A Practical Methodology for Quantifying Random and Systematic Components of Unexplained Variance in a Wind Tunnel

Richard DeLoach[1], Clifford J. Obara[2], and Wesley Goodman[3]
NASA Langley Research Center, Hampton, Virginia, 23681

This paper documents a check standard wind tunnel test conducted in the Langley 0.3-Meter Transonic Cryogenic Tunnel (0.3M TCT) that was designed and analyzed using the Modern Design of Experiments (MDOE). The test designed to partition the unexplained variance of typical wind tunnel data samples into two constituent components, one attributable to ordinary random error, and one attributable to systematic error induced by covariate effects. Covariate effects in wind tunnel testing are discussed, with examples. The impact of systematic (non-random) unexplained variance on the statistical independence of sequential measurements is reviewed. The corresponding correlation among experimental errors is discussed, as is the impact of such correlation on experimental results generally. The specific experiment documented herein was organized as a formal test for the presence of unexplained variance in representative samples of wind tunnel data, in order to quantify the frequency with which such systematic error was detected, and its magnitude relative to ordinary random error. Levels of systematic and random error reported here are representative of those quantified in other facilities, as cited in the references.

Nomenclature

ANOVA	Analysis of variance
AoA	Model angle of attack, deg
CD	Coefficient of drag
CL	Coefficient of drag
Critical values	Quantitative references for objectively determining the significance of an observed effect
df	Degrees of freedom
Design space	A coordinate system with each axis corresponding to one independent variable. Each "site" in this space represents a unique combination of factor levels. Also called the "inference space."
F-statistic	Ratio of effects variance to error variance
Fcrit	Critical F: Criterion for significant effect
Factor	A variable for which levels changes are planned in the course of an experiment
Level	A particular set-point for a factor
psia	Pounds per square inch absolute
Reference	
Distribution	Probabilities of an experimental instance of some effect under a specified hypothesis
Site	A specific combination of factor levels
t-statistic	Ratio of estimated effect to standard error in estimating it

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

[2] Manager, 0.3-Meter Transonic Cryogenic Tunnel, NASA Langley Research Center, MS 237, Hampton, VA 23681, Member AIAA.

[3] Senior Test Engineer, 0.3-Meter Transonic Cryogenic Tunnel, NASA Langley Research Center, MS 237, Hampton, VA 23681, Member AIAA.

I. Introduction

Wind tunnel tests have been conducted historically for the purpose of constructing aerodynamic databases that characterize some test article of interest. A popular methodology known in the literature of experiment design as One Factor At a Time (OFAT) testing is often used to acquire these data. By this method, all independent variables ("factors") are held constant except one, which is monotonically varied over some prescribed range with increments that are uniform, at least over specified sub-ranges. So, for example, factors such as sideslip angle, Mach number, Reynolds number, and all control surface deflections might be held constant while angle of attack (AoA) is varied from, say, -5° to +2° in 1° increments, then from 2° to 4°, in 0.25° increments, then from 4° to 10° in 1° increments, and finally, from 10° to 20° in 2° increments. A second factor previously held constant is then incremented and the entire AoA sequence is repeated with that second factor held constant at its new level and all other factors held constant at previous levels. This process continues until the second factor has been systematically incremented through its designated range, at which point a third factor is incremented for the first time, and the entire process is repeated until all combinations of all factor levels of interest have been physically set in the tunnel.

Notwithstanding its popularity in experimental aeronautics, the OFAT method is widely characterized in the literature of formal experiment design as inefficient, and also as vulnerable to certain experimental errors that can be avoided when other testing methods are employed[1-4]. This paper is focused on examining the nature and likelihood of such OFAT-specific errors.

The inefficiency of OFAT testing in experimental aeronautics stems from a mismatch between how many factor combinations are of potential interest in a typical wind tunnel test, and how many can be physically set within the limits imposed by typical resource constraints. For example, the prior example of a representative pitch polar featured 27 AoA set-point levels, but consider a relatively small, six-factor test with only 10 levels per factor. Even in such a modest test it would still be theoretically possible to set $10^6 = 1{,}000{,}000$ unique data points. Yet in the most generously supported wind tunnel test there is seldom time to acquire more than a few thousand points when they are set one factor at a time (a few tenths of 1% of the total in this example). The actual number of possible factor combinations obviously varies from test to test, but it is safe to say that the total number of points that could be acquired without resource constraints dwarfs the number that resource constraints impose when one attempts to acquire them one factor at a time.

A common OFAT response is to assert that the entire design space is not necessarily interesting, and certain factor combinations would not be set even if there were no resource constraints. The aircraft is unstable for certain control surface combinations, for example.

Setting aside the plausible counter-argument that pathological flight conditions might actually be among the most interesting and potentially important to understand, especially for tactical aircraft that might be forced into a variety of unanticipated flight configurations as a result of damage suffered in combat, it can be conceded that there is a certain rational prioritization of factor combinations in any wind tunnel test. The issue is one of degree; is it realistic to claim that 99+% of the design space is completely uninteresting? And is it likely that the test engineer, who has sufficiently little knowledge of the test article that the considerable expense of a wind tunnel test is justified, nonetheless is so knowledgeable that he can glean all necessary additional information from the tiny fraction of the design space that can be examined one factor at a time? Even if 90% of the design space were truly uninteresting, a claim that would strain credulity, a test method that permits only a few tenths of 1% of the entire design space to be examined would only allow a few percent of even the "good stuff" to be studied.

While not every OFAT practitioner develops such a quantitative appreciation for the gulf that typically separates what he wants from he has time to get, all are likely to agree that the pressure of time is consistently sensed in OFAT wind tunnel testing. This is because of the inherent inefficiency of the OFAT testing methodology, which requires every site of interest in an enormous design space to be physically visited during the relatively brief interval of a typical tunnel entry. The result is a data collection strategy that places a very high premium on data acquisition rate.

A practical result of this general sense of urgency is that test matrices are designed to maximize speed, and since it is generally the fastest way to acquire data, factor levels are therefore typically changed monotonically. While this strategy of making systematic factor changes accomplishes the goal of maximizing data acquisition rate, it also ensures that the factors are changed systematically with time.

The difficulty with changing independent variables systematically with time is that if some unknown factor is also causing responses such as forces and moments to change systematically with time, the known and unknown effects cannot be separated. This phenomenon is a common occurrence in experimental investigations generally, a fact that has been widely recognized outside the experimental aeronautics community since Ronald Fisher first introduced standard quality assurance tactics to defend against it early in the 20th century[5]. These unknown factors, which change measured responses such as forces and moments just as if the independent variables had been changed, are called "covariate effects."

In wind tunnel testing, a partial list of candidate covariate effects would include such phenomena as flow angularity changes, differential thermal expansion of strain gages and the balance metal to which they are bonded, changing wall geometry effects due to expansion induced by frictional heating caused by high-speed flow, strain gage desiccation effects as moisture gradually leaches out of the gages in a dry tunnel causing sensitivity changes, gradual trip-dot and/or grit ablation, "sting creep" caused by changes in sting bending due to the temperature dependence of Young's modulus, ordinary instrument drift, drift in the data system, and even operator performance variations over time consisting of learning effects—the tendency to improve set-point precision over time with repetition, for example—and fatigue effects that can impair the performance of test personnel and also test equipment because of wear over time. All these covariate effects and countless others not listed here share a common characteristic whenever they are in play: they induce variations with time that are systematic, and not random, and so they cannot be distinguished from the effects of changing ordinary independent variables when those variables are also changed systematically with time.

When covariate effects are in play, they are responsible for what is in effect a time-varying bias error. This error is often difficult to detect because the changes are so gradual, requiring some interval of time before a significant cumulative effect is noticeable. Even then, covariate effects cannot generally be detected without some special effort to do so, which requires time and operating expenses that are seldom budgeted in conventional OFAT wind tunnel testing. Nonetheless, it is possible for these systematic variations to eventually dwarf the random fluctuations that garner so much attention in conventional uncertainty analyses. The most common defense against experimental error in conventional OFAT testing—replication—is ineffective against covariate effects because they are systematic, not random. Furthermore, because these effects do not reproduce exactly in magnitude or sign from test to test, experiments that do not defend against them are unlikely to generate results that can be consistently reproduced within precision limits predicted on the assumption that only random error is in play.

Because of the potentially severe impact of covariate effects, the difficulty in detecting them in conventional wind tunnel testing, and a general unawareness of this class of error within the experimental aeronautics community, it is useful in a check standard test to examine the measurement environment for evidence of covariates. This paper documents the results of a check-standard test designed to detect any covariate effects that might be in play in the 0.3-meter transonic cryogenic tunnel (0.3M TCT) at Langley Research Center.

The formal objective of this experiment was to test a null hypothesis that sample means do no vary with time and that covariate effects are therefore not in play. The a-priori intent was to reject this hypothesis only if compelled by the data to do so, in which case there would be an inference that covariates are in fact in play, and should be taken into account in future wind tunnel testing in this facility. If the null hypothesis could not be rejected with at least 95% confidence, it was to be inferred that significant covariance effects were not in play, which would suggest that standard defenses against covariate effects may not be necessary in this facility. The test was also designed to compare the magnitude of ordinary random error with the magnitude of any detected covariate effects, and to quantify how often covariate effects were observed in this test, if they ever were.

The paper begins with a brief description of the facility and test article. The next section describes the design of the experiment and the analysis methodology, which includes a discussion of unexplained variance, the approach used in this study, data structures adopted to facilitate the approach, and a description of how the test was scaled to ensure a volume of data consistent with precision requirements and inference error risk tolerance. This is followed by a section that details analysis and results. The paper concludes with a general discussion and a summary of principal findings.

II. Facility and Test Article

This test was conducted in the 0.3-meter Transonic Cryogenic Tunnel (0.3M TCT) at Langley Research Center. This facility is especially well-suited to test two-dimensional airfoil sections, featuring adaptable floor and ceiling sections that can be made to conform to flow streamlines, significantly reducing wall effects on the test article. Honeycomb and anti-turbulence screens in the settling chamber further improve flow quality. Figure 1 is a schematic drawing of the facility. One of its unique characteristics is that the flow circuit lies in a vertical plane, with the return path located below the test section.

Higher Reynolds numbers can be achieved in the tunnel's cryogenic mode of operation, but the present test was conducted at non-cryogenic conditions over a Reynolds number range of approximately 4 to 12 million per foot. Figure 2 displays the operating envelope of the tunnel, showing four sites within the envelope where data were acquired in this test. These sites were selected to provide representative coverage within the non-cryogenic subspace of the operating envelope, with specific sites defined by convenient settings of Mach number and total pressure.

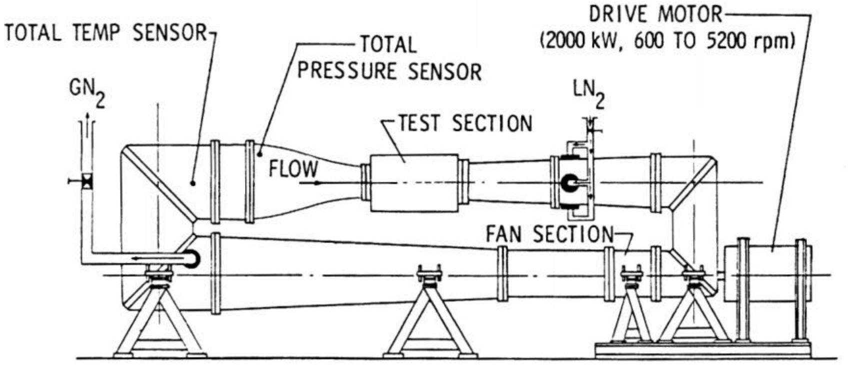


Figure 1. Schematic drawing of the Langley Research Center 0.3M Transonic Cryogenic Tunnel.

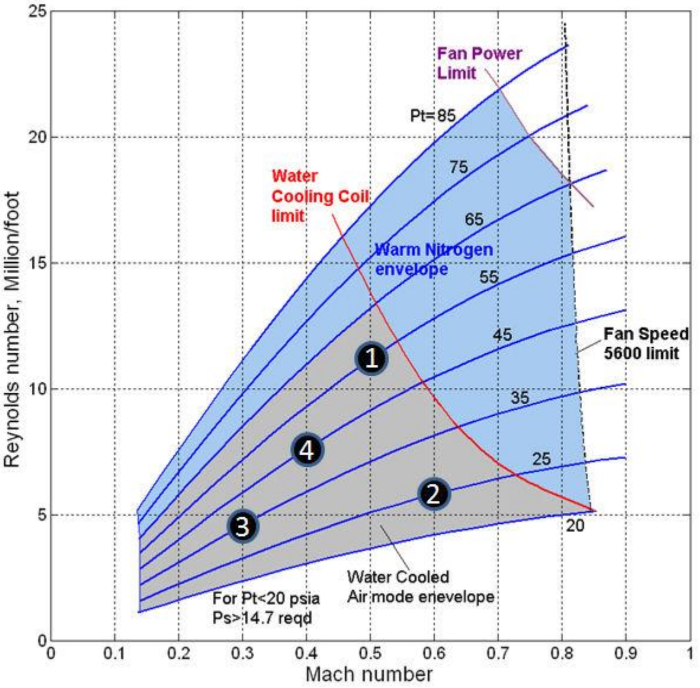


Figure 2. Operating envelope of the Langley 0.3M TCT, showing four sites where data were acquired.

Table 1 summarizes the Non-cryogenic (warm air) operating characteristics of the tunnel.

Table 1. Non-cryogenic (warm air) operating characteristics of the Langley 0.3M TCT.

Mach Number	0.15 – 0.80
Reynolds Number	1 - 13x10^6 per ft
Total Pressure, psia	14.7 – 65
Total Temperature, °F	120

The facility offers a variety of instrumentation for flow visualization, including focused Schlieren, infrared transition detection, and laser velocimetry. Hot films and wires are also available, and pressure and temperature sensitive paint techniques are supported for boundary-layer transition studies. However, in this test the primary measurement system was the facility's traversing wake survey rake, a row of precision pressure transducers attached to a dedicated electronically scanned pressure measurement system. Coefficients of aerodynamic lift and drag were determined from pressure measurements in the test section, and model pitch attitude was measured using an angular encoder.

Flow control is achieved in the 0.3-meter Transonic Cryogenic Tunnel by computer-generated solutions to a multivariable nonlinear control problem. The controller provides control of pressure to ±0.07 psia, and Mach number to ±0.002 during aerodynamic data acquisition in the presence of intrusive geometrical changes like flexwall movement, angle-of-attack changes, and drag rake traverse[6]. The repeatability of drag coefficient is cited as ±0.001 (10 counts)[7].

The test article was an NACA 0012 airfoil section with a 6.5-inch chord length (Fig. 3). This airfoil is symmetrical. The selection of test article was more or less arbitrary as the essence of the experiment is to simply observe any response variations with time when the model attitude and all tunnel state variables are set as consistently as possible to the same levels.

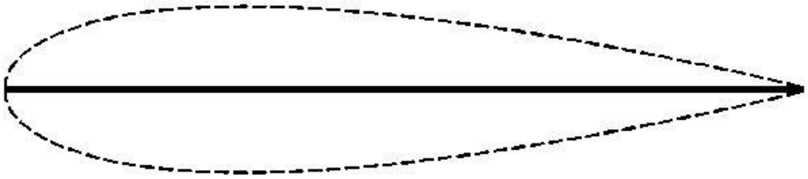


Figure 3. NACA 0012 airfoil section used as test article. Chord length is 6.5 inches.

III. Experiment Design and Analysis Methodology

This experiment was designed as a wind tunnel check standard test rather than a typical research or production wind tunnel test. The chief difference is that while a particular test article is the subject of a normal wind tunnel test, in a check standard test the tunnel itself is the subject of study. The check standard test reported in this paper was developed according to the formal precepts of a testing methodology first applied to experimental aeronautics at Langley Research Center in the mid-1990s. Called the Modern Design of Experiments (MDOE), it was introduced to provide an alternative to traditional OFAT testing that would yield higher quality wind tunnel results (lower uncertainty) in less time and with lower operating costs[8-10].

The Modern Design of Experiments proceeds from a general framework for the analysis of wind tunnel data that is centered on the concept of *variance.* Variance is a general property of any sample of data for which two or more data points differ, and because of ordinary experimental error, it can be reliably assumed that every sample of real data has this property, even if the sample is comprised entirely of ostensibly identical replicates. Unfortunately, within the experimental aeronautics community variance has had a rather restrictive historical association with random experimental error. Variance has therefore been widely regarded as an undesirable property of data, but this relatively unsophisticated view of variance not only fails to account for the effects of changes in the independent variables by which variance is intentionally induced in the data, its focus on random error also understates the role of changing bias errors in determining the quality of experimental results.

Most of the variance in a typical sample of wind tunnel data is unrelated to experimental error, and is in fact intentionally induced. We change angle of attack, Mach number, and control surface deflections among other things, precisely in order to induce changes in test article responses such as forces and moments that can be related to the changes we make in the independent variables. It would be impossible to learn anything about the aerodynamics of a wind tunnel test article absent such planned, or *explained*, variance, and in a perfect world, 100% of the variance in a wind tunnel data set would be of this kind. As it is, the extremely high precision of modern wind tunnel instrumentation ensures that almost all of the variance in a sample of wind tunnel data is in fact explained—*almost* all, but not quite all.

After accounting for all the *explained* variance, there is always some residual, or *unexplained,* variance that remains. We can regard the net observed response changes in a sample of data to be comprised of the algebraic sum of those we intended to induce, and those we did not intend to induce. Much of MDOE analysis is devoted to understanding what portion of the total observed variance is intentional and what portion is due to unknown causes. That is, much of the analysis involves partitioning the total variance into explained and unexplained components that can each be quantified and studied further.

It is because some of the total variance is unexplained that we have uncertainty in experimental results. We must therefore quantify the unexplained variance carefully in order to assess the quality of wind tunnel test results. In the case of check standard testing, the unexplained variance can be used to characterize the general state of the tunnel and its measurement systems over time, serving as a tracer to detect either improvements or deteriorating circumstances.

A. Unexplained Variance and Its Impact

We have noted that the tunnel itself is the focus of a check standard test rather than some particular test article. Another way that a check standard test differs from a conventional force/moment wind tunnel test is that in a check standard test, the focus is more on the *unexplained* variance than the *explained* variance. The unexplained variance can be further partitioned into *random* and *systematic* components. It is important to understand the distinction between these two forms of unexplained variance, and to quantify each of them.

The random component of unexplained variance is commonly regarded as "experimental error" in conventional wind tunnel testing. It imposes its presence through slightly different values that are always recorded when ostensibly identical replicates are acquired. These values are distributed as chance variations about some mean, and there are reliable theoretical foundations upon which to base an assumption that this distribution is normal, or Gaussian.

It is commonly believed—one might more accurately say it is commonly *hoped*—that the random fluctuations observed in replicated data occur about mean values that are stable with time, so that the unexplained variance has only a random component, and no systematic component. A tunnel with such behavior would be said to be in a state of "statistical control." Unfortunately, statistical control is elusive in the real world; random fluctuations in wind tunnel response measurements are often found to occur about mean values that change systematically (not randomly) with time[10-13]. *Box, Hunter, and Hunter* address this in the second edition of their seminal text on experiment design[3] where they say, *"The idea of a process in a perfect state of control contravenes the Second Law of Thermodynamics: Thus an exact state of control is unrealizable, and must be regarded as a purely theoretical concept."* (The reader will recall that the Second Law of Thermodynamics asserts that entropy increases monotonically, so that systems do not remain indefinitely in the kind of highly ordered state that statistical control represents, but tend inexorably from such a state of order to states of ever-increasing disorder).

