AIAA 2000-0825

Improved Quality in Aerospace Testing Through the Modern Design of Experiments

R. DeLoach
NASA Langley Research Center
Hampton, VA

38th Aerospace Sciences Meeting & Exhibit
10–13 January 2000
Reno, Nevada

For permission to copy or to republish, contact the American Institute of Aeronautics and Astronautics, 1801 Alexander Bell Drive, Suite 500, Reston, VA, 20191-4344.

AIAA 2000-0825

IMPROVED QUALITY IN AEROSPACE TESTING THROUGH THE MODERN DESIGN OF EXPERIMENTS

R. DeLoach*
NASA Langley Research Center
Hampton, VA 23681-0001

Abstract

This paper illustrates how, in the presence of systematic error, the quality of an experimental result can be influenced by the order in which the independent variables are set. It is suggested that in typical experimental circumstances in which systematic errors are significant, the common practice of organizing the set point order of independent variables to maximize data acquisition rate results in a test matrix that fails to produce the highest quality research result. With some care to match the volume of data required to satisfy inference error risk tolerances, it is possible to accept a lower rate of data acquisition and still produce results of higher technical quality (lower experimental error) with less cost and in less time than conventional test procedures, simply by optimizing the sequence in which independent variable levels are set.

Introduction

The testing technology community at NASA Langley Research Center has been examining a "modern design of experiments" (MDOE) approach to wind tunnel testing since January 1997.[1] MDOE methods differ in many substantive ways from conventional One Factor at a Time (OFAT) test methods that have traditionally been used in wind tunnel testing. OFAT practitioners attempt to hold all variables constant while sequentially changing a single independent variable of interest over a range of levels. The common procedure of holding constant such variables as Mach number, control surface deflections, Reynolds number, etc., while monotonically varying angle of attack, is an illustration of this method. OFAT experiment designs are popular in wind tunnel testing because of the intuitive appeal of their structure, and because of a common conviction among test personnel that such designs provide early and unambiguous indications of emerging difficulties in a test. Perhaps the most compelling reason that OFAT practitioners embrace this method is that it maximizes data acquisition rate and therefore total data volume, popular productivity metrics in late-20th century wind tunnel testing.

Proponents of MDOE at Langley have made the case that productivity is maximized when the tactical objective is to acquire ample data to meet the strategic research requirements of the test, rather than to simply acquire as many data points as resources permit. Data volume per se becomes a secondary consideration from this perspective, cast in the context of inference error risk management. The cost of data collection is likened to a premium one pays for insurance against improper scientific and engineering conclusions that may result from an inadequate volume of data. One wishes sufficient insurance that improper inferences will be unlikely, but one is hesitant to purchase more insurance than is necessary to drive inference error probabilities significantly below acceptable risk thresholds. This low data volume perspective permits the consideration of alternative experiment designs with attractive features that are impractical under operating constraints that require maximum data acquisition rates.

Specifically, relaxing the requirements for high-volume data collection permits the researcher to consider alternative structures for the test matrix that offer significant protection from experimental error, while still providing data in ample volume to meet specific test objectives. This error protection is achieved through the processes of *blocking* (especially *orthogonal* blocking) and *randomization*, techniques which are common to formal experiment design and which will be described and illustrated in the context of wind tunnel testing in this paper. This enhanced protection from experimental error and the associated potential for inference error can produce a higher quality research result while often also reducing the volume of data in an experiment and the attendant costs and cycle time. Blocking and randomization, along with the more familiar technique of replication, will be described as techniques to enhance the quality of research, and illustrated in the sections that follow.

Block Effects and Orthogonal Blocking

"Block effects" arise in wind tunnel testing when response variables (forces, moments, etc.) measured in one "block" of time differ significantly from measurements made in another block of time under circumstances expected to yield identical results within experimental error.

The most elementary form of a block effect is illustrated in figure 1, in which there is simply a discontinuous bias shift at the block boundary. In more commonly occurring circumstances there could be a continuous drift due to the superposition of systematic errors on an otherwise unchanging response. Such a condition could be modeled with a regression equation

---

* Senior Research Scientist

Copyright © 2000 by the American Institute of Aeronautics and Astronautics, Inc. No copyright is asserted by the United States under Title 17, U. S. Code. The U. S. Government has a royalty-free license to exercise all rights under the copyright claimed herein for Government Purposes. All other rights are reserved by the copyright holder.

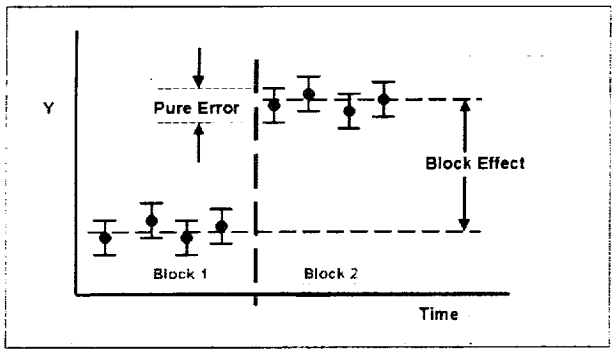


Figure 1. A block effect.

for which the intercept term is not constant but rather is a function of time. After a sufficiently long period of time, the mean value of the intercept term could be significantly different in the two blocks, resulting in a block effect.

Blocking is a method used in formal experiment designs to defend against block effects by altering the run order to impose a certain balance or symmetry between the set of points acquired during one block of time and another. This makes it possible to attribute potentially large components of the unexplained variance to block effects, essentially converting these components from "unexplained" to "explained" sources of variance. This in turn diminishes the residual unexplained variance, increasing the precision of experimental results. If the blocking is performed in a particularly clever way ("orthogonal blocking," to be described presently), it is possible to recover the coefficients of a response function regression model (within a possible shift of the intercept term) as if the block effect had not occurred. Absent such precautions, the shape of the regression curve will be impacted by the block effect and could lead to improper inferences about the true nature of the dependence of response variables on independent variables of interest. Such undetected block effects can result in inference errors that are especially difficult to understand or correct in a conventional OFAT experiment.

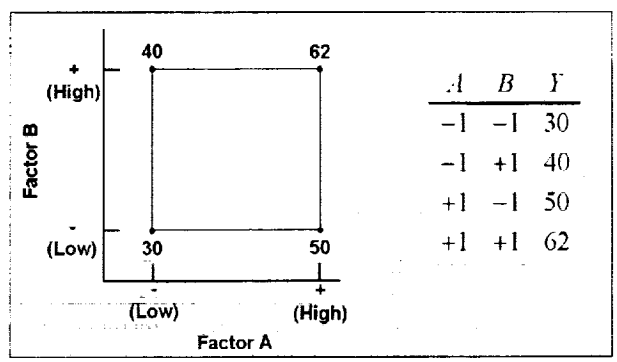
There are many potential causes of block effects in wind tunnel testing. Consider a two-shift wind tunnel operating schedule in which there are between-shift differences in operating procedures or skill levels that could cause systematic differences in the data. To block such an experiment by shift, a new independent variable could be introduced that takes on the value of "–1" for the first shift, say, and "+1" for the second shift. If a conventional regression model relating response variables to independent variables is augmented with this blocking variable, its coefficient is a measure of the response change that occurs from one shift to the next. A statistically insignificant blocking variable coefficient would provide unbiased evidence of between-shift uniformity, a desirable process feature.

Blocking the test matrix requires the independent variables to be set in a prescribed fashion that may (usually does) depart from the sequence that would achieve the highest data collection rate. However, if the experiment has been scaled so that the planned volume of data ensures sufficiently low inference error probabilities for the specific objectives, then maximum data collection rates add little to the quality of the result. On the contrary, insofar as a high-rate strategy precludes blocking options, the policy of maximizing data volume may significantly degrade the potential quality of the research result, while adding significantly to the cost by acquiring more data than required to make a proper inference. Ironically, depending on the magnitude of the block effects, it is possible for the unexplained variance eliminated by a proper blocking scheme to result in a greater increase in precision than the higher data volume would have delivered. It is therefore possible to achieve higher precision and a lower cost of data at the same time.

The result of a blocking can be illustrated with a simple example. Consider an experiment to quantify how rolling moment strength depends on left aileron deflection (change in rolling moment per unit deflection in aileron) for some model of interest. Assume that we have reason to believe that the rolling moment is approximately a linear function of aileron deflection over a sufficiently small range of deflections. Assume further, for the sake of this example, that we wish to "widen our inductive basis" by averaging our measurements over two levels of a second control surface, say flap deflection. That is, we would like to know how induced rolling moment changes depend on a unit deflection in aileron on average over some small range of flap deflections.

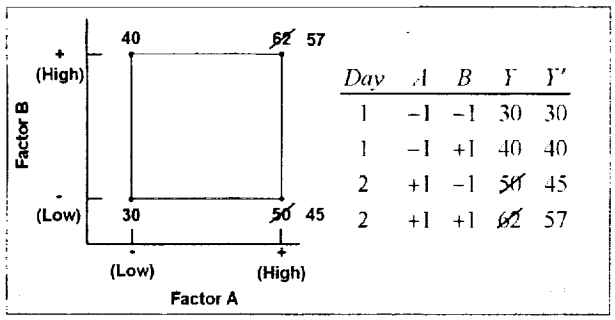
A very simple experiment design will suffice to make an initial estimate of aileron rolling moment strength averaged over two flap deflection angles. Since, for the sake of this example experiment, we are only setting two levels of each of our two variables, it is convenient to represent the smaller and larger levels by the coded variables "+1" for the larger value and "-1" for the smaller. So, for example, if we plan to change aileron deflection from 0 to 5 degrees, and flap deflection from –3 to +3 degrees, we would assign coded variable levels of –1 and +1, respectively, to aileron deflections of 0 and 5 degrees, and flap deflections of –3 and +3 degrees.

Assume that we have now measured rolling moment for the four combinations of independent variables that we are setting in this simple experiment. That is, we have measured the rolling moment for all four combinations of low and high aileron deflection, and low and high flap deflection. Figure 2 is a schematic representation of the results of the experiment, where the rolling moment response is also represented in some scheme of coded units to simplify the illustration.



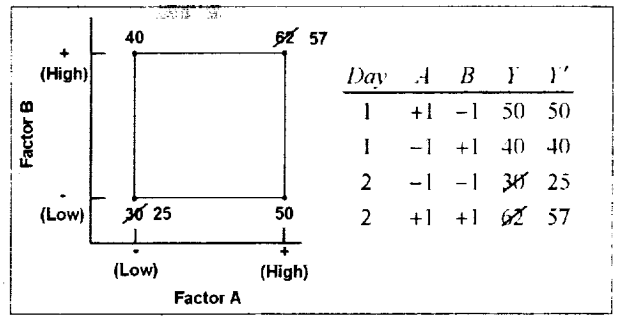
A	B	Y
–1	–1	30
–1	+1	40
+1	–1	50
+1	+1	62

Figure 2. A two-level experiment in two variables.



Day	A	B	Y	Y'
1	−1	−1	30	30
1	−1	+1	40	40
2	+1	−1	~~50~~	45
2	+1	+1	~~62~~	57

Figure 3. Block Effect affecting variable A.



Day	A	B	Y	Y'
1	+1	−1	50	50
1	−1	+1	40	40
2	−1	−1	~~30~~	25
2	+1	+1	~~62~~	57

Figure 4. Block Effect with *no* effect on variable A.

Let us assume in figure 2 that factor "A" is aileron deflection and factor "B" is flap deflection. To estimate the average effect (averaged over flap deflections) of changing aileron deflection from low to high levels, we can simply compute the average rolling moment change at low and high flap deflection. From the figure we see that the rolling moment changes from 40 to 62 units when the aileron is changed from low to high level with the flap deflection *high*. The same aileron change causes the rolling moment to change from 30 to 50 units with a low flap angle. In the former case the aileron effect is 22 units (62 – 40), while in the latter the effect is 20 units (50 – 30). We thus report an average effect of 21 units.

Now consider a slight complication. Assume that there is only enough time to acquire two of our four points before the end of the day. Flap deflections are changed in this model by means of actuators that are controlled remotely; but to change the aileron deflection it is necessary to terminate flow, open the tunnel, and physically change to an aileron machined at a different angle, a time-consuming process compared to changing the flap deflection.

Because of the difficulty in changing the aileron deflection angle, we decide to set the aileron deflection low and acquire data for the low and high flap configurations today. We will set the high aileron deflection angle tomorrow morning before we start the tunnel.

Now assume further that unbeknownst to us, the rolling moment balance sensitivity changes over night. This might be due to temperature changes, or any of an uncountable number of other reasons. The effect of this change is that rolling moment measurements on the second day read five units lower than they would have done without the sensitivity shift in the balance. The resulting situation is illustrated schematically in figure 3. Based on this data set, the aileron effect at high flap setting is now 17 units (57 – 40) instead of 22 units, and at low flap setting the effect is now 15 units (45 – 30) instead of 20 as before. We therefore report an average aileron effect of 16 units, oblivious to the five-unit change induced by the sensitivity shift in the balance.

