Response Surface Modeling Tolerance and Inference Error Risk Specifications: Proposed Industry Standards

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

This paper reviews the derivation of an equation for scaling response surface modeling experiments. The equation represents the smallest number of data points required to fit a linear regression polynomial so as to achieve certain specified model adequacy criteria. Specific criteria are proposed which simplify an otherwise rather complex equation, generating a practical rule of thumb for the minimum volume of data required to adequately fit a polynomial with a specified number of terms in the model. This equation and the simplified rule of thumb it produces can be applied to minimize the cost of wind tunnel testing.

I. Introduction

Direct operating costs and cycle time both increase with the volume of data acquired in a wind tunnel test. This suggests that low-cost wind tunnel testing would entail acquiring the smallest volume of data sufficient to accomplish the objectives of the test. The opposite approach, however, has been the standard practice in 20th-century and early 21st-century wind tunnel testing. Experimental aerodynamicists are generally inclined to acquire as much data in a given tunnel entry as resources permit. That is, rather than minimizing the cost of achieving a fixed objective, the typical approach is to maximize the volume of data that can be acquired within a specified budget. Under such circumstances the test exit strategy tends to be defined by the exhaustion of resources rather than by whether specific technical objectives have been satisfactorily achieved. The test ends when direct operating resources are consumed, or the scheduled tunnel occupancy time comes to an end. Whatever other virtues it may possess, a process that is explicitly designed to consume all available resources is not likely to be "low cost."

This focus on high-volume data collection stems from a collision between what is commonly perceived as the desired outcome of a wind tunnel test, and what is actually achievable with realistic resource constraints. The desired outcome is typically an aerodynamic database describing some test article, constructed by physically setting the tunnel to every combination of the independent variables (factors) of interest. Unfortunately, the number of potentially interesting factor combinations is often, for practical purposes, infinite. For example, a relatively modest six-factor wind tunnel test (angles of attack and sideslip, Mach number, Reynolds number, and deflections of flap and aileron, say), with only ten levels per factor, will have 10^6 = one million possible factor combinations. There may be time during a typical tunnel entry to set about 5000 of these, which is only 0.5% of the total.

Some factor combinations will be less interesting than others so there is generally some prioritization of the test matrix, but even if 90% of the design space is of no interest at all (a supposition that strains credulity), 5000 data points would only represent 5% of what is of actual interest. In this example, resource constraints would require that 99.5% of the entire design space (and even 95% of "the good stuff") be left on the table. This contrast between an ideal outcome and achievable reality is even greater in the not-uncommon case in which there are more than six factors under study, and/or more than 10 levels of each factor. It has been the author's experience over several years of making such calculations that only about 0.3% of the total number of possible factor/level combinations is ever actually set in a typical wind tunnel test.

The author has for some time advocated an alternative approach to experimental aeronautics that recognizes the limits imposed by how costly and how time consuming it is to make direct, physical measurements in a wind tunnel[1-9]. The essence of this approach is to make a relatively small number of costly, physical measurements, but to distribute them throughout the design space in such a way that they can be used to adequately estimate responses everywhere else within the design space. These response estimates are typically made by fitting low-order polynomials over limited ranges of the independent variables, much as is commonly done today to correct raw data

[1] Senior Research Scientist, NASA Langley Research Center, MS 238, 4 Langley Blvd, Hampton, VA 23681, Associate Fellow, AIAA.

for factor set-point errors. The quality of these estimates is assessed by how much they differ from physical response estimates made at the same combination of factor levels.

The planning for a wind tunnel test conducted in this way must include an estimate of the number of data points to acquire. The process of making this estimate is called "scaling" the experiment. The less risk that is acceptable in projecting a response at some design space location for which no physical measurement was made, the greater the volume of data that must be acquired. Likewise, the smaller the acceptable difference between projected and empirical response estimates at a given site in the design space, the greater the volume of data that must be available to make the projection.

Unfortunately, there is no industry consensus for the quality metrics and inference error risk standards that determine scaling requirements. Individual wind tunnel customers trained to acquire as much data as resources permit often have difficulty conceptualizing a data-minimal test strategy at all, much less the individual parameters needed to specify the fewest points needed to execute such a strategy. The purpose of this paper is therefore to review the parameters that determine scaling requirements in a wind tunnel test, and to propose a set of nominal values that can serve as generally acceptable defaults in the absence of any other customer specifications. The end result will be a relatively simple rule of thumb that relates data volume requirements to the number of terms in the polynomial response model used to fit the data.

