple, low-order model will be fitted that underestimates the complexity of the true underlying response. No amount of data can rescue such a situation if it occurs, because the resulting lack of fit error is not random but is instead a systematic bias that cannot be cancelled by simple replication. The only remedy for this type of error is to fit a more appropriate model.

It should be noted that the experimentalist is usually able to minimize the probability of serious LOF error by using experience and subject matter expertise to propose models that are generally realistic. Even for the relatively unlikely case in which he has absolutely no idea what to expect, the experimentalist can exploit the fact that an unknown response function can be approximated arbitrarily well by retaining enough terms in a Taylor Series representation of that function, assuming certain relatively mild conditions hold that are usually met (e.g. that the function and its derivatives exist for the domain of independent variables over which the function is approximated). Retaining an infinite number of such terms would permit the function to be represented exactly, but in practice one need only retain a finite number of lower-order terms to approximate typical response functions within acceptable tolerances. The more limited the range of independent variables, the better the fit provided by a low-order polynomial, so another way to minimize systematic lack of fit bias errors is to fit a few separate models over limited

subspaces. Over such limited ranges the underlying responses are likely to exhibit milder variations that can be modeled with simpler models than would be required to span the entire design space.

Notwithstanding the fact that LOF bias errors can be minimized by tactics such as design space truncation and by bringing experience and subject matter expertise to bear in the candidate model selection process, it is still necessary to assess the degree to which LOF bias errors are in play in a proposed model. Figure 5 illustrates how such errors can be detected by acquiring data at independent validation sites within the design space. These sites are represented by vertical dotted lines in Fig. 5. Validation response data acquired at these sites are indicated by the triangles. Residuals constructed at these sites suggest LOF error if they are significantly larger than can be explained by ordinary random error. Further details are given in the section below on quality assessment.

Equation 7 prescribes the minimum volume of data needed to achieve a specified level of precision, γ, with a given level of confidence, $(1 - \alpha) \times 100\%$, in a measurement environment characterized by a standard error of σ_0. The general quality assurance strategy for dealing with LOF errors in an MDOE experiment is to acquire a sufficient volume of data to ensure enough precision that LOF errors large enough to be of concern can be detected with an acceptable level of confidence. We proceed by first augmenting Fig. 4 as follows:

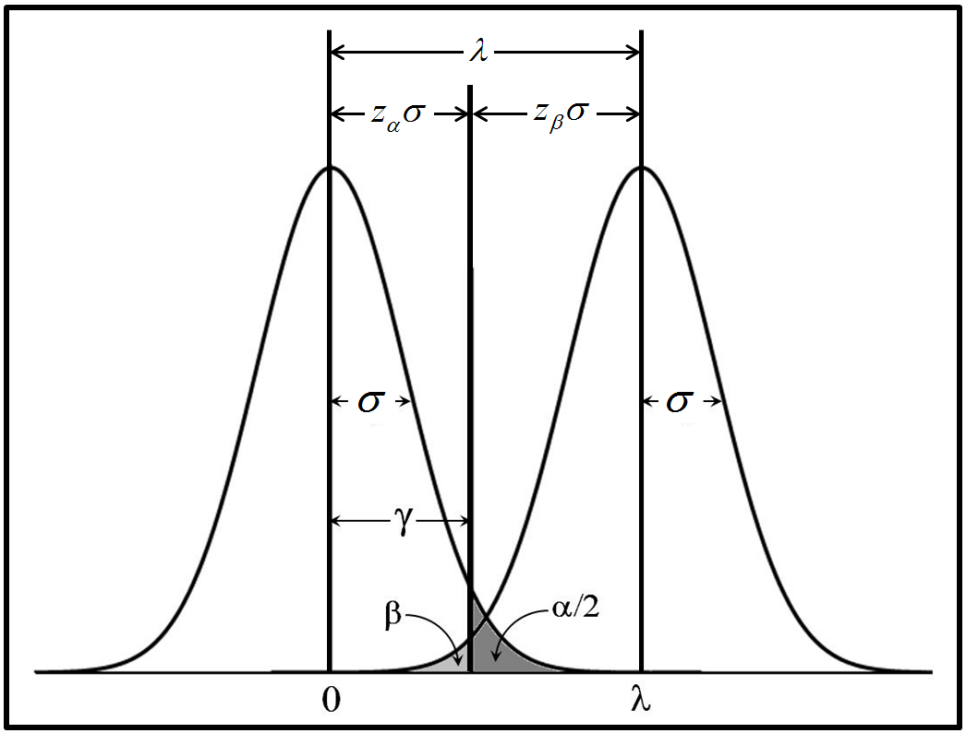

Figure 6. Distribution of response predictions. No LOF bias error (left) and LOF bias error of λ (right).

The normal PDF on the left in Fig. 6 is the same as in Fig. 4. To minimize confusion, only one of the shaded areas from Fig. 4 is retained in Fig. 6, but in all respects the PDF on the left and the one in Fig. 4 are identical. The PDF on the left in Fig. 6, centered on 0, represents the normal distribution of response model predictions that one would expect in the absence of LOF bias error. It corresponds to a null hypothesis that asserts no difference between the response estimate and the true response. The distribution on the right displays the same normal random error variance as the one on the left, but the mean of the distribution is biased by a lack of fit error of λ. It corresponds to an alternative hypothesis that asserts a significant bias between the response estimate and the true response.

For scaling purposes, we assume that λ is the smallest systematic bias error that is too large to be acceptable. That is, bias errors greater than or equal to λ are unacceptable and bias errors less than λ are within tolerance. While γ is a precision specification, λ is then an *accuracy* specification. We accept the inevitability of random fluctuations about the mean, but our accuracy specification requires that the *mean* of the distribution of model predictions be within λ of the true response.

The shaded area labeled $\alpha/2$ represents the probability that random error will cause a response estimate drawn from the distribution on the left to differ from the true response by more than γ, the specified level of precision. This random error is equally likely to be positive as negative, so the total probability for all random errors is α. A response estimate actually drawn from the unbiased distribution on the left but which falls to the right or the left of the mean of that distribution by more than γ due to random error, will be erroneously said to be biased. This would constitute a Type-I or alpha inference error. One makes such an error when one erroneously rejects an adequate response model. There are consequences of such an error in terms of time and money wasted to try to improve a model that was already adequate and only appeared to predict out of tolerance because of random error in the measurement environment, not because of any shortcomings in the model itself.

The shaded area labeled β represents the probability that random error would cause a response estimate from the biased distribution on the right to lurch within the prediction internal specified by $\pm\gamma$ about the mean of the unbiased distribution on the left. This would result in the validation of a model that actually generates response predictions with unacceptable lack-of-fit bias errors, which would constitute a Type II or beta inference error. The magnitude of such a random error would have to exceed a minimum multiple of the distribution's standard deviation, designated z_β, which is dictated by β and tabulated for a normal distribution in standard statistical tables for specified levels of β. This quantity can also be determined with functions built into standard statistical software.

Note that the distribution on the left, associated with the null hypothesis, can be one-sided or two-sided. In this example it is two-sided, meaning that we reject the null hypothesis if the response estimate is either too large or too small to be within γ of the true response. The alternative hypothesis is always one-sided; we reject it if the magnitude of the difference between the response estimate and the true response is less than λ, no matter what the sign of the difference is. These distinctions are important because the tabulated z values are different for one-sided and two-sided distributions.

For a given maximum acceptable bias error, λ, Fig. 6 makes it clear that the probability of a Type I inference error can be reduced by increasing γ; that is, by settling for less precision. However, Fig. 6 also shows that this would increase the probability of a Type II inference error. Likewise, the probability of a Type II inference error can be reduced by decreasing γ, at the expense of an increase in the probability of a Type I inference error. Both Type I and Type II inference error probabilities can be reduced, however, by reducing the prediction variance (reducing the width of the distributions), which is achieved by increasing the number of fitted points.