Now consider the identical situation, except for a change in the order in which we set the flap and aileron deflection angles. That is, assume that we acquire only two configurations at the end of the day, and then acquire the other two the next day, just as before. And just as before, assume that the balance sensitivity has changed by exactly the same amount. But this time on the first day we set low flap plus high aileron for one of the configurations we test and high flap plus low aileron for the other day-one configuration. On day two we set the high flap plus high aileron configuration and the low flap plus low aileron configuration, as in figure 4. Referring to figure 4, the aileron effect at high flap is now 17 units (57 – 40), and at low flap it is 25 units (50 – 25). The aileron effect averaged over flap deflections is thus 0.5 × (17 + 25) = 21 units, *exactly as if the balance sensitivity had NOT shifted!*

This very simple demonstration has important implications for the experimenter. In one case a significant change in the measurement system resulted in a serious error in the response variable estimate and in the other case the identical change had no effect on the result at all. The only difference in the two cases was the order in which the independent variable levels were set. The second case resulted in a better answer, notwithstanding the fact that it may have taken longer on average per data point to set the control surfaces in that order.

The implication is clear: not all test matrices are equivalent, even if they both have the identical set points. When there are systematic errors, the order of execution is critical. The next section describes an experiment to determine whether systematic errors that can induce block effects are present or not in a wind tunnel experiment, and what their magnitudes are relative to ordinary random error. But before we seek evidence of systematic error, let us return to our simple example to examine how the set-point order defends us from such errors.

Figure 5 shows the two test matrices, but with a third variable added to "A" (aileron) and "B" (flap.) The third variable, labeled "Day," is called a blocking variable. In this example the blocking variable takes on two levels just as the aileron and flap variables did, and just as we did for those variables, we can conveniently code the

Block Effects Confounded with Main "A" Effect

Day	A	B	AB
−1	−1	−1	+1
−1	−1	+1	−1
+1	+1	−1	−1
+1	+1	+1	+1

Block Effects Confounded with Interaction

Day	A	B	AB
−1	+1	−1	−1
−1	−1	+1	−1
+1	−1	−1	+1
+1	+1	+1	+1

Figure 5. Confounding of block effects.

blocking variable. Here, "−1" corresponds to the first "block" or "day" of the experiment and "+1" corresponds to the second block/day.

One of the reasons that it is convenient to represent experiments such as this one in terms of coded values is that it facilitates the calculation of the effects. To quantify the "A" (aileron) effect, for example, we simply multiply the column of signs under the "A" variable in the matrix with the corresponding response variable measurements. To account for the fact that we have multiple estimates (in this case, two estimates) of each effect, we divide the product of signs and responses by the number of "+" signs in each column. The "A" effect for the matrix in figure 2 is then, simply:

$$A = \frac{(-1\times 30)+(-1\times 40)+(+1\times 50)+(+1\times 62)}{2} = 21$$

Note in figure 5 that the "A" effect in the matrix on the left has identically the same sign pattern as the blocking variable. This means that it is impossible to distinguish between effects due to a change from one level of "A" to another, and effects due to a change from one block to another. We say that in such circumstances the aileron effect is "confounded" with the block effect—the two cannot be distinguished. There is no way to know with this matrix whether a change in rolling moment was due to the change in aileron deflection angle or just due some change (such as the sensitivity shift) that occurred from one day to the next. So if aileron effects are important to us, this is a very poor design unless we have some reason to know that there will be no change overnight of a magnitude that will be important to us. This is virtually impossible to guarantee.

In the matrix on the right of figure 5, the sign pattern for the block effect is not confounded with either of the main effects. Instead, it is confounded with the AB interaction effect, the sign pattern for which is obtained by multiplying corresponding signs for the A and B main effects. This means that we are unable to quantify the AB interaction unambiguously in this instance, but we are able to estimate the main effects clear of any block effects. The AB interaction is a measure of how the level of one variable affects the response change in another. For example, if a unit change in aileron deflection caused a different change in rolling moment for one flap setting than another, we would say that there was an interaction between the flap and aileron deflections. The interaction in a simple two-level design such as this one is defined as the average difference between the "A" effect measured at "high B" and the "A" effect measured at "low B." The effect is the same if A and B are reversed in the definition. So from figure 2, for example, it is easy to see that the AB interaction has a numerical value of "1" in this example, which is small compared to the main effects. If our discipline knowledge had informed us that the interaction effect was likely to be small or unimportant as in this case, we might be willing to surrender our ability to quantify it unambiguously in order to free the more interesting main effects from confounding by block effects. Note that this decision is not in the least bit dependent on our knowledge that block effects actually do exist, or what their magnitude will be if they do exist. By confounding the block effects with a single (relatively uninteresting) column in our design matrix, we guarantee that the effects represented by the other columns are free of block effects.

The reason that confounding with one column frees all the others from block effects has to do with a particular property of the design used in this example. We can represent each column in the design matrix as an n-dimensional vector, where "n" is the number of data points acquired (four in this case). From analytical geometry we know that the cosine of the angle included between two vectors is proportional to the sum of the products of corresponding elements. For the design in this example, the sum of the products in all corresponding rows is zero. The geometric interpretation is that the vectors represented by the corresponding columns are at right angles. That is, they are *orthogonal* to each other. This is the case for all pairs of columns in the example design, so the design is said to be orthogonal. In such a design, changes in an effect corresponding to one column will have no influence on any of the other effects. This orthogonality property is extraordinarily useful in the design of experiments, providing protection from block effects that may or may not be present, and may or may not be serious if they are present. Yet wind tunnel researchers seldom utilize orthogonal blocking to account for otherwise unexplained variance.

Block effects can occur at any time, but there are certain recurring situations in which the prudent designer will probably want to consider orthogonal blocking whenever it is possible. For example, it is a good idea to arrange shift changes to occur on block boundaries. There are innumerable potential block effects associated with shift changes, including procedural differences from one crew to the next (notwithstanding good faith efforts to ensure uniformity). There are also differences in skill level and experience, differential learning and fatigue effects, and so on. It is important to realize that between-shift blocking does not entail an indictment of one shift or the other, nor does it represent a declaration that important differences actually do exist from one shift to the next. It is simply a means of removing this concern, insofar as possible, from the long list of possible sources of inference error in an experiment.

In formally designed experiments conducted at Langley Research Center, we schedule as many planned disruptions as possible to coincide with block boundaries. For example, if an all-hands meeting occurs during the day, we schedule it to coincide with block boundaries to defend against subtle response shifts that may occur from the time the tunnel operation is stopped until it is started again. Wind-off zeros and model inversions to quantify flow angularity changes all occur on block boundaries in our formally designed experiments, and of course end-of-day operations are scheduled to coincide with block boundaries so that any overnight effects can be minimized.

Block effects can occur over arbitrary periods of time and are not limited to such obvious units as "shifts" and overnight breaks. We therefore incorporate orthogonal blocks in our formal designs at Langley even if they do not correspond to some obvious break point. The intent is to minimize the effects of changes that may

have occurred from one block of time to the next, independent of any specific reason we might have for suspecting such changes. As we will see in the next section, the systematic variations that occur in a facility as complex and as energetic as a wind tunnel tend to be large compared to the random errors associated with our very high-precision measurement systems. Their existence and magnitude justifies orthogonal blocking over relatively short periods of time.

Two important points are appropriate before leaving this simple illustration of orthogonal blocking. The first is that in an actual experiment as simple as this one, in which only two variables are in play, it would be unusual for the interaction effects to be so uninteresting that we would be willingly confound them with block effects. We would probably rely on other tactics to defend against systematic error, to be discussed below. There are practical two-level factorial designs in which many more than two variables are in play, however, and the potential exists to confound block effects with interactions of much higher order than two-way, which truly are unlikely to be of interest. One example is a configuration study in which the objective is simply to identify which among a very large number of candidate variables has an important influence on responses of interest to us. Even a modest configuration study will frequently have a half dozen or more variables to investigate. In general, the higher the order of the interaction, the less likely it is that such an interaction exists. When they do exist, as a general rule the highest-order interactions may be relied upon to have magnitudes small enough that they can be safely ignored in the presence of main effects and lower-order interactions, which tend to be larger. In a six-variable experiment, for example, main effects and some two-way interactions will almost certainly be of interest, and certain three-way interactions may be as well. Some four-way interactions might also be detectable. It is likely that five-way interactions and the single six-way interaction will be small or non-existent, and the designer may therefore wish to consider using them in blocking schemes to defend against systematic errors.

The second point to be made before leaving this example is that there are designs that are structurally more complex than the simple two-level factorial designs discussed here. These designs involve multiple levels of each independent variable, for example, as is necessary to develop response models higher than simple first-order. As will be illustrated, orthogonal blocking schemes are available for these more elaborate designs as well.

Systematic Errors

Block effects resulting from systematic errors have the consequence that the result of an experiment is dependent on the order in which the independent variable levels are set, as illustrated with the simple example of the previous section. The order of independent variables is virtually never selected in late 20th century wind tunnel testing to defend against systematic errors. Rather, the order is set sequentially to maximize data acquisition rate and therefore total data volume, two popular productivity measures (popular, but misguided, by modern experiment design standards). The implicit assumption, therefore, is that conditions exist in wind tunnel testing for which the result is independent of the order in which the data points are acquired, so that it is not necessary to optimize run order to ensure research quality. That is, the common assumption is that wind tunnels can be considered to be in a state of statistical control in which sample means are time invariant, at least for short periods over which important data sets are acquired. In this section we present the results of an experiment that was designed to examine this assumption, introduced by some comments about the general nature of systematic errors.

There are practical distinctions between systematic errors and random errors that are especially important in the design of experiments. Because random errors tend to be distributed equally above and below some (possibly biased) estimate of the true mean value, their effects can be canceled by replication over periods of time for which the mean is stable. As discussed earlier, the common practice of setting independent variables sequentially implicitly assumes a state of statistical control. Absent a stationary environment, however, this sequential set-point strategy maximizes the confounding of independent variable effects with systematic variations, as suggested by the simple block effects demonstration in the previous section.

Systematic errors have other perverse influences on research quality besides their resistance to replication and their penchant for confounding the true effects of independent variable changes. Most of the theory of errors, upon which we depend to quantify precision intervals and to make unbiased estimates of population parameters, is based upon certain assumptions about the independence of errors that are simply invalid in the presence of systematic variations. For example, if some unknown source of error persists over time and has the effect of increasing the numerical value of measurements above the true value, then two sequential measurements will not be independent of each other. This is because the source of systematic error ensures that whenever there is a positive error in the first measurement, it is much more likely that the error in the second measurement will be positive than negative.

This, incidentally, is a key weakness in the OFAT strategy of "holding all else constant" while independent variables are examined one at a time. There will be inevitable experimental errors associated with all of the variables that are "held constant." Even though the test matrix may specify a particular Mach number during an OFAT polar, for example, in truth the Mach number will always be set either slightly above or slightly below the true specified Mach number, due to inevitable set point errors and other causes. If the net effect of these errors is positive for a particular response variable—drag, for example—then the drag measurements made *at each and every* angle of attack will be biased high because of the constant Mach error caused by "holding Mach number constant." There is no opportunity for this particular contribution to the error to cancel out as additional data points are acquired. The variance is not inversely proportional to the total number of data points under such circumstances, for example, and there are other such assumptions underlying the computation of standard

precision intervals that are no longer valid. Since these errors will not be independent, the independent measurement assumptions upon which regression and other curve-fitting strategies are based will be invalid also.

Given the significant potential of systematic errors to make mischief in experimental research, it ought to be surprising that they tend to receive so much less attention than random errors. There are two practical explanations for why this is so. The first is that by their nature systematic errors are much more difficult to detect than random errors, whose presence is forced upon our consciousness with every replicate that we acquire. The low-profile nature of systematic errors aids and abets the darker angels within us, supporting tendencies for some researchers to avoid going out of their way to find trouble. "See no evil" can be a seductive philosophy when options for dealing with potential problems are limited. Much more attention also tends to be lavished upon random errors than upon systematic errors because of the relatively greater arsenal of established defense measures available to counter random errors, or at the very least, to quantify them reliably. The normal distribution, the Central Limit Theorem, the power of replication—all of these factors provide aid to the experimentalist in coping with random errors for which there are few systematic error counterparts. Too often systematic errors are removed from the analysis "by assumption." Absent any hard evidence that systematic errors are afoot, the OFAT practitioner often simply "assumes" statistical stationarity. After all, what other practical alternatives are available in an OFAT design? Even if such an assumption is not made explicitly, it is made implicitly whenever there are no overt actions taken to defend against systematic errors.

An experiment was recently conducted to examine the relative contributions of systematic and random errors to the total unexplained variance in a wind tunnel. This experiment was one element of a larger effort to demonstrate modern experiment design methods in a landing stability configuration test. The principal results of the parent landing stability experiment will be reported elsewhere, but the analysis of variance incorporated in the design is relative to the data quality issues of the present paper and will be summarized here.