Wind-off zeros and other common tactics to maintain stability in conventional wind tunnel testing cannot necessarily guarantee the degree of statistical control necessary to ensure that typical 21st-century fractional drag-count precision requirements are achieved at all times throughout the test. For example, while wind-off zeros might be acquired hourly, a change of parts per million in total variance within the intervening interval is sufficient to overwhelm typical fractional drag count error budgets. Such small changes could be attributed to a slight systematic instrument drift that might be induced by temperature changes, for example, or to countless other sources of systematic (not random) variation, some of which were listed in the introduction. Note that the non-random nature of such systematic error ensures that it is impervious to replication as a quality assurance tactic.

A systematic component to the unexplained variance in a sample of wind tunnel data has deleterious effects beyond the simple addition of another source of "noise" that would tend to mask the true "signal" in the data. Ordinary random error has that masking effect, but systematic experimental error is potentially even more troublesome.

The fundamental assumption of experimental research is that statistics such as the means and variances that characterize a sample of experimental data are in fact unbiased estimators of the corresponding population parameters they represent. The population parameters—sample statistics that would be computed for an *infinite* volume of unbiased data—represent "truth," which can only be approximated with a finite volume of data. For example, if a sample of 10 replicated measurements of lift coefficient are acquired under ostensibly identical cruise conditions and averaged, the resulting mean value would only be an estimate of the true lift. It cannot be assumed to represent the true lift exactly because of experimental error in each of the finite number of individual measurements that were averaged, but we would like to be able to assume that the sample mean is at least an unbiased estimate of the population mean.

Unfortunately, this crucial assumption about the unbiased nature of finite sample statistics is hostage to another important assumption; namely, that each measurement is independent. This means that the experimental error associated with each measurement must be just as likely to bias it high as low. Such errors can be expected to cancel to some degree when the measurements are averaged, with the sample mean just as likely to be a little high as a little low with respect to the population mean, and thus an unbiased estimator of it.

Statistical independence is implicitly declared in OFAT wind tunnel testing but standard quality assurance tactics designed to ensure it, which are widely used outside the aeronautics industry, usually consume more time than the high-speed data collection imperative of OFAT testing will permit. Unfortunately, the ubiquitous presence of covariate effects of the kind described in the introduction virtually guarantees that factor levels changed systematically with time will feature errors that are correlated with each other and are therefore not independent. This will bias the sample statistics, as Fig. 4 illustrates.

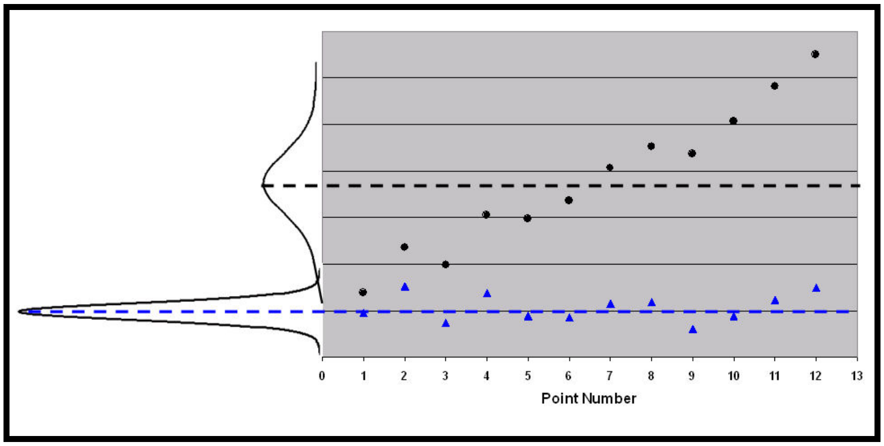


Figure 4. Replicates acquired with systematic covariate effects in play (black dots) and without systematic covariate effects in play (blue triangles). The covariate effects cause a bias in both the mean and the variance of the sample.

Because of covariate effects that bias the black-dot sequential measurements ever higher in Fig. 4, it is not just as likely that the next measurement will be biased high as low. This figure illustrates that the effect of correlated errors is to bias both the mean and the variance of the sample relative what they would be absent any systematic covariate effects, as illustrated by the blue triangle data sample.

The adverse impact of covariates can be understood in less statistically rigorous terms by simply noting that when independent variables are changed systematically with time while covariate effects are in play, it is impossible to say how much of the observed response change is attributable to the changes made in the independent variables, and how much is due to covariate effects. For example, if covariates gradually inflate estimates of lift coefficient over an interval of time during which the angle of attack is also changed systematically with time, then later lift estimates, acquired at higher AoA settings, will be biased higher. The result could be a lift polar that is rotated counter-clockwise relative to the polar that would have been acquired without covariate effects. Likewise, covariate effects that are in play entirely between two ostensibly identical polars will result in a shift of one polar relative to the other. If covariates occur at some varying rate while a polar is being acquired, they can induce artificial structure into the polar that is not observed when the polar is acquired again later on.

Most experienced wind tunnel practitioners can attest to the fact that such differences are not uncommon when OFAT data are repeated, and that it is difficult to obtain wind tunnel results that are consistently reproducible within the tight tolerances that are often specified (and claimed) for them based on considerations of random experimental error alone. This is because the necessary prerequisite for high-quality test results when independent variables are changed systematically with time is that nothing else that influences system response can be changing simultaneously.

Unfortunately such stability cannot always be reliably assumed, notwithstanding how much effort is devoted by OFAT practitioners to marketing this crucial assumption. Absent such stability, OFAT sample statistics are virtually guaranteed to differ from their corresponding population parameters, which is to say that OFAT testing makes high and perhaps unrealistic demands of the measurement environment as a condition for generating reliably reproducible results. Part of the objective of this test has been to quantify how often such idealized conditions exist in a typical wind tunnel test. We note in passing that the MDOE method is designed to achieve reproducible results even when unknown covariate effects are in play and high stability cannot be assumed.

B. Experimental Approach

We seek to quantify the total variance in a check standard test and then to partition the unexplained component into random and systematic constituents that can be quantified and further examined. The *random* component of the unexplained variance can be quantified directly if the check standard test design includes genuine replicates. It can also be quantified without genuine replicates under conditions for which there are no significant "block-factor interactions," as will be detailed presently.

The *systematic* component of the unexplained variance can be quantified by acquiring the same data sample at different points in time to test for changes in the sample mean. Changes in the sample mean over time that are too great to associate with ordinary random error can be attributed to covariate effects of the type described in the introduction, which are responsible for the kind of slowly varying bias error that results in a systematic component of the unexplained variance.

These covariate effects may be associated with some cause that can be inferred from the behavior of the data and if so, some action might be suggested that could reduce the effect. The more common situation is one in which the specific causes of systematic experimental error remains as much a mystery as the specific causes of random experimental error, in which case results may be more reproducible if standard quality assurance tactics used for decades in other industries to cope with ubiquitous covariate effects are incorporated into the design of wind tunnel tests in this facility.

C. Data Structure and Scaling the Experiment

Perhaps the most conspicuous distinction between OFAT and MDOE testing methods is the difference in attitude toward data volume. OFAT testing is fundamentally a high-volume data-collection process, for which the goal is always to *maximize* the volume of data that can be acquired within constraints dictated by finite resources. MDOE testing is oriented toward achieving specific technical objectives at low cost and in as little testing time as possible, which usually translates into *minimizing* the volume of data to be acquired. In practice, this means that establishing the smallest volume of data adequate to a particular objective, called "scaling" the experiment, is one of the first tasks in the MDOE experiment design process.

The current test was designed to quantify the unexplained variance in representative samples of wind tunnel data, to quantify the frequency with which a significant component of systematic error was detected, and to determine the magnitude of any such systematic error relative to ordinary random error. While random error can be easily detected through genuine replicates and by other means to be discussed, systematic error is harder to quantify. We can justify an inference that systematic error is present only if we can detect a difference in the means of ostensibly identical data samples acquired at different points in time. For this reason, the basic data structure for this test consisted of the same sample of data acquired at different times throughout the tunnel entry.

To facilitate a direct measure of the *random* component of unexplained variance, a decision was taken to replicate the base data sample each time it was acquired. Scaling the test then reduces to determining how often the sample is to be acquired throughout the tunnel entry, and how many points are to be acquired in each sample, half of which would be unique and the remaining half of which would replicate the first half. The unique and replicated points within each sample were acquired in random order.

In a typical MDOE scaling exercise, one determines the smallest (least costly) volume of data necessary to ensure an acceptable probability of detecting response modeling errors that are large enough to be of concern. Specifying how large the error can be before it is regarded as troublesome, and specifying the required probability of detecting an error of such magnitude, are both part of the hard work of an MDOE experiment design.

The scaling of a check standard test differs from a typical MDOE scaling exercise in that a common data sample is simply reproduced after different time intervals, and the only basis for determining how many times to reproduce the sample is that it should occur often enough that any resulting estimate of systematic unexplained variance will be based on a reasonable number of degrees of freedom. The samples should also be reproduced at intervals of time that are likely to be long enough for cumulative effects of practical concern to develop.

For the present test, it was more or less arbitrarily decided to acquire the same sample of data in the morning and the afternoon of each test day, and that the test would be comprised of 10 test days. This would ensure 20 ostensibly identical samples of data, so an ample 19 degrees of freedom would be associated with any estimate of the variance in sample means over the entire duration of the test. Systematic variance estimates over shorter periods could also be estimated with a reasonable number of degrees of freedom. The test was essentially structured as four independent experiments, each consisting of 20 data samples that differed only by the date and time they were acquired, with each of the four experiments conducted at a unique combination of Mach number and total pressure (hence Reynolds number) as displayed in Fig. 2.

The scaling task now reduces to deciding how large a sample to acquire in the morning and afternoon of each of the 10 test days. This part of the scaling task proceeded along typical MDOE planning lines, in which the basic structure of the test matrix is influenced by the specific analysis to be performed on the data. Specifically, the design of the present experiment was heavily informed by the analysis of variance (ANOVA) planned for the data.

In a typical ANOVA, the data are arranged in rows and columns and the total variance of the data is partitioned into components. In a simple one-way ANOVA in which each column is comprised of genuine replicates of ostensibly identical points, the total variance of the data is partitioned into a component attributable to variations in the data across columns (the so-called "treatment variance") and a component of within-column variation that is attributable to ordinary random error. We declare the treatment variance to be significant if it is sufficiently large compared to the within-column (or "error") variance that it can be detected with some prescribed level of confidence such as 95%. A 95% level of confidence means that if two or more column means are declared different from each other, the probability is less than 0.05 that this will be an erroneous inference based on an unlucky combination of random errors.

A two-way ANOVA can also be applied to a sample of data organized into rows and columns to partition the total variance into components. The difference between a one-way ANOVA and a two-way ANOVA is that in a two-way ANOVA, the total variance of the data sample is partitioned into a component attributable to variations across *rows*, as well as the components attributable to variations across columns and the residual error variance attributed to ordinary chance variations in the data. A two-way analysis of variance accommodates data samples in which the columns are not simply composed of replicates of the same point, but can consist of different points. This would be the case, for example, when the rows correspond to different angles of attack, say, and the columns correspond to different blocks of time. In such a case, we would explain the row-wise variations by attributing them to changes in the angle of attack. A two-way ANOVA can therefore be applied to analyze a set of ostensibly identical pitch-polars acquired at different times, for example. The row-wise variation would be expected to be the dominant component of the total sample variance, but it would also be the least interesting in an analysis of unexplained variance. The more interesting question would be whether the column-wise variance, associated with variations in time, is large enough to be attributed to anything other than ordinary random error. If so, this would constitute objective evidence that covariate effects were in play.

The intent was to design the current test to accommodate a two-way ANOVA in which measurements of lift coefficient and drag coefficient were organized into rows and columns, with rows corresponding to angles of attack and each column corresponding to the point in time when the acquisition of that data sample began, for example. It has already been established that the number of columns was rather arbitrarily selected to be 20, based on the number of available test days and a resolution to acquire one data sample in the morning and one in the afternoon. It remains to quantify how many data points to acquire in each of the 20 ostensibly identical samples that are to differ only by when they were acquired.

We wish to acquire enough data in each of the 20 samples so that if the ANOVA does indicate that two or more sample means differ by more than random error can explain, we can infer whether this difference is large enough to be important for practical purposes. This naturally leads to certain related questions that must be answered first. We must decide how large a difference is "large enough." We must also recognize that whether or not two samples actually do differ by this amount, we will have to make an inference by overtly declaring that they do or that they do not, and in either case we might be right or we might be wrong.

There are two ways to make an inference error in this case: We can declare a significant difference between two sample means when in fact there is none (a so-called "Type I" or "alpha" inference error), or we can declare that there is no significant difference when in fact there is one (a so-called "Type II" or "beta" inference error). The probability of making inference errors of either type approaches zero asymptotically as the sample size approaches infinity, and can be driven to arbitrarily low non-zero levels by acquiring a sufficiently large but finite volume of data. Typically during the experiment design process we declare our inference error risk tolerance by specifying acceptable probabilities of committing Type I and Type II inference errors, and then calculate how much data is required to ensure that the risk is no greater than we have specified.

It can be shown[1] that in a measurement environment characterized by irreducible random experimental error with a known standard deviation of σ, the volume of data in each of two samples that is necessary to resolve a difference in sample means of δ, with Type I and Type II inference errors that do not exceed α and β, respectively, is given by this formula:

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

The quantities z_α and z_β are normal deviates that are tabulated in as a function of α and β in standard statistical tables. They can also be calculated with the NORMSINV function of Excel. In this experiment, *n* represents the number of rows in the ANOVA data structure that features 20 columns corresponding to different points of time.

The quantity δ represents the smallest difference in two sample means that is too large to attribute to random error. A convenient metric for this difference is the 95% LSD (Least Significant Difference), which represents the smallest difference between two replicates that can be resolved with 95% confidence. The 95% LSD can be expressed as a multiple of σ as follows:

$$95\%\ LSD = \left(2\sqrt{2}\right)\sigma \tag{2}$$

Insert Eq. (2) into Eq. (1):

$$n = \left(\frac{z_\alpha + z_\beta}{2}\right)^2 \tag{3}$$

Thus, given the 95% LSD criterion, the sample size depends only on inference error risk tolerance.

For this test, a value of 0.05 was specified for α, the maximum acceptable probability of erroneously rejecting a null hypothesis asserting that the difference in sample means is too small to attribute to anything other than random error. We would commit such an error if we claimed that some shift in response measurements occurred over time that was too large to attribute to random error, when in fact such a shift had *not* occurred. Stating it another way, we wish to be 95% confident of any assertion we make claiming more change in measured responses over time than can be attributed to random error.

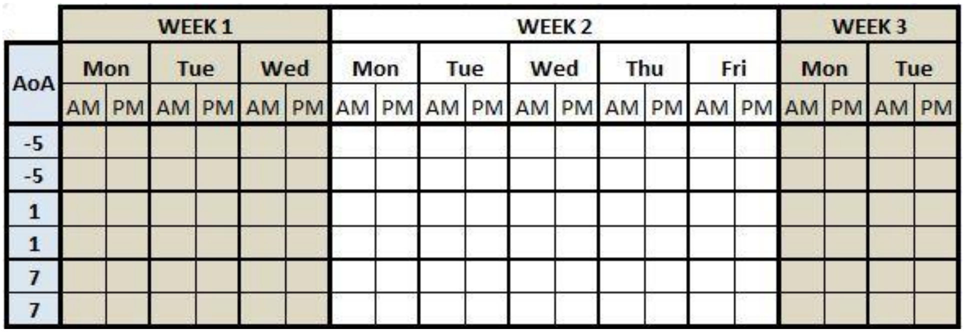
The quantity β represents the maximum acceptable probability of erroneously rejecting an *alternative* hypothesis by asserting that no systematic response shift had occurred over time when it fact it had. This is judged to be the more serious error, since erroneously rejecting the *null* hypothesis simply results in a somewhat more conservative statement of the quality of the experimental outcome (a claim of more experimental error than is actually in play), while erroneously rejecting the *alternative* hypothesis results in a claim of greater quality than we are entitled to assert. In addition to misrepresenting our results, this could also give rise to experimental claims that could not be reproduced later within specified tolerance levels. Since erroneously rejecting the alternative hypothesis has greater consequences, we specified for this experiment a maximum acceptable inference error probability of $\beta = 0.005$, 10 times smaller than the acceptable probability of erroneously rejecting the null hypothesis.

The corresponding normal deviates for $\alpha = 0.05$ and for $\beta = 0.005$ are 1.960 and 2.576, respectively. (The former corresponds to a double-sided normal probability distribution while the latter is for a single-sided distribution). Substituting these values for z_α and z_β in the formula for *n* above:

$$n = \left(\frac{z_\alpha + z_\beta}{2}\right)^2 = \left(\frac{1.960 + 2.576}{2}\right)^2 = 5.143 \approx 6 \tag{4}$$

We therefore required each ostensibly identical sample of data to consist of six measurements. Because we wished to incorporate genuine replicates in the design, the final data structure consisted of three unique angles of attack, each replicated, that were acquired in the morning and the afternoon of 10 test days at each of the four sites in the operating envelope that are indicated in Fig. 2. The angles of attack selected for this test were -5°, 1°, and 7°. The final data structure was then as in Fig. 5, with coefficients of lift and drag measured in each cell and the angle of attack levels set in a different random order for each column of data.

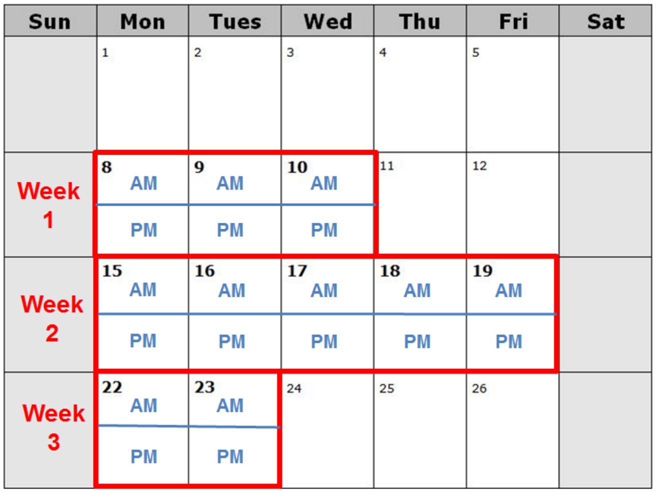
The original plan was for the 10-day tunnel entry to span two five-day work weeks, but technical difficulties that occurred on Thursday of the first week were not resolved until the end of the next day. The result was that the tunnel entry actually spanned parts of three work weeks. The first week consisted of the initial Monday–Wednesday interval, the second week consisted of the following Monday–Friday interval as planned, and a third "week" was added consisting of the subsequent Monday and Tuesday to make up for the two days that were missed in the first week.



