Check-Standard Testing Across Multiple Transonic Wind Tunnels with the Modern Design of Experiments

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

This paper reports the result of an analysis of wind tunnel data acquired in support of the Facility Analysis Verification & Operational Reliability (FAVOR) project. The analysis uses methods referred to collectively at Langley Research Center as the Modern Design of Experiments (MDOE). These methods quantify the total variance in a sample of wind tunnel data and partition it into explained and unexplained components. The unexplained component is further partitioned in random and systematic components. This analysis was performed on data acquired in similar wind tunnel tests executed in four different U.S. transonic facilities. The measurement environment of each facility was quantified and compared.

I. Introduction

This paper presents an analysis of data acquired in support of the Facility Analysis Verification and Operational Reliability (FAVOR) project, in which similarities and differences among four US transonic wind tunnels were studied by executing nominally similar test matrices in each facility on the same test article, balance, and sting. The participating tunnels were the National Transonic Facility at Langley Research Center (LaRC), the 11-Ft Unitary Plan wind tunnel at Ames Research Center (ARC), the 16T wind tunnel at the Arnold Engineering and Development Center (ARDC), and the 8x6-Foot supersonic wind tunnel at Glenn Research Center (GRC). The test article was the AEDC 16T check standard model, a 5% scale model of an F-111.

The stated objective of the FAVOR project was to compare test methods, techniques, and procedures, as well as data reduction methods, flow quality, and aerodynamic data acquired across the four facilities in nominally identical wind tunnel tests. In support of these objectives, the NASA Aeronautics Test Program Office requested an independent analysis of the FAVOR data featuring techniques that are commonly employed in formal experiment design applications. The specific request is for an analysis of the data that utilizes methods referred to collectively at Langley Research Center as the Modern Design of Experiments[1-4] (MDOE). While the word "design" features prominently in the name of this experimental methodology, it actually consists of unified experiment design, execution, and analysis processes. The FAVOR tests were not designed or executed according to MDOE principles, but aspects of the MDOE analysis method can still be applied to the data.

The objective nature of MDOE analytical methods is especially attractive when the analysis could be influenced to some degree by subjective a-priori expectations of the result. It is more difficult for such expectations to impact an analysis based on prescribed computations and quantitative inference rules as an MDOE analysis is, than it is for them to impact a conventional analysis that may be more open to subjective interpretation. All parties using the specific MDOE methods employed in the FAVOR analysis reported in this document, with the same sample of data, will produce the identical result.

Multiple factor effects are partitioned by MDOE through an analysis of variance (ANOVA), which will be demonstrated in this report using the FAVOR data. While FAVOR was executed as a conventional One Factor At a Time (OFAT) test, the ANOVA method of partitioning factor effects can still be illustrated using replicated polars, for which time serves as a hidden second variable. Response changes that occur with time are especially relevant quality considerations that represent a key to improving facility performance, as will be discussed presently. Such changes can be responsible for a systematic component of the unexplained variance in a wind tunnel test that can dominate the more widely recognized random component. This systematic unexplained variance is attributable to factors that do not always reproduce precisely from facility to facility.

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

Not all between-facility differences can be attributed to systematic unexplained variance. It can also be difficult to achieve the identical result in multiple facilities because of such differences as scale effects when the same test article is installed in test sections that are of different size, for example, notwithstanding wall-effects corrections (that differ from facility to facility), which are applied to account for these differences. Different procedures, different instrumentation, and different levels of operator skill, training, and experience from one facility to the next can also make it difficult to precisely reproduce results across facilities. Nonetheless, systematic components of the unexplained variance and the resulting loss of statistical independence can be a significant contributor to irreproducibility.

The effects of systematic error variations are compounded by the OFAT convention of varying factors such as AoA in a systematic way. If only random variations occur while a conventional polar is acquired, the result is a certain amount of "fuzz" around what is essentially the right answer. However, if systematic variations are also in play as a polar is acquired, the shape of the polar can be altered, or if such variations occur between polars, the polars can be displaced from each other. The result in either case will not simply be random fluctuations about the right answer, but the wrong answer altogether. For this reason, an examination of the relative magnitude of random and systematic components of the unexplained variance will constitute a key element of the MDOE analysis of FAVOR data reported here.

Unfortunately, MDOE quality assurance tactics that are designed to defend against systematic variation were not employed in the FAVOR tests, so any systematic error effects that might have been in play are now firmly embedded in the data. While it is now too late to minimize systematic error effects in the FAVOR data, it is still possible to detect the presence of such effects. The MDOE analysis of the FAVOR data uses ANOVA to partition the total variance in sets of polar replicates into an *explained* component due to AoA, and into two components of *unexplained* variance, one due to time (systematic error) and one due to ordinary chance variations (random error). The various facilities are compared on the basis of these two error components.

In summary, this document describes an MDOE analysis of the FAVOR data that progressed through the following four phases:

1. **Identification of Replicated Polars**. Each polar acquired more than once, either within a single facility or across two or more facilities was identified.
2. **Data Reduction**. Replicated polars were clustered, and truncated to the largest angle of attack common to all replicates in a given cluster.
3. **Analysis of Variance**. Within- and between-facility ANOVA calculations were performed on each set of replicated polars included in the analysis, with a view to assessing the significance of any within- or between-facility systematic unexplained variation.
4. **Facility Comparisons.** The participating facilities were compared on the basis of random and systematic components of the unexplained variance in each set of polar replicates. At the request of the ATP program office, specific facilities are not identified by name, but rather are referred to as tunnel A–D.

I. Test Article and Instrumentation

Figure 1 shows the planform of the test article, the AEDC check standard model consisting of a 5% model of the F-111. The wings of the check standard model were modified to provide a 48-inch span at a fixed wing sweep angle of 35 deg. The wing span represents a compromise needed to accommodate test sections of significantly different size. It is nearly the largest span model that the NTF and the Glenn 8×6 tunnel can accommodate, but somewhat small for AEDC and Ames.

Trip dots of the same size were applied at the same location in all four tests. They were located on the nose and upper and lower surfaces on the wing strake, wing, and horizontal and vertical tails. The intent was not to remove any of the trip dots once they were applied, but in the case of the test at Glenn Research Center, temperatures in the test section were great enough to cause the trip dot adhesive to fail. The trip dots were replaced with grit at this facility.

Two control surface configurations were tested, designated Configuration 0 and Configuration 1. The horizontal tail was not deflected in Configuration 0, but it was deflected 10^0 in Configuration 1.

The NTF-115 single-piece moment-type balance was designed for use with the F-111 test article. A requirement was for all tests to use the same balance and calibration. Once the sting, balance, model, and instrumentation was built up it remained as one unit for the completion of the four tests. This was to ensure that the bridging of the balance, the installation of the balance to the model and sting, and the build-up of the model did not changed from test to test.

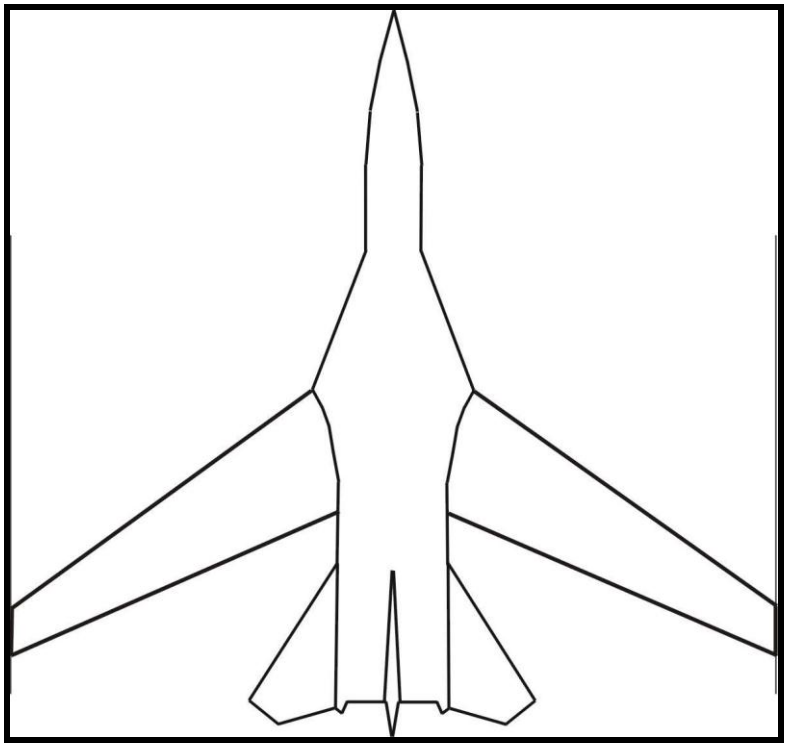

Figure 1. Planform of test article

Figure 2 shows the test article as mounted for testing at AEDC.