The design of the landing stability configuration test entailed selecting the smallest number of configuration combinations necessary to quantify the effects of a number of configuration variables on certain landing stability response variables of interest for a flight configuration described simply as a "Generic Winged Body." The test was conducted during November 1999, in the ViGYAN Low Speed Wind Tunnel in Hampton, VA.

A nominal landing configuration was replicated at randomly selected times a total of seven times during the eight-day period in which data were acquired in this experiment. For each of the seven replicated configurations, a total of 30 randomly ordered combinations of angle of attack and sideslip angle were acquired in the range of 10° to 14° for angle of attack and −4° to +4° for sideslip. Eight of these model attitude combinations were replicates of the 12° angle of attack and 0° sideslip angle that represented a nominal landing attitude for this vehicle. There were thus a total of 56 replicates of each of the response variables measured. These variables were the stability axis and body axis force and moment coefficients of a standard six-element balance. The seven nominally identical groups of eight replicates were displaced in time by periods of a few hours to a few days, while the eight within-configuration replicates were all acquired within about a 15-minute period. Table I represents the lift coefficient data for these 56 replicates.

A one-way analysis of variance (ANOVA) was conducted to compare the variance of column means with within-column variance to quantify the ratio of "relatively long-term" between-column variance (hours/days) with "relatively short-term" within-column variance (15 ± 1 minutes). Such an analysis was performed for all six standard force and moment components in both stability axis and body axis coordinate systems. Table II is the ANOVA table for the analysis of the lift data appearing in Table I.

Within-column sums of squares are computed by squaring the difference between each point and its column mean and summing all 56 of these squared deviations. Between-column sums of squares are computed by squaring the difference between each column's mean lift coefficient and the average of all 56

Table I: Seven groups of eight lift coefficient replicates

Run 2	Run 7	Run 31	Run 34	Run 36	Run 58	Run 64
0.1918	0.1902	0.1911	0.1915	0.1920	0.1951	0.1894
0.1882	0.1893	0.1889	0.1894	0.1892	0.1931	0.1878
0.1902	0.1902	0.1917	0.1913	0.1920	0.1948	0.1879
0.1904	0.1909	0.1921	0.1913	0.1907	0.1986	0.1912
0.1903	0.1904	0.1908	0.1921	0.1902	0.1974	0.1889
0.1913	0.1913	0.1919	0.1916	0.1927	0.1990	0.1909
0.1915	0.1921	0.1921	0.1917	0.1922	0.2001	0.1909
0.1907	0.1910	0.1927	0.1908	0.1910	0.2024	0.1897

Table II. One-way ANOVA table for lift coefficient data

Source of Variance	SS	df	MSE	F	F_{crit} (0.01)	p
Between-Column	3.36E-04	6	5.59E-05	23.2	3.2	8.88E-13
Within-Column	1.18E-04	49	2.41E-06			
Total	4.54E-04	55				

lift coefficients, summing those seven squared deviations, and multiplying the sum by the number of rows (8) to reflect the entire sample size of $7 \times 8 = 56$ lift coefficient replicates. There are $8 - 1 = 7$ degrees of freedom within each column (one df lost to compute the column mean) and there are 7 columns, for a total of 49 within-column degrees of freedom. Similarly, there are $7 - 1 = 6$ between-column degrees of freedom.

Mean square errors (MSE in Table II) are computed by dividing the sum of squares, SS, by the degrees of freedom, df. The mean square error, or variance, is a measure of the variability associated with a given source, averaged over the degrees of freedom associated with that source. The degrees of freedom represent the minimum number of points required to estimate the sum of squares, given the mean.

The ratio of between-column to within-column variances is recorded in the ANOVA table as the F statistic. In Table III the F statistic is seen to have a value of 23.2 for these lift coefficient data. That is, the variance associated with differences in time of the order of hours to days is over 23 times greater than the average variance experienced over a 15-minute period.

While this suggests a significantly greater variance over the longer period of time than the shorter, it is not impossible that this large ratio might be due to chance variations in the data that caused the denominator to be smaller than usual and the numerator to be larger than usual for this one data set. The F statistic follows a known probability distribution, which facilitates an upper-tailed significance test for the computed F statistic. The exact shape of the F distribution is different for each unique combination of numerator and denominator degrees of freedom, but critical values are tabulated for various df combinations and various significance levels. For six between-column df and 49 within-column df, the critical F statistic associated with a significance level of 0.01 is 3.2. That is, under the null hypothesis that no difference exists between within- and between-column (i.e., short- and long-duration) variances, there is only a 1% probability that their ratio would exceed 3.2 by pure chance when there were as many as 6 between-column df and 49 within-column df. Since the computed F statistic of 23.2 exceeds this critical value, we infer that there is less than a 1% probability that no difference exists between long-term and short-term variance for these data, and we reject the null hypothesis, accepting the alternative hypothesis that longer-term variations are in fact greater than shorter term variations. Confidence in our inference is bolstered by the p-statistic listed in the ANOVA table, which represents the probability of an F ratio as large as 23.2 occurring by chance under the null hypothesis. The very small number for this probability gives us further confidence to reject the null hypothesis and conclude that long-term variations are significantly greater than short-term variations in the data set we have analyzed.

Similar analyses of variance were made for all six standard force and moment components both in body axis and stability axis coordinate systems. The results are displayed in figures 6 and 7. Because there was the same number of within- and between-column degrees of freedom in all cases (49 and 6, respectively), the critical F statistic for 0.01 significance is 3.2 in all cases.

Note in figures 6 and 7 that without exception, every response variable displayed an F statistic greater than 3.2 and therefore significantly more variation over the longer

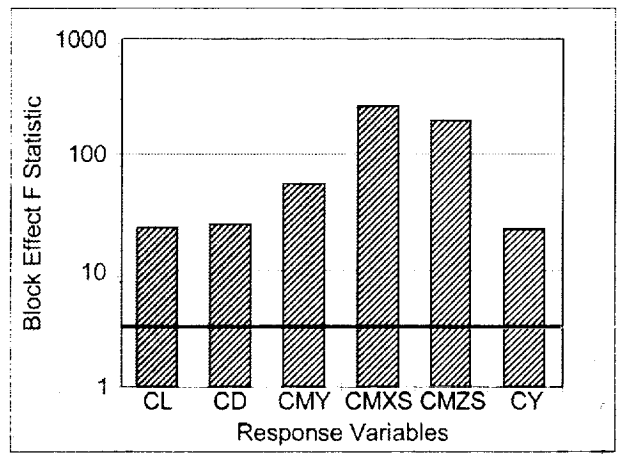


Figure 6. Stability axis block effect F statistics. Blocks of hours/days. F = 3.2 critical at 0.01 level.

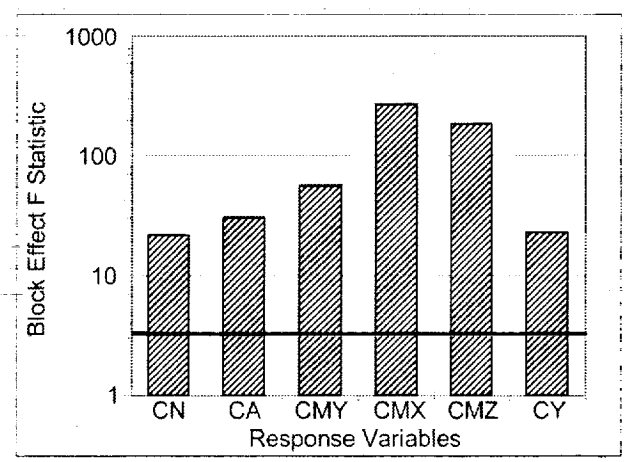


Figure 7. Body axis block effect F statistics. Blocks of hours/days. F = 3.2 critical at 0.01 level.

term (hours/days) than the shorter term (15 minutes). Unless both the short-term and the long-term variations occur symmetrically about the same mean over the period for which replicates are acquired, the mechanisms responsible for the greater long-term variance will cause simple replication to have a different impact than if only short-term variations were present. A likely outcome is that not as much of the longer-term variance will be canceled by replication. The uncanceled long-term variance component will comprise a kind of rectification error, resulting in a net positive or negative bias.

Having partitioned the variance of the entire ensemble of data into long-term and short-term components, it is possible to estimate the relative contribution of each. The within-column and between-column sums of squares add to produce the total sum of squares that would be obtained by adding the squared differences between each of the 56 points and the 56-point mean. Likewise, the within-column and between-column degrees of freedom sum to the total degrees of freedom given the mean (49 + 6 = 56 – 1 = 55). The variances, representing the ratios of the sums of squares to the degrees of freedom, do *not* add, however. That is, the total variance of the 56-point ensemble of replicates is not simply the sum of the within- and between-column variances. This complicates what would otherwise be a straightforward calculation of the relative contribution of each variance component to the total, but numerous measures exist for estimating this. One of the most common metrics is "ω^2", the "statistical power" metric.[2] It is computed by the following formula:

$$\omega^2 = \frac{SS_{effect} - \left(df_{effect} \times MSE\right)}{MSE + SS_{total}} \qquad (1)$$

The statistical power ranges from 0 to 1 and is interpreted as the portion of the total variance that is explained by a particular effect (e.g., block effects). The contribution of block effects on the order of hours-to-days to the total unexplained variance is represented graphically in figures 8 and 9 for stability axis and body axis forces and moments.

Strictly speaking, the statistical power formula is intended for "fixed effects" models, in which the levels of the between-column variable are specifically selected values. The current experiment involves a "random effects" design in which those levels (elapsed time from start of experiment) were selected at random. While the mathematical calculations are the same in fixed effect and random effect designs, the interpretation of the results are different. In this case we are not free to generalize the specific numerical values of our variance component computations to all other cases of block effects, but must apply them only to the specific block durations of this experiment. Nonetheless, the trend across all response variables of significant block effects suggests that short-term and long-term variance differs significantly, which suggests further that systematic variations cannot be discounted.

A further refinement of the one-way ANOVA study was made possible by the fact that the eight replicates acquired within each column in Table I were actually acquired in two blocks of four. For reasons dictated by an element of the experiment to be reported elsewhere than this paper, the 30 combinations of angle of attack and sideslip acquired for each model configuration were grouped into three blocks containing 9, 9, and 12 points. The first four of the within-column replicates in each column of Table I were acquired at random in the first block of nine points and the second group of four was likewise acquired at random within the second nine-point block. There was no delay in the acquisition of the first block and the second, which required an average of 15 minutes total. The elapsed time between block centers was therefore on the order of 5–10 minutes.

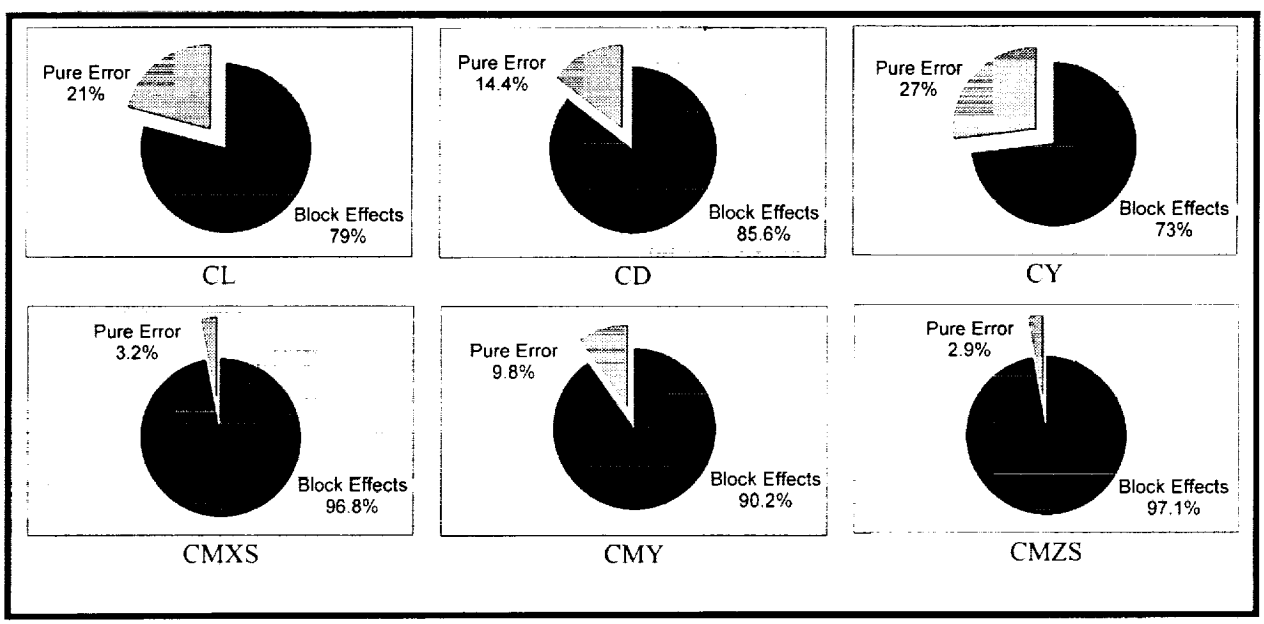


Figure 8. Long term (hours/days) stability axis block effect contributions to unexplained variance.