Figure 6 shows that λ, α, β, and σ are all related through the following equation:

$$\lambda = z_\alpha \sigma + z_\beta \sigma = \left(z_\alpha + z_\beta \right) \sigma \tag{8}$$

Square both sides of Eq. 8 and insert the prediction variance displayed in Eq. 5 since the distributions of Fig 6 represent response model predictions:

$$\lambda^2 = \frac{p\left(z_\alpha + z_\beta \right)^2}{N} \sigma_0^2 \tag{9}$$

Solve for N:

$$N = \frac{p\left[\left(z_\alpha + z_\beta \right) \sigma_0 \right]^2}{\lambda^2} \tag{10}$$

The quantity in square brackets in Eq. 10 can be seen in Fig. 6 to be the value of λ when $\sigma = \sigma_0$. Designate this λ_0:

$$N = p\left(\frac{\lambda_0}{\lambda} \right)^2 \tag{11}$$

Eq. 11, developed to scale the experiment to satisfy accuracy specifications for prescribed levels of inference error risk, is analogous to Eq. 6, which was developed to scale the experiment to satisfy precision specifications for

the same levels of inference error risk. Achievable precision and accuracy are proportional for a given volume of fitted data, so either approach can be used to scale a test. If one scales for precision, one must accept the corresponding accuracy, and if one scales for accuracy, one must accept the corresponding precision; the two quantities cannot be independently specified because they are related as follows (see Fig 6):

$$\gamma = \left(\frac{z_\alpha}{z_\alpha + z_\beta} \right) \lambda \tag{12a}$$

$$\lambda = \left(\frac{z_\alpha + z_\beta}{z_\alpha} \right) \gamma \tag{12b}$$

For typical values of $\alpha = 0.05$ and $\beta = 0.01$, say, the corresponding *z*-values are $z_\alpha = 1.960$ and $z_\beta = 2.326$. In this case,

$$\gamma = \left(\frac{1.960}{1.960 + 2.326} \right) \lambda = 0.457 \times \lambda \tag{13a}$$

$$\lambda = \left(\frac{1.960 + 2.326}{1.960} \right) \gamma = 2.187 \times \gamma \tag{13b}$$

Analogous to Eq. 3, the quantity in parentheses in Eq. 11 represents an accuracy ratio. The numerator, λ_0, is the accuracy that can be achieved in the existing measurement environment by acquiring no more than *p* data points, the minimum needed to fit a *p*-term response model. The denominator, λ, is a more stringent accuracy specification that can be achieved if more than *p* data points are fitted to the model. We will call this ratio the "Accuracy Gain" required to satisfy our quality requirements and designate it as G_A, analogous to the G_P term defined in Eq, 3 for precision. We then have a very simple formula to compute how much data is required to satisfy our accuracy requirements:

$$N = pG_A^2 \tag{14}$$

where

$$G_A = \frac{\lambda_0}{\lambda} \tag{15}$$

As an example, consider the 95% Least Significant Difference (LSD) as an accuracy specification. The 95% LSD is the smallest difference between two single-point response estimates that can be resolved with 95% confidence. We therefore say that if our response model is biased by an amount, λ, that is as large as the 95% LSD, we wish there to be no more greater probability than $\beta = 0.01$ that random error about a distribution mean biased by that amount will cause a response estimate to fall between $\pm z_\alpha \sqrt{(p/N)} \sigma_0$ of the true response.

For $\alpha = 0.05$ and $\beta = 0.01$ we have $\lambda_0 = (1.960 + 2.326) \times \sigma_0 = 4.286 \times \sigma_0$. The 95% LSD is $2\sqrt{2} \times \sigma_0$, so

$$G_A = \frac{4.286 \times \sigma_0}{2\sqrt{2} \times \sigma_0} = 1.515 \tag{16}$$

and from Eq. 14:

$$N = pG_A^2 = p\left(1.515\right)^2 = 2.30p \tag{17}$$

Therefore, to satisfy a 95% LSD accuracy requirement with the inference error risk probabilities that have been prescribed requires a sample of data to be fitted to a *p*-term model that has 2.3 data points for every term in the model.

A d^{th}-order polynomial in *k* factors has *p* terms, where p is computed as follows:

$$p = \frac{\left(d+k\right)!}{d!k!} \tag{18}$$

For example, a 4^{th}-order polynomial in 3 factors has (4+3)!/(4!3!) = 35 terms, including the intercept. By Eq. 17, to fit such a model so that a bias error of magnitude $2\sqrt{2}\sigma_0$ can be detected with a probability of $(1 - \beta) \times 100 = 99\%$ would require a fitted data sample of 2.30 × 35 = 81 points, rounded to the nearest higher integer. From Eq. 13a the precision required to ensure such a probability is ±0.457 × $2\sqrt{2}\sigma_0$ = ±1.3 counts with $(1 - a) \times 100 = 95\%$ confidence.

If this precision is not satisfactory to satisfy other criteria, the scaling can be repeated with a tighter precision specification, which would call for a larger sample of fitted data and result in an accuracy specification more stringent than the 95% LSD. Say, for example, that it was desirable to be able to use the model to predict responses with a precision of ±1 count with $(1 - a) \times 100 = 95\%$ confidence. We have that $\gamma = z_\alpha\sigma_0 = 1.960 \times \sigma_0$, and if $\gamma = 1$ count, then by Eq. 3 we have

$$G_P = \frac{z_\alpha\sigma_0}{\gamma} = \frac{1.960\sigma_0}{1} = 1.960\sigma_0 \tag{19}$$

so that by Eq. 7 we have

$$N = pG_P^2 = p\left(\frac{z_\alpha}{\gamma}\right)^2\sigma_0^2 \tag{20}$$

Eq. 20 underscores that the minimum volume of fitted data needed to satisfy given precision and inference risk specifications is proportional to the complexity of the response model (i.e. the number of terms, *p*) and the random error variance of the measurement environment, σ_0^2. Inserting values for the precision and inference risk specifications of this example into Eq. 20 ($\gamma = 1$ count and $\alpha = 0.05 \rightarrow z_\alpha = 1.960$), and using the same number of terms in the model, $p = 35$, we have

$$N = \left(1.960\right)^2 p\sigma_0^2 = \left(3.842 \times 35\right)\sigma_0^2 = 134.5\sigma_0^2 \tag{21}$$

In a measurement environment for which $\sigma_0 = 1$ count, for example, this tighter precision specification would require 135 data points (rounding to the next highest integer), compared to 81 points required for the less stringent precision requirement associated with a 95% LSD accuracy specification in the same measurement environment. By Eq. 13b, this larger sample size would tighten the accuracy specification from $\lambda = 95\%\ LSD = 2\sqrt{2}\sigma_0 = 2.828$ counts to *2.187* counts, a 29% increase in accuracy achieved at the cost of a 67% increase in data volume (54 additional points to be fitted).

The Modern Design of Experiments is silent as to whether such a tradeoff is advisable; it simply provides the experimenter with a quantitative means of evaluating it. The benefit of the increase in precision in this example may or may not be considered worth the cost of acquiring the additional data needed to achieve it, but the experimenter would have good estimates of both the cost and the benefit upon which to base a decision

III. Impact of Truth Surrogate Variance on Quality Assessment

The distributions displayed in Fig. 6 feature dispersion attributed to response model prediction variance. This is the same dispersion one would observe in a sample of residuals constructed by subtracting the *true* response (a constant) from model predictions at each of a number of model validation sites in the design space. The true response would display no variance, so all of the residual variance would be due to random error in the model predictions. This is the implicit assumption that is made when confirmation points are rather imprecisely assumed to represent "true responses" against which model predictions are assessed.

Some interesting new insights can be achieved by explicitly accounting for variance in the direct measurements used to assess model predictions at various validation sites in the design space. These direct physical measurements are in fact subject to random experimental error; and while each measurement may represent an unbiased estimate of the true response at a given validation site, it is only an imperfect surrogate for the true response, which is not generally known.

A. Single-Point Physical Measurement as Truth Surrogate

In this section we will assume that the truth surrogate is characterized by variance. We will also assume that the truth surrogate variance is the same as the variance in the sample of data used to fit the response model, with each characterized by a standard error of σ_0.