This paper coalesces selected scaling concepts applied by the author elsewhere[2] to serve as a basis for proposing certain quality-related extensions of those concepts. Section II of this paper briefly reviews the derivation of a basic formula for estimating data volume requirements. Section III proposes certain criteria for tolerance and inference error risk. Section IV quantifies uncertainty reductions available through the application of selected quality assurance tactics in conjunction with the scaling standards proposed in Section III. A few concluding remarks are offered in Section V.

II. Scaling Fundamentals

Scaling a wind tunnel test begins by recognizing that while there may also be other specific objectives associated with the particular circumstances of a test, the general objective is always the same for tests of the type that are intended to produce a database of force and moments or pressures for various combinations of independent variables. The objective of such tests is to learn enough about the test article that it is possible to adequately predict its future behavior over the ranges of independent variables that were examined in the test.

The scaling function is particularly impacted by what we mean by "adequately." We have an intuitive understanding that adequate predictions will not be possible unless enough data are acquired, and while it is not intuitively clear precisely how many data points this entails, the need for an adequate database is keenly felt, even when it is not quantitatively defined. The result is a high-volume data collection imperative, in which the wind tunnel practitioner often feels compelled to ere on the side of the angels by acquiring as much data as possible while there is an opportunity to do so.

It is useful to give some thought to the characteristics of an "adequate" response prediction. Because of experimental error we do not expect any response estimate to perfectly represent "truth," but we can specify some range centered on the estimate within which the true response is expected to lie with some level of confidence. If such a range were centered on the true response, then adequate response estimates would fall within it a specified percentage of the time if they were replicated many times. So the concept of adequacy reduces simply to a specification of how small such an interval needs to be to satisfy our adequacy requirements, and how often we require our response estimates to fall within that interval. As an example of such an adequacy requirement, we might require estimates of drag coefficient to lie within plus or minus one drag count of the true drag, with 95% confidence.

The strategy for minimizing cost is to fit a regression model to a relatively small sample of data, which is then used to predict responses elsewhere in the design space. Resource savings are achieved because the cost of estimating responses throughout the design space with such a model is so much less than the cost of empirical estimates that would in some sense cover the design space. The model, of course, covers the design space with infinite resolution in that response estimates can be made for any site within the design space. Empirical response estimates can only be made at discrete sites, no matter how many are made. So in that sense, the response model provides superior coverage, as well as costing less.

We assess the adequacy of a response model by evaluating the residuals, or differences between empirical response estimates (direct measurements) and response estimates based on the regression model. We must infer from the size of the residuals whether the response model is adequate or not.

Note that for any given residual that we are considering, our inference might be right or it might be wrong, regardless of what we infer about the adequacy of the model. We might erroneously reject an adequate model because some measurement outlier resulted in an unacceptably large residual, or we might erroneously validate an inadequate model because experimental error resulted in an artificially small residual.

Figure 1 illustrates this evaluation process graphically. There are two probability distributions represented in this figure, only one of which can exist at a given time. The distribution on the left is centered on zero, illustrating how response model predictions might be distributed about some mean if there is no true difference between computational response estimates and "truth." We do not know where the truth actually lies, so we use an empirical response estimate as a surrogate. Thus, the distribution on the left represents how response model predictions would be distributed due to ordinary model prediction uncertainty, under a null hypothesis that no systematic difference exists between the response model's prediction and a measurement made at the same design space site.

The distribution on the right represents how response model predictions would be distributed if, instead of matching the data on average, they were displaced by an amount that is just barely too great to be acceptable by our adequacy criteria. That is, we say that response model predictions that differ from measurements by less than δ are adequate, but if the differ by δ or more, they are not.

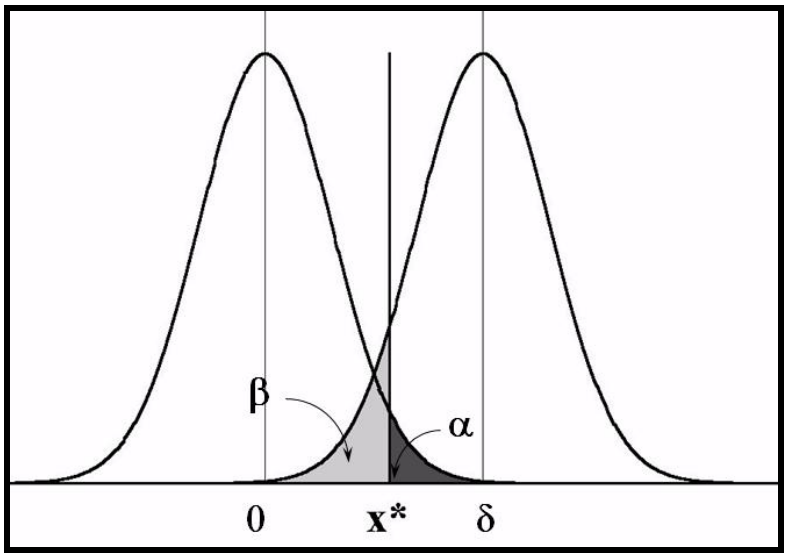

Figure 1. Risks of erroneously rejecting null and alternative hypotheses.