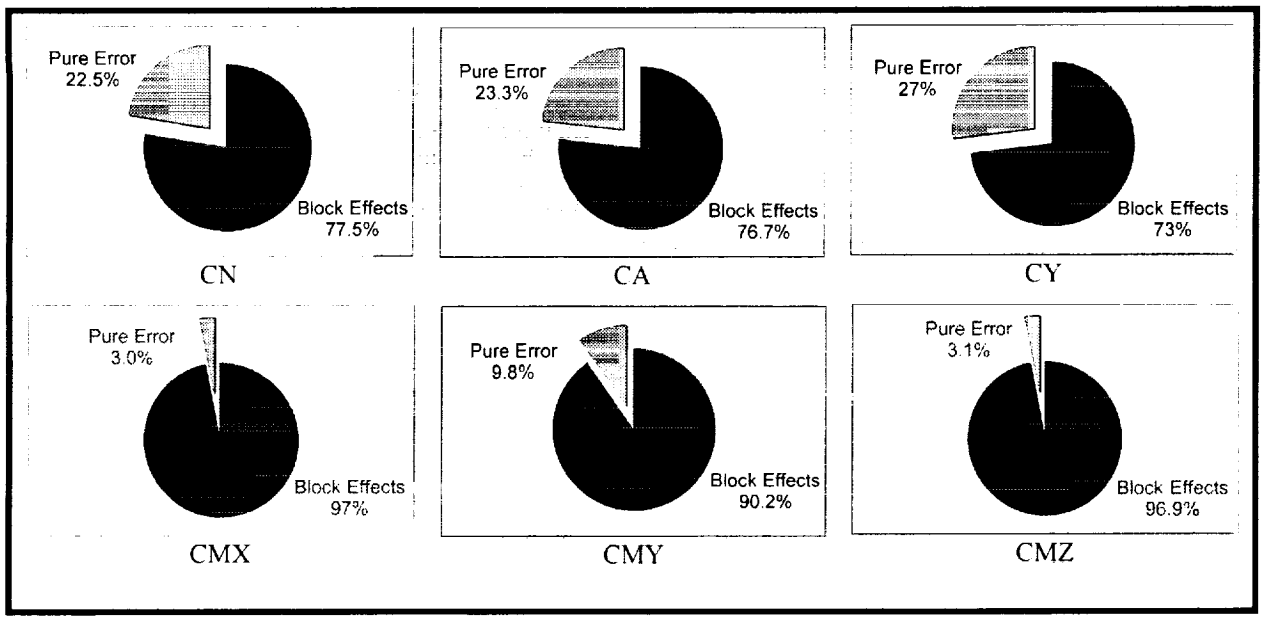


Figure 9. Long term (hours/days) body axis block effect contributions to unexplained variance.

Since each column in Table I features two blocks, it is possible to perform a two-way analysis of variance on this data set. This enables us to partition the variance of the entire 56-point ensemble into further components. In addition to the between-column block effects studied previously, it is possible with a two-way ANOVA to quantify the within-column block effects as well. These are the contributions to the total variance due to any systematic ("DC") shift that might occur in the approximately 5–10 minute period that elapsed between the two within-column blocks. Furthermore, since there are replicates within each of the in-column blocks, it is possible to estimate a third component of the total variance due to the interaction of within-column and between-column blocks.

The computations required in a two-way ANOVA are straightforward extensions to the one-way ANOVA computations described above. They are, however somewhat more involved and tedious and will not be described here. Any introductory statistics text can be consulted for the computational details; see, for example, reference 3.

The two-way ANOVA table for the lift coefficient data of Table I appears in Table III. Interpretations are analogous to those provided above for the one-way ANOVA. Note that the F statistic for between-column blocks is 32.8 for the two-way ANOVA, compared to 23.2 for the one-way case. In the two-way ANOVA the unexplained variance is partitioned into three component parts. One is the variance component due to short-term block effects, the second is the interaction component, and the third is the remainder, unattributable to any specific source. Thus, the two-way ANOVA has explained part of the previously unexplained variance, resulting in a smaller unexplained variance component. The between-column variance is a larger multiple of this reduced unexplained variance. That is, the short-term blocking has increased the precision with which effects such as the between-column block effects can be detected. Note that the F statistic involves a ratio of mean square errors (variances) and recall that the mean square error is inversely proportional to the number of data points. This means that the increase in F from 23.2 to 32.8 achieved by short-term blocking represents an increase in precision that would require a data volume

Table III. Two-way ANOVA table for lift coefficient data

Sources of Variance	SS	df	MSE	F	F_{crit} (0.01)	p
Between-Column Blocks	3.36E-04	6	5.59E-05	32.8	3.3	2.46E-14
Within-Column Blocks	2.54E-05	1	2.54E-05	14.9	7.3	0.0004
Interaction	2.10E-05	6	3.50E-06	2.1	3.3	0.08
Total Block Effects	3.82E-04	13				
Pure Error	7.16E-05	42				
Total	4.54E-04	55				

increase by a factor of 32.8/23.2 = 1.41 to achieve by simple replication. This is one way that formally designed experiments achieve equivalent precision results with significantly less data than conventional OFAT designs.

Also note that the F statistic for the within-column block effect of 14.9 exceeds its critical value of 7.3, implying that for this data set, even short-term block effects are statistically significant at the 0.01 level. That is, we can say with at least 99% confidence that the change in mean lift coefficient values between two blocks of replicates acquired 5–10 minutes apart is greater than the within-block differences due to ordinary chance variations. The implications of this result are very important for the design of experiments. It suggests that sample means are not stable even over periods as short as a few minutes. Under such circumstances, the quality of experimental results depends on the order that independent variable levels are set. Setting independent variable levels in sequential order, even over a period of a few minutes, will result in an efficient confounding of independent variable effects with the systematic errors responsible for the block effects.

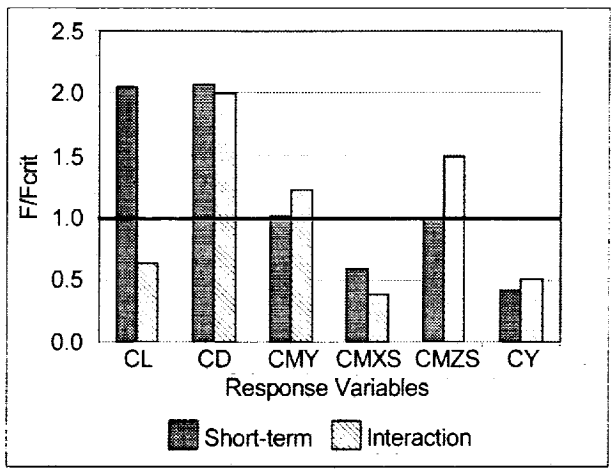


Figure 10. Short term (5-10 min) and short/long interaction F statistics for stability axis block effects.

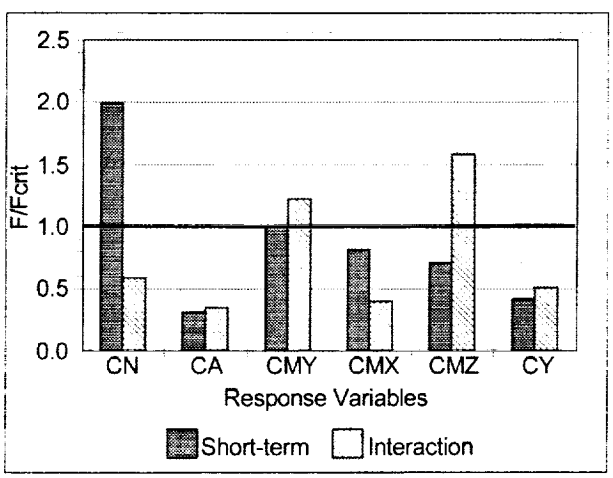


Figure 11. Short term (5–10 min) and short/long interaction F statistics for body axis block effects.

The interaction F statistic is not greater that its 0.01 significance critical value. This means that for lift coefficient, we cannot say with at least 99% confidence that the short-term (5–10 minute) block-to-block shifts are different in such short-term block pairs acquired hours-to-days apart.

The one-way ANOVA results have already shown that long-term (hrs/days) block effects exist for all of the response variables examined. See figures 6 and 7. Figures 10 and 11 display the ratio of measured F statistic to critical F statistic for the short-term block effects (5–10 minutes) and the short/long interactions, for stability axis and body axis forces and moments. Statistically significant effects have ratios greater than one.

Figures 10 and 11 reveal that significant short-term block effects (5–10 minutes) were detected for the coefficients of lift, drag, and normal force. Marginally significant short-term effects (at the 0.01 level) were observed for pitching moment and stability axis yawing moment. No significant block effects were observed in this data set for axial force, side force, body-axis yawing moment, or rolling moment in either coordinate system.

Significant (0.01) interaction effects were observed for drag, for pitching moment, and for yawing moment in both coordinate systems. This means that short-term block effects were larger at certain times during the experiment than at others for these response variables. No significant interactions were observed for lift or for any of the body-axis forces, nor were there interactions for rolling moment in either coordinate system.

There are a large number of possible sources of the systematic variation detected in an experiment. It is not really practical to identify all of them, or to correct all that are identified. The blocking techniques described in this paper, along with randomization methods to be described presently, provide substantial defenses against the effects of such systematic errors, whether they are known or unknown.

Figures 12 and 13 illustrate for stability axis and body axis forces and moments the relative contributions of short-term block effects to the total variance observed in the nominally 15-minute period in which the eight within-column points were acquired in each column of Table I. These were computed using equation 1 for the statistical power of each block effect.

The longer-term block effects (hrs/days) of figures 8 and 9 clearly dominate the total variance. The short-term block effects in figures 12 and 13 contribute much less to the total short-term variance, with pure error due to ordinary chance variations representing the majority of the variance.

The question arises as to how great the contribution must be to have an important influence on overall variance estimates. Interpretations of statistical power calculations (equation 1) generally follow Cohen[2], who provides the following rules of thumb according to reference 4: Large effects have an ω^2 value of 15% or greater. Medium effects have an ω^2 value of about 6%. Small effects have ω^2 value of 1%. By these standards, the short-term block effects found to be statistically

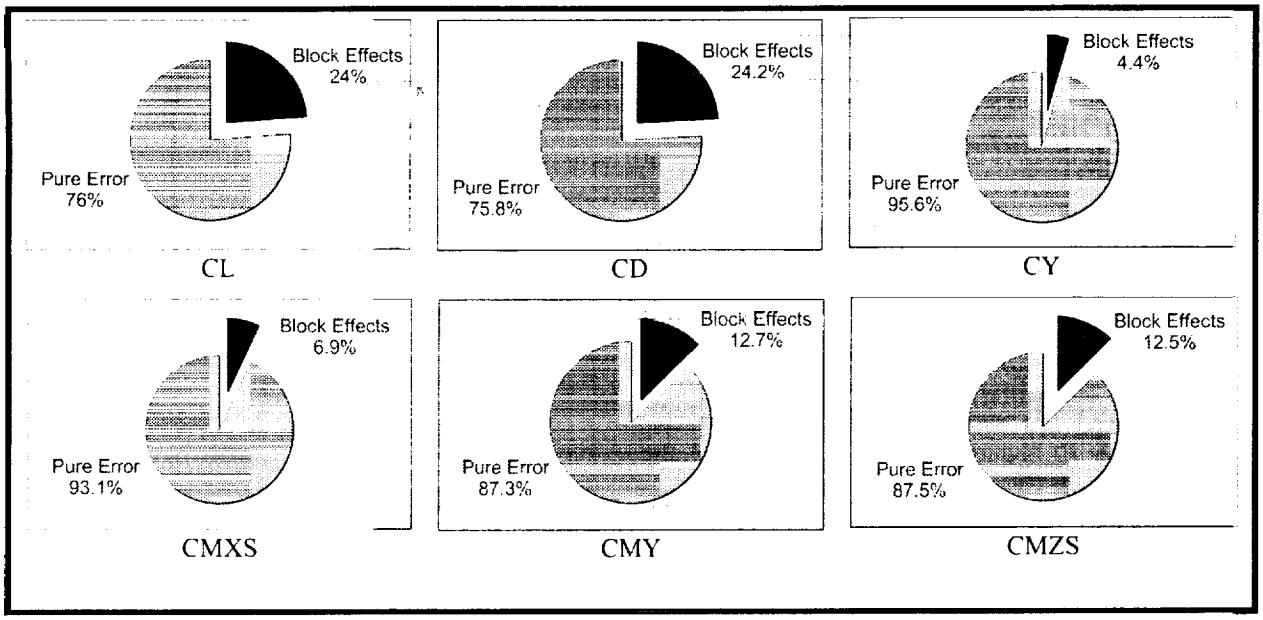


Figure 12. Short term (5–10 min) stability axis block effect contributions to unexplained variance.