The residual variance consists of the sum of the response model's prediction variance as given in Eq. 5 and the variance of the truth surrogate, which is just the square of the standard error, σ_0:

$$\sigma^2 = \sigma_0^2 + \frac{p}{N}\sigma_0^2 \rightarrow \sigma = \sigma_0\sqrt{1+\frac{p}{N}} \tag{22}$$

Insert Eq. 22 into Eq. 8:

$$\lambda = \left(z_\alpha + z_\beta\right)\sigma_0\sqrt{1+\frac{p}{N}} \tag{23}$$

Square both sides of Eq. 23:

$$\lambda^2 = \left(z_\alpha + z_\beta\right)^2\left(\frac{N+p}{N}\right)\sigma_0^2 \tag{24}$$

Multiply both sides by N:

$$N\lambda^2 = N\left(z_\alpha + z_\beta\right)^2\sigma_0^2 + p\left(z_\alpha + z_\beta\right)^2\sigma_0^2 \tag{25}$$

Gather terms in N:

$$N\left[\lambda^2 - \left(z_\alpha + z_\beta\right)^2\sigma_0^2\right] = p\left(z_\alpha + z_\beta\right)^2\sigma_0^2 \tag{26}$$

Solve for N:

$$N = p\left[\frac{\left(z_\alpha + z_\beta\right)^2\sigma_0^2}{\lambda^2 - \left(z_\alpha + z_\beta\right)^2\sigma_0^2}\right] \tag{27}$$

Recall the definition of λ_0:

$$\lambda_0 = \left(z_\alpha + z_\beta\right)\sigma_0 \tag{28}$$

Insert Eq. 28 into Eq. 27:

$$N = p\left[\frac{\lambda_0^2}{\lambda^2 - \lambda_0^2}\right] \tag{29}$$

Divide the numerator and the denominator of Eq. 29 by λ^2:

$$N = p\left(\frac{\frac{\lambda_0^2}{\lambda^2}}{1 - \frac{\lambda_0^2}{\lambda^2}}\right) \tag{30}$$

Insert Eq. 15 into Eq. 30:

$$N = p\left(\frac{G_A^2}{1 - G_A^2}\right) \tag{31}$$

B. Replicated Truth Surrogates

Compare Eq. 31, which accounts for variance in the truth surrogate, with Eq. 14 reproduced here for convenience, which assumes a constant truth surrogate:

$$N = pG_A^2 \tag{14}$$

At first glance, the difference between Eq. 14 and Eq. 31 seems unremarkable. Both suggest that the number of fitted points required to satisfy accuracy specifications and inference error risk requirements is represented by the product of p, the number of terms in the response model, and an elementary function of G_A. Recall that G_A is the "Accuracy Gain" representing the factor by which the accuracy associated with a minimum volume of fitted points, λ_0, must be increased by acquiring more data to achieve a tighter accuracy specification, λ.

It is instructive to compare minimum data volume requirements for the two cases: 1) when the truth surrogate is constant as in Eq. 14, and 2) when the truth surrogate is subject to random error as in Eq. 31. The former case was considered in an earlier example for which maximum acceptable inference error probabilities were specified as $\alpha = 0.05$ and $\beta = 0.01$, and the fitted data volume was scaled for an accuracy specification consisting of the 95% Least Significant Difference. In that example, G_A had a value of 1.515 per Eq. 16. For a 4th-order polynomial in three factors for which $p = 35$, a minimum of 81 fitted points was specified.

When we try to use the same $G_A = 1.515$ in Eq. 31we encounter a problem, because the numerator requires G_A to be less than 1 in order for the minimum volume of fitted data to be a positive number. This requires that λ be greater than λ_0, which means that bias errors smaller than λ_0 cannot be detected with specified inference error risks. In this example, λ is the 95% LSD, which is $2\sqrt{2}\sigma_0$. The smallest bias error that can be detected with Type I and Type II inference error risks of α and β, respectively, is $\lambda_0 = (z_\alpha + z_\beta) \times \sigma_0$, which for this example with $\alpha = 0.05$ and $\beta = 0.01$ is $4.286\sigma_0$. This is larger than the 95% LSD, so there is no volume of data that can provide sufficient precision to ensure that random errors superimposed upon predictions biased by as little as the 95% LSD would not fall within the $\pm\gamma$ tolerance interval, erroneously validating the biased model.

The effect of variance in the truth surrogate, then, is to establish a floor for the magnitude of bias errors that can be detected without incurring more than specified levels of inference error risk. Systematic LOF bias errors larger than $\pm 4\sigma_0$ as in this example, which can evade detection more than $\beta \times 100\%$ of the time, would not be acceptable in the typical example considered here.

The reason that variance in the truth surrogate causes so much trouble is that it limits options for reducing dispersion in the residual distribution by increasing data volume. Even if the fitted data sample were to approach

infinity so that the prediction component of the residual variance approached zero, this would have no effect on the truth surrogate variance. We can decrease the residual variance by reducing the truth surrogate variance through replication, as we will now consider.

1. *Data Volume Requirements with Replicated Truth Surrogates*

Consider a response model that is to be validated at a given site in the design-space using the mean of an *m-point* sample of truth surrogate replicates, where $m > 1$. Equation 22 then becomes

$$\sigma^2 = \frac{\sigma_0^2}{m} + \frac{p}{N}\sigma_0^2 \rightarrow \sigma = \sigma_0\sqrt{\frac{1}{m} + \frac{p}{N}} = \sigma_0\sqrt{\frac{N+mp}{mN}} \tag{32}$$

Insert Eq. 32 into Eq. 8:

$$\lambda = \left(z_\alpha + z_\beta\right)\sigma_0\sqrt{\frac{N+mp}{mN}} \tag{33}$$

Equation 33 represents the smallest systematic lack of fit modeling error that can be resolved in an experiment in which N points are fitted to a p-term polynomial response model, given a truth surrogate consisting of an m-point sample mean, if in assessing the residuals Type-I and Type-II inference error probabilities are not to exceed α and β, respectively. Let λ_{max} represent the largest acceptable tolerance for modeling error. Then given α, β, and p, we seek values of m and N for which $\lambda \le \lambda_{max}$.

It is convenient to express λ_{max} as a multiple of σ_0. That is, let $\lambda_{max} = \kappa\sigma_0$. Then given α, β, and p, we seek values of m and N for which the following inequality is valid:

$$\lambda = \left(z_\alpha + z_\beta\right)\sigma_0\sqrt{\frac{N+mp}{mN}} \le \kappa\sigma_0 \tag{34}$$

so the limit on λ is defined by

$$\left(z_\alpha + z_\beta\right)^2\left(\frac{N+mp}{mN}\right) = \kappa^2 \tag{35}$$

or

$$m\kappa^2 N = N\left(z_\alpha + z_\beta\right)^2 + mp\left(z_\alpha + z_\beta\right)^2 \tag{36}$$

Solving for N:

$$N = \frac{mp\left(z_\alpha + z_\beta\right)^2}{m\kappa^2 - \left(z_\alpha + z_\beta\right)^2} \tag{37}$$

Divide the numerator and denominator of Eq. 37 by $(z_\alpha + z_\beta)^2$:

$$N = \frac{mp}{m\left(\frac{\kappa}{z_\alpha + z_\beta}\right)^2 - 1} \tag{38}$$

Eq. 38 is then

$$N = \left(\frac{m}{\frac{m}{G_A^2} - 1} \right) p = \left(\frac{m}{\frac{m - G_A^2}{G_A^2}} \right) p \tag{39}$$

or

$$N = \left(\frac{m G_A^2}{m - G_A^2} \right) p \tag{40}$$

The term in parentheses in Eq. 38 is just $1/G_A = \lambda/\lambda_0$, which can be seen by multiplying the numerator and denominator of that term by σ_0 and noting that $\lambda = \kappa\sigma_0$ and $\lambda_0 = (z_\alpha + z_\beta)\sigma_0$.