AoA	WEEK 1						WEEK 2										WEEK 3			
	Mon		Tue		Wed		Mon		Tue		Wed		Thu		Fri		Mon		Tue	
	AM	PM	AM	PM	AM	PM	AM	PM	AM	PM	AM	PM	AM	PM	AM	PM	AM	PM	AM	PM
-5																				
-5																				
1																				
1																				
7																				
7																				

Figure 5. Basic data structure, supporting a two-way ANOVA, executed at four unique sites within the operating envelope of the tunnel.

Figure 6 illustrates the test schedule. This unanticipated alteration to the original schedule was something of a blessing in disguise, as it made for three natural intervals of differing durations over which to examine the tunnel for covariate effects. The experiment thus permitted tests for the presence of covariates to be executed for four tunnel states by three time intervals by three model attitudes, for a total of 36 cases for lift and 36 for drag.



Sun	Mon	Tues	Wed	Thu	Fri	Sat
	1	2	3	4	5	
Week 1	8 AM PM	9 AM PM	10 AM PM	11	12	
Week 2	15 AM PM	16 AM PM	17 AM PM	18 AM PM	19 AM PM	
Week 3	22 AM PM	23 AM PM	24	25	26	

Figure 6. Three intervals of time, November 2010.

IV. Analysis and Results

The analysis of experimental results from this test proceeds along two lines. A check standard test is an opportunity to assess the state of a wind tunnel by quantifying the random and systematic components of unexplained variance in representative samples of data acquired in that facility. However, while today most experimental aerodynamicists take for granted the reality of ordinary *random* unexplained variance, regarding it as a natural and inevitable element of wind tunnel testing, much of the industry takes a rather different view of *systematic* unexplained variance. The systematic component of unexplained variance generated by covariate effects is no less natural and inevitable than the random component, but there is an inclination among wind tunnel practitioners to ignore it completely, notwithstanding its potential to be the dominant source of uncertainty in a wind tunnel test.

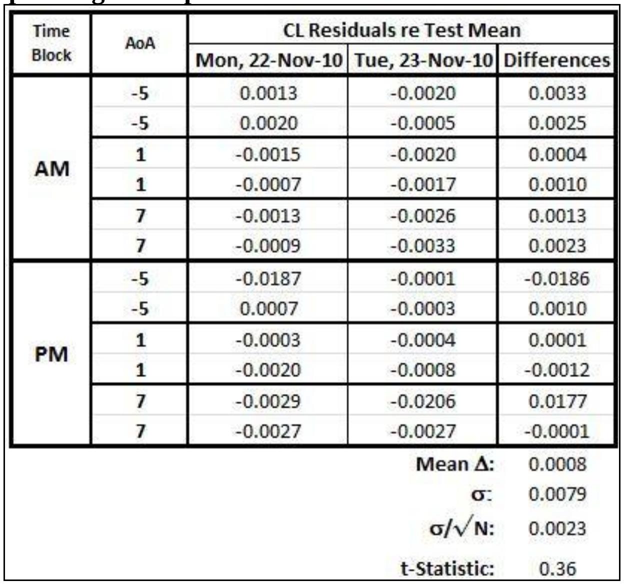
Detailed speculation about motives for the short shrift given to systematic unexplained variance is beyond the scope of this paper, but it may be related in part to a visceral appreciation of the adverse impact that bias errors changing slowly with time would have if all independent variables were also varied systematically with time, as required by the high-speed imperative of OFAT testing. The argument seems to go that 1) OFAT testing is popular, 2) reliably reproducible OFAT results could not be consistently achieved if systematic covariate effects were in play, ergo 3) they must not be in play. In fairness, many OFAT practitioners probably rely on quality assurance tactics such as wind-off zeros to protect them from covariate effects, but the fact that consistently reproducible OFAT wind tunnel test results can at times be elusive remains a dot that is yet to be connected in this argument.

Because of the reluctance of the OFAT community to even recognize systematic unexplained variance, much less defend against it, part of the present study is devoted to formal tests of a null hypothesis stating that there is no significant difference between the means of data samples that are ostensibly identical except for the date and time they were acquired. It is only when compelled by the data to reject such hypotheses that systematic unexplained variance will be claimed.

A. Hypothesis Testing with Paired t-test

Formal hypothesis testing entails the same basic elements regardless of the details of the hypothesis being tested, which can be illustrated with a simple example that utilizes data acquired at Site 1 of the tunnel's operating envelope (see Fig. 2), during what is labeled "Week 3" of this test in Fig. 6. These data are displayed in Table 2, which presents ostensibly identical lift coefficient data acquired on Monday and Tuesday of the same week.

Table 2. Ostensibly identical samples of lift coefficient data acquired on successive days at Site 1 in the tunnel operating envelope.



Time Block	AoA	CL Residuals re Test Mean		
		Mon, 22-Nov-10	Tue, 23-Nov-10	Differences
AM	-5	0.0013	-0.0020	0.0033
	-5	0.0020	-0.0005	0.0025
	1	-0.0015	-0.0020	0.0004
	1	-0.0007	-0.0017	0.0010
	7	-0.0013	-0.0026	0.0013
	7	-0.0009	-0.0033	0.0023
PM	-5	-0.0187	-0.0001	-0.0186
	-5	0.0007	-0.0003	0.0010
	1	-0.0003	-0.0004	0.0001
	1	-0.0020	-0.0008	-0.0012
	7	-0.0029	-0.0206	0.0177
	7	-0.0027	-0.0027	-0.0001
			Mean Δ:	0.0008
			σ:	0.0079
			σ/√N:	0.0023
			t-Statistic:	0.36

If the Monday and Tuesday data were identical and the differences recorded in the third column of Table 2 were therefore zero for all 12 pairs of data points, it would be difficult to make the case that any systematic change occurred between the times that the Monday and Tuesday data samples were acquired. Absent any consideration of random experimental error, one might therefore interpret the non-zero (0.0008) average difference in lift coefficient as evidence of some systematic, albeit relatively small, overnight change in lift.

Unfortunately, the fact that each data sample was comprised of a finite and not particularly large number of individual measurements, each susceptible to ordinary chance variations, complicates the interpretation of a non-zero difference in the two sample means. Because of the random error we can no longer use "zero difference" as the criterion for no systematic change. If we construct a null hypothesis that there is no difference between the Monday and Tuesday samples, we are only entitled to reject that hypothesis (and infer a true difference) if the observed difference is too large to attribute to the random error of the data sample.

It is convenient to construct a test statistic that represents the difference in sample means in terms of the standard error in estimating that mean. The sample standard deviation for the 12 differential lift coefficient values was 0.0079, which we divide by the square root of the sample size (12) to compute the standard error in the estimate of the mean. Thus, we would report a mean difference of 0.0008, with a value of 0.0023 as the standard error in the estimate of that mean. Since the estimate of the mean is less than the standard error in estimating it, we are unable to reject the null hypothesis and in this instance we are not entitled to infer any systematic difference between the samples acquired on Monday and Tuesday.

If we normalize the estimate of the mean by the standard error in estimating it, we can create a dimensionless test statistic (a so-called "t-statistic") that expresses the mean as a multiple of the standard error: t = 0.0008/0.0023 =

0.36. The formal hypothesis testing process entails comparing such a test statistic with some criterion established a-priori. We either reject the null hypothesis or not, depending on whether the t-statistic is greater than this criterion.

The criterion is based on a "reference probability distribution" that quantifies how often, because of random error, a t-statistic of a given non-zero magnitude can be expected under the null hypothesis, when its true value should actually be zero. For a sufficiently large sample size, the Central Limit Theorem assures us that this distribution is Gaussian. However, for smaller samples a t-distribution more accurately reflects how often one should expect to see a t-statistic of a given size under the null hypothesis. The t-distribution approached a Gaussian distribution as the size of the sample increases. For a sample size of 12 as in the current example, there is very little difference between the two.

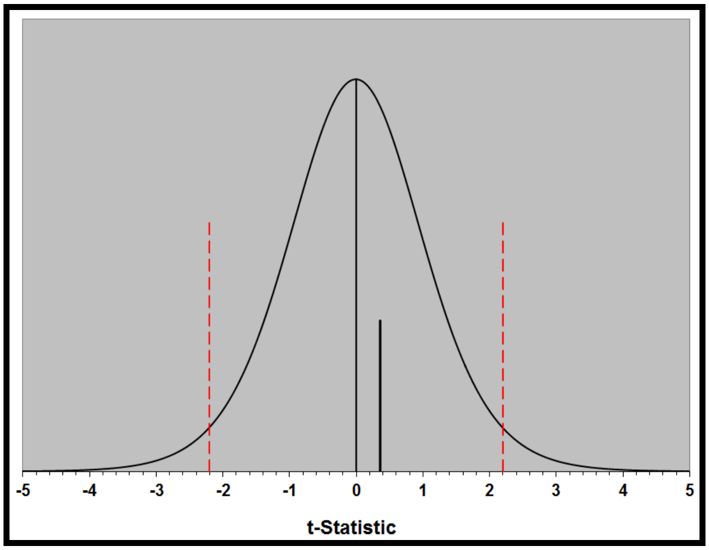


Figure 7. Eleven-df reference t-distribution to test null hypothesis claiming no difference in lift coefficient from Monday to Tuesday of Week 3 at Site 1 of operating envelope. Solid line: test statistic. Dashed lines: test criterion.

Figure 7 displays a t-distribution for a 12-point sample size that therefore has 11 degrees of freedom available for estimating the standard error (one degree of freedom is consumed in estimating the mean). Because this is a probability distribution, the total area under the curve is "1," but the area under the curve that is bounded by the two dashed lines is 0.95. This means that under the null hypothesis, for which the true value of the t-statistic is zero, any non-zero estimate of the t-statistic attributable to random error will have a 95% probability of falling between the two dashed lines. Note that the t-statistic just represents how many "sigma" the sample mean differs from zero. If the reference distribution in Fig. 7 corresponded to a large enough sample of data to be Gaussian, the 95% confidence limits represented by the two dashed lies would be at ±1.960 (colloquially expressed as "plus or minus two sigma"). The 95% confidence interval for an 11 degree of freedom t-distribution is roughly 10% wider instead, bounded by ±2.201 to reflect the fact that the variance in this distribution is less certain because it was estimated from a relatively small data sample.

We require the mean of a 12-point sample of response differences to differ from zero by more than 2.201 sigma before we are entitled to say with 95% confidence that the two samples actually do differ. The vertical line in Fig. 7 corresponds to the t = 0.36 value computed for Site 1 lift differences from Monday to Tuesday of "Week 3." There may in fact be a small systematic difference between the means of the two data samples compared in this analysis, but because the test statistic is well within the ±2.201 criterion for claiming a real difference, we are unable to reject the null hypothesis with 95% confidence, and we therefore infer that there is no significant difference in the two samples of data.

The same paired t-test described here in detail for lift measurements made on successive days at Site 1 of the operating envelope was also executed for the lift data acquired on those same two days at Sites 2, 3, and 4 (See Fig. 2). Results for all four sites are presented in Fig. 8.

Just as was the case for Site 1, the difference in sample means between ostensibly identical samples of lift data acquired at Site 4 on the Monday and Tuesday of "Week 3" (see Fig. 6) was too small to resolve, given the intrinsic variability of the data. Therefore, for both Site 1 and Site 4, the null hypothesis cannot be rejected with 95%

confidence, and we infer that at these two sites in the operating envelope there was no significant difference in lift data sample means for the data acquired on those two days.

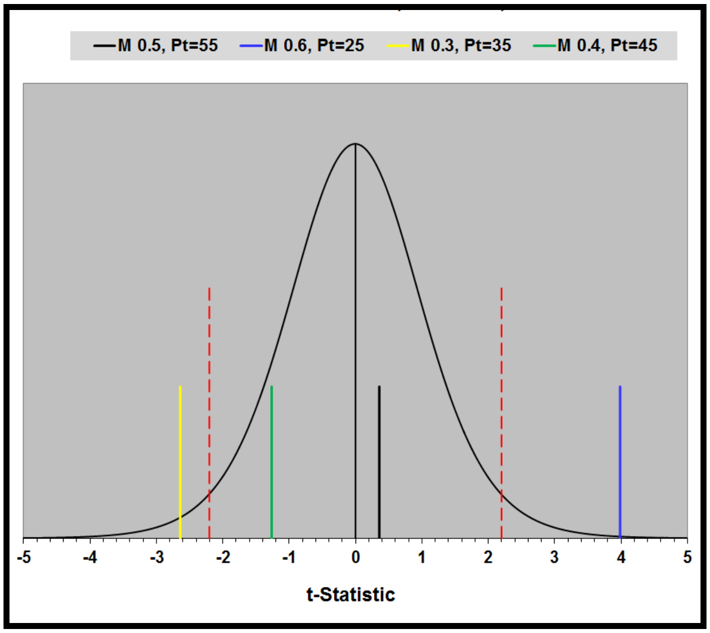


Figure 8. Eleven-df reference t-distribution to test null hypothesis claiming no difference in lift coefficient from Monday to Tuesday of Week 3 at four sites of operating envelope. Dashed lines: test criterion.

For Sites 2 and 3 the difference in sample means from Monday to Tuesday was in fact larger than could be attributed to random error. The test statistics for those two sites fall outside the 95% confidence interval limits, and the null hypothesis of no difference in sample means could be rejected for those two sites with less than a 5% chance of an inference error.

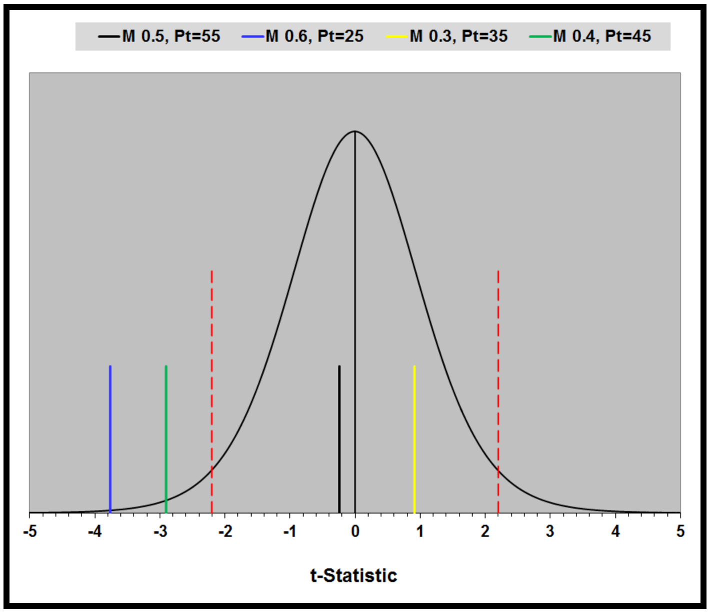


Figure 9. Eleven-df reference t-distribution to test null hypothesis claiming no difference in drag coefficient from Monday to Tuesday of Week 3 at four sites of operating envelope. Dashed lines: test criterion.

There are a number of implications in the results displayed in Fig. 8. The first is that since sample means drifted over a two-day period by more than could be attributed to random error, there is evidence of covariates effect in play. This implies that experimental errors for individual data points acquired in sequence are correlated to some degree, and are thus not independent, which means that the corresponding sample statistics are not likely to be unbiased estimators of their corresponding population parameters unless independence is overtly restored by quality assurance tactics designed for this purpose. Furthermore, unless there is reason to believe that the systematic variation in one test will reproduce exactly the next time such a test is executed, it is unlikely that the same experimental results will be obtained within limits expected on the basis of random error alone.

None of these issues is especially serious when quality assurance tactics are employed that are commonly used outside the experimental aeronautics community, such as randomizing the set-point order and blocking the data by time[2–5, 10, 12]. However, they can present problems in OFAT testing when high data-rate requirements foreclose options to take the time that is necessary to ensure statistical independence.

Figure 9 is similar to Fig. 8 except that it involves drag data instead of lift data. Again, for two of the four sites examined in the operating envelope, ostensibly identical data samples acquired a day apart had sample means that differed by more than could be attributed to random error. As with the lift data, this shifting of sample means in drag data implies a systematic component of the unexplained variance, which can be problematic unless overt steps are taken to ensure the independence of experimental errors.

B. Hypothesis Testing with Two-Way Analysis of Variance

Thus far, the analysis has been limited to the two-day interval of "Week 3," for which significant systematic error seems to be in play in half the cases examined. To examine other time intervals will require a slightly more powerful analytical method than the paired t-test used so far, although the basic idea will be the same. That is, we will still develop numerical test statistics which we will compare with criteria developed from reference distributions to either reject a null hypothesis claiming no systematic error, or not.

Consider the basic data structure of Fig. 5. We could in principal compare the means of any two columns of data using a paired t-test, but with 20 columns there are 20!/((20-2)!2!) = 190 possible pairs. Besides being tedious, such an analysis could be misleading. Imagine that inferences are made with 95% confidence in each of the 190 comparisons. While the probability of any one such inference being valid is 0.95, the probability that all of them are valid is a vanishingly small $0.95^{190} = 0.00006$. The probability that even a subset of 10% of the inferences would all be valid is only $0.95^{19} = 0.38$.

Fortunately, the analysis of variance (ANOVA) techniques described earlier were developed to cope with just this situation. A detailed treatment of the mathematics of ANOVA is beyond the scope of this paper but such details are readily accessible in standard textbooks such as Ref. 14. Applications to wind tunnel testing are described in Ref. 15. Without belaboring all of the details, ANOVA quantifies various components of the variance in a sample of data by computing two quantities for each of them, the "sum of squares" and the "degrees of freedom." Variance is simply the ratio of these two quantities. The sum of squares for the total variance of an n-point sample of data, for example, is computed by subtracting a convenient reference (the sample mean) from every data point in the sample, squaring each difference, and summing up all n of the squared differences.

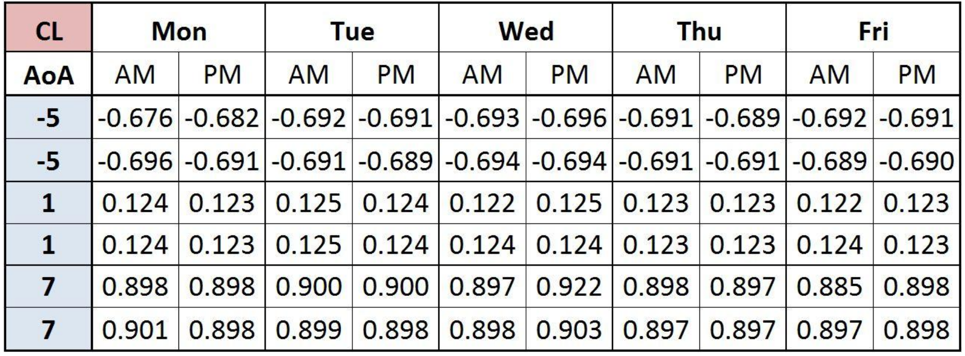
The degrees of freedom simply represent the smallest number of points needed to compute the sum of squares when the mean is known, which for an n-point sample is just n - 1. (The n^{th} point can always be inferred by subtracting the sum of $n – 1$ points from the sum of all n points, which is just n times the known mean.)