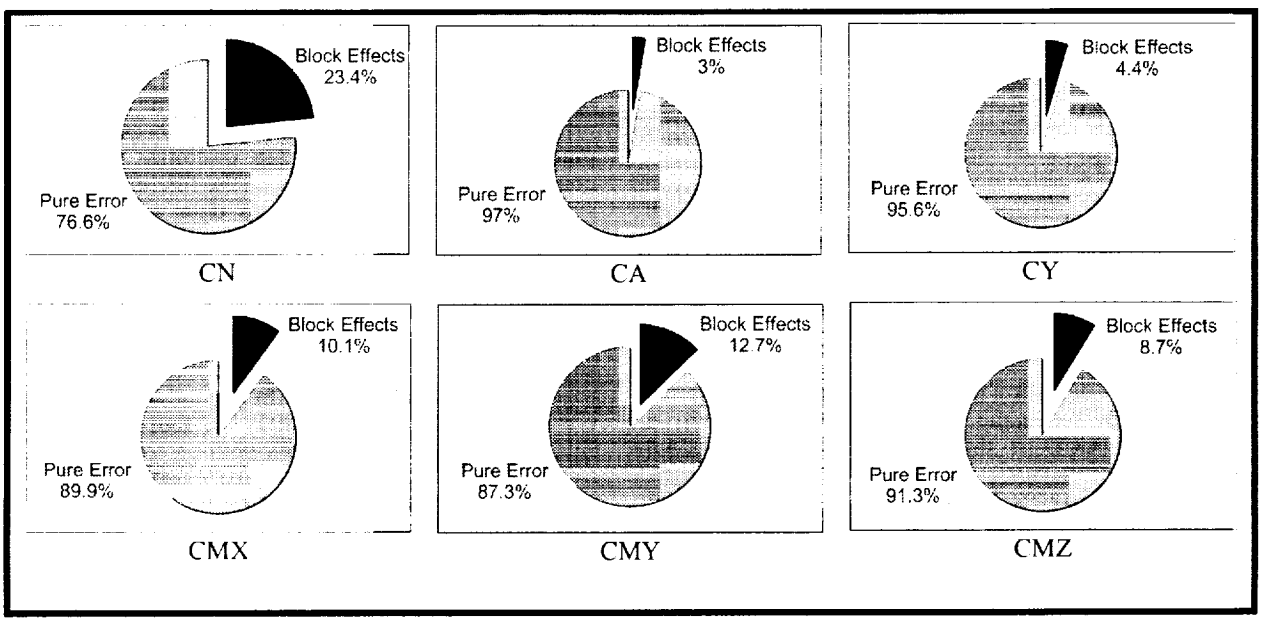


Figure 13. Short term (5–10 min) body axis block effect contributions to unexplained variance.

significant at the 0.01 level are all classified as medium to large effects.

This investigation of systematic errors can be summarized as follows: Significant block effects were detected for the full spectrum of stability axis and body axis forces and moments over periods of hours-to-days. A substantial number of response variables displayed significant block effects over much shorter periods—periods as short as 5–10 minutes. Insofar as the order that independent variables are set can influence the quality of experimental results when block effects are present, these results suggest that run order is important even over periods of time as short as a few minutes. While the details of block effects are likely to vary from tunnel to tunnel, and possibly from test to test within a tunnel or even from time to time within a test, these results suggest clearly that systematic errors cannot simply be assumed not to exist. Their presence motivates the development of explicit measures to defend against them, such as the orthogonal blocking technique illustrated earlier in this paper. The concept of orthogonal blocking illustrated in the simple two-level factorial design of the previous section will be extended in the next section to more complex designs required for wind tunnel response surface modeling.

Blocking for Higher-Order Models

The two-level factorial design considered earlier accommodated orthogonal blocking by confounding uninteresting high-order interaction terms with block effects. Such designs are useful in a broad array of applications but are limited in one important respect: The fact that they feature only two levels of each independent variable means that response models developed from data acquired with such designs can only accommodate first-order and mixed first-order terms. For example, in the two-variable, two-level factorial design described earlier, with variables A and B, it is possible to develop response models featuring the linear A and B terms and a second-order interaction term, AB. The pure quadratic terms (A^2 and B^2) cannot be quantified, however, because to do so requires at least three levels of the independent variables.

A useful extension to the two-level factorial design, accommodating full second-order response models, is due to Box and Wilson.[5] The Box-Wilson, or *Central Composite Design (CCD),* is illustrated in figure 14 for the case of two variables.

The four data points comprising the "square" in figure 14 are two-level factorial design points of the kind considered earlier. They are at the four points defined by all combinations of the "+1" and "–1" coded variables. There are also a number of "center points" in this design, which are replicates of the point at coordinate (0,0) in the coded variable inference space of figure 14. Finally, there are four so-called "star points" arranged on the axes of the coded-variable coordinate system. While there is no restriction in principal on the distance these points can be from the center, there are important advantages if they are at the same distance as the corner points in the "square." In that case the "square points" and "star points" lie on a circle with a radius of square root of two in coded units.

One of the advantages of the equiradial design established by adjusting the star point distances from the center to be the same as the corner points in a two-variable CCD is that this facilitates orthogonal blocking with two blocks that we will refer to as the "star block" and the "square block." The star block consists of the four star points plus half the center points, and the "square block" consists of the four corner points plus the other half of the center points. One can add or subtract a constant from every response variable measurement in the star block, and add or subtract a constant of the same or different magnitude to every response measurement in the square block, and the regression coefficients obtained by fitting a second-order model will be identical to the ones computed without the bias shifts.

This property means that a blocking variable that takes on a value of –1 for the square block, say, and +1 for the star block will be orthogonal to all the regressors in the second-order design matrix. The consequence of this is that any shift in response measurements made in one block relative to those made in the other will have no impact on the numerical value of the regression coefficients. In other words, an experiment blocked in this way will result in the same regression coefficients in the presence of block effects as would have been obtained with a statistically stationary measurement

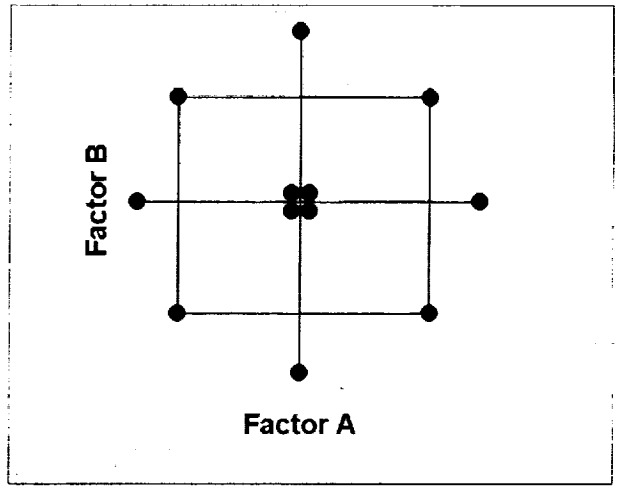


Figure 14. Box-Wilson design in two variables.

environment (no block effects). This has important practical consequences for the design of experiments. Oftentimes the scale of an experiment is simply too great to permit all data to be acquired under homogeneous conditions. As was illustrated in the previous section, it may not even be possible to assume that homogeneous conditions will exist much longer than a few minutes in real-world wind tunnel environments. The possibility of orthogonal blocking in such circumstances is of great practical utility in that it can minimize the adverse consequences of block effects on the fidelity of response surface models produced from data acquired under non-stationary conditions.

This orthogonal blocking feature of the Box-Wilson design is largely responsible for its great popularity as a design for second-order models. We utilized a slight variation of this design in the Generic Winged Body study for which systematic errors were estimated, as reported earlier in this paper. In our variation there were eight center points rather than the four illustrated in figure 14; however, this does nothing to diminish the orthogonal blocking of the Box-Wilson. The only requirement for orthogonal blocking in an equiradial two-variable Box-Wilson is that we assign an equal number of center points to each of the two blocks. (This imposes the obvious constraint that there be an even number of center points in the design.) In our case there were four center points assigned to each block, as described previously.

Table IV displays the Box-Wilson design we employed with angles of attack and sideslip as the two variables. This design was implemented for each configuration studied in the Generic Winged Body experiment. Lift coefficient values are included in Table IV for one particular configuration.

Alpha and beta in Table IV correspond to angle of attack and angle of sideslip. Both are represented in engineering units and in coded values. Block 1 is the square block, with its blocking variable coded with a value of "–1." Block 2 is the star block, with its blocking variable coded as "+1."

An ordinary linear regression was performed on the data in Table IV, fitting the measured lift coefficient data

Table IV. Box-Wilson design with lift coefficient data

SET POINT			CODED			C_L	ELAPSED TIME, Min
BLOCK	ALPHA	BETA	BLOCK	ALPHA	BETA		
1	12	0	-1	0	0	0.53788	0.0
1	10	4	-1	-1	1	0.45030	1.1
1	12	0	-1	0	0	0.53735	1.4
1	14	4	-1	1	1	0.63367	2.0
1	12	0	-1	0	0	0.53987	3.2
1	14	-4	-1	1	-1	0.62933	3.4
1	12	0	-1	0	0	0.53904	5.0
1	10	-4	-1	-1	-1	0.44622	7.1
2	9.17	0	1	-1.414	0	0.41021	8.2
2	12	0	1	0	0	0.53767	8.9
2	12	-5.66	1	0	-1.414	0.54388	10.0
2	12	0	1	0	0	0.53956	11.2
2	12	5.66	1	0	1.414	0.54906	12.3
2	12	0	1	0	0	0.53930	13.8
2	14.83	0	1	1.4142	0	0.66577	14.1
2	12	0	1	0	0	0.54066	15.3

to a full second-order response surface model in the two model attitude variables, as follows:

$$C_L = b_0 + b_1 A + b_2 B + b_{12} AB + b_{11} A^2 + b_{22} B^2 \qquad (2)$$

In this case the b values are the regression coefficients and A and B represent angle of attack and angle of sideslip, respectively. Table V.a. displays the numerical values of the coefficients and the uncertainty in estimating them.

The last two columns in Table V.a. give equivalent information about the quality of the regression coefficient estimate. The column of t-statistics describes the coefficients as a multiples of the standard error ("one-sigma" value) in estimating them. This indicates how many "standard deviations" the estimate of the regression coefficient is away from zero. Larger values impart higher confidence that the regression coefficient is "real"; that is, that the non-zero value of the regression coefficient is not due simply to experimental error. Note the very large value for the t-statistic corresponding to the linear angle of attack term, for example, indicating a very strong linear component in the dependence of lift coefficient on angle of attack.

The right-most column represents the probability that a t-statistic as large as the one determined for a given term in the model could have occurred just by chance, given the uncertainty in estimating the regression coefficient. Values less than 0.05 suggest less than a 5% probability of a chance occurrence due simply to noise, or conversely, greater than 95% confidence that the regression coefficient is non-zero. We will adopt the popular convention of retaining in the model only those terms for which we have at least 95% confidence that the regression coefficient is non-zero. In this case we would drop the AB interaction term and the pure quadratic term for angle of attack and fit a *reduced* second-order model with terms as in Table V.b. This represents our best estimate of a second-order response model for C_L.

We can partition the variance of the entire ensemble of lift coefficient data in Table IV into *explained* and *unexplained* components via an analysis of variance, where "explained" variance refers to differences in the lift data that we can predict, or "explain", by a model with terms as in Table V.b. If we have included genuine replicates in the design of the experiment, as we have done in the case of our CCD via the center points, we can further partition the unexplained variance into "pure error" and "lack of fit" components.

The unexplained variance results in uncertainty in the estimates made using the regression model. The pure-error component of the unexplained variance is due to ordinary chance variations in the data—"experimental error." The lack of fit component results from an inadequate model. If the true C_L response model contained significant third-order terms, for example, then

Table V.a. Regression coefficients for full unblocked C_L response model.

Factor	Coefficient Estimate	DF	Standard Error	t for H_0 Coeff=0	Prob > |t|
Intercept	5.389E-01	1	5.48E-04		
A-AoA	9.099E-02	1	5.48E-04	165.970	< 0.0001
B-Sideslip	1.967E-03	1	5.48E-04	3.587	0.0050
A^2	-1.051E-03	1	5.48E-04	-1.917	0.0842
B^2	3.188E-03	1	5.48E-04	5.815	0.0002
AB	6.350E-05	1	7.75E-04	0.082	0.9363

Table V.b. Regression coefficients for reduced unblocked C_L response model.