Equation 40 reveals that in order for N to be positive, the number of truth surrogate replicates per validation site must exceed the square of the Accuracy Gain Factor, $G_A = \lambda_0/\lambda$:

$$m_{\min} > G^2 = \left(\frac{\lambda_0}{\lambda} \right)^2 \tag{41}$$

Return now to the previous example, in which $\alpha = 0.05$ and $\beta = 0.01$ so that $z_\alpha = 1.960$ and $z_\beta = 2.326$, and for which λ = the 95% LSD $= \kappa\sigma_0$, where $\kappa = 2\sqrt{2}$. For an unreplicated truth surrogate it was not possible to resolve a bias error this small with $(1 - \beta) \times 100\%$ confidence because of the noise floor established by the variance in the truth surrogate. Equation 41 represents the minimum number of truth surrogate replicates required to reduce the variance in the mean of those replicates to a level that would ensure that this bias error could be resolved with acceptable inference error risk. That is, this is the number of truth surrogate replicates required to ensure that random prediction error occurring about a model biased to this degree would not generate residuals small enough to erroneously validate the model. Inserting values from this example into Eq. 41:

$$m_{\min} > G^2 \rightarrow m_{\min} > \left(\frac{\lambda_0}{\lambda} \right)^2 = \left(\frac{z_\alpha + z_\beta}{\kappa} \right)^2 = \left(\frac{z_\alpha + z_\beta}{2\sqrt{2}} \right)^2 = \frac{1}{2}\left(\frac{z_\alpha + z_\beta}{2} \right)^2 \tag{42}$$

so:

$$m_{\min} > \frac{1}{2}\left(\frac{1.960 + 2.326}{2} \right)^2 = \frac{18.370}{8} = 2.296 \tag{43}$$

or, rounding to the next highest integer, $m_{min} = 3$.

To review, we assume that we have fitted a sample of data sufficiently large to ensure that there is no more than an $(\alpha \times 100)\%$ probability of a Type-I inference error at a given validation site. In this example, the truth surrogate at that site must consist of the mean of at least three genuine replicates to also ensure that if a systematic LOF bias error is greater than or equal to a specified limit (in this example, λ = the 95% LSD), there will be no more than a $(100 \times \beta)\%$ chance of random error causing the model to be erroneously validated due to a Type II inference error. This would occur if a biased model prediction were to fall within $\pm\gamma = z_\alpha\sigma$ of the truth surrogate due to random error.

2. *Optimum Number of Truth Surrogates per Validation Site*

In the example just considered, if one uses the mean of a three-point sample as the truth surrogate, one can detect modeling errors as small as the 95% LSD while incurring no more inference error risk than $\alpha = 0.05$ and $\beta = 0.01$ for Type I and Type II inference errors, respectively. Note, however, that Eq. 37 reveals a tradeoff between the number of truth surrogate replicates, m, and the number of fitted data points, N. The former determines the variance in the truth surrogate while the latter determines the response model prediction variance, so increasing one

decreases the other and conversely. Insert $m = 3$ into Eq. 37 with $p = 35$, $z_\alpha = 1.960$, and $z_\beta = 2.326$ from the current example:

$$N = \frac{3 \times 35 \times (1.960 + 2.326)^2}{(3 \times 8) - (1.960 + 2.326)^2} = 343.3 \approx 344 \tag{44}$$

In the earlier example in which the truth surrogate was constant, $N = 81$ points was sufficient to ensure that errors the size of the 95% LSD could be detected within the $\alpha = 0.05$ and $\beta = 0.01$ inference error risk limits. A constant truth surrogate; that is, one with a variance of zero, can be regarded as the mean of an m-point sample in the limit as m approaches infinity. From Eq. 37:

$$\lim_{m \to \infty} N = p\left(\frac{z_\alpha + z_\beta}{\kappa}\right)^2 \tag{45}$$

Inserting values from the current example:

$$\lim_{m \to \infty} N = \frac{35 \times (1.960 + 2.327)^2}{8} = 80.4 \to 81 \tag{46}$$

which demonstrates that the earlier result is a special case of the more general expression given in Eq. (37).

Compared to the 81 fitted points required to satisfy tolerance and inference error risk constraints when the truth surrogate is a constant, it seems excessive that 344 fitted points would be required to satisfy the same constraints when the truth surrogate is a random variable consisting of the mean of three truth surrogate replicates. One might wish to consider a few more truth surrogate replicates if it would significantly reduce the number of fitted points required to achieve the stated quality specifications. Figure 7 is a plot of Eq. 37 for a response surface model with $p = 35$ terms. It displays the tradeoff between the number of truth surrogate replicates, m, at a given validation site and the corresponding minimum number of fitted points, N, required to resolve a systematic error of size $\lambda = \kappa\sigma_0$ with inference error risk tolerance levels of $\alpha = 0.05$ and $\beta = 0.01$. Here, the systematic error tolerance, λ, is the 95% LSD for which $\kappa = 2\sqrt{2}$.

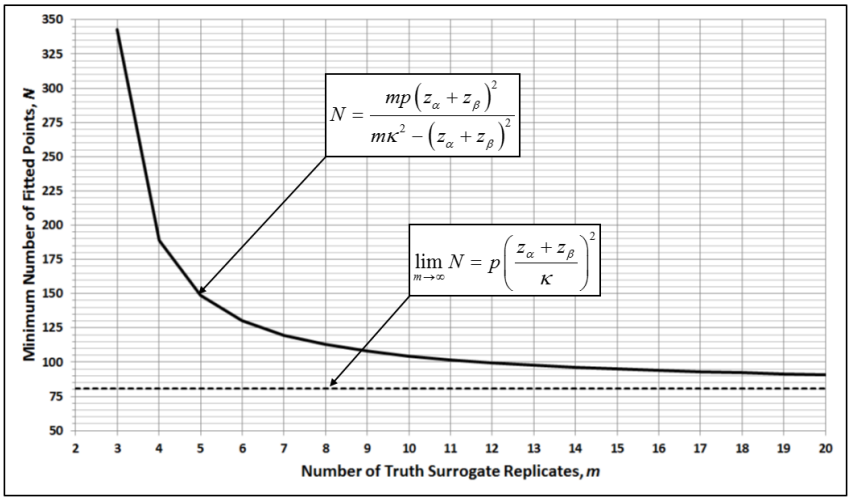

Figure 7 Tradeoff between truth surrogate replicates and minimum fitted sample size. *$\alpha = 0.05$, $\beta = 0.01$, λ = 95% LSD so $\kappa^2 = 8$, $p = 35$*

The sample size required to resolve a given systematic error with specified inference error probabilities decreases as the number of truth surrogate replicates increases, approaching as an asymptotic limit the sample size

requirement for a truth surrogate with no variance. For the current example in which α and β are 0.05 and 0.01, respectively, at least three truth surrogate replicates are necessary to resolve a systematic error, λ, as small as the 95% LSD, but if there are only three replicates we have seen that the minimum number of fitted points must be 344. By Eq. 37 or Fig. 7, adding only one more truth surrogate replicate (changing m from 3 to 4) reduces the fitted sample size requirement from 344 to 189, a reduction of 155 points. Each new truth surrogate replicate reduces the minimum volume of data that must be fitted in order to resolve a given systematic error within specified inference error limits, but the rate of reduction decreases monotonically as the number of replicates increases.

Let us assume that all validation sites have the same number of truth surrogate replicates, m, and that there are S such validation sites. Specifying one more truth surrogate replicate at each validation site therefore adds S points to the total test matrix. If this results in a corresponding reduction in fitted points that is greater than S, then adding that truth surrogate replicate at each validation site is cost-effective. A point of diminishing returns occurs when there are already so many replicates that adding one more at all S validation sites would increase the total replicates by more than the reduction it would facilitate in fitted points. Figure 8 illustrates this tradeoff.

Figure 7 is augmented in Fig. 8 by the addition of two more curves. As in Fig. 7, the solid curve shows how the number of fitted points, N, decreases as the truth surrogate variance is reduced by adding replicates at each site. The dashed curve shows how the total number of truth surrogate replicates across S validation sites, mS, increases with m, the number of replicates per site. Each new truth surrogate replicate adds S points to the total, where $S = 20$ in this example. The curve with markers is the sum of the first two curves, representing the total sample size, $T = N + mS$.