Variance components associated with row-wise and column-wise variation are computed similarly. The column-wise sum of squares is computed by squaring the difference between each column mean and the grand mean of all the data, and for "c" columns, adding all "c" such squared values. Likewise, the row-wise sum of squares is computed by squaring the difference between each row mean and the grand mean of all the data, and for "r" rows, adding all "r" such squared values. The sums of squares are normalized by multiplying the row-wise value by the number of columns and the column-wise value by the number of rows so that the results are not biased by the geometry of the array of data being analyzed, and then an error sum of squares is computed by subtracting the row-wise and column-wise sums of squares from the total sum of squares.

Degrees of freedom (df) are similarly computed for the total sample of data, the rows, the columns, and the residual error. The total df is just $n – 1$ for an n-point sample, as noted above. If that sample is structured as r rows and c columns, there are $r – 1$ and $c – 1$ degrees of freedom for rows and columns, respectively. The residual error df, like the error sum of squares, can be determined by subtraction.

A certain amount of bookkeeping is required in an analysis of variance to keep track of all of the constituent sums of squares and degrees of freedom. This is facilitated by an analysis of variance table that will be constructed for illustration purposes using the data in Table 3. The n = 60 numbers in this six-by-ten table are lift coefficients acquired in the current test over five consecutive days at Site 1 in the tunnel operating envelope (Fig. 2). That is, they were all acquired at the same Mach number and total pressure set points of Mach 0.5 and 55 psia, respectively. Each number in a given row was recorded at the same angle of attack set point as indicated by the left-most column, and each column was acquired at a different date/time. Under a null hypothesis asserting no significant change with time, the 10 numbers in any one row should be the same within random error. Likewise any differences in the means of the 10 columns should be attributable to ordinary random error under this hypothesis.

Table 3. Ten ostensibly identical six-point samples of lift data acquired over five consecutive test days at Site 1 in the operating envelope of the tunnel.



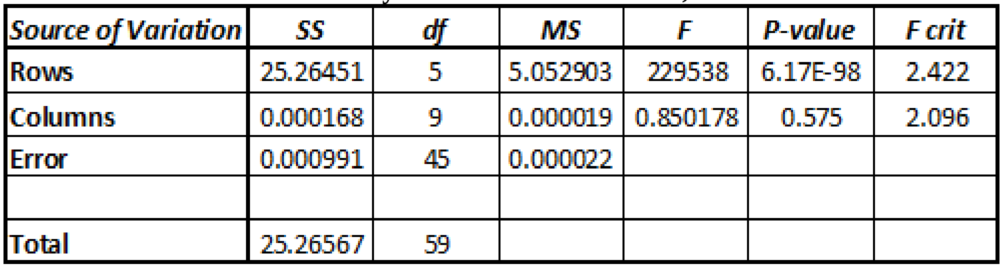
CL	Mon		Tue		Wed		Thu		Fri	
AoA	AM	PM	AM	PM	AM	PM	AM	PM	AM	PM
-5	-0.676	-0.682	-0.692	-0.691	-0.693	-0.696	-0.691	-0.689	-0.692	-0.691
-5	-0.696	-0.691	-0.691	-0.689	-0.694	-0.694	-0.691	-0.691	-0.689	-0.690
1	0.124	0.123	0.125	0.124	0.122	0.125	0.123	0.123	0.122	0.123
1	0.124	0.123	0.125	0.124	0.124	0.124	0.123	0.123	0.124	0.123
7	0.898	0.898	0.900	0.900	0.897	0.922	0.898	0.897	0.885	0.898
7	0.901	0.898	0.899	0.898	0.898	0.903	0.897	0.897	0.897	0.898

By inspection it is clear that the lift numbers throughout any one column vary substantially. This simply reflects the strong dependence of lift coefficient on angle of attack and is not particularly interesting for the purposes of the current study, in which we are much more keen to discover whether the numbers change substantively across columns; that is, with time. As noted above, we attack this problem by quantifying the total variance in the 60 numbers and partition that into a row-wise component we associate with AoA changes and a column-wise component we associate with changes from one block of time to another. A common occurrence in any ANOVA is that the column-wise and row-wise components of variance never account for the total variance. That is, there is always some residual variance that is neither associated with rows nor columns, but is something that is simply left over, that we attribute to ubiquitous chance variations in the data. The ANOVA calculations therefore enable us to account for *explained* variance (row-wise changes attributable to changes in angle of attack) and both components of *unexplained* variance, one component associated with column-wise changes attributable to unknown covariate effects (which may or may not be in play, as is to be objectively determined) and one component associated with residual random error that can be reliably assumed to be always present.

Table 4 is a "two-way ANOVA table" that summarizes elements of the total variance and each of its components. The first column identifies sources of variation and the next two columns list sums of squares and degrees of freedom calculated as outlined earlier. Note that with n = 60 data points in the sample, the total number of degrees of freedom is just n – 1 = 59. Likewise, with r = 6 rows and c = 10 columns there are r – 1 = 5 row degrees of freedom and c – 1 = 9 column degrees of freedom. Subtracting 5 + 9 = 14 from 59 yields 45 error degrees of freedom. The error sum of squares is similarly determined by subtracting the row and column values from the total.

The fourth column, labeled "MS" for "Mean Square", is the actual variance and is computed by dividing the sum of squares by the corresponding degrees of freedom: MS = SS/df. Note that the constituent sums of squares and degrees of freedom add to their corresponding totals but the variances, constructed as ratios of these additive components, do not add.

Table 4. Lift Coefficient Two-Way ANOVA Table for Site 1, Week 2.



Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Rows	25.26451	5	5.052903	229538	6.17E-98	2.422
Columns	0.000168	9	0.000019	0.850178	0.575	2.096
Error	0.000991	45	0.000022			
Total	25.26567	59				

The error mean square is particularly interesting, because this is just the square of the random error standard deviation. That is

$$\sigma = \sqrt{MS_{error}} = \sqrt{0.000022} = 0.0047 \tag{5}$$

An ANOVA can provide a relatively reliable estimate of random error because of the large number of degrees of freedom typically associated with the estimate compared, say, to the number of genuine replicates that can be acquired with resources that are allocated for such purposes in a typical OFAT test matrix. In this case the value of 0.0047 that is estimated for the standard deviation in lift coefficient is based on 45 degrees of freedom, as the ANOVA table indicates.

Returning to the ANOVA table, the column labeled “F” contains the test statistic that will be used to test the null hypothesis that there is no significant variation across columns (with time). It is analogous to the t-statistic developed for a similar purpose when the paired t-test was used earlier. Computing F is trivial. The row F is computed by dividing the row MS by the error MS and the column F is likewise computed by dividing the column MS by the error MS. The F-statistics therefore simply express the row and column variance components as multiples of the irreducible random error variance.

Note that the row F tells us that row-wise variance in this sample of data is 229,538 times larger than the variance associated with ordinary chance variations in the data. It is the ratio of two variance estimates, each based on finite subsets of data that are subject to random fluctuations, so it is possible that a particularly large F-statistic might simply be the chance result of a waxing numerator and a waning denominator, when in fact the true ratio, absent all the “jiggling” from random error, might be closer to 1. If that were the case, we would be less inclined to attribute the variation to anything beyond random error. The probability that a given F value is due only to chance is given in the P-value column of the ANOVA table.

The P-value column indicates how likely it is that an F value as large as the one in the F column could occur as the result of random variations in the numerator and denominator variances, if there actually was no significant difference in the numerator and denominator (true F close to 1). In the present case of row-wise variation, that probability is 6.17×10^{-98}, a number that is almost comical in its minuteness. The probability is therefore essentially zero that row-wise variation as large as that seen in Table 3 could simply reflect random variations in data that are actually independent of angle of attack. We are left with the HIGHLY likely inference (confidence level of 1 - 6.17×10^{-98}!) that these changes are not random at all, but are attributable to the systematic changes made intentionally in the angle of attack.

Such statistical confirmation of well-known aeronautical engineering principles is superfluous in the case of row-wise variation (The aeronautical engineer may be relieved to learn that statistics support his long-held suspicion that angle of attack influences lift, but he hardly needs anything so elaborate as an analysis of variance to reach this unremarkable conclusion!) In the case of column-wise variance, however, there is no basis for making a-priori inferences, and in such a case the ANOVA results will be more useful.

There is a common convention that inferences should not be made with less than 95% confidence, so by this convention we would not infer significant row-wise or column-wise variance unless the corresponding P-value is less than 0.05. Surely with a P-value of 6.17×10^{-98}, the row-wise variation qualifies! However, the column P-value in Table 4 is 0.575 with a corresponding F of only 0.850178. This suggests that the column-wise variance is very nearly the same as the variance attributable to random error, and there is therefore little basis in this data sample for inferring any systematic variation of lift with time. The P-value indicates that there is a relatively high probability that the observed F-stat is attributable to nothing more than random fluctuations in the data. We are unable, then, to reject a null hypothesis asserting no systematic change with time, at least for this site in the operating envelope and for this particular interval of time. There are of course other time intervals and other Mach/pressure combinations to examine as well, but before proceeding, we comment on the last column in the ANOVA table, labeled “Fcrit.”

Fcrit is a critical F value that is analogous to the critical t-statistic represented graphically by dashed lines in Figs. 7. Like the critical t, it is based on a specified probability criterion for a given reference distribution. The t-statistic for the prior 11-df variance estimate follows the probability distribution of Fig. 7, with 95% of the distribution lying between the critical t-values located at ±2.201. The F-statistic has a probability distribution that is uniquely determined by the number of numerator and denominator degrees of freedom in the estimate of F.

The F reference distribution is bounded on the low end by zero. The area under the F distribution between zero and the critical F value in the ANOVA table corresponds to some specified probability that, for a given number of numerator and denominator degrees of freedom, determines the value of Fcrit. For the Fcrit values in the current ANOVA table, that probability is 95%. This means that if we conclude that the corresponding effect is real, there is less than a 5% chance of an inference error attributable to unlikely chance variations in the data if the observed F-statistic exceeds the critical F. Because the row-wise F of 229,538 exceeded its corresponding Fcrit value of 2.422, we reject the null hypothesis of no significant difference between row-wise variance and variance due to random error, concluding that there is some systematic (non-random) variation in lift across rows. Because the column-wise F of 0.850178 did not exceed its corresponding Fcrit value of 2.096, we cannot reject the null hypothesis of no significant difference between column-wise variance and variance due to random error, and we conclude that there is no evidence of systematic (non-random) variation in lift across columns in this data sample.

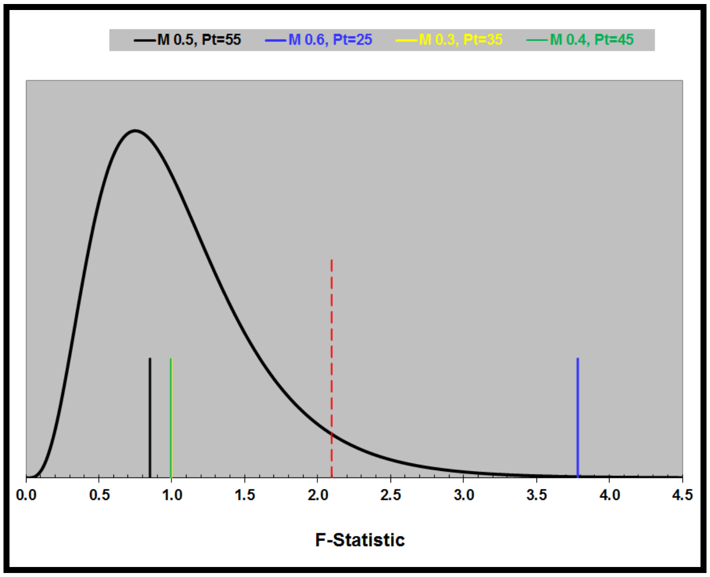


Figure 10. F reference distribution for 9 numerator df and 45 denominator df to test null hypothesis claiming no difference in lift coefficient for Week 2 at four sites of operating envelope. Dashed lines: test criterion.

Figure 10 displays the reference distribution for F statistics with numerator Mean Squares based on 9 df (corresponding to the 10 columns of Table 3) and denominator Mean Squares based on 45 df, corresponding to the number of error degrees of freedom from ANOVA Table 4. The dashed line represents the corresponding critical F-value of 2.096. We reject the null hypothesis for any F values larger than this.

Estimated F-statistics for ostensibly identical samples of lift data acquired over the five-day time interval of Week 2 are displayed in Fig. 10 for the four sites of the operating envelope that were evaluated in this test (See sites displayed in Fig. 2). The F-statistic for Site 2 exceeded the critical value, suggesting that there were differences in lift data sample means acquired at this Mach/Pt combination over the five day period that were too large to attribute to ordinary random error. No significant differences in overall lift sample means were detected for the other three sites that were examined in the tunnel's operating envelope.

Figure 11 is similar to Fig. 10 except that it displays results for drag coefficient data. The F-statistic for Site 1 is less than the critical F, indicating that no significant variation in drag sample means were detected over the five-day test period of Week 2 at this site in the tunnel's operating envelope. However, the F statistics for all three of the other sites were greater than the critical value, suggesting that the sample means varied with time by more than can be explained by random error. This is evidence of one or more non-random sources of variation.

Certain patterns begin to emerge when Figs. 10 and 11, representing the results of an ANOVA for multiple blocks of time in Week 2, are compared with Figs. 8 and 9, representing the paired t-test used to compare data acquired during two days in Week 3. Note that the largest test statistic (t-statistic for Figs. 8 and 9 and F-statistic for Figs. 10 and 11), indicating the greatest instability of sample means, occurs for Site 2, whether it is lift data or drag data, and whether it is a t-test of two different days or an F test of 10 different time blocks in a different week. Likewise, the smallest test statistics for lift or drag, and for either time period, indicating the most stable sample means, occurred at Site 1 in the design space.

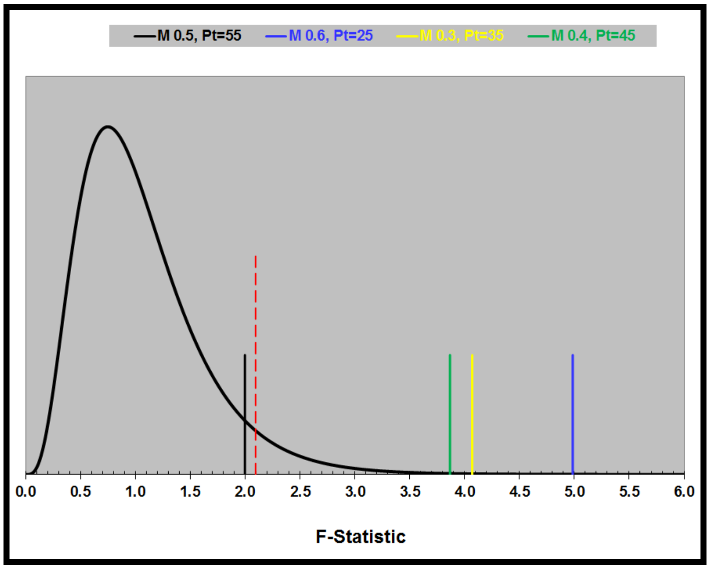


Figure 11. F reference distribution for 9 numerator df and 45 denominator df to test null hypothesis claiming no difference in drag coefficient for Week 2 at four sites of operating envelope. Dashed lines: test criterion.

A glance at Fig. 2 reveals that Site 1 had the highest total pressure of the four sites that were examined, and Site 2 had the lowest total pressure. Furthermore, the lift results of both the ANOVA and paired t test indicate that the test statistics for Sites 3 and 4 follow the same pattern. For all four sites, the lift test statistics increased monotonically with decreasing total pressure, indicating that sample means for lift data tend to be less stable with time for lower total pressures than higher total pressures. The same pattern is observed with the drag data acquired over the 10 time blocks in Week 2. For the two-day interval of Week 3, the highest and lowest drag test statistics do correspond to the lowest and highest pressures, but the test statistics for the two intermediate pressures were reversed. This might have been due to other unknown systematic effects in play, or it might simply reflect an anomaly due to the short duration of the Week 3 time interval. In any case, the general trend seems to suggest less stability for the means of ostensibly identical samples acquired at lower pressures than at higher pressures.

For the demonstration of a paired t-test conducted to test for changes from Monday to Tuesday of Week 3, there were two responses and four sites within the Mach/pressure operating envelope of the tunnel, so eight cases that were examined for evidence of variation with time. In four of the eight, differences were detected between the means of two ostensibly identical samples acquired a day apart, that were too large to attribute to ordinary random fluctuations in the data.

Likewise, there was a demonstration of a multiple-comparison analysis involving nominally identical samples of lift and drag data acquired in 10 different blocks of time over five consecutive test days in what is identified in Fig. 6 as Week 2. An analysis of variance was conducted to test a null hypothesis claiming no significant difference among the means of the 10 lift samples or among the means of the 10 drag samples. Again, there were eight independent tests of this null hypothesis, corresponding to the two responses and four operating envelope sites. As with the tests conducted over the two-day interval of Week 3, the tests applied to the five-day interval of Week 2 resulted in four cases out of eight in which the null hypothesis could be rejected with at least 95% confidence. Similar results were obtained by an ANOVA applied to the three-day interval of Week 1: differences among the means of nominally identical samples that are too great to attribute to random error, which occurred at all four sites during Week 1 for drag, and at half the sites for lift.

C. Two-Way Analysis of Variance with Replication

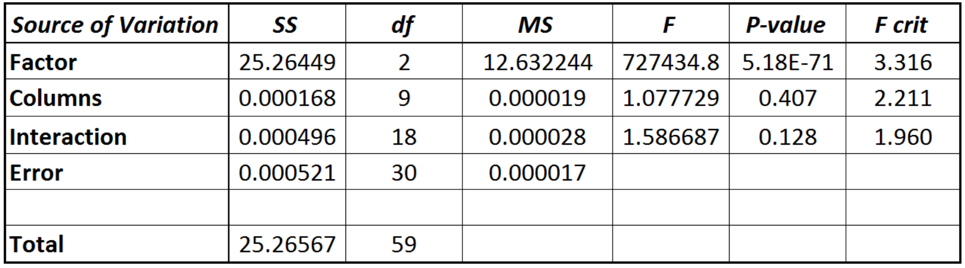
Results presented in the precious section were obtained with a simple two-way ANOVA. The fact that each of the three unique angles of attack in every column of Fig. 5 was replicated facilitates a somewhat more complex ANOVA, known as a two-way ANOVA *with replication*. Replication of each unique angle of attack permits a direct estimate of random error that is independent of the estimates of total variance, column-wise variance, and row-wise variance that are used to determine random error in an ordinary two-way ANOVA (by subtraction, as described above). To make this direct estimate we proceed in the usual way when the intent is to quantify variance using replicates. A sum of squares is first computed. For sample subsets of two replicates each as in this case (Fig. 5), Eq. (6) can be used to compute the contribution to the error sum of squares made by each pair of replicated AoA points, where y_1 and y_2 are responses at the two nominally identical AoA settings:

$$SS = \frac{1}{2}\left(y_2 - y_1\right)^2 \tag{6}$$

These sums of squares are then added for all pairs of AoA settings (three per column) to obtain the sum of squares for random error. There is one degree of freedom ("*m – 1*", with *m = 2*) for each pair, so 3 df for each column. These are recorded in the ANOVA table in the usual way. The two-way ANOVA with replication allows the component of variance not attributed to rows or columns, previously assumed to be entirely due to random error, to be partitioned further, into two components. One part is the random error as before, but the second part is a component of variance attributable to interactions between factors (in this case, AoA settings) and columns, or "blocks" (of time).