Because of the uncertainty in response model predictions, we cannot simply evaluate the adequacy of a given response prediction on the basis of whether it exceeds δ or not. Any residual that is to the right of δ in this figure is unlikely to have been drawn from the distribution on the left that corresponds to a model with no lack of fit. If the model prediction is being compared to a representative measurement (not an outlier), residuals that large would indicate an inadequate model. However, consider the more interesting case when a residual is greater than zero, but less than δ. We cannot simply assume the model is adequate because the residual is less than δ, because it could have been drawn from the distribution on the right, with some negative prediction error. That is, it could have been drawn from a distribution actually centered on a mean that is too great to be adequate. Note also that there is some non-zero probability, however small, that an exceptionally large residual is still drawn from the left-hand distribution, and is simply as large as it is due to an unlikely prediction error that just happens to be several standard deviations away from a true mean of zero.

The vertical line at x* is intended to resolve these ambiguities. Any residual that is large enough to lie to the right of x* indicates an inadequate model, and the null hypothesis that there is no systematic difference between empirical and measured response values at that design space site will be rejected for such cases. A residual that lies between 0 and x* will be interpreted as confirming the model. In that case we will reject an alternative hypothesis claiming that the true residual is at least of magnitude δ.

We will calculate where to place x* in a moment, but first notice that the dark shaded area under the left-hand distribution that is to the right of x*, labeled α, represents the probability of erroneously rejecting the null hypothesis by invalidating a model that is actually adequate. This can occur when a residual exceeds x* only because of random prediction error. Likewise, the lighter shaded area labeled β represents the probability of erroneously rejecting an

alternative hypothesis to the effect that the regression model predicts responses that differ from measured values by at least δ. It that case, the model actually would be inadequate, but the residual would pass muster simply because of experimental error.

We now express the distance of x* from each mean in multiples of the standard deviation. We say that x* is a distance of z_α standard deviations to the right of the left distribution's mean of zero, and it is a distance of z_β standard deviations to the left of the right distribution's mean of δ. These z values, called standard normal deviates, are tabulated in standard statistics tables for normal distributions, and depend only on α and β.

If we let σ_m be the standard error ("one sigma") for model prediction represented by the standard deviation of the distributions in Fig. 1, then we have

$$\delta = z_\alpha \sigma_m + z_\beta \sigma_m = \left(z_\alpha + z_\beta \right) \sigma_m \tag{1}$$

It can be shown[10] that while prediction uncertainty is site-dependent, the average uncertainty across all points used to fit the response model is as follows:

$$\sigma_m = \left(\sqrt{\frac{p}{n}} \right) \sigma \tag{2}$$

where *p* is the number of parameters in the response model and *n* is the number of data points used to fit the model. This result is true for any order of polynomial in any number of independent variables. Note that because there must be at least as many points fitted in the regression as there are parameters in the model, $n \geq p$ and therefore $\sigma_m \leq \sigma$. That is, the standard error of a response model prediction will never be greater than the standard deviation of an empirical estimate, and will generally be smaller to the extent that $n > p$. This reduction is attributable to a characteristic of response surface modeling called *hidden replication*, which causes the errors to partially cancel for points distributed above and below a well;-fitted curve faring through the data.

If we insert Eq. (2) into Eq. (1), we get

$$\delta = \left(z_\alpha + z_\beta \right) \left(\sqrt{\frac{p}{n}} \right) \sigma \tag{3}$$

Solving for *n*:

$$n = \left[\left(z_\alpha + z_\beta \right)^2 \left(\frac{\sigma}{\delta} \right)^2 \right] p \tag{4}$$

As noted earlier, a minimum of $n = p$ points is required to fit a polynomial response model with *p* terms, but Eq. (4) indicates that some multiple of p, represented by the term in brackets, is necessary to meet adequacy requirements. The bracketed term is a proportionality constant that is a function of four quantities. One is σ, the standard deviation in replicated response measurements, which can be obtained via replicates specified in the experiment design or can be estimated from the residuals of a well-fitted response model. It can also be estimated from historical data external to the experiment.