The m-axis starts with $m = 3$ in this example because it is not possible to resolve a systematic lack of fit error as small as the 95% LSD with Type I and Type II inference error probabilities no greater than $\alpha = 0.05$ and $\beta = 0.01$ respectively, as noted above. For $m = 3$ truth surrogate replicates at each of $S = 20$ validation sites, 60 points are added to the 343 fitted points needed to ensure that the 95% LSD can be resolved with prescribed inference error probabilities, for a total of 403 points.

Adding one more truth surrogate replicate changes m from 3 to 4, which increases the total number of truth surrogate replicates by 20 from 60 to 80, but decreases the required number of fitted points by 154 from 343 to 189. This would result in a net reduction of 134 points, with attendant reductions in wind-on minutes and associated data acquisition costs.

Adding a fifth truth surrogate replicate at each validation site increases the number of truth surrogates by 20 from 80 to 100, but reduces the requirement for fitted points from 189 to 149 for a decrease of 40 points—a benefit that is twice the "cost" of the 20 extra truth surrogate replicates. The total of fitted points plus truth surrogate replicates required to resolve a 95% LSD with Type I and Type II inference error probabilities no greater than $\alpha = 0.05$ and $\beta = 0.01$ respectively, drops from 269 to 249 when the number of truth surrogate replicates at each validation site increases from 4 to 5. Further increases in m continues to reduce the required number of fitted points, but the cost in additional truth surrogate replicates is progressively greater than the reduction in fitted point as m increases, suggesting that $m = 5$ is the optimum number of replicates in this example, as indicated in Fig. 8.

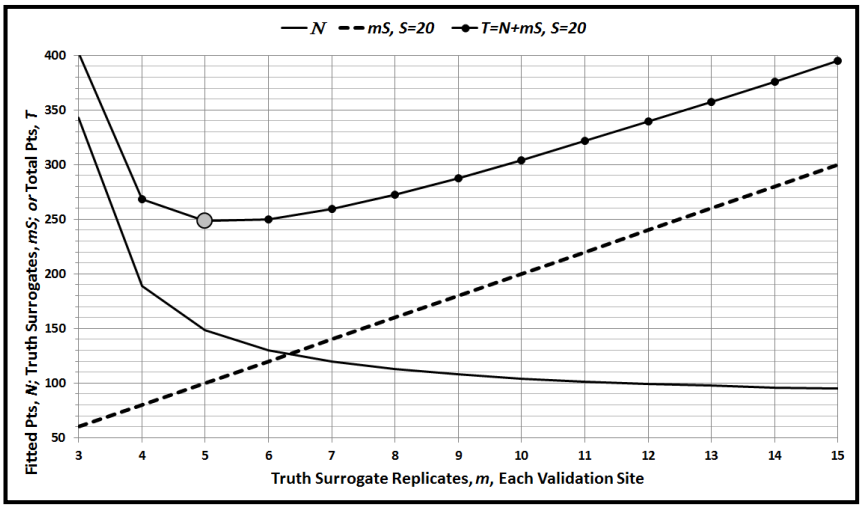

Figure 8 Total sample size, *T*, is the sum of mS truth surrogate replicates, plus N fitted data points: $\alpha = 0.05$, $\beta = 0.01$, $\lambda = 95\%$ LSD so $\kappa^2 = 8$, $p = 35$

Let m_o represent the optimum number of truth surrogate replicates at each validation site. This is the value of m that corresponds to the smallest total data volume needed to resolve a given effect within specified inference error risk limits. This total includes all fitted points and all validation replicates.

To compute m_o, take the derivative of Eq. 40 with respect to m.

$$\frac{dN}{dm} = -\left(\frac{G_A^2}{m - G_A^2} \right)^2 p \tag{47}$$

This derivative represents how many fewer points are needed to fit the response model within given tolerance and inference error specifications when one more truth surrogate replicate is added at each validation site in the design space. If an additional truth surrogate replicate is added to a sample of m such replicates at a given validation site, Eq. 47 describes the corresponding change (reduction, because of the minus sign) in the minimum number of points that must be fitted in a measurement environment with a standard random error of σ_0, in order to resolve a systematic response modeling lack of fit error of magnitude $\lambda = \kappa\sigma_0$, without Type I and Type II inference error probabilities exceeding specified limits α and β, respectively. Note in Eq. 47 that the benefit of fewer fitted points decreases rapidly—as the square of the number of truth surrogate replicates, m.

Adding a truth surrogate replicate at each of S validation sites adds S points to the total volume of acquire data. If this reduces the required number of fitted points, N, by more than S, then the total data volume will be reduced. The optimum number of truth surrogate replicates corresponds to the following condition:

$$S \le \left| \frac{dN}{dm} \right| \tag{48}$$

The equality holds in Eq. 48 when $m = m_0$. Combine Eq. 47 and Eq. 48 to compute m_o.

$$\left(\frac{G_A^2}{m_o - G_A^2} \right)^2 p = S \tag{49}$$

Solve for m_o:

$$m_o = G_A^2 \left(1 + \sqrt{\frac{p}{S}} \right) \tag{50}$$

By Eq. 50, the optimum number of truth surrogate replicates at each validation site depends on the number of terms in the model, p, and the number of validation sites, S, as well as the Accuracy Gain Factor, G_A. The Accuracy Gain Factor depends on the standard error of the measurement environment's irreducible random error, the acceptable probabilities for Type I and Type II inference errors when evaluating the residuals, and the smallest systematic lack of fit bias error in the response model, λ, that is too large to tolerate.

3. *Minimum Sum of Fitted Plus Validation Points*

Equation 40 represents the minimum number of fitted data points with a random standard error of σ_0 that are needed to achieve enough precision to assure a given accuracy specification, λ, without incurring more than prescribed levels of Type I or Type II inference error when the fitted model is assessed by analyzing residuals at validation sites not used to fit the model. It is a function of the number of truth surrogate replicates used at each of the validation sites. Equation 50 represents how many such replicates are needed at each validation site to optimize the tradeoff between decreases in fitted data and increases in validation data. Insert Eq. 50 into Eq. 40 to estimate the smallest volume of fitted data required when the optimum number of truth surrogate replicates, m_o, is acquired at each validation site:

$$N\big|_{m=m_o} = N_0 = \left[\frac{G_A^2\left(1+\sqrt{\frac{p}{S}}\right)G_A^2}{G_A^2\left(1+\sqrt{\frac{p}{S}}\right)-G_A^2} \right] p \tag{51}$$

which reduces to

$$N_0 = G_A^2\left(p+\sqrt{pS}\right) \tag{52}$$

Let T_0 represent the total volume of data to be acquired, including fitted points and validation points. It will be the sum of N_0 from Eq. 52 plus m_o times the number of validation sites, S:

Let T_o represent the total volume of data to be acquired. It consists of N_0, given in Eq. 52, which is the minimum number of fitted points needed to ensure that an effect as small as λ can be detected with inference error risks no greater than α or β; plus m_oS, where S is the number of validation sites and m_0 is the optimum number of truth surrogate replicates per validation site, given in Eq. 50.

$$T_0 = N_0 + m_o S \tag{53}$$

Insert Eq. (52) and Eq. (50) into Eq. (53):

$$T_0 = G_A^2\left(p+\sqrt{pS}\right) + G_A^2\left(1+\sqrt{\frac{p}{S}}\right)S \tag{54}$$

This reduces to

$$T_0 = G_A^2\left[\left(p+S\right)+2\sqrt{pS}\right] \tag{55}$$

The quantity $p + S$ in Eq. 55 represents the absolute fewest data points that can be acquired in an RSM experiment in which a p-term response model is fitted and then assessed at S validation sites. Extract this term from the square bracket of Eq. 55:

$$T_0 = \left(p+S\right)G_A^2\left[1+\left(\frac{2}{p+S}\right)\sqrt{pS}\right] \tag{56}$$