The implications of block/factor interactions will be described in the discussion section to follow. For now, Table 5 displays the two-way ANOVA results with replication for the data of Table 3, which can be compared with the two-way ANOVA without replication performed for the same data and displayed in Table 4.

Table 5. Lift Coefficient Two-Way ANOVA with Replication. Site 1, Week 2.



Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Factor	25.26449	2	12.632244	727434.8	5.18E-71	3.316
Columns	0.000168	9	0.000019	1.077729	0.407	2.211
Interaction	0.000496	18	0.000028	1.586687	0.128	1.960
Error	0.000521	30	0.000017			
Total	25.26567	59				

The ANOVA table of Table 5 includes the additional source of variation that an ANOVA with replication facilitates; namely, the block/factor interaction. Also, instead of "Rows" as a source, there is a source called "Factor", which refers in this case to angle of attack. This is because we are partitioning the row-wise variation into one component that is explicitly attributable to factor (AoA) changes, and another component attributable to the random error revealed by the replication of AoA. Because there are *m = 3* AoA levels, there are *m – 1 = 2* degrees of freedom for factors.

The block/factor interaction sum of squares is determined by subtracting the factor, column, and error sums of squares from the total sum of squares. The interaction degree of freedom can also be determined by subtraction, or by simply multiplying the degrees of freedom for factors (2) by the degrees of freedom for columns (9), to obtain the 18 interaction degrees of freedom. Note that because there are three pairs of replicates per column, each supplying one pure error degree of freedom, and because there are 10 columns in Table 3, there are a total of 30 error degrees of freedom.

Mean Square (MS) values are computed as before, by dividing the SS by the df in each row of the ANOVA table. Likewise, F-statistics are computed as before, by dividing the MS for each row in the ANOVA table by the error MS. So again, the F values represent a kind of signal to noise ratio, in which the variance associated with each source is expressed as a multiple of the irreducible random error variance.

There were more error degrees of freedom and a larger error sum of squares when the ANOVA was executed without replication than with replication (compare Tables 4 and 5), because the former case lumped the interaction component together with the random error component. Replication therefore permits the random error to be quantified with greater precision, and reveals that it is smaller than might have been supposed without isolating the interaction component of the total variance. The practical effect of reducing the noise in this way is to increase the signal to noise ratio for evaluating the significance of other sources of variance. That is, the F statistics will tend to be greater because the error MS, now devoid of interaction effects, is smaller.

For the 60-point data sample of Table 3 that we have been using to illustrate the application of ANOVA methods, the interaction effects are rather small and the impact of isolating them is minimal. For example, Tables 4 and 5 reveal that divesting the error mean square of its interaction component only resulted in a reduction in the error sum of squares from 0.000022 to 0.000017, with a corresponding reduction in the estimate of standard random error from 0.0047 per Eq. (5) to 0.0041 by a similar calculation. This simply means that because the block/factor interaction was so small at Site 1 during Week 2, random error estimates made with and without accounting for it are similar. However, for other sites in the tunnel operating envelope and for other time periods, the interaction was larger, and accounting for it had a significant effect.

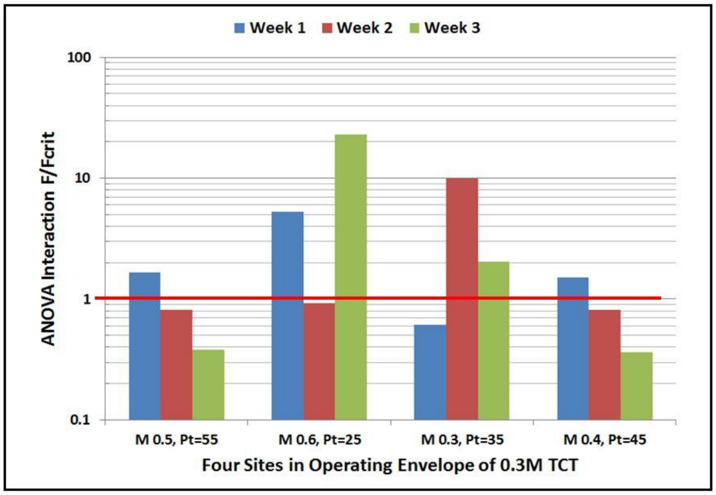


Figure 12. Lift Block/Factor Interactions. Ratio of ANOVA F-statistic to Critical F exceeds 1 when interaction is detected with at least 95% confidence.

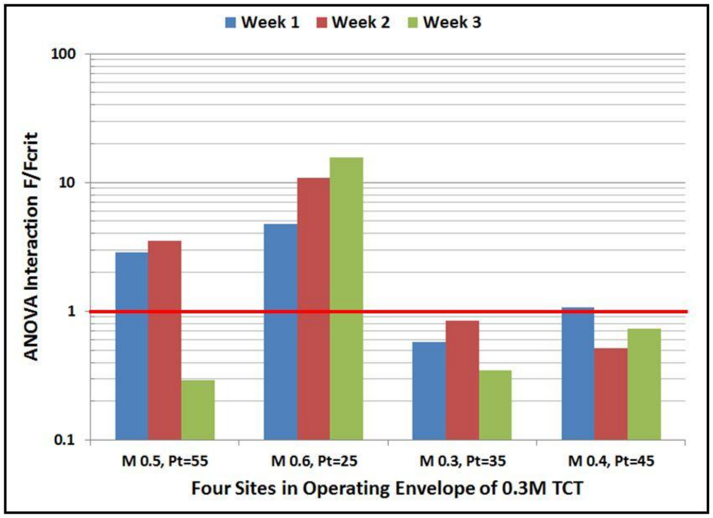


Figure 13. Drag Block/Factor Interactions. Ratio of ANOVA F-statistic to Critical F exceeds 1 when interaction is detected with at least 95% confidence.

Figures 12 and 13 display the ratio of interaction F-statistics to their corresponding critical F values for all four sites in the operating envelope and for the three time intervals displayed in Fig. 6. Note that these ratios are represented logarithmically. The bars in these figures extend above the horizontal red line when ratios are greater than one. This indicates that a block/factor interaction is detected with at least 95% confidence. Interactions are detected with at least 95% confidence in 12 of the 24 independent cases examined (two responses by four sites by three time intervals).

Large block/factor interactions reduce the random error estimate significantly when they are independently taken into account using ANOVA with replication. For reasons presented in the Discussion section below, such interactions are often assumed to be negligible, as they in fact were for the lift data acquired at Site 1 during Week 2

(Tables 4 and 5). If interactions are not anticipated, an ANOVA without replication permits more degrees of freedom to be assigned to the estimate of random error and is preferable. For this reason, and because it is easier in a paper that has a substantial tutorial dimension such as this one to explain two-way ANOVA when the added complication of block/factor interactions is omitted, the data were first analyzed as a two-way ANOVA without replication. These were the results presented in Figs. 10 and 11. However, a subsequent ANOVA with replication demonstrated that the initial impulse to ignore interactions, while justifiable in some cases such as the Site 1 lift data acquired in Week 2, was not justified in about half of the cases examined, as Figs. 12 and 13 suggest.

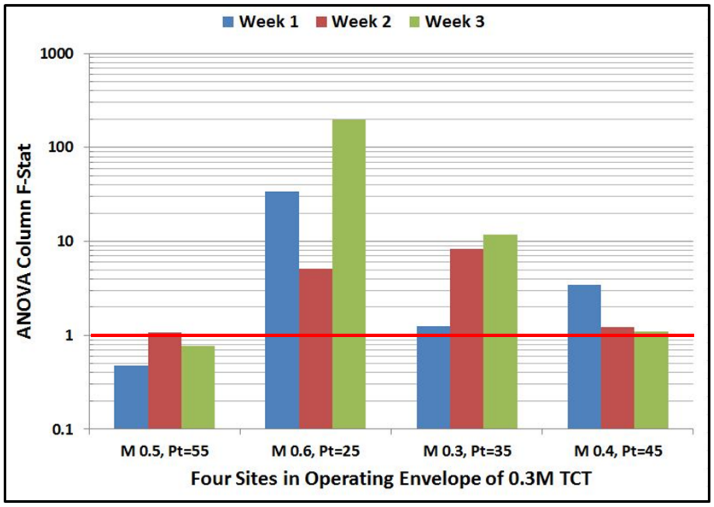


Figure 14. Ratio of systematic to random unexplained variance for lift: F statistics from a Two-Way ANOVA with replication.

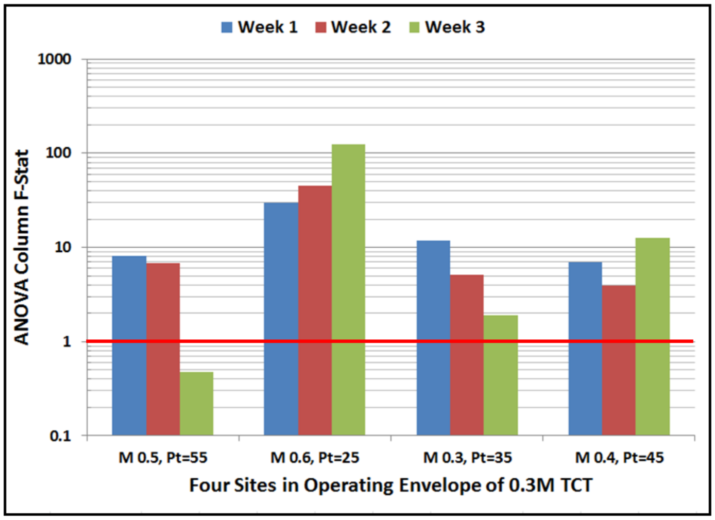


Figure 15. Ratio of systematic to random unexplained variance for drag: F statistics from a Two-Way ANOVA with replication.

Smaller error mean squares that result from separating out a significant block/factor interaction component lead to larger F statistics (great signal to noise due to reduced noise). That is, the higher precision associated with smaller noise levels enables more subtle effects to be detected. Figures 14 and 15 summarize the results of two-way ANOVA with replication for lift and drag, respectively.

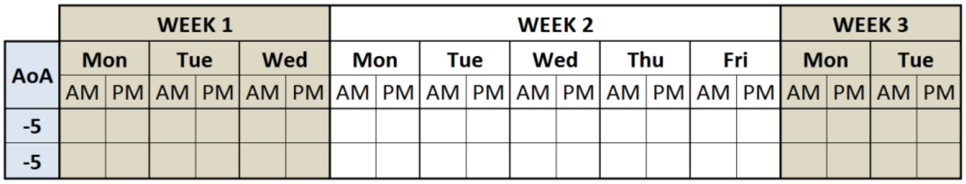
These figures display the F statistics that express the systematic component of unexplained variance as a multiple of the random component. All bars in Figs. 14 and 15 that exceed the horizontal red reference line at “1” represent cases in which the systematic component is larger than the random component. Note the logarithmic axis. Again there are 24 cases represented in these two figures, consisting of two responses by four sites in the tunnel’s operating envelope (Fig. 2), by the three time periods of Fig. 6. In 16 of these cases (67%), the systematic component of unexplained variance exceeded the random component by at least a factor of 2. In four other cases, the systematic error exceeded the random error by a factor of between 1 and 2. In eight of the cases (33%), the systematic component of unexplained variance exceeded the random component by more than a factor of 10.

Note that the variances are "sigma squared" values, so the error ratios in engineering units would be proportional to the square roots of these F statistics. Even so, the systematic error is clearly non-negligible and can be several multiples of the random error, up to an order of magnitude for Site 2 in Week 3.

D. One-Way Analysis of Variance for Individual Angles of Attack

Results reported in prior sections have reflected the stability of sample means for which the samples consisted of data points acquired at different angles of attack. These results are therefore in some sense integrated across the three angle of attack set points examined in this test. It is possible that systematic covariate effects could be in play that cause systematic changes in responses acquired at positive AoA settings to have one sign while responses acquired at negative AoA settings have the opposite sign. If that is the case, then sample means might be relatively stable with time only because of systematic changes occurring in opposite directions for data acquired at different angles of attack within the sample. To eliminate this potential source of confusion, independent analyses were conducted on data samples with the same AoA set-point.

The data structure required to support such an analysis is particularly simple because the columns contain only replicates, without responses acquired at multiple factor levels. The total variance of the data sample can thus be partitioned into at most two components, the ordinary random error that is always present, and a column-wise component of variance that may or may not be significant. Gone from this analysis are any considerations of dominating factor effects and complications related to block/factor interactions. Because there is only one potential source of variance besides the ubiquitous random error, an ANOVA on such a data sample is called a one-way ANOVA. Figure 16 is a representative data structure supporting such an analysis, in this case for data acquired at a constant AoA set-point of -5°.



	WEEK 1						WEEK 2										WEEK 3			
AoA	Mon		Tue		Wed		Mon		Tue		Wed		Thu		Fri		Mon		Tue	
	AM	PM	AM	PM	AM	PM	AM	PM	AM	PM	AM	PM	AM	PM	AM	PM	AM	PM	AM	PM
-5																				
-5																				

Figure 16. Representative data structure for One-Way ANOVA.

The F-statistic for a Two-Way ANOVA describing the variation of sample means with time for data acquired at Site 1 (Mach 0.5, Pt – 55 psia) during Week 2 was less than its critical F value, as Fig. 11 shows. However, those samples consisted of six data points each, two replicates of three unique angles of attack. If we revisit this data sample and perform a one-way ANOVA on only the data acquired at an angle of attack of 7°, we get the results displayed in Table 6.

Table 6. Drag Coefficient One-Way ANOVA. Site 1, Week 2. AoA = 7°.

Source of Variation	***SS***	***df***	***MS***	***F***	***P-value***	***F crit***
Columns	3.42E-06	9	3.80E-07	7.691	0.002	3.020
Error	4.94E-07	10	4.94E-08			
Total	3.92E-06	19				

There are 20 data points in this sample for each of the two response variables, lift and drag. This is because there were two points acquired at an angle of attack of 7° in the morning and in the afternoon of the five consecutive test days of Week 2. One degree of freedom is lost to the estimate of the mean, as usual, leaving a total of 19 df available to assess the variance, as ANOVA Table 6 indicates. These are m – 1 = 9 column df for the 10 columns, and the remaining 10 df are available to estimate the random error, since there is one random error df from each pair of 7° replicates acquired in the morning and in the afternoon of the five test days.

The F statistic representing the ratio of column-wise (systematic) variance to random variance is 7.791, indicating considerably more systematic error than random error. This F value exceeds the critical F of 3.020 by enough to drive the corresponding P-value as low as 0.002, which suggests that the null hypothesis of no significant difference between systematic and random unexplained variance can be rejected with 99.8% confidence. We therefore infer that there is a column-wise component of unexplained variance attributable to changes with time that exceeds to variance due to ordinary random error. Note also that the column-wise sum of squares, 3.42×10^{-6}, is 87.2% of the total sum of squares, with only 12.8% due to random error. Clearly in this instance the variance with time exceeds the random experimental error. This is the norm in wind tunnel testing when covariate effects are in play, causing sample means to wander with time. The key question at this stage is how often such significant systematic error is actually observed.

To answer this question, independent one-way analyses of variance were conducted for lift and drag data acquired for all 36 combinations of three AoA set-points by four distinct Mach/Pt test sites within the tunnel's operating envelope (Fig. 2), by the three time intervals of Fig. 6. The cases were counted in which column F statistics exceeded the critical F for 95% confidence, permitting the null hypothesis of no difference between variations in time and random error to be rejected with less than a 5% probability of an inference error. The results are displayed in Fig. 17.

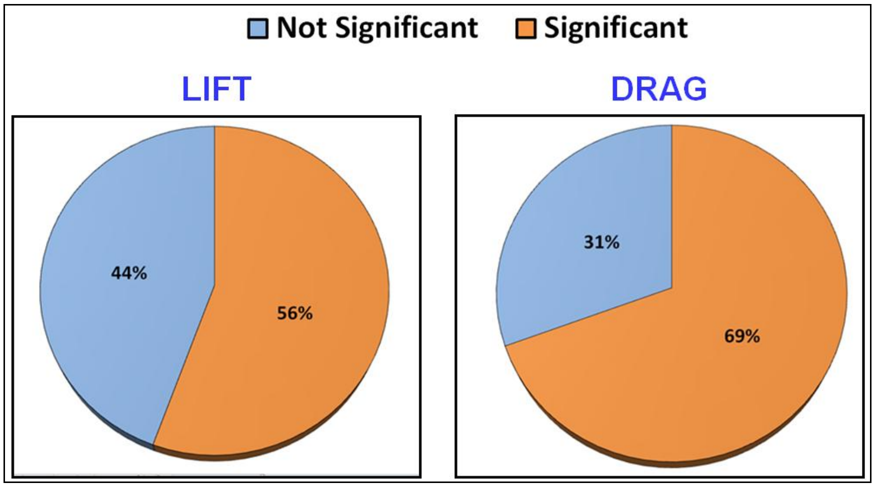


Figure 17. Frequency of significant unexplained systematic variance.

These results indicate that systematic unexplained variance is in play more often than not—56% of the time for the lift cases examined and 69% of the time for drag. On the other hand, evidence for statistical control can be found in almost half the lift cases and about a third of the drag cases, so allocating relatively small data samples to this issue can generate misleading results.

E. Systematic Trend Analysis

The analysis of variance results reported in previous sections provide objective and unambiguous evidence (95% confidence) that sample means are not stable with time as is widely assumed in OFAT testing, but rather, that those means vary with time by more than can be attributed to ordinary random error. However, the ANOVA results are silent as to the nature of the variation. For example, a significant column F-statistic simply means that sample means differed in at least two blocks of time. This could be due to effects limited to a single morning or afternoon, for example. On the other hand, significant time-wise variation might indicate a systematic trend in the response errors, or possibly some other pattern.

To address this question, lift and drag data for the 36 unique combinations of three angles of attack, four sites in the operating envelope, and three time periods (Weeks 1, 2, and 3) were further examined. There were 16 cases out of 36 for which no significant unexplained systematic variance was detected for lift and 11 such cases for drag (44% and 31%, respectively, per Fig. 17). Data for the remaining 20 lift cases and 25 drag cases were plotted as a function of time and the following second-order polynomial function of time was fitted to each of these data sets:

$$y = b_0 + b_1 t + b_2 t^2 \tag{7}$$

Each of the three polynomial coefficients in Eq.(7) was evaluated by linear regression, and the standard error in estimating the first- and second-order coefficients, b_1 and b_2, was also estimated. These two coefficients were examined to determine if they were statistically significant, meaning large enough compared to the standard error in estimating them to infer with at least 95% confidence that they were non-zero. Each of the 20 lift cases and 25 drag cases displaying systematic variation were then assigned to one of three categories, depending on which of the polynomial coefficients in Eq. (7) was significant.

The first category corresponded to the case in which neither b_1 nor b_2 were significant so that the fitted model was simply $y = b_0$. In this case there was no evidence of a trend in the systematic variation, and we assume that a significant ANOVA column F statistic could be attributed to a fairly short-duration phenomenon that caused sample means to change abruptly, rather than trending steadily over extended periods. The AoA = -5° drag data acquired at Site 2 during Week 2 displays this behavior.