The other three parameters describe how good the response model is required to be. A "good" model in this context is one that reproduces physical response measurements adequately; that is, within a specified tolerance that is represented by δ in the data volume formula. Note that we cannot require the response model to predict *true* responses within some tolerance, as the true forces, moments, pressures, etc., experienced by the test article with a given combination of independent variable levels always remains unknown absent an infinite volume of data. The best we can realistically expect is that the response model will estimate measured responses within some tolerance, δ. If the physical measurements are a good surrogate for the true response, then response estimates based on the model will represent the true response well.

III. Specification of Tolerance and Inference Error Risk

Equation (4) quantifies the impact on data volume requirements of five quantities that dictate curve-fitting adequacy requirements. There must be a minimum of one unique data point for every term in the model so the data volume requirement is proportional to *p*, the number of such terms. The parameter count, or number of terms in a d^{th}-order polynomial in *k* factors, can be computed as follows:

$$p = \frac{(d+k)!}{d!k!} \tag{5}$$

For example, a 1st-order polynomial in one factor (a straight line) has 2!/1!1! = 2 parameters (a slope and an intercept) and so requires a minimum of two data points to fit. However, *p* is just the minimum number of points that would permit a particular polynomial to be developed. Such a minimal data volume does not speak to the adequacy of fit issues represented graphically in Fig. (1). We may say that there are *p location* decrees of freedom that define the value of a response estimate, but additional *dispersion* degrees of freedom are also necessary to address quality requirements. These are dictated by the bracketed term in Eq. (4).

There are four quantities in the bracketed term that address model fitting adequacy. Note that *n* is proportional to σ^2, the unexplained variance in the fitted data sample. This means that the poorer the measurement environment, the more data are required to fit a given polynomial with a given quality of fit.

The data volume requirement is also inversely proportional to the *square* of δ, the fitting tolerance. This means that fitting tolerance is a very sensitive determinant of data volume requirements in a response surface modeling experiment. There is a good opportunity here to save significantly on the cost of an experiment. If there is no practical reason in a particular application to require, say, a half drag count tolerance instead of a one-count tolerance (that is, if model predictions that agree with empirical estimates within one count are for all practical purposes acceptable), then designing the experiment for the one drag count tolerance you need instead of the half-drag count tolerance you might simply like to have can reduce data volume requirements by a factor of four. This is the difference between executing a test in one week instead of a month, for example. If the half-count precision is truly needed, of course the additional resources can be justified. But there is considerable potential for cost savings by designing to your actual requirements, but nothing substantially more stringent.

The bracketed term in Eq. (4) also contains the quantities α and β, which represent maximum tolerable inference error risk probabilities for erroneously rejecting an adequate model or erroneously validating an inadequate one from an analysis of residuals. Note that these are the only two inference errors we can make in evaluating the adequacy of a model, and fortunately, they are mutually exclusive. If we recognize that an adequate model is indeed adequate, or even if we have developed an inadequate model but are successful in detecting this, then we will not have made an inference error. In the former case, we have successfully finished our task. In the latter case we know we must fit a different model, perhaps requiring an augmentation of the data sample already fitted.

For simplicity, the inference error risk has been cast in terms of the *z* values that represent standard normal deviates. These assume that variance estimates are based on a statistically significant number of data points. Standard references such as Coleman and Steele[11] cite 10 data points as adequate for this assumption, and as few as 20 points would probably be considered luxurious by most experts. However, if fewer points are available to estimate unexplained variance, then the *z* values can be replaced by *t* values that are also tabulated in standard statistical tables. They approach the *z* values as the data volume increases, but for smaller data samples they are progressively larger than *z* the smaller the sample size. This simply reflects the greater uncertainty in variance estimates that are not based on very much data. The *t* values also depend on the number of points used to estimate the unexplained variance, as well as the quantities α and β.

The author first presented Eq. (4) at the 20th AIAA Ground Testing Conference in 1998, although it was derived in that paper from a consideration of regression coefficient uncertainties rather than model prediction uncertainties. Both approaches yield the same result, but while Eq. (4) does reveal the relationships among various factors that dictate minimum data volume requirements, it has certain practical shortcomings. For example, for reasons that are not entirely clear, unexplained variance estimates do not seem to be readily available for wind tunnels. This seems to be a quantity that is expected to be generated from scratch with every new tunnel entry, instead of being quantified and documented in advance by facility personnel. So using a value of σ^2 in pre-test scaling activities usually entails a bit of guesswork.

The customer of a wind tunnel test often has only an imprecise idea of what the tolerance specification, δ, ought to be. Seldom does he state with conviction that it must be "x counts"; he is more likely to ask for something rather

more qualitative, such as a "best effort" or a "state of the art" result, whatever, exactly, those terms may actually mean.