Factor	Coefficient Estimate	DF	Standard Error	t for H_0 Coeff=0	Prob > |t|
Intercept	5.384E-01	1	5.07E-04		
A-AoA	9.099E-02	1	5.85E-04	155.428	< 0.0001
B-Sideslip	1.967E-03	1	5.85E-04	3.359	0.0057
B^2	3.188E-03	1	5.86E-04	5.445	0.0001

a second-order model would not provide an adequate fit to the data. This would be revealed in the analysis of variance in the form of a lack-of-fit component of the unexplained variance that was large compared to the pure-error component. Insignificant lack of fit is a necessary (although not sufficient) condition for an adequate response model.

Table VI is the ANOVA table corresponding to the model in Table V.b. The model F statistic is a measure of signal to noise ratio. A large model F implies that the explained variance is large compared to the unexplained variance and that the range of response values that were fit in the regression is large compared to the experimental error. Small F values occur when the regression simply fits noise. In Table VI the explained variance exceeds the unexplained or residual variance by a factor of more than 8000, indicating ample signal to noise ratio. Numbers in the right-most column represent the probability of obtaining the corresponding F statistic simply by chance, given the level of experimental error. Small values for the model term imply adequate signal-to-noise.

The lack-of-fit F statistic represents the ratio of lack-of-fit error variance to pure error variance. In this case there is over three times as much lack-of-fit as pure error, a troubling result that suggests the model is not entirely adequate. This conclusion is reinforced by the p-statistic in the right-most column, which is only marginally above our 0.05 significance threshold. This indicates that we may not be on very firm ground in claiming an adequate fit to our C_L model.

We now repeat the regression analysis, but this time we fit the lift coefficient data to a full second-order model

Table VI. ANOVA table for unblocked C_L model.

Source	Sum of Squares	DF	Mean Square	F Value	Prob > F
Model	6.63E-02	3	2.21E-02	8066.23	< 0.0001
Residual	3.29E-05	12	2.74E-06		
Lack of Fit	2.33E-05	5	4.65E-06	3.38	0.072
Pure Error	9.63E-06	7	1.38E-06		
Cor Total	6.64E-02	15			

Table VII.a. Regression coefficients for full blocked C_L response model.

Factor	Coefficient Estimate	DF	Standard Error	t for H_0 Coeff=0	Prob > |t|
Intercept	5.39E-01	1	4.47E-04		
Block 1	-0.0008				
Block 2	0.0008				
A-AoA	9.10E-02	1	4.47E-04	203.76	< 0.0001
B-Sideslip	1.97E-03	1	4.47E-04	4.40	0.002
A^2	-1.05E-03	1	4.47E-04	-2.35	0.043
B^2	3.19E-03	1	4.47E-04	7.14	< 0.0001
AB	6.35E-05	1	6.31E-04	0.10	0.922

in A and B that is augmented with the blocking variable, Z, and its regression coefficient, d:

$$C_L = b_0 + b_1 A + b_2 B + b_{12} AB + b_{11} A^2 + b_{22} B^2 + dZ \quad (3)$$

The regression coefficients and their uncertainties are presented in Table VII.a.

As in the unblocked case, the coefficient for the AB interaction term is not statistically significant at the 0.05 level and is dropped from the model. Fitting the data to a model without this interaction term produces the coefficients in Table VII.b.

Note that the pure quadratic angle of attack term that was not statistically significant at the 0.05 level in the unblocked model is now significant and is retained in the model. The reason is that the portion of the total variance that is due to the block effect was "unexplained" in the previous model, while in the current model it has been explained by attributing it to the block effect. By removing a portion of the unexplained variance, the blocking has resulted in a smaller residual unexplained variance and thus a higher precision in the estimates of candidate regression model coefficients. In this particular case, this has resulted in our being able to infer with the requisite 95% confidence that a relatively small pure quadratic angle of attack term is in fact real. In the unblocked case we lacked the precision to confidently identify this small effect as real.

The coefficients for blocks 1 and 2 in Table VII.b. indicate that a block effect has produced an average change in the intercept term of 0.0008 between blocks. That is, the estimated response surface is riding on a "DC term" that can differ by 0.0016, depending on which block of time you wish to consider. Whether such an uncertainty in the intercept term of the model is important depends on the precision requirements of the researcher. In the relatively common situation in which the total budget for all sources of error in lift coefficient is 0.001, this block effect would represent 160% of the entire error budget.

Note from the last column in Table IV that the elapsed time to acquire both blocks of data was only about 15 minutes, so that the two block "centers" were on the order of 5–10 minutes apart. This suggests that systematic errors of unknown origin can induce substantial error in a relatively short period of time if there are no defenses provided against them in the design of the experiment.

Table VII.b. Regression coefficients for reduced, blocked C_L response model.

Factor	Coefficient Estimate	DF	Standard Error	t for H_0 Coeff=0	Prob > |t|
Intercept	5.39E-01	1	4.47E-04		
Block 1	-0.0008	1			
Block 2	0.0008				
A-AoA	9.10E-02	1	4.24E-04	214.66	< 0.0001
B-Sideslip	1.97E-03	1	4.24E-04	4.64	0.001
A^2	-1.05E-03	1	4.24E-04	-2.48	0.033
B^2	3.19E-03	1	4.24E-04	7.52	< 0.0001

Table VIII. ANOVA table for blocked C_L model.

Source	Sum of Squares	DF	Mean Square	F Value	Prob > F
Block	9.68E-06	1	9.68E-06		
Model	6.63E-02	4	1.66E-02	11541.20	< 0.0001
Residual	1.44E-05	10	1.44E-06		
Lack of Fit	5.91E-06	4	1.48E-06	1.05	0.456
Pure Error	8.47E-06	6	1.41E-06		
Cor Total	6.64E-02	15			

Table VIII is the ANOVA table corresponding to the model in Table VII.b.

The impact of blocking the experiment can be assessed by comparing the ANOVA table for the blocked and unblocked cases. The lack of fit F statistic for the unblocked case was 3.38, indicating that lack of fit was a considerably greater source of unexplained variance than the pure error due to ordinary chance variations in the data. After blocking this ratio drops to 1.05. The corresponding p-statistic of 0.456 is comfortably above our 0.05 significance threshold for goodness of fit. This suggests that much of the lack of fit in the unblocked case can be attributed to efforts to fit the lift data only to angle of attack and angle of sideslip, when in fact a third important variable—the blocking variable—was also changing during the data acquisition. That is, we tried to fit across a block boundary for which there was a significant shift in the lift measurements due to some unknown systematic error, without taking that shift into account in the model.

Note the substantial increase in the model F statistic—from 8066 for the unblocked case to 11541 for the blocked case. The ratio of these F statistics is proportional to the ratio of unexplained variances, which vary inversely with the number of data points. The ratio of F statistics—11541/8066 = 1.43—thus represents the increase in data volume necessary to achieve the same level of precision by replication that was achieved by simply blocking the experiment. We would have had to acquire 43% more data to achieve the same precision that the blocked experiment provided, or, conversely, by blocking the experiment we could achieve the same precision as the unblocked case with only about 70% of the data.

The enhanced precision provided by blocking allows the regression terms of the model to be identified with much higher confidence. This is reflected in the larger t-statistics (or equivalently, the smaller p-statistics) for the blocked vs. unblocked cases. As noted previously, the increased precision made the difference in being able to even resolve one of the model terms. By being able to say with confidence that a small pure quadratic angle of attack term exists, not only are we able to make better lift predictions with our model, but we gain potentially valuable insights into the underlying physics as well.

The enhanced precision due to blocking is also reflected in the root mean square residual errors. The blocked/unblocked ratio is 70%, meaning that all precision-interval error bars attached to predictions made in the blocked case will be 70% of the corresponding unblocked case. The tighter error bars mean the quality of the prediction model is greater in the blocked case.

Residuals were computed by subtracting predicted values based on the blocked and unblocked models from the actual measured values for lift in Table IV. The average magnitude of the unblocked residuals was 0.00130, compared to 0.00077 for the blocked case. That is, the blocked residuals were only 77/130 = 59% the size of the unblocked residuals on average.

To summarize this illustration of orthogonal blocking for a second-order response surface model, the act of blocking the experiment resulted in a substantial increase in the precision of the experimental result, and an improved insight into the structure of the underlying response relationship. These substantive gains were achieved largely at "no extra cost." That is, the precision enhancement did not require the specification of additional replicates, but merely a reordering of the sequence in which the data were acquired. Moreover, these improvements did not rely upon "assigning causes" to the systematic error; indeed, the cause for the systematic variation from block to block remains unknown. Needless to say, it was not necessary to eliminate the cause of the systematic error, or to correct for it through some calibration scheme that depends on understanding and predicting the systematic variations. The same orthogonal blocking technique would have eliminated the block effect no matter how large it was, or what the cause of it was.

The block effect translates into an uncertainty in the absolute accuracy of the response estimates, in that it essentially shunts what would otherwise be distortions in the shape of the response surface into some uncertainty in the true intercept term of the model. Blocking results in a properly shaped response surface positioned above a constant "DC" reference level that splits the difference between the reference levels in the two separate blocks. Absent any a-priori knowledge of which block represents the "true" reference level, this is about the best one can hope to do in the presence of systematic variations. The coefficient of the blocking variable contributes to an understanding and quantification of the bias error

associated with systematic variations in the facility. The total uncertainty associated with a model prediction should include both the reduced precision error due to blocking, plus a bias error component that reflects the uncertainty in the intercept term.

Randomization

Blocking, and in particular, orthogonal blocking, has been introduced as an effective defense against systematic variations that may occur in an experiment. However, for blocking to be the most effective, it is necessary to identify the block boundaries. Certain obvious candidates such as change of shift or end of day have been discussed. Blocking on arbitrary time intervals has also been demonstrated as a defense against block effects that may or may not occur over those intervals. The question arises as to how to cope with systematic variations when block boundaries are unknown, or that may occur *within* the boundaries of a particular block. Randomization—the act of setting the levels of independent variables in random order—is an effective technique for dealing with within-block systematic variations, as will be illustrated in this section.

Randomization is used in modern experiment design to ensure that changes in response variables are related unambiguously to changes in the independent variables that influence them and that extraneous sources of variation are eliminated. Proponents of the OFAT method try to meet this requirement by holding everything constant except one variable, but this approach relies on assumptions of statistically stationary test conditions that are often violated at the precision level required by modern wind tunnel tests. That is, even though the OFAT practitioner changes no other variable except the one under investigation at a particular phase of the test, other covariate variables are often in fact changing with time. Temperatures vary, instruments drift, flow angularity wanders, the data system gradually requires recalibration, and so on. As a result, the OFAT practitioner does not "hold everything constant," despite the best of intentions, and substantial systematic errors occur in time throughout the experiment.

The OFAT practitioner is made vulnerable to these systematic errors when he implements the high data-rate strategies dictated by current notions of "productivity" in wind tunnel testing. High data rate operations require independent variable levels to be changed sequentially in time in order to maximize data volume—it generally takes less time to move to another independent variable a short distance away and in the same direction as the last change than to change independent variables in any other pattern. However, sequentially changing the independent variables causes responses of the system under study to vary systematically with *time* just as the systematic errors do, *guaranteeing* an efficient confounding of the effects of interest with the systematic errors.

Given the perceived need for sequential level changes in the independent variables to ensure "productivity," the OFAT practitioner is limited to a data quality assurance strategy that depends on identifying and either eliminating or correcting for significant sources of systematic variation that occurs during an experiment. That is, the OFAT practitioner depends upon "statistical control"—a state in which sample means are stable (time-independent)—to obtain high quality research results. Absent statistical control, response changes induced by systematically varying the independent variables will be correlated with any other time-varying response change, including all systematic errors. Under those circumstances, true but subtle independent variable effects induced by systematic set-point changes cannot be distinguished from experimental error components that are also changing systematically with time.

Unfortunately, statistically stationary conditions are difficult to achieve in a research environment as complex and energetic as a modern wind tunnel, as the analysis of variance in representative wind tunnel test results presented earlier in this paper demonstrates. Even in the unlikely event that every detected component of systematic error could be identified as to its source, and every source could be either removed or accounted for by some correction factor, there is no way to know if other sources exist that have simply not yet been detected. Furthermore, there can be no assurances that the existing corrections are stable with time. On the contrary, many almost surely will not be.

Ronald Fisher was the first to recognize that the simple device of randomizing the order in which independent variable levels are set in an experiment liberates the researcher from the need to establish a state of statistical control as a prerequisite for high-quality research results. He notes that randomization procedures provide relief "from the anxiety of considering and estimating the magnitude of the innumerable causes by which . . . data may be disturbed."[6] Numerous subsequent researchers have commented on the virtues of randomization since Fisher introduced it in the 1920s. For example, Brown and Melamed[7] say, "Randomization procedures mark the dividing line between modern and classical experimentation and are of great practical benefit to the experimenter . . . for they provide relief . . . from the classical difficulties of trying to hold everything constant without ever being certain that this has been achieved. [Randomization] provides no absolute guarantee, of course, but it has been judged superior to any alternative yet devised, and its possibility in experimental work is part of what distinguishes experiments, in the strict sense of the word, from quasi-experiments and surveys." Cochran and Cox[8] compare randomization to an insurance policy, describing it as "a precaution against disturbances that may or may not occur and that may or may not be serious if they do occur."