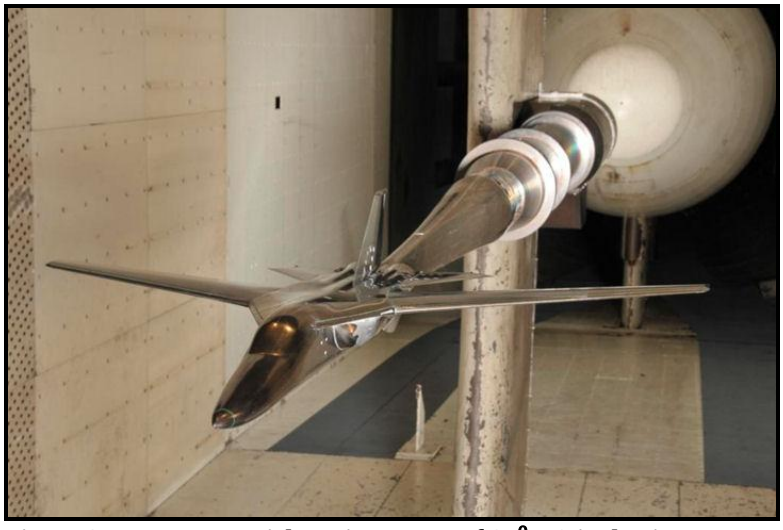

Figure 2. F-111 test article. Wing sweep of 35^0, 48-inch wing span

II. Favor Data Description

This section presents a brief description of the FAVOR data structure. It documents the design spaces examined at each facility and the volume of data acquired, including an identification of the data replicated within and between facilities.

Nominally similar test matrices were executed in all four tunnels, but with minor variations from facility to facility. The variations reflect differences in the Mach/Re design space from tunnel to tunnel, and how model inversions were executed. For example, at Tunnels A, C, and D inversions were achieved by rolling the model in the positive direction, while at Tunnels B, C, and D the model was rolled in the negative direction to invert it. That is,

inversions were executed with both positive and negative roll at tunnels C and D, with only positive roll at Tunnel A, and with only negative roll at Tunnel B. Tunnel D acquired data at both $+90^0$ and -90^0 of roll, and Tunnel C acquired data at $+90^0$ but not at -90^0 of roll. Neither Tunnel A nor Tunnel B acquired data with a 90^0 roll angle.

Beta sweeps were replicated at Tunnels A and B, but not at Tunnels C and D, and only Tunnel B replicated beta sweeps both with a model roll angle of 0^0 and 180^0. At tunnel A, all replicated beta sweeps were executed with the model inverted.

There were a total of 39,772 force, moment, and pressure polars acquired in the FAVOR study. Of these, 32,300 were pressure polars, acquired from 22 taps on each wing plus 6 fuselage taps. There were thus 32300/(22+22+6) = 646 sweeps across all four tunnels, or an average of 161.5 runs per facility. The 39,772 – 32,300 = 7,474 force and moment polars were acquired in the two conventional coordinate systems (body axis and stability axis), for 3,736 such polars in each coordinate system. The average of 3,736/646 = 5.78 force/moment polars per sweep in each coordinate system is between five and six because one of the tunnels did not report stability axis side force data.

The number of alpha and beta settings varied from polar to polar and the precise total of individual force, moment, and pressure values recorded in this test was not determined, but it was something on the order of a half million such numbers. The FAVOR test shares with typical OFAT experiments the property that massively more data were acquired than were necessary to achieve a particular purpose. In the case of the FAVOR study, the volume of acquired data significantly exceeds what would have been needed to reliably establish such differences and similarities as may exist in the data acquired among the four tunnels, and to adequately characterize the measurement environment of each tunnel. It is also likely that procedural differences from one tunnel to another could have been detected without the need to acquire so much data.

One downside to the acquisition of high volumes of data is the prodigious labor that would be required if all of it were analyzed. To make the initial analysis manageable, the analysis of the pressure data was postponed, and the current paper is limited to a comparison of the four FAVOR tunnels on the basis of stability axis force and moment data only. Of the total of 646 alpha and beta sweeps acquired from the four tunnels, 113 or 17.5% were neither replicated across two or more tunnels, nor replicated within the same tunnel in which they were acquired, rendering these data of limited utility in comparing facilities, or in assessing the measurement environment in a specific facility. Of the remaining 533 polars that were either replicated between or within selected facilities, 496 or 93.1% were alpha sweeps. The remaining 37 (6.9% of the replicated polars) were beta sweeps.

In all four tunnels the great majority of data were acquired with a roll angle of 0^0. Two configurations were tested at all four tunnels, designated Configurations 0 and 1. The two configurations differed by control surface deflections in the tail, with most of the data acquired at Configuration 0. A small number of runs acquired at a "Configuration 2" were also acquired at one of the tunnels. With the exception of a relatively small number of beta sweeps, OFAT data were acquired in each tunnel by changing angle of attack in the conventional way, while holding Mach number, Reynolds number, and configuration constant at nominal sideslip and roll angles of 0^0. To facilitate comparisons among all four facilities, only Configuration 0 alpha sweeps at zero roll angle are considered in this preliminary report, which is intended to focus primarily on the analysis methodology. A more comprehensive report describing results acquired at other test conditions is being prepared for delivery to the NASA Aeronautics Test Program Office.

Figures 3–5 display the Mach/Re design space for alpha sweeps at Configurations 0 and 1, and with Roll = 0^0 and $\pm 180^0$. Beta sweeps were also acquired at a small subset of these design space sites, as were a few sweeps with roll angles of $\pm 90^0$. Alpha and beta sweeps were acquired at a total of 94 unique combinations of all other variables examined in the FAVOR test. Unfortunately, despite the effort revealed in these figures to replicate similar conditions across tunnels, data were reported from all four tunnels for only one of the 94 conditions, a Configuration 1 Mach 0.4 beta sweep with a Reynolds number of 2.5 million/ft and a roll angle of 0^0. (The FAVOR team also made independent preliminary data comparisons across all four tunnels for Mach/Re combinations of [0.60, 3.85] and [0.85, 4.5], although the design space plots suggest that Reynolds number for Tunnel A was set slightly below levels common to Tunnels B-D at these two design space sites).

The fact that data were reported from all four tunnels at so few common set-point conditions restricts what can be objectively inferred about between-tunnel differences and similarities in measured response levels. In statistical parlance, we say that the opportunity to compare *location* variables (sample means) across facilities is limited. We are therefore prevented from answering such relevant questions as, "How do measured drag values compare across all facilities at any other combination of variables than the one (or possibly a few) common to all four tunnels?" However, several polars were replicated within each individual facility, and these within-facility polar replicates facilitated a reasonably detailed investigation of *dispersion* variables (variance and standard errors) that revealed much about the measurement environments in each tunnel, as will be described. Within-tunnel measurement environment metrics (standard errors) will be compared across facilities in this report.