As for the quantities *p*, α, and β, it has been the author's experience that it is essentially hopeless to try to extract estimates from a wind tunnel customer. In its current state of evolution, the experimental aeronautics community has had insufficient experience coping with more than one factor at a time to have developed any reliable experiential basis for estimating what order polynomial is necessary to fit a given response variable over some specified range of multiple factors. The typical customer is also likely to specify "zero" as the only acceptable probability for any type of inference error risk in his experiment, failing to grasp that such an outcome cannot be achieved short of acquiring an infinite volume of data.

It is for these practical reasons that the author now proposes some modifications to the scaling equation that entail what will hopefully be perceived as reasonable standards for an acceptable response modeling result. For example, while it is difficult to obtain independent estimates of δ and σ, it may be less difficult to achieve a consensus on an acceptable tolerance as expressed as a multiple of the standard error in unexplained variance. For example, consider two empirical response estimates that represent replicates separated by an amount that can just barely be resolved with 95% confidence. We call this the 95% Least Significant Difference, and if we adopt this as our response model prediction tolerance, then we have that $\delta = 2\sqrt{2}\sigma$ and the quantity $(\sigma/\delta)^2$ in Eq. (4) reduces to a constant, namely 1/8. Such a tolerance specification requires that the regression model produce a response estimate that does not differ from the corresponding empirical response estimate by enough to resolve with 95% confidence. It seems reasonable to regard this as a practical definition of a response model that gets the same answer as a physical measurement, which is really all that can be asked of such a model. With this 95% LSD tolerance specification, Eq. (4) reduces to this:

$$n = \left[\frac{1}{2} \left(\frac{z_\alpha + z_\beta}{2} \right)^2 \right] p \tag{6}$$

That is, given a 95% LSD tolerance for regression model response predictions, the multiplier of *p* reduces to half the square of the average z value. The research engineer is relieved of any responsibility to articulate a specific numerical tolerance, and the facility engineer does not have to supply a value for σ.

We now consider candidates for a default specification of α and β that might be used for scaling purposes in the not uncommon situation in which the customer has no alternative preference. If the customer does have in mind specific (non-zero!) inference error tolerance specifications, he can always use Eq. (6).

It should be rather easy to achieve a consensus on a default value for α. This is just the probability of erroneously rejecting a valid model. Alternatively, there is a (1 – α) × 100% probability that a valid model will indeed be declared valid. The author proposes an α value that is already commonly used in other contexts within the industry; namely, 0.05. In that case, there will be a 95% probability that a valid model will be recognized as valid, which is not likely to be regarded as a particularly controversial standard. In that case, the x* reference line of Fig. 1 would be located at "two sigma," again indicating that a valid model will produce residuals no greater than x* 95% of the time.

The beta value is a little less intuitive, although a consensus on some reasonable value for this quantity should also be fairly easy to achieve. Begin by recognizing that we set values of α and β in order to minimize certain types of risk. The alpha risk, also known as the probability of committing a "Type I" inference error, is set to a relatively low level to reduce the probability of rejecting an adequate model. Erroneously rejecting a model that is actually valid would probably entail the expenditure of unnecessary additional resources wasted in trying to recover a model that had already been properly fitted to the data. Such an outcome would be undesirable, to be sure, but the consequences would not be catastrophic.

Contrast this with the beta risk, also known as the probability of committing a "Type II" inference error by erroneously validating an inadequate model. This is, by any reasonable standard, a very serious error; we do not want to get the wrong answer. Because the consequences of such a Type II inference error are so much greater than the consequences of a Type I inference error, it will almost always be desirable to set beta rather smaller than alpha. The author often uses a value of beta of 0.01 in conjunction with an alpha value of 0.05. This corresponds to one chance in 20 of rejecting a valid model, but only one chance in a hundred of validating an inadequate one.

Figure 2 displays how data volume as a multiple of parameter count varies with the specification of beta for three alpha values. The smaller beta is, the less risk one is willing to assume and therefore the more data one must acquire

to defend against such risk. The red dot corresponds to $\alpha = 0.05$ and $\beta = 0.01$, for which case roughly 2.6 data points must be acquired for each parameter in the model. A novel interpretation of this result was proposed by a colleague[12] who noted the similarity with the Nyquist condition of signal conditioning, in which the highest frequency component of a signal must be sampled more than twice per cycle, with most practical implantations requiring at least 2.5 samples. It appears that a similar condition exists in response surface modeling, in which the basic response model parameter set must be estimated a minimum number of times to achieve an adequate result.