When independent variable set points are randomized, this de-couples the true effects of the independent variable changes from response changes due simply to systematic error. The systematic errors now cause the response variables to change one particular way in time, while the randomized set-point changes cause the response variables to change in a completely uncorrelated way. The systematic error variations are thus segregated from the true effects of independent variable changes.

Randomization physically induces the state of statistical independence among data points that is a

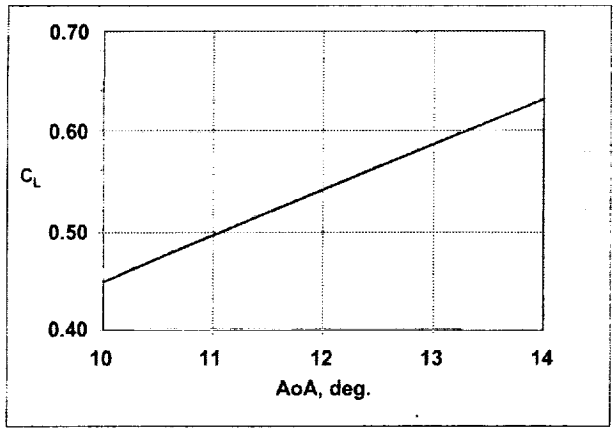


Figure 15. Lift coefficient response function, assumed to be error free for sake of discussion.

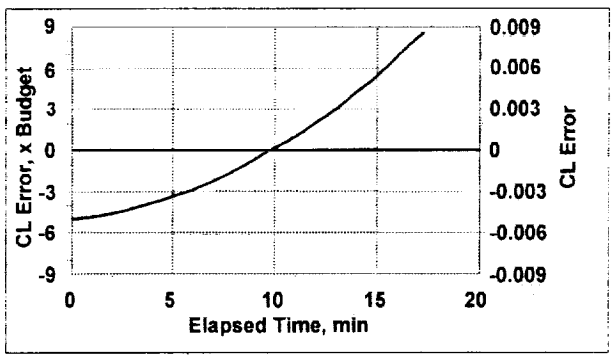


Figure 16. Systematic error in lift coefficient.

prerequisite for so many simplifications and conveniences in analyzing experimental data. Specifically, important distributional properties assumed to characterize certain sample statistics such as the mean and the standard deviation are only truly imparted to those statistics when the data are independent. For example, the mean of a finite sample of data is an unbiased estimate of the general population mean only if the individual observations in the sample are independent. Likewise, the variance in the distribution of sample means is only inversely proportional to the volume of data acquired in the special circumstance of independence among the individual data points. This latter condition is the foundation of all error control strategies that are based upon replication. This is one of the greatest practical reasons that systematic errors are so difficult to eliminate. They do not feature the conveniently symmetrical fluctuations about a mean of zero that enable us to cancel them by replication, as we can do with random errors that are identically and independently distributed. However, randomizing the order in which the independent variables are set tends to induce in systematic errors the same convenient features that make it so much easier to cope with random errors. The remainder of this section will illustrate the effect of randomization for a case in which substantial systematic error is present (several multiples of the entire error budget).

We start with figure 15, which is a graph of lift coefficient as a function of angle of attack at zero sideslip angle based on the model with regression coefficients given in Table VII.b. The fit of this model to the data is believed to be quite good, displaying no significant lack of fit error, ample signal to noise ratio, small root mean square residual errors, and small measured residuals. A detailed analysis of residual patterns was performed that is beyond the scope of this paper. However, that analysis also suggests the model is adequate for estimating lift coefficient over the range of angles of attack and sideslip for which data were acquired. The apparent good quality of this model notwithstanding, there will certainly be *some* error in model predictions, however small. Nonetheless, for the purpose of this discussion of randomization, we will assume that figure 15 represents that most elusive of all commodities in experimental research—Mother Nature's true response function. That is, we will assume that figure 15 represents what the researcher would observe in the absence of all error.

Now imagine that over the relatively short period of time it takes to acquire these data, a significant systematic error manifests itself as in figure 16. The effect of this error is that measured lift coefficients acquired earlier are biased below their true value (negative error component) and measurements made later are biased higher than their true value. This error could be the result of drift in the instrumentation or the data system or it might be the result of flow angularity changes induced by thermal expansion or contraction of the facility, or it could be the algebraic sum of these plus an uncountable number of other causes. The source or sources of the error is unimportant for the purpose of this discussion; all that is relevant is that the error exists and is substantial. In this case we assume that the entire error budget for lift coefficient is ±0.001. The root-mean-square systematic error over the period of data acquisition is 0.0043, which is greater than the entire error budget by more than a factor of four on average, and approaches nine times this error budget for some points acquired late in the series.

Figure 17 shows the original, error-free, lift curve and the lift curve adjusted for the systematic error of figure 16, side by side. This comparison reveals one of the most insidious features of systematic error, which is the difficulty in detecting it. Even though one of these figures is error-free and the other contains errors that exceed the entire budget by several hundred percent, it is difficult to identify which curve contains the error and which one does not. This is true even though the magnitude and specific time dependence of the error is known. It would be virtually impossible to estimate the magnitude of such a systematic error, or even to detect it, without this knowledge. This tends to refute arguments advanced by OFAT practitioners that setting sequential independent variables is a necessary prerequisite for identifying subtle errors as an experiment progresses. The errors in one of the curves in figure 17 are anything but subtle, and yet they are virtually undetectable. (It is the curve on the right in figure 17 that has the systematic errors, incidentally.) The only hope that the OFAT practitioner has is that systematic errors will have the courtesy not to visit themselves upon him during the data acquisition period. The earlier analysis of variance in the

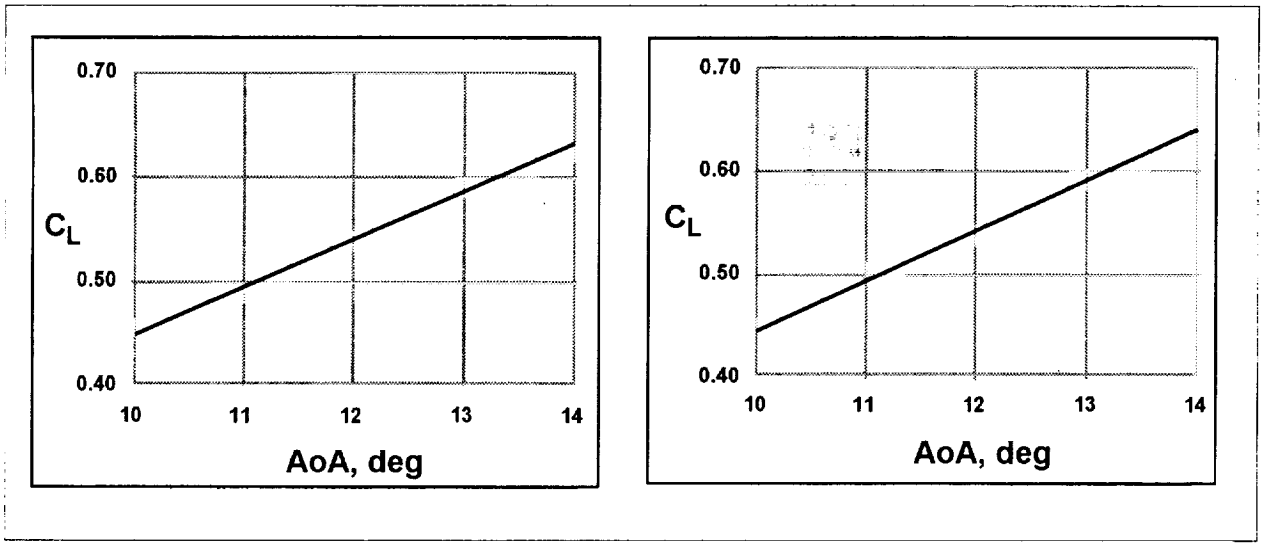


Figure 17. C_L with and without substantial systematic error (over times entire error budget on RMS basis and approaching a factor of nine for some individual points).

Generic Winged Body study suggests that such hopes are likely to be in vain, unfortunately. Systematic errors are in fact a common feature in experimental research, and conditions that are statistically stationary at the precision levels required in modern wind tunnel testing seem to be especially elusive.

As mentioned earlier in this section, Ronald Fisher first proposed randomization as a solution to the insidious problem of substantial but virtually undetectable systematic errors in experimental research. He made this proposal in the 1920s while performing research at the Rothampsted Agricultural Research Station outside of London. Fisher was engaged in experiments to assess the effectiveness of various candidate materials and methods for improving agricultural yield. While it was clear when a plot of ground dressed with some experimental fertilizer produced a higher yield of potatoes than another plot dressed in the conventional way, for example, it was never entirely clear that the fertilizer difference was the cause of any observed difference in yield. Despite the best efforts to enforce uniformity, one could never be certain over the course of a growing season that both plots of land received precisely the same rainfall, or that the slopes of the two fields caused precisely the same amount of water to be retained. Uniformity in soil richness could not be guaranteed, nor that infestations of various pests would affect both the experimental plot and the control plot equally. In short, there was no way to attribute a difference in yield to the treatment under study, or to myriad other factors that could systematically influence it.

Fisher's simple but effective solution was to assign both control and experimental treatments to a number of different plots at random, rather than to a single plot each. While plot-to-plot variability in clay content was inevitable, for example, there was no reason to believe that all of the plots dressed with experimental fertilizer would be sandy and all of the control plots heavy with clay or vice versa, if the two treatments were assigned completely at random to a number of different fields. The same could be said for all other factors. The solution was not to enforce uniformity ("statistical control"), but rather to ensure balance in the design. If the fertilizer treatment comprised the only systematic difference between two equal numbers of otherwise randomly assigned plots, then while some plot-to-plot variance was inevitable, the only explanation for a difference in *average* yield between the two treatments would be the treatments themselves. All other factors would average out.

Wind tunnel research involves units of time as a perfectly analogous counterpart to Fisher's plots of ground. By assigning levels of the independent variables that interest us (angle of attack, say) to different blocks of time at random, factors that might decrease a response early and increase it later on (such as the systematic error of figure 16) will be just as likely to increase the response at a particular independent variable reading as decrease it there. The net effect averaged over a sufficient number of measurements is that the errors will balance and cancel out.

To illustrate this, we will examine how Fisher's randomization idea affects the contaminated lift coefficient data in the right side of figure 17. Figure 18 represents a number of samples taken from the contaminated curve with the angle of attack levels set at random. Each sample is comprised of a fixed component due to the true lift coefficient corresponding to the angle of attack at which the measurement was made, plus a component due to the error varying systematically with time. If a particular angle of attack were set relatively early, the variable component of the measurement would subtract from the fixed component, and the measurement would lie somewhat below the true curve. This is because the systematic error causes negative errors early. If the next highest angle of attack happened to be set late, say, then it would lie somewhat above the true curve because late in the data acquisition period the systematic error is causing positive errors. By assigning angles of

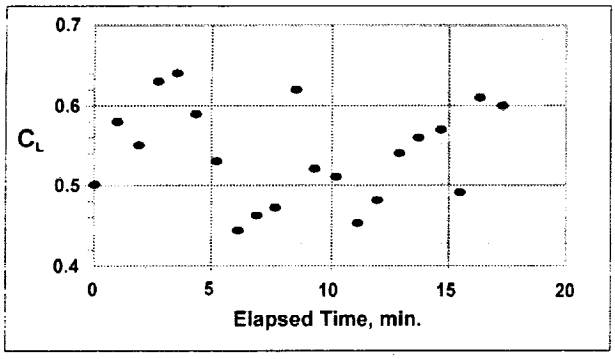


Figure 18. C_L as a function of time. Angle of attack levels set in random sequence in the presence of systematic error.

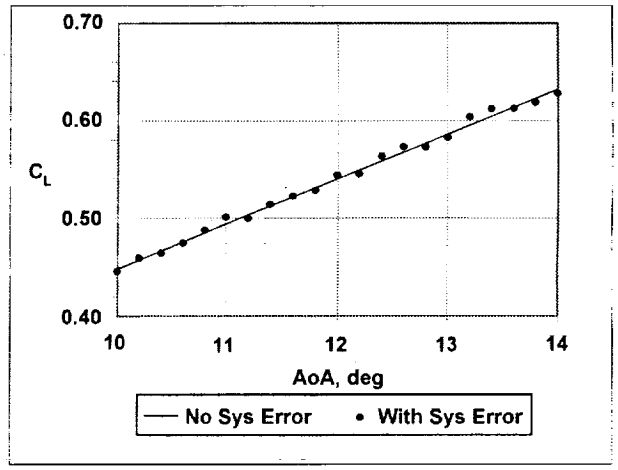


Figure 19. C_L as a function of angle of attack. Angle of attack levels set in random sequence in the presence of systematic error.

attack to periods of time selected at random, we assure that everywhere along the curve, the measured values are just as likely to overshoot as undershoot the true curve.