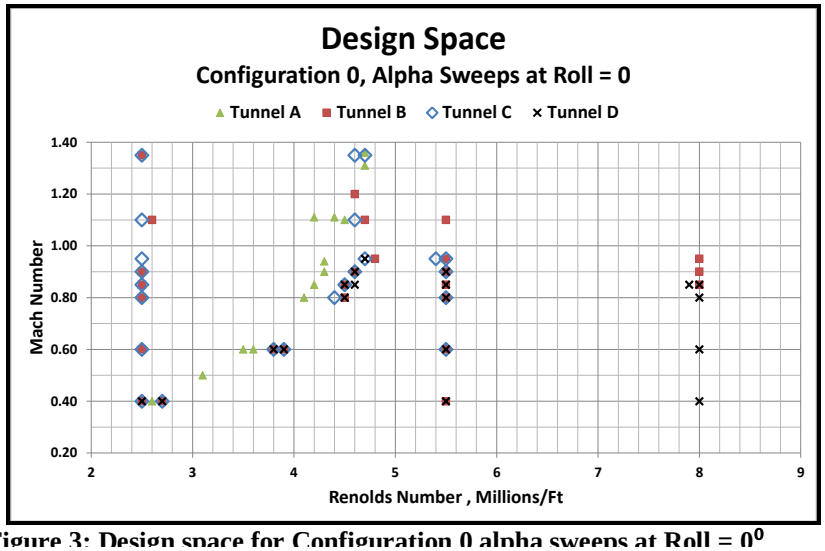

Figure 3: Design space for Configuration 0 alpha sweeps at Roll = 0^0

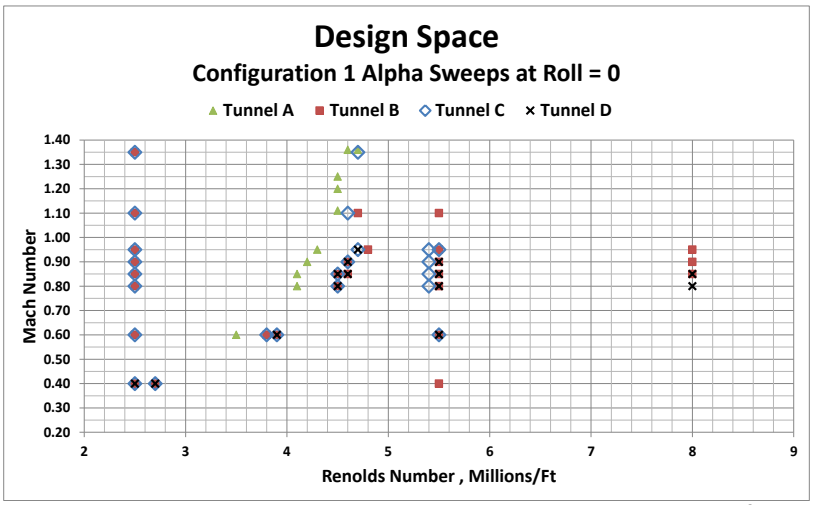

Figure 4: Design space for Configuration 1 alpha sweeps at Roll = 0^0

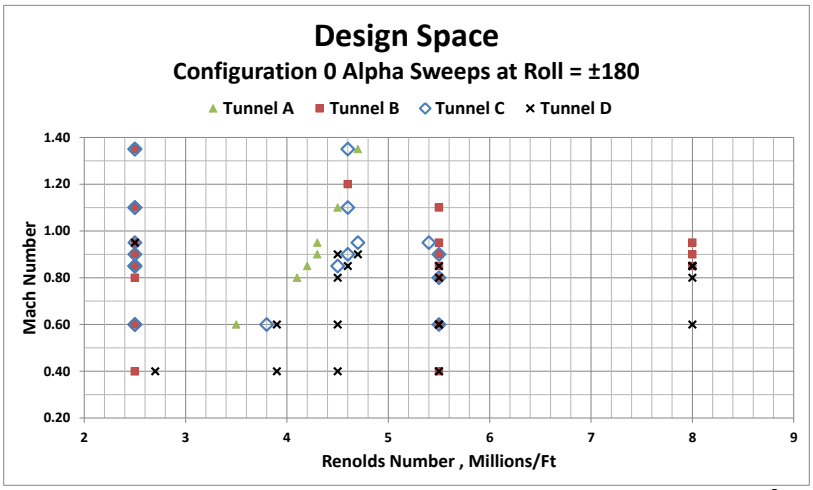

Figure 5: Design space for Configuration 0 alpha sweeps at Roll = $\pm 180^0$

III. Analysis Method and Representative Results

This section focuses on the FAVOR data analysis methodology. The general framework for the analysis is then described, highlighting the role of unexplained variance in the analysis.

A. General Analysis Framework

Every data set, including the FAVOR data, features the property of variance, owing to inevitable differences among individual data points. This would be due to experimental error if nothing else, but usually the greatest contributor to variance is the change made intentionally to levels of the various independent variables, or factors, of the experiment from one point to the next. We rely upon these changes to learn something from the test; obviously if the response variables did not change from one point to the next, nothing could be learned about the test article. We say that such purposely induced variance is *explained* by the changes known to have been made in the various factor levels. All of the information obtained about the test article is contained within this explained variance, and in a perfect world, 100% of the variance in a sample of experimental data would be intentionally induced and explained by factor changes.

Unfortunately, after explaining all of the variance that was intentionally induced by factor changes, there always remains some residual *unexplained* variance. That is, the total variance of the data sample always exceeds the explained variance by some amount. The unexplained variance causes uncertainty in the experimental result; we are unable to attribute it to any of the changes that we intentionally made to the system under study. The implication is that other agents besides us also influence the system response we are measuring, so that any response change we observe from one data point to the next is the algebraic sum of effects we can explain and effects that we cannot explain. There is uncertainty in the experimental result because we are unable to precisely allocate observed response changes among these explainable and unexplainable causes.

The aerodynamicist is primarily concerned with *explained* variance. He seeks to understand how changes in independent variables cause changes in the forces, moments, and pressures that interest him. The facility engineer, however, is more concerned with *unexplained* variance, since this is what impacts the quality of information produced in his facility. In either case, a useful start to the analysis is to quantify the total variance in a given data sample, and then to partition that variance in to explained and unexplained components.

B. Analysis of Variance

Consider the data sample in Table 1. This sample represents eight Tunnel-D yawing moment polars at Mach 0.85 and Reynolds number per foot of 4.5E06, ostensibly identical except for the dates and times they were acquired. The first six polars were acquired within a two-hour interval, and the last two were acquired four days later, within a half hour of each other. We ask if there is any significant difference between one polar and another in this data sample. That is, we wish to know if the polars change with time by more than can be attributed to ordinary random error.

Table 1: Replicated drag polars for Tunnel D, Configuration 0, Roll = 0^0, Re 4.5E06, Mach 0.85

AoA	CD: 1CDS_SLC							
Set-Point	Run 1	Run 2	Run 3	Run 4	Run 5	Run 6	Run 7	Run 8
-2.8	0.0354	0.0353	0.0355	0.0353	0.0354	0.0351	0.0352	0.0350
-1.8	0.0303	0.0304	0.0303	0.0304	0.0304	0.0303	0.0302	0.0303
-0.8	0.0289	0.0289	0.0289	0.0289	0.0290	0.0289	0.0288	0.0288
0.2	0.0300	0.0299	0.0299	0.0299	0.0299	0.0298	0.0299	0.0298
1.2	0.0337	0.0337	0.0336	0.0337	0.0336	0.0337	0.0337	0.0338
2.2	0.0412	0.0415	0.0412	0.0411	0.0410	0.0410	0.0414	0.0411
3.2	0.0526	0.0528	0.0524	0.0526	0.0527	0.0526	0.0528	0.0526

We are generally unable to answer this question by simply examining the drag polars graphically, as in Fig. 6. This is because the dynamic range of the data sample is large compared to differences in drag coefficient that would be large enough to be of concern. For example, these data span approximately 250 drag counts, while polars typically have to repeat within something on the order of one drag count or less to meet typical wind tunnel precision requirements for drag. It would be quite difficult to detect such small differences visually by examining graphed data as in Fig. 6.

Other graphical methods are available that yield clearer insights into the reproducibility of replicated polars. For example, the mean drag coefficient can be computed for each angle of attack, and departures from that mean can then be graphed as a function of angle of attack to remove dynamic range effects. However, graphical analysis can be rather subjective, and while the plotting of data often offers interesting insights and will always be a crucial element of any examination of experimental data, a more quantitative and objective method to detect significant differences among replicated polars would have some advantages. The analysis of variance (ANOVA) is such a method.

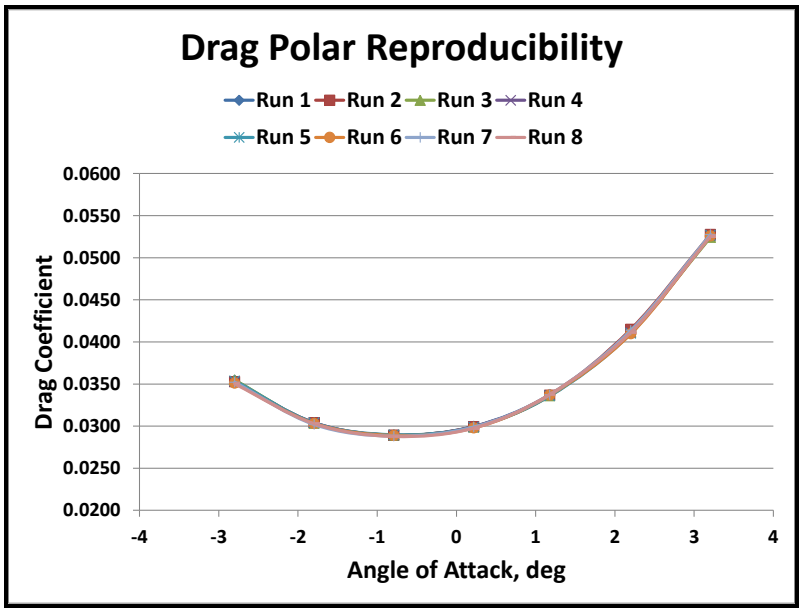

Figure 6: Drag Polar Replicates: Tunnel D, Configuration 0, Roll = 0^0, Re 4.5E06, Mach 0.85

ANOVA has been used extensively in the analysis of the FAVOR data as reported in this paper. ANOVA methods are more commonly used outside the experimental aeronautics community than within it, but the basic concept is easy enough to for any wind tunnel practitioner unfamiliar with this technique to implement with a relatively mild learning curve. The reader is referred to standard textbooks for the computational details of ANOVA[5], and Ref. 6 describes specific applications to wind tunnel testing; but the drag data from Table 1 can be used to illustrate the general concept.

There are 56 values of drag coefficient presented in seven rows and eight columns of Table 1. We can compute the total variance of these 56 numbers in the usual way, by summing the squared deviations of each number from the mean of all 56 numbers, and dividing this so-called "sum of squares" by the smallest number of points needed to compute the sum of squares, given the mean. This latter number, called the degrees of freedom, is just $n - 1$ for an n-point sample, because we can always multiply the sample mean by the number of points to compute the sum of all n numbers, and if we have n – 1 of them we can obtain the n^{th} point by subtraction. The total variance computed by dividing the sum of squares by the degrees of freedom is just the square of the standard deviation.