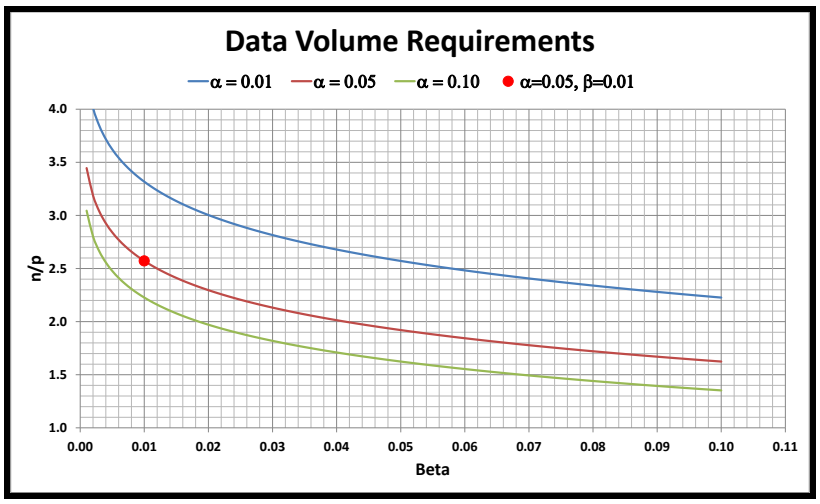

Figure 2. Data volume requirements for a 95% LSD tolerance specification

In the spirit of practical simplifications to Eq. (4) that are being proposed in this paper, Fig. 3 is a repeat of Fig. 2 that also indicates the beta value that would correspond to $n = 3p$ points rather than the $n = 2.6p$, points corresponding to $\beta = 0.01$ for $\alpha = 0.05$. This figure indicates that by acquiring a round three data points for each model term, the type II inference error protection increases, because the inference error risk drops from 0.01 to a little more than 0.003, affording roughly three times the protection against erroneously validating an invalid model.

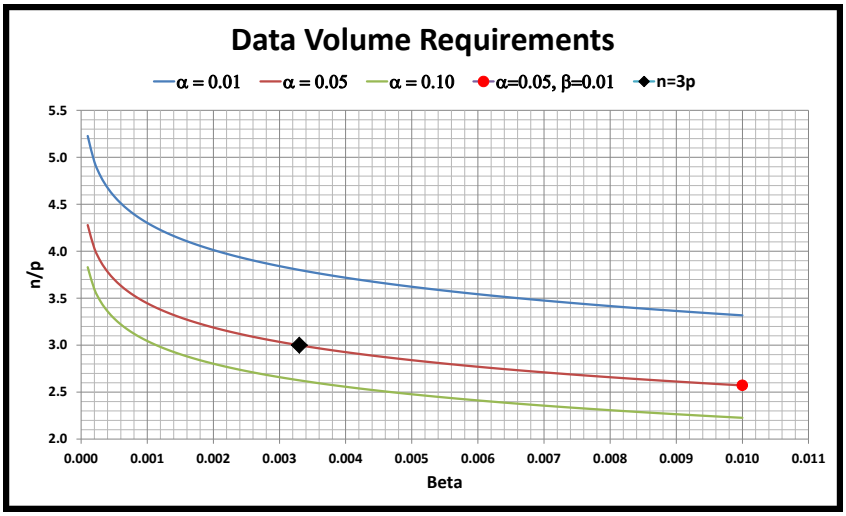

Figure 3. Data volume requirements for a 95% LSD tolerance specification

Figure 4 displays the sensitivity of beta risk to the model prediction tolerance specification for a given Type I inference error risk ($\alpha = 0.05$), and a given data volume, $n = 3p$. We have proposed a 95% LSD tolerance standard for response surface model predictions, for which δ is roughly 2.8σ. Other tolerance standards would afford different levels of beta risk protection, which this figure indicates. In general, the greater the acceptable prediction tolerance, δ, the less the beta risk associated with a given volume of data. So, for example, if one were willing to

declare adequate a prediction model that generated responses within 3σ of the corresponding measured values instead of the 2.8σ that is associated with a 95% LSD tolerance, the risk of erroneously validating a model with those somewhat reduced standards would drop from just over 0.003 to just over 0.001, roughly a factor of three.