Examine figure 19 to see this with the lift data of this example. The very same data points acquired with angle of attack set in random order and plotted as a function of time in figure 18 are plotted as a function of angle of attack in figure 19. Each angle of attack had the same chance of being acquired early, when systematic errors were negative, as late, when they were positive. The result is that there is a random distribution of errors above and below the true curve. When angles of attack were acquired systematically with time, the errors were not independent. If the error suffered at the previous angle of attack was positive, the error of the next angle of attack is likely also to be positive. In the case of randomized angle of attack settings, however, the direction of the previous error provides no information about the direction of the next error. That is, randomization enforces the all-important state of statistical independence on the errors that they would not otherwise have.

The difference between setting angle of attack in random order when there are systematic errors present and setting it sequentially in time can be seen in figure 20. On the left is the result of setting angle of attack sequentially in time. You can see that systematic error effects are completely indistinguishable from angle of attack effects. This is because both are correlated with time. The actual shape of the response function is distorted. The slope is wrong and the quadratic term will likewise be corrupted by the systematic error, which generates a general misrepresentation of the response dependence on the variables changed sequentially in time.

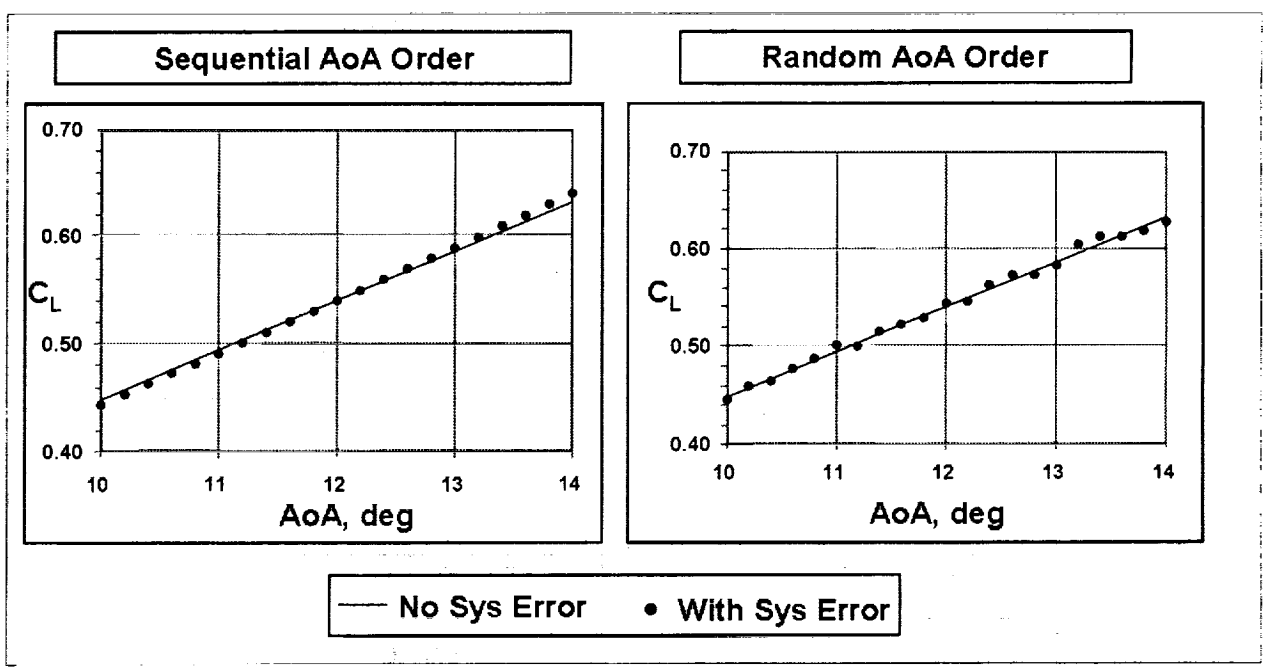


Figure 20. C_L as a function of angle of attack. Angle of attack levels set sequentially vs. randomly.

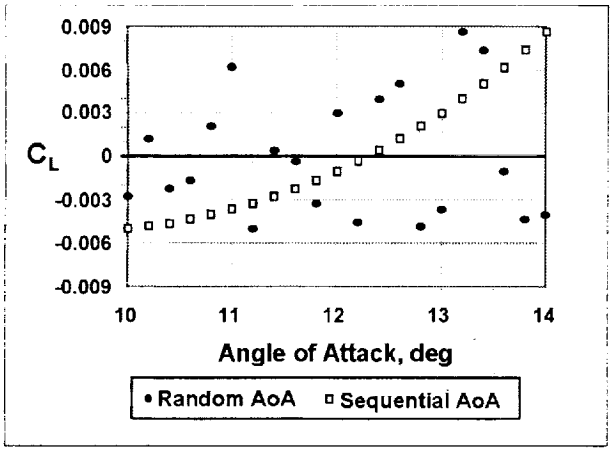


Figure 21. Distribution of errors about True C_L curve. Angle of attack levels set sequentially and in random order.

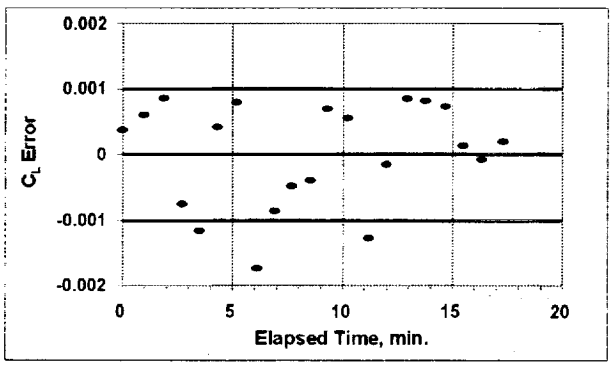


Figure 22. Response model lack of fit error. Comparison of fitted model with true response curve.

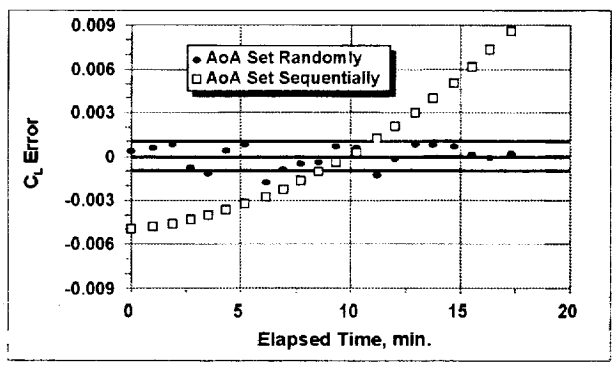


Figure 23. Effect of large systematic error with and without randomizing the independent variable.

On the right of figure 20, randomizing the setting order for angle of attack has produced a set of data with some scatter, but about the true curve. The basic underlying form of the response function is faithfully reproduced, albeit with some scatter about the true line. The scatter results from the effective conversion of systematic error to an additional component of random

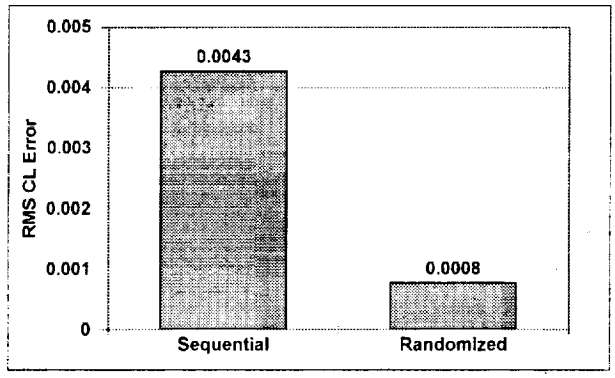


Figure 24. Root mean square lift coefficient errors, sequential and randomized angle of attack run order.

error. If the resulting precision is insufficient to satisfy design criteria, ordinary replication can be used to achieve arbitrarily high precision, since statistical independence of the errors has been induced by randomization.

Figure 21 displays the residuals from figure 20; that is, the difference between measured points and the true lift curve. The effect of randomization is most clear in this figure. For the case in which angle of attack levels were set sequentially in time, the error is obviously systematic and there is no statistical independence from one level to the next. For the randomized angle of attack case the errors are randomly distributed and independent of each other.

The random nature of the distribution of errors in figure 21 implies that arbitrarily high levels of precision can be achieved simply by replication. The independent nature of the errors is a necessary prerequisite for assuring that fundamental assumptions upon which regression analysis depends, for example, are satisfied.

Figure 22 shows the error associated with the best fit of the data acquired by randomizing angle of attack. Because the randomization converted the systematic errors to random errors distributed more or less uniformly around the true lift curve, the difference between the best fit and true lift curves are very small compared to the original systematic error. The bulk of the residuals are within the 0.001 error budget for lift, notwithstanding the fact that the pre-randomization systematic error featured a root mean square error of 0.0043. Figure 23 compares the pre- and post-randomization error directly. In addition to being much smaller in magnitude, the fitting error is random, with all of the attendant conveniences.

Figure 24 compares the RMS systematic error with the RMS fitting error after randomizing the independent variables. The 0.0043 RMS systematic error is a factor of five times greater than the 0.0008 RMS post-randomization fitting error.

Concluding Remarks

Experimental results can depend sensitively on the order in which data are acquired when systematic errors are present. This provides opportunities for increasing the quality of research by optimizing the test matrix

design to take advantage of this fact. An analysis of variance in a recent wind tunnel test revealed substantial systematic errors over periods so short that statistically stationary conditions could not be said to exist for periods longer than a few minutes. Under these conditions, blocking, randomization, and replication were seen to increase research quality significantly. In a recent wind tunnel test precision levels were increased by blocking over periods as short as 5–10 minutes, by amounts that would have required over 40% more data to achieve by simple replication alone. Conversely, blocking provided the means to achieve comparable precision levels with approximately 70% of the data volume. Randomization was shown to reduce RMS systematic errors significantly, and to induce statistical independence in the experimental data. The latter is a prerequisite for regression theory and for certain distributional assumptions to be valid, including the assumption that sample means represent unbiased estimates of general population means.

This paper also suggests that late 20th century concepts of productivity in wind tunnel testing, based on maximum data collection rate, work at cross-purposes to high research quality objectives. The sequential setting of independent variable levels demanded by high-rate data collection strategies guarantees the greatest possible confounding of independent variable effects with systematic errors. The only way for high-speed data collection methods to address this problem is to achieve a state in which experimental results are independent of the order that data are acquired. Such a statistically stationary state is difficult to guarantee, simply because of the impossibility of proving a negative. No matter how much effort is devoted to ridding the experimental environment of systematic variations, it can never be possible to know that there are no more such errors still in play.

The techniques of randomization, blocking, and replication have been used successfully in a broad range of applications since these methods were codified in the formal experiment design methodology first proposed by Fisher and his colleagues in the early part of the 20th century. They are rooted in a sound theoretical basis and have had the benefit of decades of practical experience in numerous industrial, medical, agricultural, scientific, and engineering applications. Adopting these methods to wind tunnel testing is believed to represent a low risk proposition by which substantial improvements can be made in the quality of aeronautical research. Significant first-mover advantages may accrue to those who embrace these methods early.

Acknowledgments

The author is pleased to acknowledge his debt to Dr. James C. Yu for early and continued support of formal experiment design methods as Chief of the Experimental Testing Technology Division. The Langley Wind Tunnel Reinvestment Program, managed by Mr. James A. Osborn, generously supported this effort. Mr. Jeffrey S. Hill, Manager of the National Transonic Facility, has been a constant source of support for high-quality, low-cost wind tunnel testing technology development.

References

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

2) Cohen, J. (1977) Statistical Power Analysis for the Behavioral Sciences (rev. ed.). New York: Academic Press.

3) Laplin, Lawrence L. (1983) *Probability and Statistics for Modern Engineering*. Boston: PWS Publishers

4) Brown, S. R. and Melamed, L. E. (1990). "Experimental Design and Analysis." Quantitative Applications in the Social Sciences, Series No. 07-074. Sage Publications. Newbury Park.

5) Box, G. E. P., and K. B. Wilson (1951). On the experimental attainment of optimum conditions, *J. Roy. Stat. Soc., Ser. B*, **13**, 1.

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

7) Brown, S. R. and Melamed, L. E. (1990). "Experimental Design and Analysis." Quantitative Applications in the Social Sciences, Series No. 07-074. Sage Publications. Newbury Park.

8) Cochran, W. G. and Cox, G. M. (1992). *Experimental Designs. 2nd ed.* Wiley Classics Library Edition. New York: Wiley.