Two drag measurements at AoA = -5° were acquired in the morning and again in the afternoon of each of the five test days of that week, at each of the four sites of the tunnel operating envelope examined in this test. The data acquired at Site 2 (Mach 0.6, Pt = 25 psia) are displayed in Fig. 18. The major vertical axis divisions span the 24-hour interval from midnight to midnight and the minor divisions denote eight-hour periods starting at midnight on each day. All data were acquired each day during the second of the three eight-hour intervals, between 0800 and 1600. Within each eight-hour interval for which data were acquired, the two left-most points were always acquired sometime during the morning (between 0800 and 1200) and the right-most pair of points was always acquired in the afternoon, sometime between 1200 and 1600.

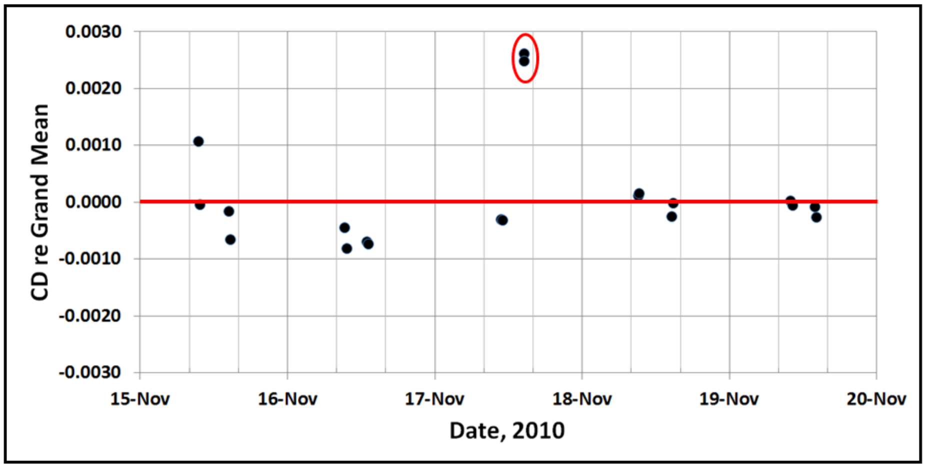


Figure 18. Drag coefficients for Site 2, Week 2, AoA = -5° relative to the grand mean of all Site 2 data acquired at AoA = -5°. Significant column-wise ANOVA F statistic but no evident trend.

Each data point represents a measurement of drag coefficient relative to the grand mean of all 40 Site 2 drag coefficient measurements made at AoA = -5° throughout the test (two in the morning and two in the afternoon of 10 test days). Clearly the two afternoon measurements acquired on Wednesday, the third day of that week (enclosed in red ellipse in Fig. 18) seem to have been drawn from a distribution with a mean that is biased by about 25 drag counts relative to the rest of the data.

Based on how little difference there is between these two points, and judging from how well every other morning and afternoon pair of replicates agrees, it does not seem plausible that this bias in the means of the Wednesday afternoon data is due to ordinary chance variations in the data. Rather, it appears as if something happened to bias those points high, which was not part of a perceptible long-term trend. There were eight similar cases out of 36 for lift (22%) and 12 out of 36 for drag (33%).

The second category of data for which the ANOVA revealed significant systematic unexplained variance is one in which b_1, the first-order coefficient of Eq. (7), is significant but not the second-order coefficient, b_2. Such data displays a first-order trend with time that is significant because the slope of the best straight line fitting the data (b_1) is large compared to the uncertainty in estimating it. Figure 19 is an example of sample means trending with time. In this case, drag measurements made at AoA = 7° and at Site 2 in the tunnel operating envelope (Fig. 2) in the two days of "Week 3" (Fig. 6) are plotted. Analogous to Fig. 18, each data point represents an empirical estimate of drag coefficient relative to the grand mean of all 40 Site 2 drag coefficient measurements made at AoA = 7° throughout the test (two in the morning and two in the afternoon of 10 test days).

Figure 19 illustrates a classic block effect, in which the means of two presumably identical samples acquired in different blocks of time are displaced by more than can be attributed to ordinary random error. All eight points in this figure were acquired under ostensibly identical conditions, and absent any experimental error, all eight should fall on the red line corresponding to no difference with the grand mean of 40 supposedly identical such measurements acquired throughout the entire test. We anticipate some departures from this red reference line due to ordinary random error, to which we attribute the scatter in the four points acquired on each day. However, the means of these two samples are clearly displaced by more than the points are scattered on each individual day. Note also that the mean of the data acquired on the second day of this "Week 3" interval is closer to the grand mean than the data acquired on the first day. Numerous possible covariate effects have been listed earlier in this paper. Any of those, or any other covariate effects that can be drawn from an enormous list of possible candidates, could be responsible for the behavior of the data displayed in Fig. 19. However, this migration in the direction of smaller departures from the grand mean is consistent with what are known as "learning effects," by which performance improves with time. Again, there are for practical purposes an infinite number of other candidate effects besides learning effects that could explain these observations, but learning effects is one example.

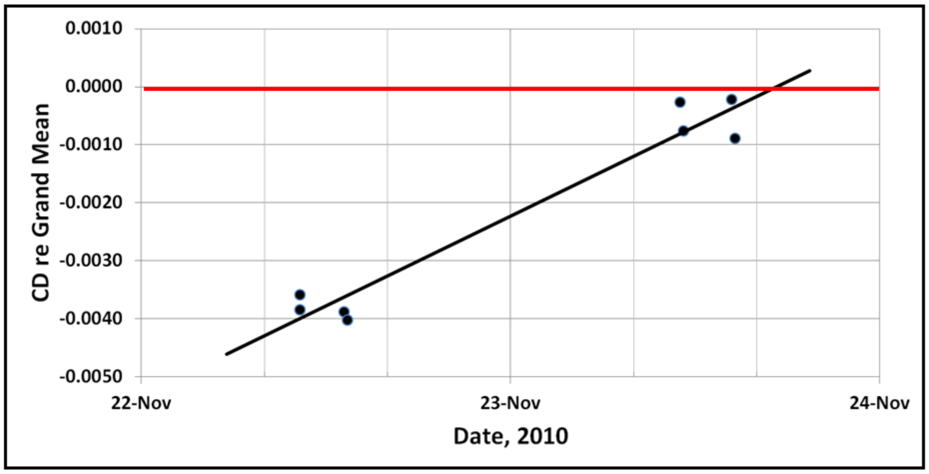


Figure 19. Drag coefficients for Site 2, Week 3, AoA = 7° relative to the grand mean of all Site 2 data acquired at AoA = 7°. An overnight systematic block effect.

Block effects are common in wind tunnel testing and in fact the phenomenon displayed in Fig. 18 can also be categorized as a block effect. Other linear trends extended over more than two blocks of time. Of the 36 unique combinations of model attitude, tunnel state, and time interval examined in this test, 10 of the lift samples (28%) and eight of the drag samples (22%) displayed this linear trending behavior. The 10 lift data sets that displayed a linear trend comprise half of the 20 lift cases that displayed any evidence of significant systematic unexplained variance, and the eight drag sets that displayed a linear trend comprise about a third (32%) of the 25 drag cases found by ANOVA to feature systematic unexplained variance.

The third and final category of trending systematic unexplained variance based on Eq. (7) is one in which regression analysis fitting the response data to time reveals a significant second-order coefficient (b_2 in Eq (7)). That is, these are the cases for which the regression estimate of b_2 is large enough compared to the uncertainty in estimating it to infer with at least 95% confidence that the coefficient is non-zero, and therefore real. A non-zero second-order or quadratic term in the regression model describing variation with time implies a slope that is changing with time; that is, curvature in the time trend. Figure 20 is an example.

As in Figs. 18 and 19, the points in Fig. 20 represent response differentials. In this case, lift measurements are plotted that were acquired at AoA = -5° and at Site 2 in the tunnel operating envelope (Fig. 2) in the three days of "Week 1" (Fig. 6). Each point representing an empirical estimate of lift coefficient relative to the grand mean of all 40 Site 2 lift coefficient estimates made at AoA = -5° throughout the test (two in the morning and two in the afternoon of 10 test days). Absent any experimental error, the 12 points in this figure should be identical, and should all lie on the red line denoting no departure from the 40-point grand mean. Allowing only for ordinary chance variations in the data as is the standard operating procedure of OFAT testing, the means of the four-point samples acquired on each day should lie near the red line, and should display no trend with time.

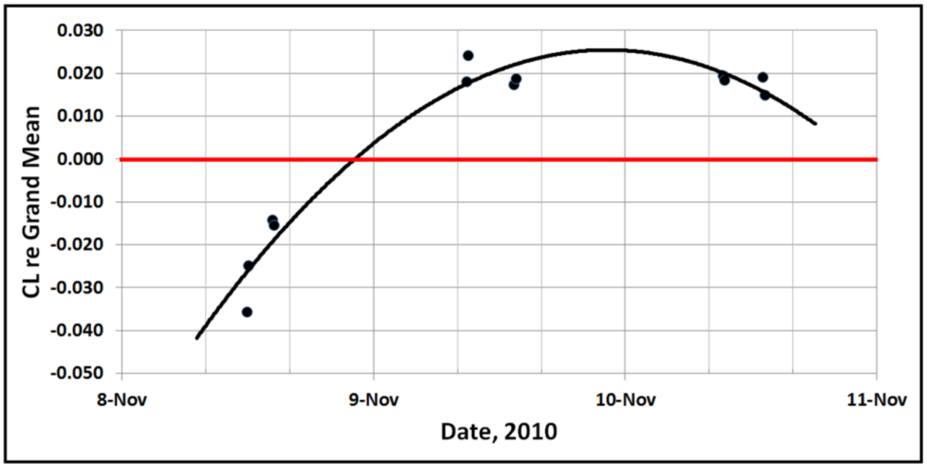


Figure 20. Lift coefficients for Site 2, Week 1, AoA = -5° relative to the grand mean of all Site 2 data acquired at AoA = -5°. Illustration of curvature in time trend.

In this case the means of the four daily samples obviously change with time, and do so systematically rather than randomly. Each daily sample mean differs from the grand mean by more than can be explained by the chance variations that occur within each daily sample, but the means of the latter two samples seem to have stabilized at a more or less constant bias of about 200 counts above the grand mean. As in all other similar cases, this behavior is attributed to ubiquitous covariate effects, but the exact nature and cause of those effects in this specific instance is unknown. Of the 36 independent combinations of model attitude, tunnel state, and time interval that were examined, two showed curvature for lift (6%) and five showed curvature for drag (14%).

It is instructive to modify Fig. 17 to incorporate the frequency with which various types of trends illustrated in this section were observed. Figure 21 indicates how likely it is that a given sample of data will feature a significant component of systematic unexplained variance with a specified behavior over time:

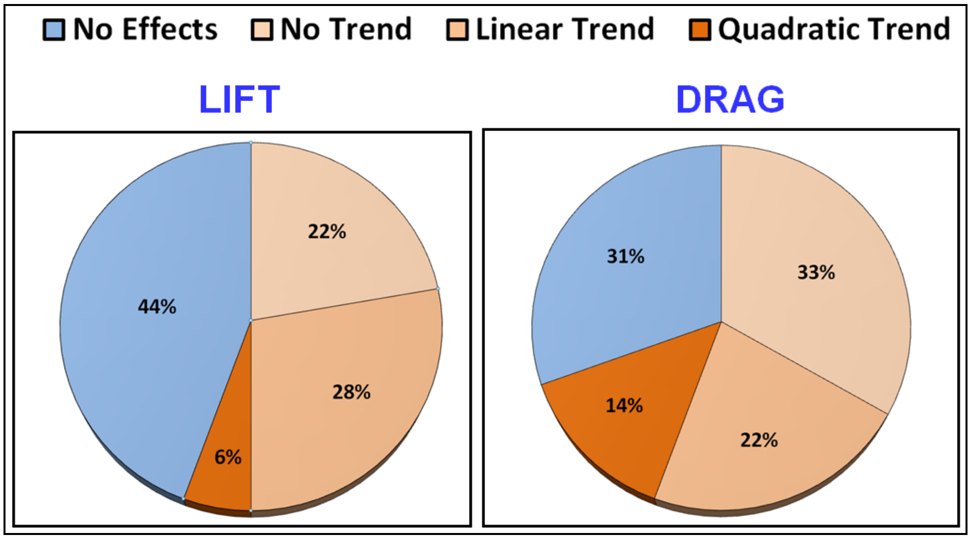


Figure 21. Various patterns of unexplained systematic variance.

F. Relative Magnitudes of Random and Systematic Error

The previous subsections under the Analysis and Results header have focused on the analytical methods of objective hypothesis testing by which a null hypothesis was tested that asserts no significant variation among the means of ostensibly identical samples that differ only by when they were acquired within a tunnel entry. Various combinations of model attitude, tunnel state, and time interval were examined, and the cases were noted in which the null hypothesis could be rejected with 95% confidence, and when the null hypothesis could not be rejected on the evidence in the data. Statistically significant levels of systematic unexplained variance were detected in over half of the lift cases examined and in more than two out of three of the drag cases. By "systematic unexplained variance" it is meant a non-random component of the total unexplained variance in a data sample that is therefore attributable to time-varying bias errors assumed to be due to unknown covariate effects.

"Statistically significant" implies nothing more than that the systematic variance component could be unambiguously detected, where we have defined "unambiguously" to mean "with at least 95% confidence." However, it is possible that systematic error can be *statistically* significant at the same time it is not significant from a practical engineering perspective. (too small to be problematic). Figure 22 illustrates such a case.

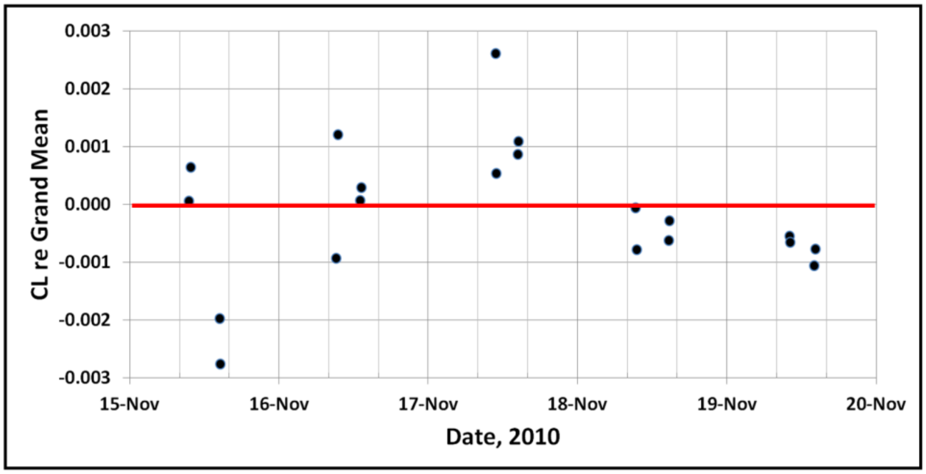


Figure 22. Lift coefficients for Site 2, Week 2, AoA = 1° relative to the grand mean of all Site 2 data acquired at AoA = 1°. Systematic variation of sample means with time that is nonetheless not physically significant.

Lift measurements displayed in Fig. 22 were acquired at AoA = 1° and at Site 2 in the tunnel operating envelope (Fig. 2) during the five days of Week 2 (Fig. 6). As with earlier presentations of this type, each point representing an empirical estimate of lift coefficient relative to the grand mean of all 40 Site 2 lift coefficient estimates made at AoA = 1° throughout the test (two in the morning and two in the afternoon of 10 test days). Absent any experimental error, the 20 points in this figure would all lie on the red line denoting no departure from the 40-point grand mean. Allowing only for random error, the means of the daily samples, as well as the means of the individual morning and afternoon samples, would lie near the red line. In this case, however, an ANOVA revealed that the variance with time exceeded the random error variance by more than a factor of four, justifying rejection of the null hypothesis of no systematic variance with more than 98% confidence. This is likely due to the afternoon replicates acquired on the first day of this interval, 15-Nov-10. Based on variance in the daily samples, these points seem to be biased lower than can be explained by ordinary random error. Likewise, at least one of the morning points acquired on the third day of the interval, 17-Nov-10, may be high enough to bias the sample mean too high to be explained by random error.

While it does appear upon inspection that some of the daily samples are biased relatively low or relatively high compared to the grand mean by amounts that cannot be attributed to the random error evident in the daily samples, a glance at the scale on the left reveals that for practical purposes these variations are of little concern. Even the most egregiously divergent sample means are within 0.003 of the grand mean, which is likely to be regarded as acceptable for all but the highest precision performance lift tests. In this case, then, the statistically significant systematic component of unexplained variance could only be detected so unambiguously because of the especially low random error in play for this particular combination of model attitude, tunnel state, and interval of time. It was not because the systematic error itself was necessarily large enough to be problematical.

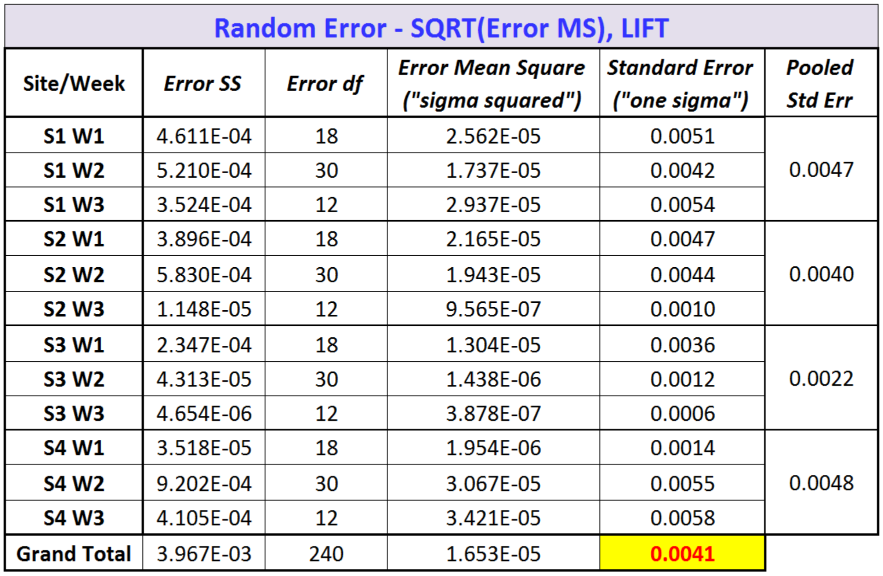
Small random error is likely to have been responsible for the "statistical significance" of systematic variance in other cases examined in the test, although as Figs. 18–20 indicate, there were also cases in which the systematic error was large enough in absolute terms to be troublesome. A total of 36 independent combinations of model attitude, tunnel state, and time interval were examined in this test. It is difficult to make an objective determination of how many of the 20 lift cases and 25 drag cases with statistically significant variations in sample means were cases in which these variations were also significant from an engineering perspective. This is because the allowable tolerance for experimental uncertainty varies, depending on the details of the specific wind tunnel test. In this study we simply report the situations in which systematic unexplained variance can be said with at least 95% confidence to be greater than ordinary random error, contravening the standard OFAT assumption of statistical control.