The quantity 2/(p+S) is just the reciprocal of the arithmetic mean of p and S. The square root term is just the geometric mean of p and S. Let $\mu_a(p,S)$ and $\mu_g(p,S)$ represent these two means, respectively. Then Eq. 55 becomes:

$$T_0 = \left(p+S\right)G_A^2\left[1+\frac{\mu_g\left(p,S\right)}{\mu_a\left(p,S\right)}\right] \tag{57}$$

IV. Quality Assurance

Inference error probabilities that play such an important role in the scaling process also have a significant and largely unanticipated effect on quality assurance. The quality of a response model has traditionally been assessed by examining residuals—differences between model prediction and some unbiased estimate of the true response at various sites throughout the design space. All of the internal information about the quality of a response model; that is, all information that is available from data acquired in the experiment without regard to external information about the test article and its expected behavior, is completely contained in the residuals. Various standard and reliable tests

of the residuals typically entail plotting them in different ways. For example, the residuals might be plotted on normal probability paper to test if they are normally distributed, as would be expected if differences between measurements and model predictions are attributable to random experimental error without any significant systematic lack of fit bias errors due to poor model formulation. Such residuals fall on a nominally straight line in a normal probability plot. The residuals are also commonly plotted against predicted response levels, against levels of each factor, against time, and in any other way that might be expected to reveal patterns in the residuals. It is generally the absence of patterns in such plots that indicate an adequate response model.

A. The Critical Binomial Number Test of Response Model Adequacy

We now describe a test of model adequacy that is somewhat more quantitative than standard residual plotting methods. It assumes that model predictions have been made for a number of validation sites randomly distributed throughout the design space, where data have been acquired but not used to fit the response model that is being evaluated. The residuals from these validation sites are regarded as a series of Bernoulli trials with the customary pass/fail binary outcome possibilities. If the model predicts adequately at a given validation site it is said to have passed that Bernoulli test, else it is said to have failed it. We begin by defining an "adequate prediction" as one that results in a residual that is within some prescribed tolerance. As we will see, this seemingly innocuous definition of prediction adequacy features an assumption that is not true in the presence of inference error risk. Ironically, the better the model fits the data, the less likely this assumption is to be true. We will address this assumption presently, but will begin by first neglecting inference error risk.

Let us assume that data have been acquired at S validation sites, and that there is an expected per-trial success probability of p_s at each of these sites. The probability of failure is p_f, which is of course, $1 - p_s$. We expect the success probability to be no less than $1 - \alpha$, where α is the maximum Type-I inference error risk for which the experiment has been scaled; say 0.05. That is, we expect $(1 - \alpha) \times 100\%$ of the validation-site residuals to be within tolerance in the absence of significant lack of fit, with no more than $\alpha \times 100\%$ of them out of tolerance due to ordinary random experimental error. A significantly higher failure rate would be interpreted as evidence of the systematic lack of fit bias error that would characterize an inadequate model.

Figure 9 displays the binomial distribution for 100 trials with a per-trial success probability of 0.95. The Critical Binomial Number (CBN) for a significance of 0.01 is 89. This means that if the per-trial success probability is in fact 95%, there is less than a 1% chance that there will be fewer than 89 successes in 100 trials. If there are 89 or fewer successes, we can infer that the per-trial success probability is less than 95%, with no more than a 1% chance of an inference error.

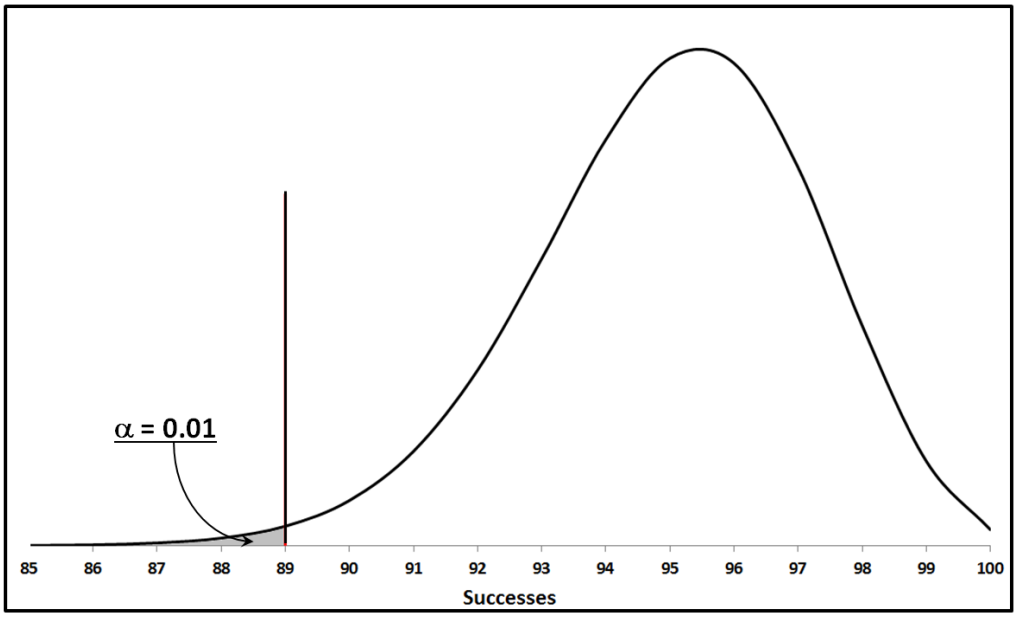

Figure 9. Binomial Distribution for 100 trials with $p_s = 0.95$. The *CBN for α=0.01 is 89.*

The test of model adequacy simply requires that the number of successful Bernoulli trials exceed the Critical Binomial Number. This appears to have a 99% chance of occurring in this example if the per-trial success probability is in fact 95% as we expect; that is, if there actually is a 95% probability that a measurement made at a given site will lie within the model's 95% prediction interval for that site.

Notwithstanding the apparent ease with which an adequate model ought to be able to pass this test, experience has shown that this happens much less often than expected, even for models that pass all other tests of model adequacy and give every indication of being perfectly valid response models. This apparent mystery can be explained by what is known in forensic science as the Prosecutor's Fallacy.

The prosecutor commits this fallacy when he demonstrates that there is a high probability that the defendant's fingerprints will be on the murder weapon if he is guilty, and then introduces evidence to show that his fingerprints are in fact on the weapon. The fallacy is in assuming there is relevance in the high probability that the defendant's fingerprints would be on the weapon *given that he is guilty*. To obtain a legitimate guilty verdict, the prosecutor actually needs to demonstrate a high probability of guilt *given that the defendant's fingerprints are on the weapon,* not the reverse of this. There may be some other explanation for the fingerprints—perhaps the defendant handled the weapon before or after the crime—or reliable witnesses may place him somewhere else at the time of the crime. The conditional probability of "A" given "B," P(A|B), is not generally the same as the conditional probability of "B" given "A," P(B|A), and to assume so is to commit the Prosecutor's Fallacy.

We commit the Prosecutor's Fallacy during the Critical Binomial Number Test (or any time we assess model prediction adequacy by counting out-of-tolerance residuals at validation sites) when we note the high probability that a residual will be out of tolerance given that the model prediction is biased, then note that a residual is out of tolerance at a given validation site and conclude from this that the model fits poorly there. We are not interested in the high probability that a residual will be out of tolerance at some site in the design space *given that the model is biased there*. We want to know the reverse of this—the probability that the model prediction is biased at some site *given that the residual is out of tolerance there*.

Consider a good model for which there is only a small probability, ε, that predictions are biased at any given site in the design space. The model predicts adequately *at* $(1-\varepsilon) \times 100\%$ of the sites. Random error is expected to cause $\alpha \times 100\%$ of the residuals to be out of tolerance even where the model is unbiased, so $\alpha \times (1-\varepsilon) \times 100\%$ of the validation sites will feature residuals that are out of tolerance only because of random experimental error, and not because of any response model bias. Let us further assume that the experiment has been scaled to ensure that there is only a small probability, β, of erroneously validating a biased model, so there is a probability of $(1-\beta) \times 100\%$ that a residual will be out of tolerance where the model truly is biased. This means that $\varepsilon \times (1-\beta) \times 100\%$ of the validation sites will feature residuals that are out of tolerance because the model is in fact biased.