We now partition the total variance into three components, one associated with row-wise variation, one associated with column-wise variation, and a third component representing whatever is left over, which we attribute to ordinary random error. The variance components are computed analogously to the calculation of total variance. That is, sums of squares and degrees of freedom are computed for each component, and the variance is computed by taking the ratio.

The sums of squares for row-wise variance is computed by squaring the difference between each row mean and the average of all row means, and then adding all of those squared differences. Likewise, the sums of squares for column-wise variance is computed by squaring the difference between each column (polar) mean and the average of all column means, and then adding all of those squared differences. Note that the average of all column means is the same as the average of all row means, which is just the average of all of the data in the table (the Grand Mean).

The row-wise degrees of freedom is just $r - 1$ for a table with r rows, for reasons explained earlier in the discussion of total variance. There are thus six row-wise degrees of freedom for the data in Table 1. Likewise, there are $c - 1$ column-wise degrees of freedom for a table with c columns, and the 8-column Table 1 therefore has seven column-wise degrees of freedom.

The sum of squares for the left over component of total variance that is neither row-wise nor column-wise, and which is therefore attributed to random error, is computed by subtracting the column-wise and row-wise sums of squares from the total sum of squares. The random error degrees of freedom can be computed the same way. That is, for *r* rows and *c* columns, we have

$$\begin{aligned} df_{Error} &= (rc-1)-(r-1)-(c-1) \\ &= rc-1-r+1-c+1 \\ &= rc-r-c+1 \\ &= (r-1)\times(c-1) \end{aligned} \tag{1}$$

So the random error degrees of freedom is just the product of the row and column degrees of freedom. For an array such as Table 1, consisting of eight 7-point polars, there are seven column degrees of freedom, six row degrees of freedom, and 55 total degrees of freedom. There are thus 55 – (6+7) = 42 random error degrees of freedom or, equivalently, 6 × 7 = 42 degrees of freedom available to assess random error.

The sum of squares depends on how many numbers are used in the calculation; each new number adds to the sum of squares, which can increase with sample size without bound. In order to make meaningful comparisons of the column-wise and error variances, each computed with different amounts of information, it is necessary to normalize them to each other. We can accomplish this by multiplying the column-wise sum of squares by the number of rows. This adjustment completes the calculation of column-wise sums of squares, which are then divided by the *c – 1* column-wise degrees of freedom to produce the column-wise variance. Likewise, the random error sum of squares, computed by subtracting the row and column sums of squares from the total sum of squares, is then divided by the random error degrees of freedom. The result is the random error component of variance. Since variance is a kind of "average sum of squares" defined as the average squared departure from the mean per degree of freedom, it is often called the Mean Square, and abbreviated MS.

We are now in a position to infer whether the column-wise component of variance, attributable to changes in the polar means with time, is greater than the random error variance. We simply divide the column-wise MS by the error MS. This ratio is called the F-statistic to honor Ronald Fisher, who introduced the analysis of variance almost 100 years ago. The F-statistic is the ratio of two variance estimates, each based on a finite and often relatively small number of random variables. It is also a random variable, and because of ordinary chance variations that can occur in a finite sample of data, the F-statistic takes on a distribution of values as both the numerator and denominator wax and wane with such variations, even under the null hypothesis of no true difference between column-wise and random error. Just as with any other random variable, while theoretically the F statistic can take on a range of values, some are more likely than others. There is thus a probability distribution associated with the F statistic, and under the null hypothesis that no significant difference truly exists between column-wise and random variance, we expect F values greater than 1 to be progressively less and less likely the larger F is. This results in a probability density function for F that is skewed rather heavily to the right.

There is actually a family of F distributions, with the precise shape of each depending on the number of numerator and denominator degrees of freedom. Fig. 7, which corresponds to the seven column degrees of freedom of the numerator and 42 error degrees of freedom of the denominator of the F statistic for the data of Table 1, displays the general behavior of the F distribution. This distribution assumes a null hypothesis that no true difference exists between column-wise and random error variance (all polar replicates drawn from the same population of measurements, with no systematic difference from polar to polar).

The vertical red line in Fig. 7 marks the location of what is called the critical F value, Fcrit. This location is defined by the area under the F distribution to the right of Fcrit, customarily designated α, which is 0.05 in this figure. We will compare our computed F value with Fcrit in order to objectively infer whether the null hypothesis should be rejected. If F > Fcrit, we will reject the null hypothesis and infer that there is some systematic difference between one or more of the polars and all the rest of them. In that case the probability of an inference error due to ordinary chance variations in the data will be no greater than α, and we will conclude with 100 × (1 – α)% confidence that the variance from column to column is too great to be explained by ordinary random experimental error. If, on the other hand, the F value for our data lies to the left of Fcrit, we will be unable to reject the null hypothesis at what is formally described as the α level of significance, and we will conclude that the data do not support with at least 100 × (1-α)% confidence an inference that systematic differences exit from one or more polars to the others in this data sample.

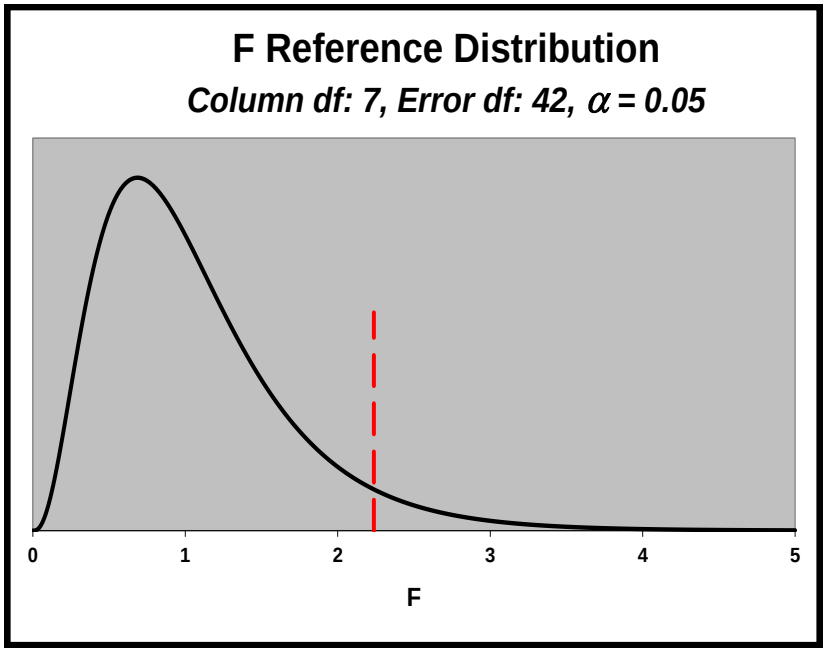

Figure 7: F Distribution for ANOVA on array of eight 7-point polars arranged in columns. Red line is critical F corresponding to $\alpha = 0.05$.

By now it should be clear that the analysis of variance entails a certain amount of bookkeeping. We have four sources of variation in the form of ANOVA we are presently discussing; the total variance and components associated with rows, with columns, and with random error. For each of these we must compute a sum of squares, the degrees of freedom, and their ratio, which is the variance or Mean Square. F statistics are then computed for row and column variance components by expressing their variances as a multiple of the random error variance. In this application the row-wise F is already known to be large because of the influence of angle of attack changes from row to row, but the column-wise variance relative to random error is of considerable interest. Critical F values are computed for a specified α based on the number of numerator and denominator degrees of freedom in the F statistic. The null hypothesis of no significant differences among polar replicates is then either rejected, with a probability of no more than $(100 \times \alpha)$% of an inference error, or it is not, depending on the relative size of F and Fcrit. In practice, the rather tedious calculations of ANOVA are automated in many commercially available software packages.

The ANOVA table is a standard structure for supporting the bookkeeping of an analysis of variance. Table 2 displays the ANOVA table for the eight ostensibly identical drag polars in Table 1. This table was created by using the "two-way ANOVA without replication" data analysis add-in of Excel.

Table 2: ANOVA table for eight ostensibly identical drag polars acquired in Tunnel D with Configuration 0, Roll = 0^0, Re 4.5E06, Mach 0.85.

Source	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Rows	0.003435	6	5.73E-04	51648	1.5E-79	2.324
Columns	1.45E-07	7	2.06E-08	1.863	0.1004	2.237
Error	4.66E-07	42	1.11E-08			
Total	0.003436	55				

The first column of the ANOVA table identifies the source of variation, and the second and third columns display their sums of squares and degrees of freedom. The fourth column, labeled MS, records the ratio of values in the SS and df columns to its left. Entries in the F column are computed by simply dividing the corresponding Mean Square by the error Mean Square. This calculation reveals that the row-wise component of variance is 51,648 times larger than the variance attributable to random error. The quantity to the right of each F value, called the P-value,

simply displays the probability that the corresponding F value would occur by chance because of ordinary random fluctuations in the data, if the null hypothesis is actually true. It represents the area under the F distribution of Fig. 7 to the right of F = 51,648, an incomprehensively small number. We are entitled therefore to make the rather unremarkable inference that drag coefficients change with angle of attack, and to do so with a vanishingly small probability of an inference error.