Having said that, one calculation that remains to be reported is the actual standard error ("one sigma" value) observed for random and systematic unexplained variance. We examine that question now.

Recall that the standard error is just the square root of the mean square from an ANOVA table. We used Eq. (5) to compute the standard random error by taking the square root of the error mean square from Table 4, the two-way ANOVA table for data displayed in Table 3. These were the lift data acquired during Week 2 at Site 1 of the operating envelope (Figs. 2 and 6). We subsequently obtained a more refined estimate of the error mean square by executing a two-way ANOVA *with replication* to remove the contribution of block/factor interactions from the error mean square. This allowed us to obtain a more pure estimate of ordinary random error. See Table 5 and discussion following it.

We performed similar two-way analyses of variance with replication for all 12 combinations of the four test sites and three "weeks" displayed in Figs. 2 and 6, respectively. The error mean squares and their corresponding degrees of freedom from these 12 ANOVA tables for lift and the 12 corresponding drag ANOVA tables are displayed in Tables 7 and 8.

Table 7: Random standard errors for lift coefficient computed from ANOVA statistics



Random Error - SQRT(Error MS), LIFT

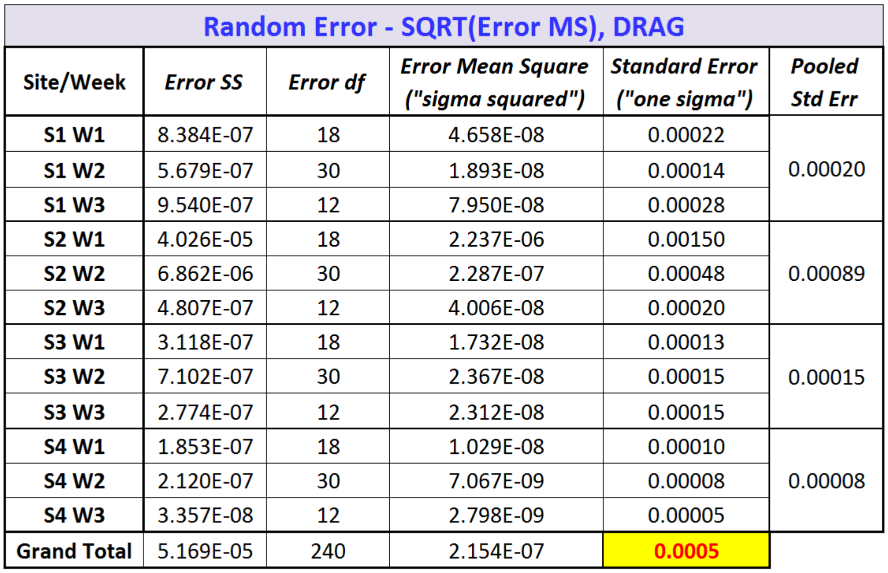
Site/Week	*Error SS*	*Error df*	*Error Mean Square ("sigma squared")*	*Standard Error ("one sigma")*	*Pooled Std Err*
S1 W1	4.611E-04	18	2.562E-05	0.0051	0.0047
S1 W2	5.210E-04	30	1.737E-05	0.0042	
S1 W3	3.524E-04	12	2.937E-05	0.0054	
S2 W1	3.896E-04	18	2.165E-05	0.0047	0.0040
S2 W2	5.830E-04	30	1.943E-05	0.0044	
S2 W3	1.148E-05	12	9.565E-07	0.0010	
S3 W1	2.347E-04	18	1.304E-05	0.0036	0.0022
S3 W2	4.313E-05	30	1.438E-06	0.0012	
S3 W3	4.654E-06	12	3.878E-07	0.0006	
S4 W1	3.518E-05	18	1.954E-06	0.0014	0.0048
S4 W2	9.202E-04	30	3.067E-05	0.0055	
S4 W3	4.105E-04	12	3.421E-05	0.0058	
Grand Total	3.967E-03	240	1.653E-05	**0.0041**	

Each row in Tables 7 and 8 corresponds to a different combination of operating envelope site and time interval ("week"). The error sum of squares, degrees of freedom, and mean square are copied directly from the corresponding tables for two-way ANOVA with replication, such as Table 5. The standard error column is generated by taking the square root of each error mean square.

The sums of squares and degrees of freedom are pooled in these tables in two ways. First, in the Grand Total row at the bottom of each chart, the sums of squares and degrees of freedom are simply added for all 12 site/week combinations. The corresponding error mean square is computed by dividing the total sum of squares by the total degrees of freedom, and a grand integrated estimate of standard error is obtained from the square root of this grand mean square. This is the integrated "one sigma" value for lift coefficient (Table 7) and drag coefficient (Table 8).

The rows in these tables are clustered into four groups of three time intervals each, with ach group corresponding to a specific site in the operating envelope. A pooled standard error is computed for each site by adding its three sums of squares and dividing this by the sum of the three corresponding degrees of freedom and taking the square root of the result. These are the pooled standard error numbers displayed in the far-right column for each of the four sites in the operating envelope.

Table 8: Random standard errors for drag coefficient computed from ANOVA statistics.

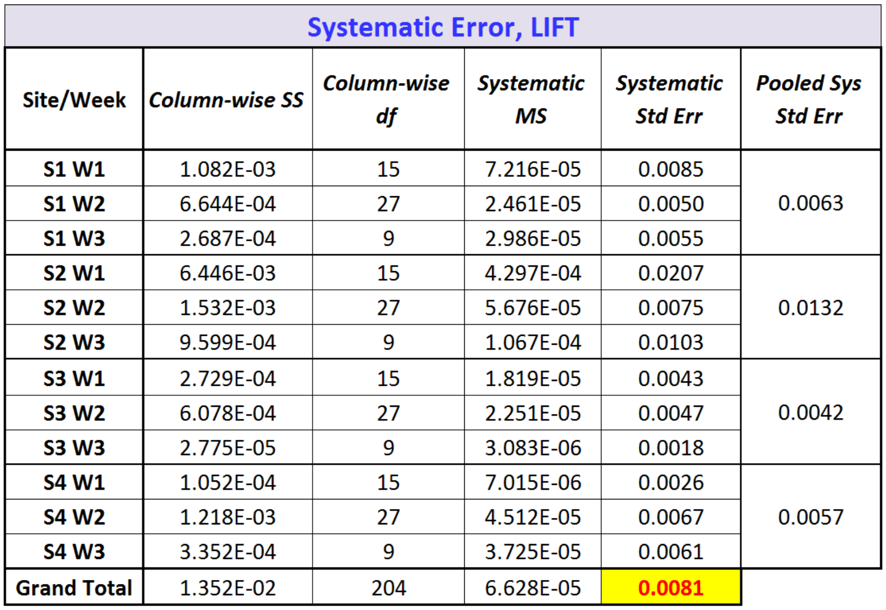


Random Error - SQRT(Error MS), DRAG

Site/Week	*Error SS*	*Error df*	*Error Mean Square ("sigma squared")*	*Standard Error ("one sigma")*	*Pooled Std Err*
S1 W1	8.384E-07	18	4.658E-08	0.00022	0.00020
S1 W2	5.679E-07	30	1.893E-08	0.00014	
S1 W3	9.540E-07	12	7.950E-08	0.00028	
S2 W1	4.026E-05	18	2.237E-06	0.00150	0.00089
S2 W2	6.862E-06	30	2.287E-07	0.00048	
S2 W3	4.807E-07	12	4.006E-08	0.00020	
S3 W1	3.118E-07	18	1.732E-08	0.00013	0.00015
S3 W2	7.102E-07	30	2.367E-08	0.00015	
S3 W3	2.774E-07	12	2.312E-08	0.00015	
S4 W1	1.853E-07	18	1.029E-08	0.00010	0.00008
S4 W2	2.120E-07	30	7.067E-09	0.00008	
S4 W3	3.357E-08	12	2.798E-09	0.00005	
Grand Total	5.169E-05	240	2.154E-07	0.0005	

Tables 9 and 10 are similar to Tables 7 and 8 except that they display the ANOVA sum of squares and degrees of freedom for column-wise (that is, systematic time-wise) variation. The standard error calculations performed in these tables therefore describe the systematic (not random) component of unexplained variance.

Table 9. Systematic standard errors for lift coefficient computed from ANOVA statistics.



Systematic Error, LIFT

Site/Week	*Column-wise SS*	*Column-wise df*	*Systematic MS*	*Systematic Std Err*	*Pooled Sys Std Err*
S1 W1	1.082E-03	15	7.216E-05	0.0085	0.0063
S1 W2	6.644E-04	27	2.461E-05	0.0050	
S1 W3	2.687E-04	9	2.986E-05	0.0055	
S2 W1	6.446E-03	15	4.297E-04	0.0207	0.0132
S2 W2	1.532E-03	27	5.676E-05	0.0075	
S2 W3	9.599E-04	9	1.067E-04	0.0103	
S3 W1	2.729E-04	15	1.819E-05	0.0043	0.0042
S3 W2	6.078E-04	27	2.251E-05	0.0047	
S3 W3	2.775E-05	9	3.083E-06	0.0018	
S4 W1	1.052E-04	15	7.015E-06	0.0026	0.0057
S4 W2	1.218E-03	27	4.512E-05	0.0067	
S4 W3	3.352E-04	9	3.725E-05	0.0061	
Grand Total	1.352E-02	204	6.628E-05	0.0081	

The random and systematic lift standard errors for the four operating envelope sites that are displayed in Tables 7 and 9 are represented graphically in Fig. 23. A total standard error is constructed as the root sum square of the random and systematic error.

The "one-sigma" standard random error for lift was less than 0.005 for all sites within the tunnel and as low as 0.002 at Site 3. It is unclear if this is because Site 3 featured the lowest Mach number of the four sites tested, but it is plausible that flow speeds below a certain threshold result in lower random error. However, there does not appear to be a strong general correlation between levels of random error and Mach number.

Table 10: Systematic standard errors for drag coefficient computed from ANOVA statistics

Systematic Error, DRAG					
Site/Week	***Column-wise SS***	***Column-wise df***	***Systematic MS***	***Systematic Std Err***	***Pooled Sys Std Err***
S1 W1	5.148E-06	15	3.432E-07	0.00059	0.00042
S1 W2	3.500E-06	27	1.296E-07	0.00036	
S1 W3	5.311E-07	9	5.901E-08	0.00024	
S2 W1	5.892E-04	15	3.928E-05	0.00627	0.00395
S2 W2	1.803E-04	27	6.677E-06	0.00258	
S2 W3	2.642E-05	9	2.935E-06	0.00171	
S3 W1	1.267E-06	15	8.449E-08	0.00029	0.00026
S3 W2	1.804E-06	27	6.682E-08	0.00026	
S3 W3	2.757E-07	9	3.063E-08	0.00018	
S4 W1	6.262E-07	15	4.175E-08	0.00020	0.00015
S4 W2	3.810E-07	27	1.411E-08	0.00012	
S4 W3	1.422E-07	9	1.580E-08	0.00013	
Grand Total	8.096E-04	204	3.969E-06	**0.0020**	

Figure 24 compares the random and systematic components of unexplained variance in drag data at the same four sites within the tunnel operating envelope as in Fig. 23. Again, the total error represents a root sum squared combination of the random and systematic components. The “one-sigma” standard random error for drag coefficient was no greater than 0.00020 (two drag counts) for all sites except Site 2, where it was 0.00089. It was 0.8 drag counts at Site 4, 1.5 counts at Site 3, and 2.0 counts at Site 1.

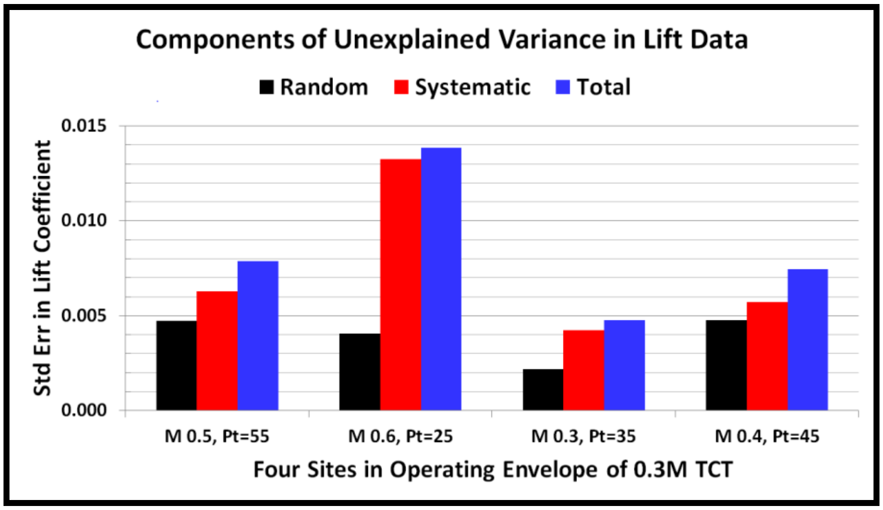


Figure 23. Unexplained variance components for lift coefficient estimates at four sites in the operating envelope of the 0.3M Transonic Cryogenic Tunnel

The systematic error, due to time-varying sample means, exceeded the random error at all four operating envelope sites for both lift and drag. However, for both lift and drag, larger systematic errors were generated at Site 2 than at any of the other sites that were examined. This site is defined by a Mach number of 0.6 (highest for the four sites) and a total pressure of 25 psia, (lowest of the four sites). While no immediate explanation suggests itself, the fact is that the means of ostensibly identical data samples varied the most with time under these conditions.

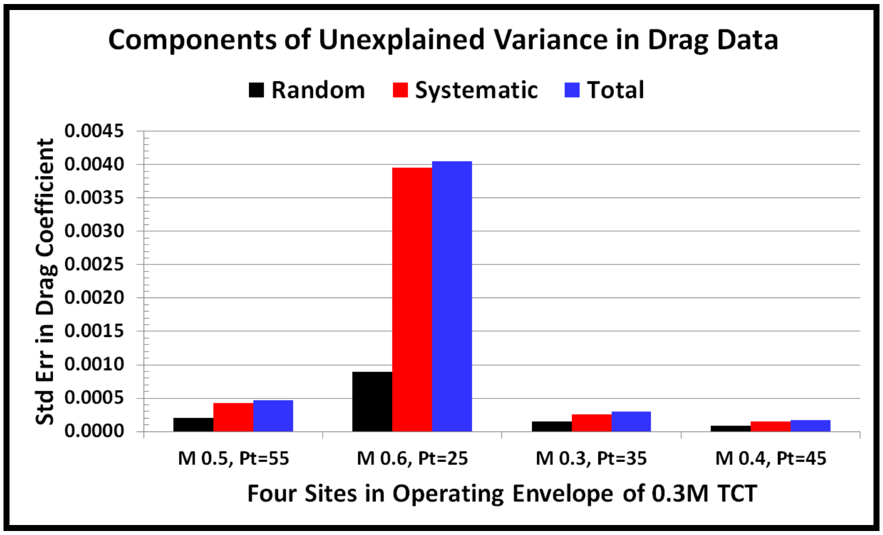


Figure 24. Unexplained variance components for drag coefficient estimates at four sites in the operating envelope of the 0.3M Transonic Cryogenic Tunnel

It is instructive to consider the ratio of the systematic unexplained variance observed in this test to the random unexplained variance, and also the ratio of the total unexplained variance to the random unexplained variance. This is because in conventional OFAT testing, the systematic error is generally neglected under an assumption of statistical control, by which the total error is widely believed to be random.

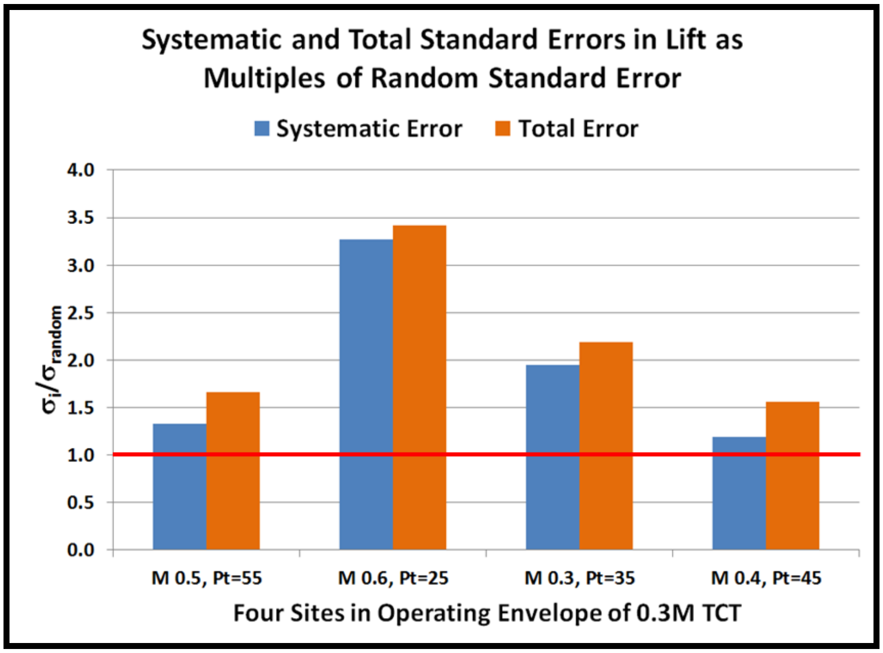


Figure 25. Ratios of systematic and total standard lift errors to random lift standard error.

Figure 25 represents the standard and total lift errors of Fig. 23 as multiples of the standard random error. The ratio exceeds 1 in all cases. It is just over one for Sites 1 and 4, where the systematic error is therefore at least as large as the random error and at Site 3 the systematic error is about double the random error. It is more than three times larger at Site 2. Note also that the larger the systematic error relative to the random error, the less impact the random error has on the total. This is because the random and systematic error components are root sum squared to form the total. For practical purposes, an uncertainty assessment that accounted for the systematic unexplained variance due to time-varying sample means could ignore the random error with relatively little adverse impact on the uncertainty assessment.

Figure 26 is the equivalent of Fig. 25, but for drag errors rather than lift errors. The same general patterns are observed for drag as were observed for lift. In all cases, the systematic error exceeds the random error. The systematic component of the total error makes the total at least twice as large as it would be absent any systematic variation of sample means with time. The total drag error is over four times larger than the random drag error for Site 2. In all cases, the systematic component of the total error is large enough that completely neglecting the random error would have little impact on the assessment of total uncertainty, as long as the systematic error was well understood.