The residual at a randomly selected validation site can thus be out of tolerance whether the model is biased or not. The probability of an out of tolerance residual is $\varepsilon(1-\beta)$ when the model is biased and $\alpha(1-\varepsilon)$ when the model is unbiased so there is a total probability of $\varepsilon(1-\beta)+\alpha(1-\varepsilon)$ that a residual will be out of tolerance, indicting the response model either rightly or wrongly.

Let "*A*" represent the case in which the model is biased and let "*B*" represent the case in which the residual is out of tolerance. Then $P(A|B)$ is the conditional probability that the model is biased, given that the residual is out of tolerance:

$$P(A|B) = \frac{\varepsilon(1-\beta)}{\alpha(1-\varepsilon)+\varepsilon(1-\beta)} \tag{58}$$

Expanding and rearranging terms:

$$P(A|B) = \frac{\varepsilon-\varepsilon\beta}{\alpha-\alpha\varepsilon+\varepsilon-\varepsilon\beta} = \frac{\varepsilon-\varepsilon\beta}{\alpha+\varepsilon-\varepsilon(\alpha+\beta)} \tag{59}$$

Assuming that the model has been fitted to a data sample scaled to ensure small inference errors, and further assuming that the residuals provide evidence of a generally adequate model (so small α, β, and ε), then we can neglect second-order terms so that Eq. 59 becomes, to a good approximation:

$$P(A|B) \approx \frac{\varepsilon}{\alpha + \varepsilon} \tag{60}$$

The effect of a non-zero Type-I inference error probability, α, is to ensure that $P(A|B)$ is not 100% as one might expect it to be. That is, because this inference error probability is not zero, an out-of-tolerance residual is not a perfect indicator that the model predicts inadequately at a given site in the design space. The reason this is so is that, neglecting bias errors in the measurement system and interactions represented by the small second-order terms in Eq. 59, there is only one way for the *model* to be biased (an inadequate model formulation, with probability ε that this occurs at a randomly selected site in the design space), but there are *two* ways for the *residual* to be out of tolerance (*random error* with probability α, and a systematic LOF *bias error* due to inadequate model formulation with probability ε).

Figure 10 shows how the probability of an *unbiased* response model prediction increases with $1-\varepsilon$, the probability of an unbiased model, *notwithstanding the fact that the residual is out of tolerance*. Plots are presented for sample sizes scaled to ensure three different Type-I inference error probabilities. A Type-II inference error probability of $\beta = 0.01$ is assumed in all three cases, but the probability of a false indictment of the response model at a randomly selected design-space site is largely independent of β for small β. See Eqns. 59 and 60.

The vertical line at $1 - \varepsilon = 0.95$ in Fig. 10 cuts through the three curves at a point representative of a model that is likely to be regarded as adequate, given a 95% probability that it produces unbiased response estimates for any randomly selected site in the design space. This figure reveals that for response models generally, but especially for models that generate unbiased predictions in a large fraction of the design space ("good" models), an out-of-tolerance residual is not an entirely reliable indicator of model inadequacy. The greater the probability that ordinary random error will cause a residual to be out of tolerance (that is, the larger alpha is), the more likely it is that an out of tolerance residual falsely indicts the model. For models that generate unbiased predictions in a large fraction of the design space (the right side of Fig. 10), the chance of a false indictment is especially high; for a typical value of $\alpha = 0.05$ .it exceeds 50% for models that predict adequately in 95% or more of the design space. This is because the better the model, the less likely it is that out of tolerance residuals will be due to model imperfections and thus the more likely it is that they will simply reflect imperfections in the measurement environment; i.e. experimental error.

In the limit where $\varepsilon = 0$ so that $1 - \varepsilon = 1$ (no lack-of-fit bias), there is by definition a 100% probability that the model is adequate at every site, including 100% of the sites where residuals may be out of tolerance. At those sites, out-of-tolerance residuals can only be attributed to experimental error, and not to model imperfections.

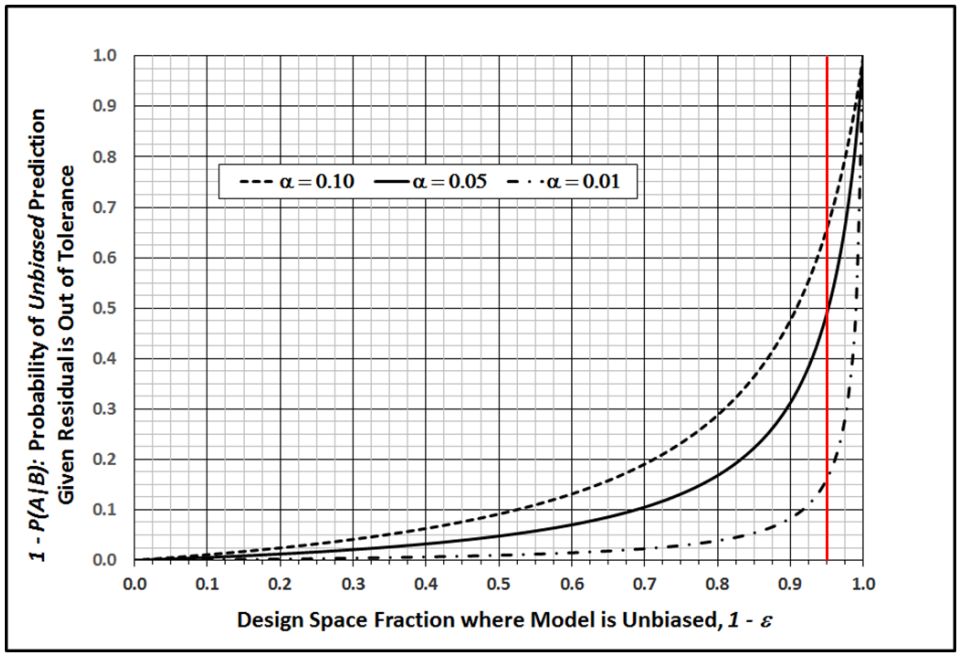

Figure 10. Probability that a response model prediction is actually unbiased at a given design space site, even though the residual is out of tolerance at that site.

B. Tolerance Metric for Response Model Adequacy Assessment

It has been demonstrated above that the probability of a residual being out of tolerance, either due to random experimental error or due to systematic LOF bias error, is $\varepsilon(1-\beta)+\alpha(1-\varepsilon)$. We therefore expect the same fraction of randomly selected validation sites to generate residuals that are out of tolerance. Let P_f represent this percentage of failed validation Bernoulli trials. Then

$$P_f = \varepsilon(1-\beta)+\alpha(1-\varepsilon) \tag{61}$$

Solving for ε:

$$\varepsilon = \frac{P_f - \alpha}{1-\alpha-\beta} \tag{62}$$

Equation 62 indicates that the fraction of the design space in which the response model actually *is* biased, ε, is the fraction of the design space, P_f, in which the response model only *appears* to be biased (due to out-of-tolerance residuals), corrected for Type-I and Type-II inference error probabilities α and β respectively. It is only when both α and β are zero that $\varepsilon = P_f$.