Aerodynamicists will no doubt derive considerable comfort from the fact that statisticians can so unambiguously confirm that drag depends on angle of attack, although this inference might not have entirely escaped their attention even without the benefit of a formal analysis of variance. While the existence of such significant row-wise variation is obviously of little interest, the column-wise variation in a sample of data such as displayed in Table 1 is rather more important. Column-wise variance that significantly exceeds random error variance would mean at the very least that uncertainty estimates based only on random error are understated. Unfortunately, it also means rather more than that. Significant variation among ostensibly identical polars means that responses such as lift and drag change with *time*, as well as with such factors as angle of attack and Mach number. Obviously time-dependent forces and moments would wreak havoc with any effort to construct an aerodynamic database by simply acquiring unreplicated data points at different times, under the assumption that all measured response differences are attributable exclusively to specified changes in the independent variables.

The ANOVA table for the drag data of Table 1 displays a column-wise F value of 1.863. This is greater than 1, to be sure, and may suggest that the true column-wise variance does indeed exceed the random error variance. However, because this value is not greater than the critical F value of 2.237, we cannot reject the null hypothesis with at least 95% confidence, which is our criterion.

We note in passing the importance of establishing such a criterion *before* the data are analyzed, to avoid succumbing to what might be colloquially described as "Shifting Goalpost Disease," in which personal preferences incline us to tune the criterion to be more in harmony with our a-priori expectations. For example, the p-value for column-wise variation of 0.1004 would entitle us to reject the null hypothesis at nominally the 0.10 significance level, concluding with nearly 90% confidence that there are in fact significant systematic differences among two or more of the polars we examined. If we had a personal preference for establishing the existence of such systematic differences among ostensibly identical polars, and if the 0.05 significance level had not been formally declared before we began the analysis, there might be some temptation to simply change the criterion on the fly. The obvious implications for research integrity require no elaboration. Suffice it to say that the objective of an analysis of variance is to *find out* if there are systematic differences among ostensibly identical polars, not to *prove* that there are or that there are not.

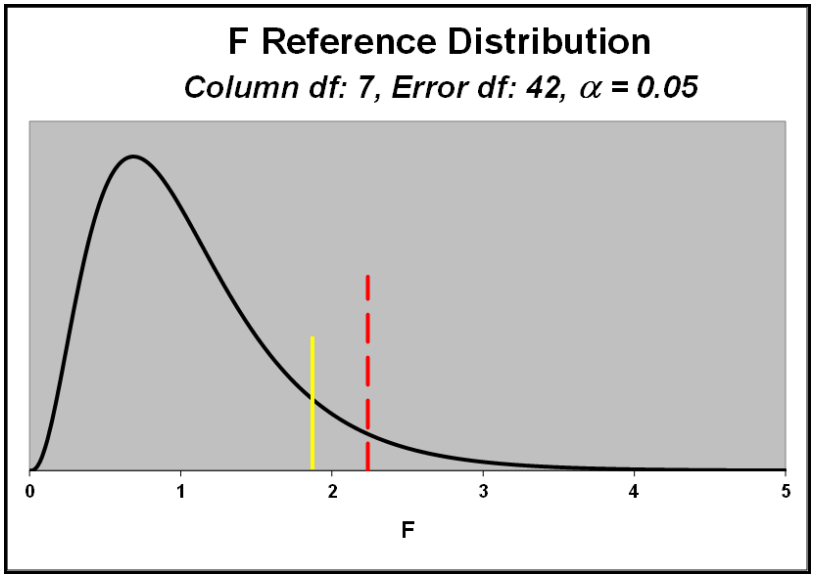

Figure 8: Measured F (yellow) < Fcrit(α=0.05) for Tunnel D drag polar replicates at Configuration 0, Roll = 0^0, Re 4.5E06, Mach 0.85.

Figure 8 superimposes on the reference F distribution of Fig. 7 the F value that was computed in Table 2. Since it is to the left of Fcrit, we are unable to reject the null hypothesis, and therefore we conclude that the data of Table 1 do not support an inference that drag polars changed with time over the four days in which these data were acquired.

Figure 9 reinforces this conclusion. The mean values of drag polars displayed in Table 1 are plotted as a function of the date and time that acquisition of the polar began. Major vertical grid lines denote midnight while the minor vertical gridlines delineate eight-hour intervals (so 8:00 AM and 4:00 PM). The first six polars were acquired within a two-hour interval, and the last two were acquired four days later, within a half hour of each other. The sample means are well within ±1 drag count, which is not significantly greater than the standard deviation for irreducible random error.

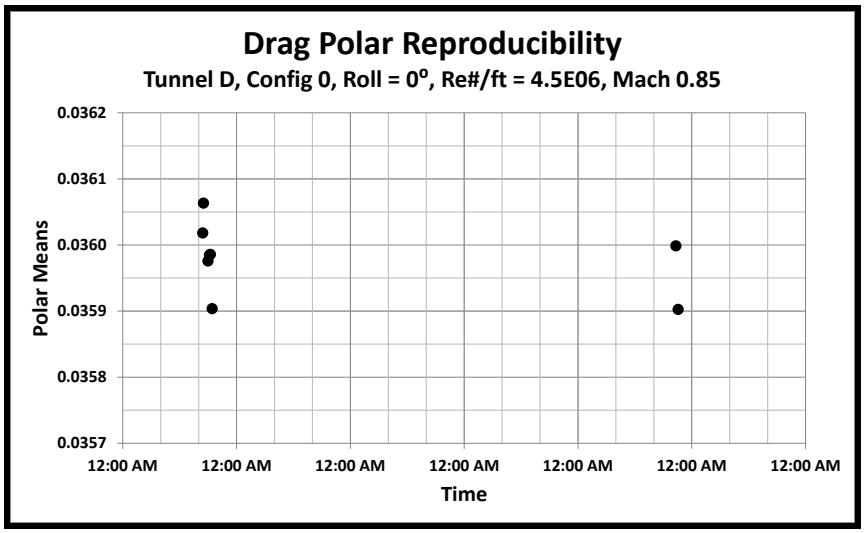

Figure 9: Time history of polar means for Tunnel D drag polar replicates at Configuration 0, Roll = 0^0, Re 4.5E06, Mach 0.85.

Now consider the data sample in Table 3. This sample presents yawing moment data for the same polars displayed in Table 1; namely, for eight Tunnel-D polars at Configuration 0, Mach 0.85, and Reynolds number per foot of 4.5E06, ostensibly identical except for the dates and times they were acquired. We ask the same question for the yawing moment polars as for the drag polars: Is there any systematic variation from polar to polar that is too large to attribute to ordinary random variations in the data?

Table 3: Replicated yawing moment polars for Tunnel D, Configuration 0, Roll = 0^0, Re 4.5E06, Mach 0.85.

AoA Set-Point	CMZ: 1CYMS_SLC							
	Run 1	Run 2	Run 3	Run 4	Run 5	Run 6	Run 7	Run 8
-2.8	-0.00003	-0.00002	0.00003	0.00003	0.00004	0.00003	-0.00002	-0.00002
-1.8	0.00000	-0.00001	0.00004	0.00005	0.00006	0.00006	0.00001	0.00001
-0.8	0.00002	0.00003	0.00008	0.00007	0.00007	0.00008	0.00004	0.00005
0.2	0.00004	0.00004	0.00007	0.00009	0.00010	0.00009	0.00006	0.00004
1.2	0.00003	0.00001	0.00006	0.00006	0.00006	0.00005	0.00003	0.00004
2.2	0.00000	0.00001	0.00002	0.00004	0.00002	0.00004	-0.00001	0.00000
3.2	-0.00008	-0.00007	-0.00006	-0.00006	-0.00004	-0.00005	-0.00010	-0.00010

These yawing moment polars are plotted in Fig. 10. They all display similar behavior over the angle of attack range tested, and while no two polars overlap perfectly, there is no clear indication that one or more polars are drawn from a significantly different population than the rest.

An analysis of variance performed on the yawing moment data yields a result that might not have been entirely unanticipated from a visual inspection of the polar plots of Fig. 10. The ANOVA results are shown in Table 4.

Both the row-wise and column-wise variance components were unexpected. The row-wise F of 194.5 is much greater than its critical value of 2.324, and the tiny level of the corresponding P-value makes it unambiguously clear that the yawing moment varied with angle of attack. This is in fact clear from the polar plots in Fig. 10. It is a mildly

surprising result because the configuration was symmetrical and the sideslip angle was nominally zero. There seems to be some coupling into angle of attack that suggests an unanticipated lateral asymmetry.