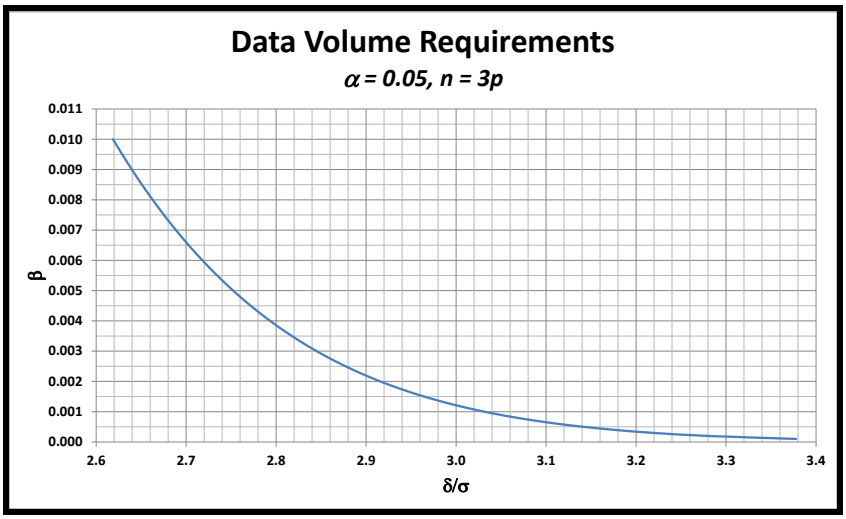

Figure 4. Risk of validating an inadequate model, vs. prediction tolerance

The chief conclusion to be drawn from these figures is that response model prediction adequacy depends on a complex interplay of multiple parameters. Understanding just what we mean by "adequate" is crucial to defining the data volume requirements necessary to ensure such adequacy, but it is proposed here that a policy of acquiring three data points for every parameter in a polynomial response model is likely to result in response models capable of reproducing measured response values within acceptable limits (less than a 95% least significant difference), and with a relatively low probabilities that an analysis of residuals would result in the validation of an inadequate model due to experimental error.

IV. Hidden Replication and Uncertainty Reduction

If we do in fact acquire *n = 3p* data points to fit a response model, then there will extra degrees of freedom beyond the minimum needed to fit the model. The process of 'hidden replication" described earlier will result in a reduced standard prediction error, by Eq. (2). For the case of n = 3p, Eq. (2) leads to the following:

$$n = 3p \rightarrow \sigma_m = \left(\sqrt{\frac{p}{n}}\right)\sigma = \left(\sqrt{\frac{1}{3}}\right)\sigma = 0.577\sigma \tag{7}$$

That is, because the response model prediction uncertainty is based on the entire ensemble of data that have been fitted and is not limited to the uncertainty associated with an individual measurement, there is considerable potential for error cancellation. The result is that a properly scaled response surface modeling experiment will minimize the resources required by ensuring that a volume of data ample to satisfy quality specifications is acquired, but no more than that, and such an experiment will also result in significantly less experimental error than is associated with a typical high-volume data acquisition exercise.

It is important to note that the unexplained variance of a typical wind tunnel test will feature a systematic component as well as the random component that is more widely recognized by the wind tunnel community[13]. The systematic component is caused by the fact that random chance variations in the data occur about sample means that are not generally stable over meaningful time intervals. A drag polar acquired later in the test might be biased significantly higher or lower than it would have been had it been acquired earlier. This reflects the effects of covariates—factors such as instrument drift and various temperature effects that influence response levels systematically rather than randomly, and are not controlled by the experimenter.

The hidden replication that results in a reduction in prediction error indicated by Eq. (2) does not apply to the systematic component of unexplained variance. No amount of replication will cause systematic (non-random) errors to cancel. However, randomizing the set-point order of the test matrix, a quality assurance tactic widely employed outside of the experimental aeronautics community to control covariate effects, converts the systematic component of unexplained variance to another component of random error. The result is that when the set-point order is randomized, the hidden replication can reduce the systematic error as well as the random error.

Figure 6 compares standard errors under a number of conditions. The blue bars labeled "OFAT" are set to "one" as a reference, and represent the standard error in an ordinary one factor at a time wind tunnel test. The rust-colored bars show the effect of hidden replication when the set-point order is not randomized. In that case, the reduction described by Eq. (2) applies only to the random component of unexplained variance and not the systematic component. The more severe the systematic unexplained variance, the less effective hidden replication is seen to be as a quality assurance tactic. The green bars display the effect of hidden replication in combination with randomizing the set-point order. When all components of the error are thus random, there is a uniform reduction of 0.577 for $n = 3p$, per Eq. (7).