For the purpose of quantifying the suitability of a response model, it may make more sense to quantify the *complement* of ε; namely, $1-\varepsilon$, as follows:

$$\varepsilon = \frac{P_f - \alpha}{1-\alpha-\beta} \rightarrow 1-\varepsilon = 1-\frac{P_f-\alpha}{1-\alpha-\beta} = \frac{1-\alpha-\beta-(P_f-\alpha)}{1-\alpha-\beta} = \frac{(1-P_f)-\beta}{1-\alpha-\beta} \tag{63}$$

Let $P_s = 1 - P_f$: represent to percentage of successes in a series of Bernoulli validation trials. Then Eq. 63 becomes

$$1-\varepsilon = \frac{P_s-\beta}{1-\alpha-\beta} \tag{64}$$

The quantity $1-\varepsilon$ in Eq. 64 is an estimate of the fraction of the design space in which the response model exhibits no significant LOF bias. It is P_s, corrected for Type-I and Type-II inference error probabilities α and β respectively, which are known because they were specified as a prerequisite for the scaling the experiment. We compare $1-\varepsilon$ with some specified minimum standard to assess whether the response model meets quality specifications. For example, assume a model that predicts without significant bias error in 95% of the design space is to be regarded as adequate. Assume further that a sufficient volume of data has been acquired to ensure maximum Type-I and Type-II inference error probabilities of $\alpha = 0.05$ and $\beta = 0.01$, respectively, and that only $P_s = 92\%$ of the validation-site residuals are within tolerance. Absent any correction for inference error, it would appear as if the model is inadequate, since only 92% of the residuals are within tolerance, which is less than the 95% criterion. However, use the above equation to correct P_s for α and β:

$$1-\varepsilon = \frac{P_s-\beta}{1-\alpha-\beta} = \frac{0.92-0.01}{1-0.05-0.01} = 96.8\% \tag{65}$$

We see, then, that an original estimate suggesting that the response model is unbiased over only 92% of the design space, which is unacceptable, has been revised to an acceptable value of 96.8% by taking inference error risk into account.

V. Summary

This paper describes how to estimate data volume requirements for a Modern Design of Experiments (MDOE) response surface modeling wind tunnel test, a process known as *scaling*. It also describes two quantitative methods for assessing response model adequacy that augment conventional methods in which patterns in plotted residuals are used to reveal possible problems. The first of these new methods entails estimating the fraction of the design space over which the model predicts adequately, and the second is based on a Critical Binomial Number analysis of validation residuals. The scaling process and model validation methods are summarized here, with details presented in the main body of the paper.

A. Scaling

1. Document the standard random error of the measurement environment, σ_0. This should be available from facility test personnel, or it can be estimated pre-test by calculating the standard deviation of a statistically significant number of genuine replicates for each response of interest.
2. Specify the number of design space sites, *S*, where the adequacy of candidate model response estimates are to be assessed by comparing them to independent direct measurements that will serve as unbiased surrogates for the true response.
3. Specify the accuracy tolerance, λ, and maximum acceptable probabilities of inference error risk, α and β.
    a. The accuracy tolerance, λ, represents the largest systematic lack of fit bias error that is still small enough to be acceptable in a response estimate.
    b. Inference errors occur when an adequate model is declared inadequate (a Type-I or alpha error, also called the *significance* of the test) and when an inadequate model is erroneously validated (a Type-II or beta error). Maximum acceptable probabilities for Type I and Type II inference errors are designated as α and β, respectively. A value of $\alpha = 0.05$ is a common industry convention, with β selected independently according to the experimenter's risk aversion. Since validating an inadequate model may have greater consequences that rejecting one that is adequate, the experimenter may wish to specify a value of β that is smaller than α.
4. Determine standard normal deviates, z_α and z_β for α and β using tabulated values or statistical software.
5. Compute λ_0, the tightest accuracy that can be achieved with specified inference error risk probabilities without any residual degrees of freedom:

$$\lambda_0 = \left(z_\alpha + z_\beta \right) \sigma_0$$

6. Compute the Accuracy Gain Factor, G_A, as follows:

$$G_A = \frac{\lambda_0}{\lambda}$$

7. Specify the order of the polynomial to be fitted, *d*, and the number of independent variables to be included, *k*. Compute *p*, the number of terms in the model including the intercept:

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

8. If $\lambda > \lambda_0$ so that $G_A < 1$, compute the minimum number of fitted points, *N*, required to satisfy accuracy and inference error risk specifications as follows:

$$N = p\left(\frac{G_A^2}{1 - G_A^2} \right)$$

9. For an accuracy specification $\lambda \leq \lambda_0$ so that $G_A \geq 1$, it will not be possible to fit a sample of data of any size that ensures such a tight accuracy tolerance can be validated within specified inference error risks. This

limit can be circumvented, however, by reducing the residual variance so that $\lambda \le \lambda_0$ for arbitrarily small λ. To do so requires that the irreducible random error variance associated with a single point truth surrogate be replaced with the smaller variance associated with a mean of m truth surrogate replicates. The minimum number of replicates, m, required at each validation site to ensure that $G_A \ge 1$, is as follows:

$$m_{\min} > G_A^2$$

and the minimum volume of fitted data corresponding to m replicates of the truth surrogate is

$$N = \left(\frac{mG_A^2}{m - G_A^2} \right) p$$

10. The more the truth surrogate is replicated at each validation site, the fewer the number of points that will have to be fitted to satisfy accuracy and inference error risk specifications, but the more validation points that will have to be acquired. For S validation sites, mS validation points will be needed. The number of replicates that optimizes this tradeoff by minimizing the sum of fitted plus validation points is m_o, computed as follows:

$$m_o = G_A^2 \left(1 + \sqrt{\frac{p}{S}} \right)$$

11. For resource planning purposes, the minimum necessary volume of data is then $T_0 = N + m_0 S$, which can be estimated as follows:

$$T_0 = \left(p + S \right) G_A^2 \left[1 + \frac{\mu_g \left(p, S \right)}{\mu_a \left(p, S \right)} \right]$$

where $\mu_g(p,S)$ and $\mu_a(p,S)$ are the geometric and arithmetic means of p and S, respectively. If the design space has been subdivided into subspaces to reduce the order of models required to span them, this number should be adjusted accordingly.

B. Fraction of Design Space for which the Model Predicts Adequately

1. Compute the standard error of the residual variance as follows:

$$\sigma = \sigma_0 \sqrt{\frac{N + mp}{mN}}$$

2. Compute $\gamma = z_\alpha \sigma$, the half-width of the $(1 - \alpha) \times 100\%$ prediction interval for response model residuals.
3. Compute P_s, the percentage of residuals from the S validation sites that are within $\pm\gamma$ of 0.
4. Estimate the fraction of the design space over which the model predicts adequately, $1 - \varepsilon$, by revising P_s for inference error risk as follows:

$$1 - \varepsilon = \frac{P_s - \beta}{1 - \alpha - \beta}$$

5. Assess the adequacy of the response model according to whether $1 - \varepsilon$ is large enough to be acceptable; say 0.95.

C. Critical Binomial Analysis

1. Estimate the ratio of the number of residuals that are legitimately out of tolerance due to biased response predictions, to the number that are out of tolerance either because of model bias or because of random error:

$$\frac{\varepsilon(1-\beta)}{\alpha(1-\varepsilon)+\varepsilon(1-\beta)}$$

2. Use this ratio to correct the number of Bernoulli trials that were successful during the model validation process to include residuals that were within tolerance plus residuals that were only out of tolerance because of random error, and not because of biased response predictions by the model under evaluation.
3. Compare this corrected number of successful model confirmations to the Critical Binomial Number associated with a binary distribution for which the per-trial probability of success is expected to be $1 - \alpha$.

References

[1]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.

[2]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

[3]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.

[4]DeLoach, R. "The Objective of Aerospace Ground Testing: An Alternative Perspective". Article for AIAA Ground Testing Technical Committee Newsletter. June, 2002.

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

[6]DeLoach, R. "Productivity Enhancement and Quality Assurance in Aerospace Testing with the Modern Design of Experiments" Invited Keynote Address, First International Aerospace Symposium of South Africa, Johannesburg, South Africa, November 2009.

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

[8]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.

[9]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.

[10]DeLoach, R. (2002) "Applications of the Modern Design of Experiments at NASA Langley Research Center (Invited)" Proceedings of the American Statistical Association, Section on Physical and Engineering Sciences [CD-ROM], New York, NY: American Statistical Association.

[11]DeLoach, R. "Formal Experiment Design as a Tool to Automate Aerospace Ground Testing (Invited)". Tenth Annual Spring Research Conference on Statistics in Industry and Technology. June 4-6, 2003 University of Dayton.

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