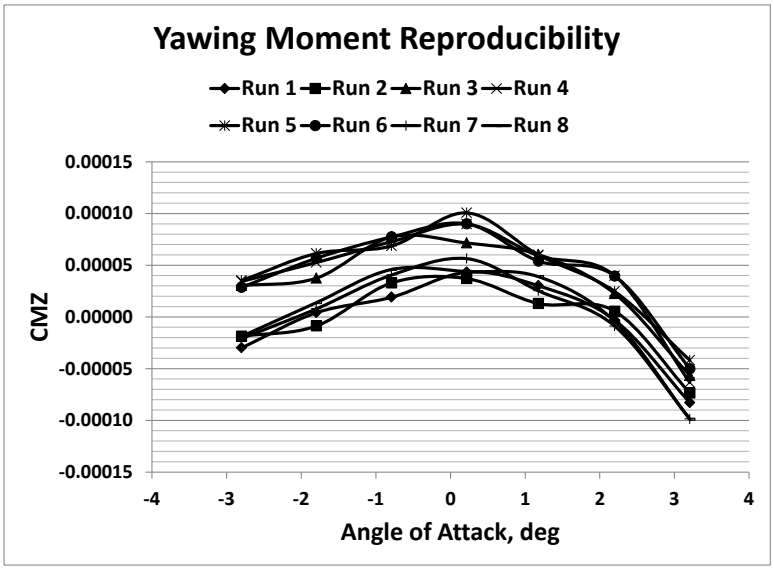

Figure 10: Yawing Moment Polar Replicates: Tunnel D, Configuration 0, Roll = 0^0, Re 4.5E06, Mach 0.85.

Table 4: ANOVA table for eight ostensibly identical yawing moment polars acquired in Tunnel D with Configuration 0, Roll = 0^0, Re 4.5E06, Mach 0.85.

Source	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Rows	9.94E-08	6	1.66E-08	194.5	5.41E-29	2.324
Columns	2.38E-08	7	3.4E-09	39.8	1.51E-16	2.237
Error	3.58E-09	42	8.52E-11			
Total	1.27E-07	55				

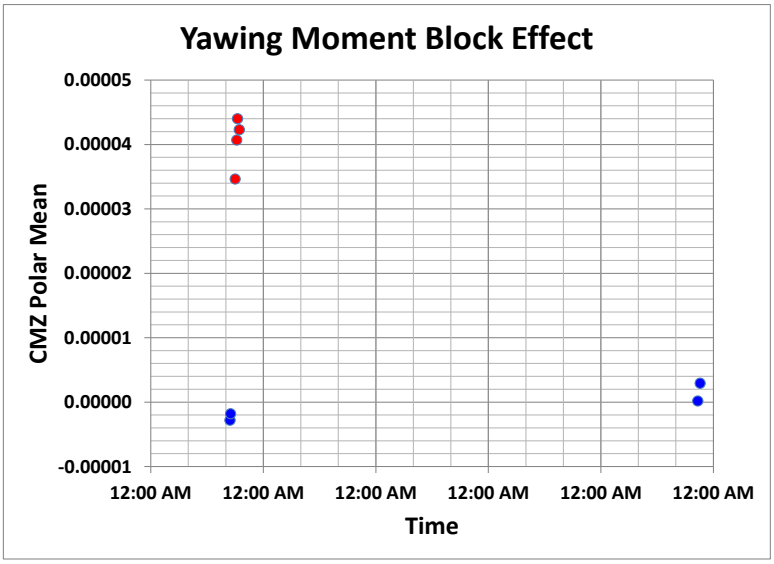

Figure 11: Four-day time history of polar means for Tunnel D yawing moment polar replicates at Configuration 0, Roll = 0^0, Re 4.5E06, Mach 0.85.

Of greater interest in the present analysis is the relatively large F for column-wise variance, indicating significant differences among two or more yawing moment polars, notwithstanding the fact that no systematic variation was observed in drag polars acquired with the same alpha sweeps. Figure 11 displays the time history of yawing moment polar means for this set of ostensibly identical polars. This figure reveals three blocks of time between which the tunnel appears to transition from one state to another and then back again. On the first day, the first two polars were acquired between a nine-minute interval. After a delay of 57 minutes, the next four polars were acquired within a 53-minute interval. These polars are biased significantly higher than the first two polars. Four days later, the final two polars were acquired, again within a nine-minute interval. They seem to have been drawn from the same population as the first two polars.

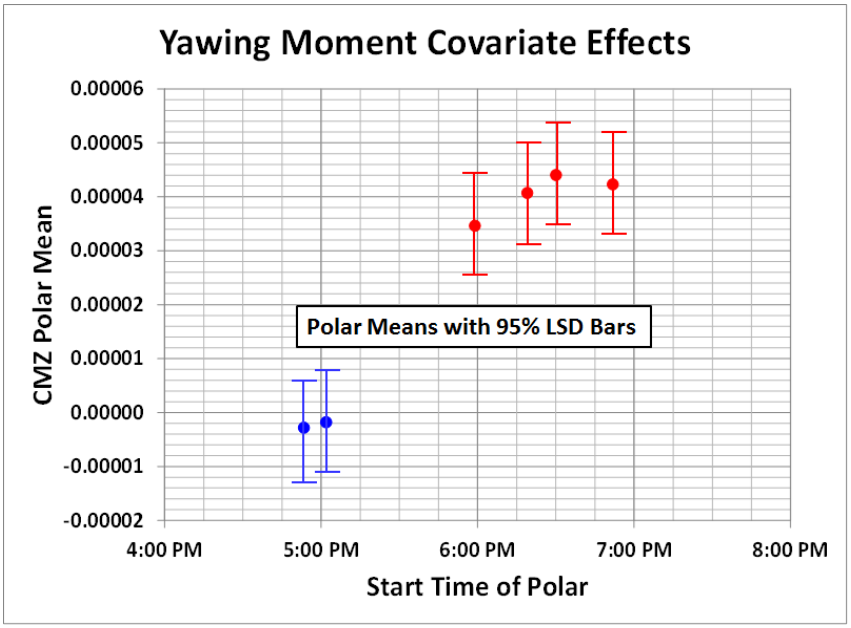

Figure 12: Two-hour time history of polar means for Tunnel D yawing moment polar replicates at Configuration 0, Roll = 0^0, Re 4.5E06, Mach 0.85.

Figure 12 is a time history of the six polars acquired on the first day of this data sample, providing greater time resolution. The polar means feature 95% Least Significant Difference bars that indicate ranges within which distinctions in yawing moment estimates cannot be made with at least 95% confidence. Clearly, means of the last four yawing moment polars, acquired roughly between 6:00 PM and 7:00 PM, are shifted with respect to the means of the first two polars, acquired around 5:00 PM. This shift is sufficiently great to be resolved unambiguously (i.e., with more than 95% confidence). The fact that more than one similar polar mean is found in each block suggests that none of the points are simple outliers and that the bias shift between groups of polars is real.

Figure 13 reproduces Fig. 10 but with the each polar group highlighted. The shift between blue and red polar groups is more obvious in this figure, but the quantitative ANOVA results are considerably less ambiguous. This illustrates the common case, which is that objective results available from a quantitative analysis of variance tend to lead to relatively less ambiguous inferences than can be typically achieved by subjective graphical means.

The ANOVA results indicate that some systematic change occurred between the time one group of polars was acquired and the time the next group was acquired, but the results provide no direct indication of what caused the change. Such bias shifts are evidence of what are called covariate effects. Covariates are factors that over time cause systematic (not random) changes in measured responses such as forces and moments, but are not controlled by the experimenter. We may assume that covariates generating unexplained systematic variance during a wind tunnel test are as varied as the error sources that cause random unexplained variance, but some examples are temperature effects, instrument drift, strain gage desiccation effects, flow angularity changes, trip-dot and grit ablation, operator learning and fatigue effects, systematic set-point errors, sting creep, component wear, and changing wall effects due to thermal expansion. There may be any number of other effects in play, including effects that are unknown but that nonetheless have their impact on measurements. The FAVOR database provided an unprecedented opportunity to quantify the magnitude and frequency of occurrence of covariate effects, and to generate objective evidence either of their importance or of their irrelevance in wind tunnel testing. This is a major focus of the present analysis.