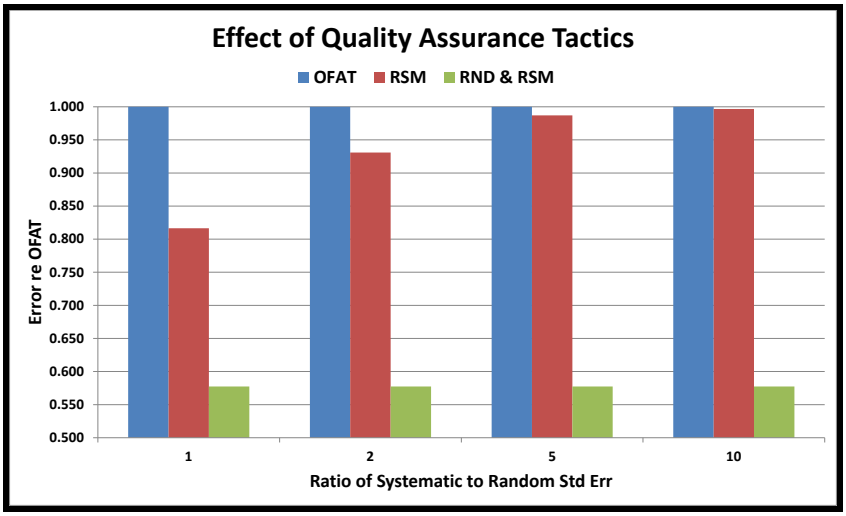

Figure 6. impact of quality assurance tactics

V. Concluding Remarks

It has been argued in this paper that the principal value of acquiring additional data in a wind tunnel test or in any other empirical investigation is to reduce the risk of making erroneous inferences about effects of a given size, or equivalently, to improve the confidence with which some given effect can be resolved. In short, the principal reason to acquire additional data is to get a better answer.

It is clear that the confidence with which a given inference can be made has an upper limit of 100%, and it is also intuitively clear that this limit is approached asymptotically as the volume of acquired data increases. What may not be as obvious is that not every data point contributes equally to this asymptotic limit. Earlier data points add much more value than points acquired later, after a relatively large volume of data are already in hand. The result is that the after some critical mass of data, the value of each new point is a monotonically decreasing function of the volume of data already in hand.

Unfortunately, even though later data points add less and less value, they are just as costly and just as time-consuming to acquire. There is thus a point of diminishing returns, beyond which the value of any further data acquisition must drop below its cost. This is true no matter how inexpensive the data acquisition process may be. All other things being equal, the point of diminishing returns simply approaches at a slower rate for a low-cost process than for one that is more expensive. But sooner or later, the point of diminishing returns must be encountered. There is thus no virtue in promiscuous data acquisition per se, notwithstanding the current focus on high-volume data acquisition in experimental aeronautics. Certainly the path to resource conservation lies in some other direction.

For this reason, there are advantages to approaches such as response surface modeling that have limited data volume requirements. It has been suggested here that as a rule of thumb, it is not necessary to acquire more than three data points for every term in a fitted response surface model.

VI. References

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

[2]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, New Mexico, June 1998.

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

[4]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.

[5]DeLoach, R. "Putting Ten Pounds in a Five-Pound Sack: Configuration Testing with MDOE". 21st AIAA Applied Aerodynamics Conference. Orlando, FL. June 23-26, 2003

[6]DeLoach, R. and Berrier, B. L. "Productivity And Quality Enhancements in a Configuration Aerodynamics Test Using the Modern Design of Experiments" AIAA AIAA-2004-1145. 42nd AIAA Aerospace Sciences Meeting & Exhibit. Reno, NV. January 5-8, 2004

[7]DeLoach, R. and Rhode, M.N. "Short-Duration, High-Quality Wind Tunnel Calibration". 1st Joint Meeting of the Supersonic Tunnel Association International and the Subsonic Aerodynamic Testing Association. Buffalo, New York. May 15-19, 2005.

[8]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.

[9]DeLoach, R. and Micol, J.R., "Comparison of Resource Requirements for a Wind Tunnel Test Designed with Conventional vs. Modern Design of Experiments Methods". AIAA 2011-1260. 48th AIAA Aerospace Sciences Meeting and Exhibit, Orlando, Florida, January 4-7, 2011.

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

[11]Coleman, H. W., and Steele, W. G., Experimentation and Uncertainty Analysis for Engineers, John Wiley and Sons, New York, 1989.

[12]Private communication, Dr. Mark Kammeyer of the Boeing Company, 50th AIAA Aerospace Sciences Meeting, 9–12 Jan 2012, Nashville, TN

[13]DeLoach, R., "Check-Standard Testing Across Multiple Transonic Wind Tunnels with the Modern Design of Experiments," AIAA 2012-3173, 28th AIAA 28th AIAA Aerodynamic Measurement Technology, Ground Testing, and Flight Testing Conference, New Orleans, LA June 2012.