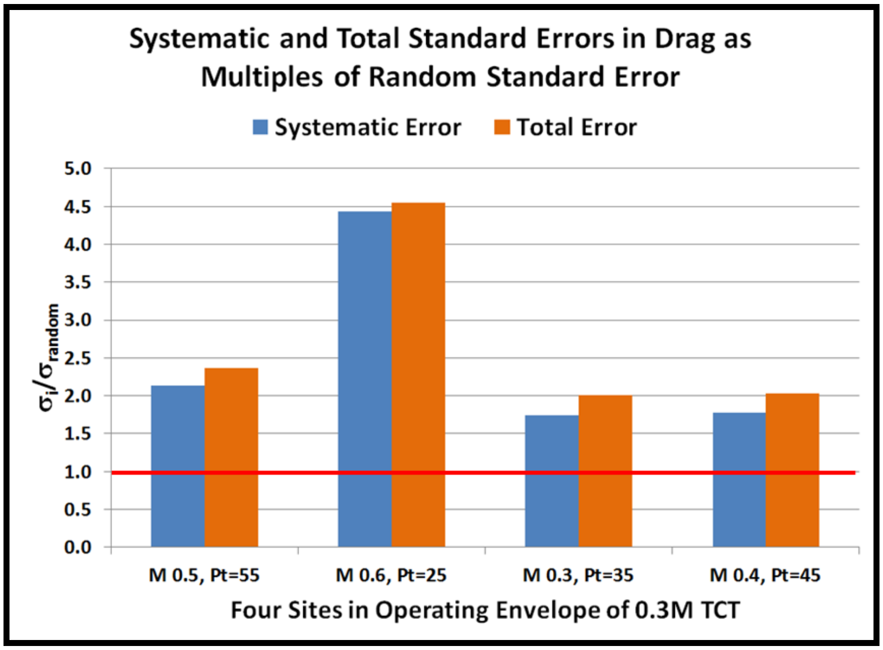


Figure 26. Ratios of systematic and total standard drag errors to random lift standard error

V. Discussion

This section elaborates briefly and in no particular order on various topics surfaced earlier in the paper. The intent is to clarify and/or accentuate a few important points.

A. Impact of Systematic Unexplained Variance

The focus of the paper has been on methods to test for objective evidence that the means of ostensibly identical data samples can vary significantly when there is some interval of time separating the acquisition of the samples. It is useful to revisit why this question is so important in experimental research generally, and in wind tunnel testing specifically.

If time-varying bias errors are in play over some interval of time, due to the kinds of covariate effects that were discussed in the Introduction, then the errors associated with data acquired in over that interval will not be independent of each other. That is, if some systematic variation (some temperature effect, say) is causing progressively higher responses over time, then it is no longer true that the next data point is equally likely to be too low as too high due to experimental error. There is a greater probability under such circumstances that it will be too high than too low. The experimental errors will be correlated to some degree, and not independent of each other.

Measurements comprising a sample of experimental data are random variables for which the exact numerical value is determined by a probability distribution. Sample statistics such as means and variances are the only metrics available to describe such variables. If the errors associated with each measurement are not independent, these sample statistics are not unbiased estimators of the population parameters ("true values") they purport to represent. Stated slightly differently, we are always limited by resource constraints to a finite number of imperfect data points, so if they display a net bias in one direction because of changes occurring systematically with time, the results of the experiment will be biased.

Consider the consequences of a bias error compared to a random error. Pure random error is responsible for some band of plausible values centered about the true response. The mean of a sample of data featuring only random error is just as likely to be higher than the true value as lower, and so represents an unbiased estimate of the true response.

If the errors are correlated and therefore occur more in one direction than another, the sample mean will no longer be an unbiased estimate of the true response. Under such circumstances, the more data that are acquired, the more likely it will be that the sample mean is displaced from the true response. Replication is not a remedy for bias errors, and high data volume simply ensures a high precision estimate of a biased result. In short, the lack of independent experimental errors ensures that the more data that are acquired, the less likely the result will fall within specified precision intervals that assume only random error. This is certainly consistent with wind tunnel testing experience generally, in which results are notoriously difficult to reproduce consistently from test to test within experimental error limits that assume random error only. This is the case despite the substantial volume of data typically acquired in a wind tunnel test.

This suggests that wind tunnel test reproducibility can be improved if quality assurance tactics are invoked to ensure independent experimental errors. This is in fact a standard operating procedure that has been practiced outside the experimental aeronautics community for nearly a century, since such tactics were first introduced by Ronald Fisher[5]. Wind tunnel facilities that incorporate such tactics into their experimental procedures are likely to see an improvement in their reputations for reproducible results.

B. Quality Assurance Tactics

We distinguish between quality *assessment* and quality *assurance*. Quality assessment entails activities designed to facilitate a quantitative estimate of the uncertainty in an experimental result. Quality assurance refers to various procedures intended to drive the uncertainty to acceptable levels. A comprehensive description of available quality assurance tactics is beyond the scope of this paper, but one such tactic designed to assure statistical independence deserves some mention.

To the extent that the independence of experimental errors is considered at all in a wind tunnel test, it is usually assumed to exist absent some overt error on the experimentalist's part to disrupt it. That is, independence is often not considered at all. When it is, borrowing from an insightful description in Ref. 3, *"[Researchers] frequently make the assumption of independence at the beginning of their writings and rest heavily on it thereafter, making no attempt to justify the assumption, even though it might have been thought that 'a decent respect to the opinions of mankind requires that they should declare the causes which impel them' to do so. The mere <u>declaration</u> of independence, of course, does not guarantee its existence."*

One standard tactic designed explicitly to ensure independence in the presence of the kinds of systematic experimental errors detected in this study is to randomize the set-point order of factor combinations prescribed in the test matrix. If angles of attack are set in monotonically increasing order while bias errors are changing with time, for example, then responses acquired at higher angles of attack will be shifted by more than responses acquired at lower angles of attack. The result will be a data structure (a polar, say) that is rotated relative to the case with no changing bias error. Randomizing the set-point order ensures that some low angles of attack are acquired early and some are acquired late. The early responses will be biased in one direction relative to the sample mean and the late responses will be biased in the direction. If the set-point order has been randomized, then each new point is as likely to have been acquired late as early, and is therefore as likely to be biased in one direction as another. The same can be said for the higher angles of attack that are acquired in random order. Randomizing the set-point order therefore has the effect of converting a slowly changing bias error into an independent random error that is just as likely to be a little too high as a little too low. The functional form of the AoA dependence is this retained, albeit at the expense of a slight increase in the random error band about that true dependence.

There can be practical considerations related to randomizing the set-point order of a test matrix, including a whole set of conditions known in the literature of formal experiment design as "restrictions on randomization." One example of such a restriction is the case in which Reynolds number is an independent variable in a cryogenic wind tunnel test. It is not practical to randomize the cooling sequence of the tunnel. Special experiment designs intended to ensure independence of measurement errors in this situation are available, including a large class of what are called "Split Plot" designs. The details of Split Plot experiment designs are beyond the scope of this paper but are easily accessible in the literature[1–5]. Other tactics are available when the restrictions are milder. For example, hysteresis due to the differences in how flow attaches to a wing that is increasing or decreasing in angle of attack necessitate some care in randomizing the AoA set-point order. Adding a low-AoA "home state" between two AoA settings that constitute a decrease in AoA ensures that both set-points are approached from below, if that is deemed desirable.

C. Block/Factor Interactions

A two-way analysis of variance with interactions partitions that component of the total variance not attributable to row-wise or column-wise variations into two components. One is the variance due to ordinary random variations in the data and the other is a component attributable to what are called block/factor interactions.

Covariate effects might be expected to result in higher or lower responses to be recorded on Tuesday than on Monday, say, when measurements are made on both days at the same angle of attack. However, one would not normally expect the effect of a one-degree *change* in angle of attack to be different on Monday than on Tuesday.

This would be unexpected on both practical and theoretical grounds. Theoretically, covariate effects act as "the tide that lifts all boats," so that both AoA measurements made on the second day would be expected to differ by the same amount from levels acquired the prior day. A differential AoA measurement should therefore yield very nearly the same result on both days. On practical grounds one would also expect this result; else the whole aerodynamic meaning of "one degree change in angle of attack" is called into question. Clearly, chaos would ensue if we had to distinguish between the Monday AoA change and the Tuesday AoA change.

If the effect of a unit change in some factor such as AoA were to be different in one block of time than in another block of time, we would say that block/factor interactions are play. As implausible as such interactions would seem to be on both theoretical and practical grounds, they were nonetheless detected in this experiment. There is both good news and bad in this observation. The good news is that when a large component of the unexplained variance previously attributed to random error is properly assigned to some other cause (block/factor interactions in this case), the remaining random error component is much smaller. By properly accounting for block factor interactions, the 0.3M TCT tunnel emerges as a much higher-precision facility. Published drag precision estimates of ±0.0010 (ten drag counts) for this facility overstate the random error displayed in Fig. 24 everywhere except the high-Mach/low-Pt corner of the tunnel's operating envelope[7]. This may be because of systematic error effects improperly assumed to be random.

The bad news about block/factor interactions, of course, is that they must be explained. **S**hort-term block effects are one plausible explanation. The systematic bias changes associated with covariate effects have been described as slowly varying and relatively long-term. But block effects—response changes that are too great to attribute to random error, which occur over some block of time—may also be in play over relatively short intervals. If that is the case, then the differential measurements made on Monday and repeated on Tuesday in the hypothetical example discussed here might each reflect some significant covariate effect in the time between successive measurements on each day. That would explain the interaction effects, and is consistent with observations made elsewhere[11] of relatively short-term block effects—changes of a multiple of the error budget in less than an hour.

There may be other explanations, of course. The unanticipated block/factor interactions observed in this test will require further study to fully understand. However, it is unlikely that such effects are limited to a single wind tunnel facility. That is, there is nothing about the 0.3M TCT tunnel to suggest that it would be more susceptible to block/factor interactions than any other tunnel. It would probably be prudent to conduct similar studies in other tunnels as well.

D. Attitudes Toward Systematic Unexplained Variance

The experimental aeronautics community is populated by experimentalists who accept the random component of unexplained variance in data as inevitable and an entirely natural reflection of reality. No one today would propose expending resources to test a null hypothesis that random error does not exist. Likewise, there are few efforts to identifying every individual cause of random experimental error with a view to perfecting the measurement environment by eliminating them. It is generally conceded that random error is the result of a large number of factors and the fluctuations associated with it are regarded as the algebraic sum of these many unknown effects. Rather than trying to eliminate random error, we focus instead on devising tactics (chiefly replication) to cope with it, and to assess the experimental uncertainty that results from its inevitable presence.

Unfortunately, many experimental aerodynamicists who are quite comfortable with the concept of random error find it troubling to believe that there might also be a non-random, systematic component of unexplained variance that is just as natural, and just as inevitable, as the random component. Researchers in other fields do not generally suffer from this peculiar blind spot with respect to systematic unexplained variance; tactics to cope with it have been a part of standard experimental operating procedures in other industries for almost a hundred years. But in OFAT wind tunnel testing, one generally feels entitled to assume that random fluctuations in the data occur about mean values that remain reliably unchanged for as long as it is convenient that they should do so.

The OFAT practitioner with a limited background in statistics can be forgiven for believing that the highest quality results can still be achieved under conditions for which covariate effects render experimental errors correlated and not independent. In any case, he can seldom afford to detect or defend against systematic unexplained variance while simultaneously attempting to cover a large design space with direct physical measurements that are executed one factor at a time. No doubt another factor is that systematic unexplained variance, unlike ordinary random error, is "stealthy," requiring advanced analytical methods and the expense of additional quality assessment data to reveal.

Furthermore, unlike random error, covariate effects are not always in play; there is a relatively high probability that any given data point is clear of their influences. An earlier study[12] estimates that the 95% confidence interval enclosing the percentage of polars afflicted with systematic covariate effects to range from 15% to 35% for that study, while the current study reports somewhat higher frequencies—56% for lift and 69% for drag. The probability that any given data sample will be substantially free of systematic effects is thus expected to be in the range of 31% to 44% by the current study and 65% to 85% by the earlier one, so it is not difficult to miss evidence of such effects, especially when few quality assessment data points are budgeted.

These frequency of occurrence numbers are consistent with what has been called the Blind Squirrel Theory for why OFAT testing has continued for so long in spite of difficulties with covariate-induced correlations among experimental errors. This theory, attributed to experiment design expert, Douglas Montgomery, notes that while the blind squirrel spends most of his time tripping over roots and falling into holes, occasionally he stumbles across an acorn lying on the ground and so does not starve to death.

More such studies will be necessary before a reliable consensus can emerge as to how likely it is that covariate effects are in play in a typical wind tunnel test but if these initial results are any guide, the frequency is likely to be inconsistent with industry quality standards that typically require results to be within prescribed tolerances with 95% confidence. Furthermore, the adverse impact of covariate effects can be substantial when they do occur; systematic errors can be several multiples of the ordinary random error commonly assumed to comprise the entire error in an OFAT test.

As results of the kind reported in this paper become more common, the experimental aeronautics community is likely to evolve into a state in which it is taken for granted that there is a systematic component of unexplained variance that is as natural as the random component, and that it is no more necessary to entirely eliminate systematic variation in order to achieve reproducible experimental results than it is to eliminate random error. The current relatively crude strategy of building aeronautical databases with a high volume of direct physical measurements is likely to yield to a more sophisticated approach in which limited resources are deployed in such a way as to enable adequate response estimates for factor combinations that cannot be physically set because of resource limitations. In that environment, it will be practical to replace the high-speed data collection imperative with a focus on quality assurance tactics that will generate substantially more reproducible experimental results by ensuring statistical independence in the measurements, while also providing greater coverage of the design space.

VI. Concluding Remarks

This paper proposes methods to quantify the random component of unexplained variance in a wind tunnel test and to objectively test a null hypothesis that the unexplained variance contains no significant systematic component attributable to the variation of sample means with time. Methods are also described for quantifying systematic variance, attributable to non-random covariate effects such as instrument drift, flow angularity changes, transducer desiccation, gradual ablation of trip dots and grit, operator learning and fatigue effects, and numerous other causes.

The effect of significant systematic variance on the independence of experimental errors is discussed, as is the resulting bias of sample statistics used to characterize experimental data. Quality assurance tactics to defend against such biases that are standard practice outside the experimental aeronautics community are briefly described.

Paired t-tests, one-way analysis of variance, two-way analysis of variance, and two-way analysis of variance with replication are all introduced as methods to objectively assess the systematic component of unexplained variance in a wind tunnel test. These methods are illustrated with data acquired for this purpose in a wind tunnel check standard test in the 0.3-Meter Transonic Cryogenic Tunnel at Langley Research Center. While results of this test only strictly apply to the facility in which they were obtained, they are representative of results that have been obtained in other facilities and are believed to be quite general. Specific findings are as follows:

- The means of nominally identical data samples acquired at different times during the test differed from each other by more than can be attributed to ordinary chance variations in the data. These variations in sample means are attributed to non-random covariate effects.

- The null hypothesis of no significant variation with time of sample means was tested for 36 unique combinations of model attitude, tunnel state (Mach number and total pressure) and time interval (two to five days). In these 36 tests
    - Significant systematic unexplained variance was detected in lift data acquired in 20 cases (56%).
    - Significant systematic unexplained variance was detected in drag data acquired in 25 cases (69%).
    - Of the 20 cases of significant systematic unexplained variance in lift
        - Ten displayed a significant linear trend with time (28% of the 36 total cases)
        - Two displayed a significant quadratic dependence on time (6% of the 36 total cases)
        - Eight displayed no particular trend with time (22% of the 36 total cases)
    - Of the 25 cases of significant systematic unexplained variance in drag
        - Eight displayed a significant linear trend with time (22% of the 36 total cases)
        - Five displayed a significant quadratic dependence on time (14% of the 36 total cases)
        - Twelve displayed no particular trend with time (33% of the 36 total cases)
- Four combinations of Mach number and total pressure were examined, generally covering the non-cryogenic operating envelope of the tunnel. Similar results were obtained for three of the four combinations but higher levels of random error for drag, and systematic error for both lift and drag, were observed at the Mach 0.6 and 25 psia combination. This was highest Mach number and lowest total pressure examined.
    - At Mach 0.6 and 25 psia
        - The standard random error in drag coefficient (“one sigma”) was 0.00089 (8.9 drag counts)
        - The standard systematic error in drag coefficient was 0.00395 (39.5 drag counts)
        - The standard random error in lift coefficient (“one sigma”) was 0.0040
        - The standard systematic error in lift coefficient (“one sigma”) was 0.013
    - For the other three combinations (Mach 0.5, 55 psia; Mach 0.3, 35 psia; Mach 0.4, 45 psia)
        - The standard random error in drag coefficient (“one sigma”) was 0.8 to 2.0 drag counts
        - The standard systematic error in drag coefficient was 2 to 4 drag counts
        - The standard random error in lift coefficient (“one sigma”) was 0.0022 to 0.0048
        - The standard systematic error in lift coefficient (“one sigma”) was 0.0042 to 0.0063
- The systematic component of the unexplained variance exceeded to random component at all four Mach/pressure sites within the tunnel operating envelope for both lift and drag.
    - For lift, the systematic error was 1.2 to 3.3 times greater than the random error
    - For drag, the systematic error was 1.7 to 4.4 times greater than the random error
    - In no cases was the average systematic error less than the average random error

The results obtained in this test support a general conclusion that systematic unexplained variance due to variations in sample means over time is as natural and ubiquitous in wind tunnel testing as ordinary random error. The identification and elimination of all sources of systematic unexplained variance is not likely to be any more practical than the identification and elimination of all sources of random unexplained variance. As is the case with random error, any sources of systematic error that can be identified and eliminated at reasonable cost should be eliminated. However, it should also be understood that a residual level of unexplained variance is inevitable, and that this unexplained variance will inevitably feature a systematic component as well as a random component, with the systematic component likely to exceed the random component. To achieve consistently reproducible wind tunnel test results requires a proactive defense against both systematic and random unexplained variance. Likewise, an accurate assessment of uncertainty requires that systematic variations be taken into account as well as random variations in the data.

Acknowledgements

This work was funded by the Aeronautics Test Program Office of the National Aeronautics and Space Administration. The support of the Ground Facilities and Testing Directorate at Langley Research Center is also gratefully acknowledged.

References

[1] Diamond, W. J., *Practical Experiment Designs for Engineers and Scientists,* 2nd Ed., Wiley, New York, 1989.

[2] Cochran, W. G., and Cox, G. M., *Experimental Designs,* 2nd Ed., *Wiley Classics Library Edition*, Wiley, New York, 1992.

[3] Box, G. E. P., W. G. Hunter, and J. S. Hunter, *Statistics for Experimenters*, 2nd Ed., John Wiley & Sons, New York, 2005.

[4] Montgomery, D. C., *Design and Analysis of Experiments*, 7th Ed., John Wiley & Sons, New York, 2009.

[5] Fisher, R. A., *The Design of Experiments*, 1st Ed., Oliver and Boyd, Edinburgh, 1935.

[6] Balakrishna, S. and W.A. Kilgore, "Microcomputer Based Controller for the Langley 0.3-Meter Transonic Cryogenic Tunnel," NASA Contractor Report 181808, Vigyan Research Associates, Hampton, VA, March, 1989.

[7] 1997-06-18_Facility_Details_for_0.3_M_Transonic_Cryogenic_Tunnel. March 20, 1996. Online PDF file linked from http://crgis.ndc.nasa.gov/historic/Transonic_Cryogenic_Tunnel.

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

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

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

[11] DeLoach, R., Hill, J. S., and Tomek, W. G. "Practical Applications of Blocking and Randomization in a Test in the National Transonic Facility" (invited) AIAA 2001-0167. 39th AIAA Aerospace Sciences Meeting and Exhibit. Reno, NV. Jan 2001.

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

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

[14] Scheffe, H., *The Analysis of Variance*, John Wiley and Sons, New York, 1959.

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