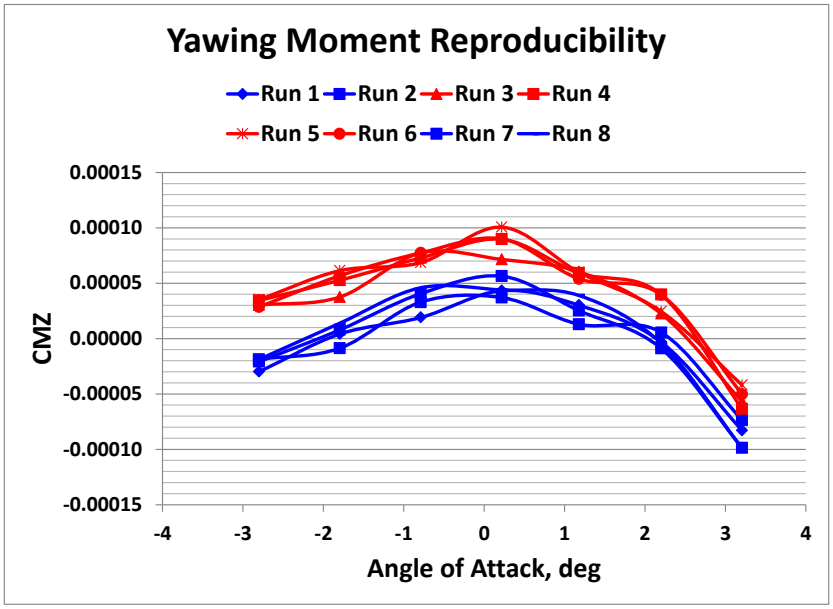

Figure 13: Grouped Yawing Moment Polar Replicates: Tunnel D, Configuration 0,

C. More from the ANOVA Table

Results from an analysis of variance are organized in an ANOVA table that reveals the significance of variance components with respect to ordinary random error variance. This is the primary function of an ANOVA. However, rather more information than this can be gleaned from the ANOVA table, as will be described here.

1. R-squared: The sums of squares column can be used to estimate how much of the total variation in the data is allocated to each of the components, since the sums of squares of all components add up to the total sum of squares. Therefore the ratio of any component SS to the total is an indication of how much variation is associated with that component.

Consider the ratio of the row-wise sum of squares to the total sum of squares. Since in this study row-wise variations are attributed to changes we induced intentionally by changing angle of attack, we say that we can *explain* this variance; unlike the column-wise variance and the random error variance that are due to unknown sources. The ratio of this explained sum of squares to the total sum of squares is called the R-squared statistic, and it would be exactly 1.0000 in a perfect world (that is, a world devoid of unexplained variance).

The R-squared value for the ANOVA table corresponding to the drag data analyzed above, Table 2, is 0.9998, which means that by this metric 99.98% of all the variation in this data sample can be explained by the angle of attack changes that were intentionally made. This also means that all of the uncertainty in this sample of drag polars is attributed to the 0.02 percent of the variations that cannot be explained.

The R-squared value computed the same way for the yawing moment data analyzed above is only 0.7843. This suggests that rather less of the total variation in this sample of yawing data can be explained compared to the drag data, which might at first seem disappointing. Actually, the yawing moment R-squared is not remarkable because of how *small* it is, but because of how *large* it is.

These data were acquired on a symmetrical configuration with nominally zero sideslip. Under such circumstances there is little reason to expect a lateral response variable such as yawing moment to change over a range of $\pm 3^0$ in angle of attack, and in that case the fraction of the total variation in the data that could be explained by changes in angle of attack should be zero. The fact that the R-squared value is *not* zero suggests that the model may not have exhibited perfect lateral symmetry, or that some set-point error in sideslip may have resulted in a sideslip different from zero. It is also possible that there might be some interaction effect in play in the control system that couples a little sideslip change into changes in angle of attack. Whatever the reason, a substantial fraction of the total variation in this sample of lateral-stability response data is attributable to angle of attack changes.

2. Standard Random Error: The error mean square from the ANOVA table is just σ^2 for random error, so the square root of this is the ordinary "one sigma" standard random error value. For the drag data analyzed above the error MS from Table 2 is 1.11E-08 and the square root of this is 0.00011, or 1.1 counts. This is a one-sigma estimate

of random error based on 42 degrees of freedom. For the yawing moment ANOVA summarized in Table 4, the error MS is 8.52E-11, for which the square root is 0.000009, or 0.09 counts.

These estimates should actually be regarded as upper limits on the random error in drag and in yawing moment because they include two other components of variation, one due to angle of attack set-point error and one due to row/column interactions, also known as "block/factor" interactions. The data were intentionally not corrected for AoA set-point error in order to facilitate a conservative evaluation of systematic errors. It is more difficult to detect small systematic differences from one polar to the next in the presence of random error that is inflated somewhat by AoA set-point variations, so any systematic between-polar differences that can be detected in this noisier environment are more likely to be real.

Block/factor interactions would exist in these data if a unit change in angle of attack caused a different change in response in one polar than another. A so-called "two-way ANOVA with replication" would permit block/factor interactions to be quantified, but such an analysis would require within-polar replicates, which were not acquired in the FAVOR study. A significant block/factor interaction would occur in this study if the same change in angle of attack caused different response changes in one polar than another. That is what it means to have a "block/factor interaction." Within-polar covariate effects could cause just such a phenomenon. For example, if some systematic effect was in play that caused pre-stall lift coefficients acquired later to be greater than pre-stall lift coefficients acquired earlier under ostensibly identical circumstances, and if angle of attack levels were set as a monotonically increasing function of time as is customary in OFAT testing, the pre-stall lift polar would be rotated counter-clockwise, featuring an artificially inflated slope. A unit increase in angle of attack would then result in a greater increase in lift for this polar than for one acquired under more stable conditions. A two-way ANOVA with replication conducted on a sample of polars that included these two could reveal significant block/factor interactions. Unfortunately, absent within-polar replication, any block/factor variance component that may exist is simply lumped in with the random error, resulting in slightly higher random error estimates than otherwise.

3. Time-varying Bias Error: The square root of the column-wise mean square from the ANOVA table is an estimate the standard error ("one sigma") for systematic time-varying bias errors. For the drag data analyzed above the column-wise error MS from Table 2 is 2.06E-08 and the square root of this is 0.00014, or 1.4 counts. This is a one-sigma estimate of the systematic component of the unexplained variance. Because it is based on a sample with eight polar replicates, there are seven degrees of freedom associated with this estimate. For the yawing moment ANOVA summarized in Table 4, the column-wise error MS is 3.40E-09, for which the square root is 0.00006, or 0.6 counts.

4. Total Uncertainty due to Unexplained Variance: The random error and the error due to systematic between-polar bias shifts can be combined to yield the total error due to unexplained variance following methods described in Coleman and Steele[7]. We will illustrate using the standard errors from the drag data we have been considering; namely, a 42 degree of freedom estimate of the random error component of the unexplained variance with a standard error of 0.00011 and a 7 degree of freedom estimate of the systematic component of the unexplained variance with a standard error of 0.00014.

Following the ISO *Guide*[8], the combined standard uncertainty due to unexplained variance would be expressed as follows.

$$u_c^2 = MS_{random} + MS_{systematic} \tag{2}$$

For the drag data of this example,

$$\begin{aligned} u_c^2 &= 1.11\times10^{-8} + 2.06\times10^{-8} = 3.17\times10^{-8} \\ u_c &= 0.00018 \end{aligned} \tag{3}$$

Coleman and Steele use the term "expanded uncertainty, $U_{\%}$" to describe the uncertainty with a specified level of confidence associated with it, which is obtained by multiplying u_c by a coverage factor drawn from the t distribution, $t_{\%}$, per ISO recommendations:

$$U_{\%} = t_{\%} u_c \tag{4}$$

The t statistic to be used as a coverage factor will depend on the level of confidence and also on a specified number of degrees of freedom. We might select 95% as a common level of confidence to specify, but we have two different values for the number of degrees of freedom, 42 and 7. These can be combined into a single value using the Welch-Satterthwaite formula[9,10]:

$$\nu = \frac{u_c^2}{\left(MS_{random}^2 / \nu_{random}\right) + \left(MS_{systematic}^2 / \nu_{systematic}\right)} \tag{5}$$

Inserting values from the drag example:

$$\nu = \frac{3.17\times10^{-8}}{\left[\left(1.11\times10^{-8}\right)^2 / 42\right] + \left[\left(2.06\times10^{-8}\right)^2 / 7\right]} = 15.8 \tag{6}$$

The t statistic for this pooled estimate of degrees of freedom at a significance of 0.05 (95% confidence) is 2.131. Multiplying this by the value computed above for u_c of 0.00018 yields a 95% prediction interval half width, $U_{\%}$, of 0.00038, or 3.8 drag counts. We would therefore expect the true value of drag to lie within an interval of ±3.8 counts of any individual drag measurement acquired in this tunnel under conditions similar to those for which these data were recorded. This interval is over 70% wider than the ±2.2 counts that would correspond to random error only. If this result is representative, it is unlikely that wind tunnel results will be reliably reproducible if uncertainty estimates ignore systematic between-polar shifts and depend only on random error estimates. We note in passing that this analysis is limited to the unexplained *variance* in a sample of data and does not include any consideration of classic static bias errors, which must also be taken into account to estimate the total uncertainty.

Before ending this discussion of methodology, we note that some polars were only replicated twice in the FAVOR data. The ANOVA methods described here are not necessary unless three or more polar replicates are to be analyzed. A simple paired t-test is sufficient when only two polar replicates are available.

The paired t-test is a common method of comparing two ostensibly identical samples of data and will not be described in detail here. The interested reader can consult any elementary statistics text for the details, but a brief overview is as follows: In our case responses from one of two polar replicates are subtracted from the other at each shared angle of attack. The result is a differential polar, consisting of response differences at each angle of attack. Under the null hypothesis of no difference in the two polars, each of these values would be zero, but ordinary random error ensures that even when there is no systematic difference between the polars, each difference will generally be non-zero except by coincidence.

The mean and the standard deviation of these differential response measurements are then computed. Again, under the null hypothesis we expect the true mean to be zero, although experimental error ensures that any specific differential mean based on a finite number of imperfect measurements will seldom be exactly that. To estimate the standard error in the mean, we divide the standard deviation by the square root of N, where N is the number of differential responses that were estimated. For the seven-point polars common in this study that spanned $\pm3^0$ in angle of attack, N = 7, for example.

A t-statistic is then formed by expressing the mean difference as a multiple of the standard error in estimating the mean. This is compared with a critical t statistic corresponding to a specified level of confidence. For reasonably large sample sizes, the critical t for 95% confidence is approximately 2, for example, so in that case any difference in the polar means that exceeds two standard deviations of the sample mean is judged to be sufficiently different from zero to claim a systematic difference in the polars with at least 95% confidence. If the polar means do not differ by at least two standard deviations, we say we are unable to reject the null hypothesis with 95% confidence, and we therefore claim no systematic difference between the polars.

All replicated polar pairs were subjected to this paired t-test for each force and moment, and the number of polar pairs that differed by more than could be attributed to random error was added to the number of samples involving three or more polars for which significant between-polar differences were detected.

Conclusions about the relative importance of random and systematic components of the unexplained variance based on the specific example presented here to illustrate the methodology must be predicated on some judgment as to how representative this example has been. Results are presented in the next section that were obtained by

applying the methods documented here to data consisting of all forces and moments acquired in all four FAVOR transonic tunnels, over a wide range of Mach and Reynolds number combinations.

IV. Cross-Facility Comparisons

This section addresses two specific questions. We wish to quantify how often there is a significant systematic component of the unexplained variance, and we would like to gauge how large it is typically, especially relative to the random error estimates that are more typically quantified when wind tunnel results are reported.

We begin be asking how much unexplained variance we encountered in the FAVOR test, and use the R-squared statistic introduced earlier as a metric. For longitudinal responses, Fig. 14 displays R-squared values for explained and total sums of squares pooled over all Mach/Re combinations for which polars were replicated within a given facility. This figure shoes that 99.9% of all the observed variance was explainable by the changes made in angle of attack, and in some cases the percentage exceeds 99.99%.

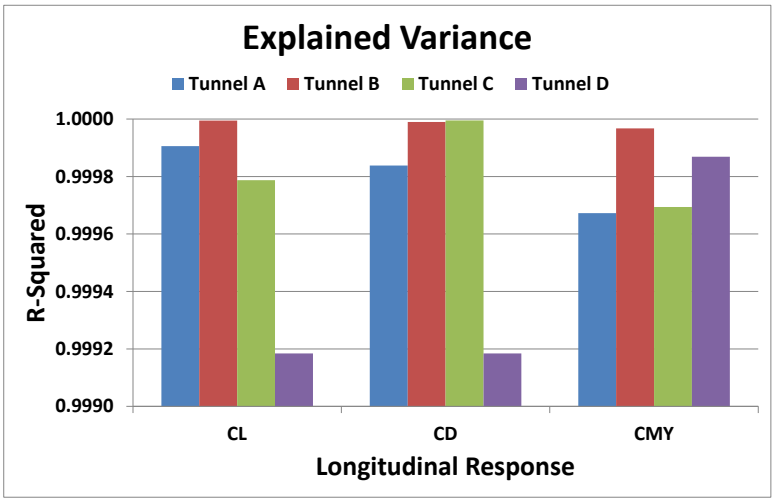

Figure 14: Explained variance for longitudinal responses, Configuration 0, Roll = 0, pooled over all replicated combinations of Mach and Reynolds number.

Figure 15 is a companion graph, for lateral/directional response variables. Rather less of the total variance is attributable to angle of attack changes in those cases, since they would not be expected to be a function of angle of attack at all absent some asymmetry or unintended sideslip set-point error, or possibly some coupling into angle of attack by the control system.

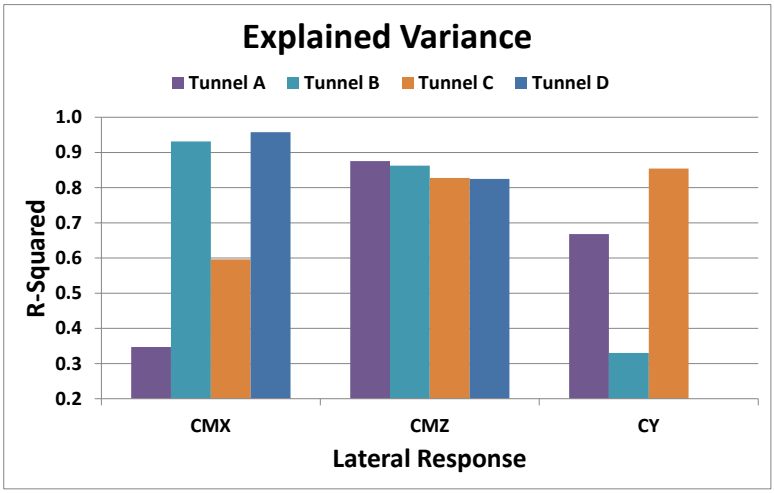

Figure 15: Explained variance for lateral/directional responses, Configuration 0, Roll = 0, pooled over all replicated combinations of Mach and Reynolds number.

Figure 16 displays what fraction of the unexplained variance is not random, but is due instead to systematic differences between polars. Note that for all responses at all facilities, it is well over half. That is, it was the general case that systematic unexplained variance exceeded random error.

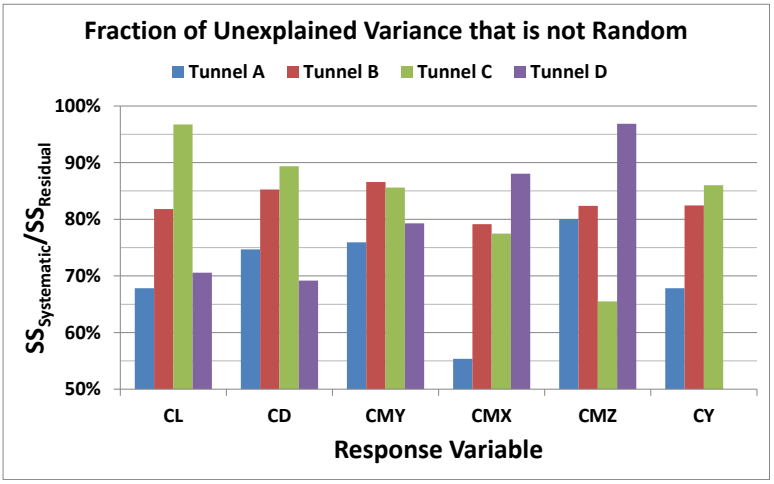

Figure 16: Fraction of unexplained variance that is systematic, not random, pooled over all replicated combinations of Mach and Reynolds number in all four tunnels.

In the previous section we developed an objective criterion for identifying a significant systematic component of the unexplained variance. We say the systematic unexplained variance is significant when a measured F value (ratio of systematic to random unexplained variance) exceeds a critical F value. The critical F is chosen so that the measured F will not exceed it more than 5% of the time due to an unlucky combination of random error fluctuations if the systematic unexplained variance is in fact insignificant.

Figure 17 indicates that significant systematic error is about as likely to occur with one force or moment as another, and that this happens too often to assume that it is a rare event.

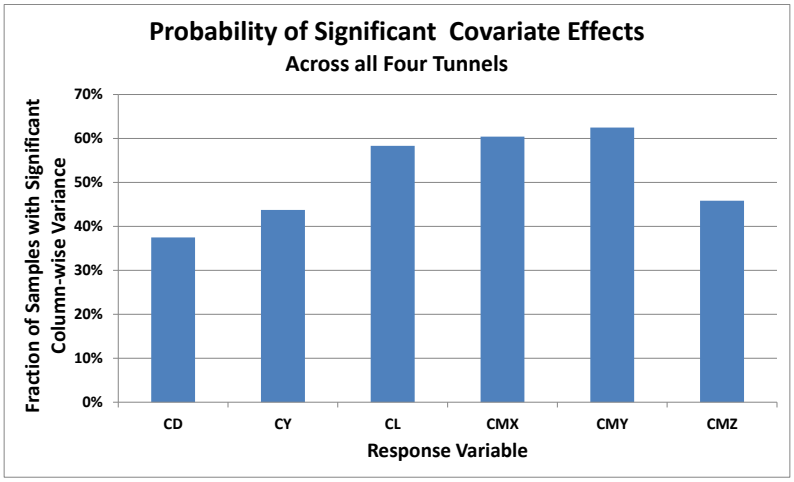

Figure 17: Frequency of occurrence of systematic unexplained variance large enough to detect with 95% confidence, over all replicated combinations of Mach and Reynolds number in all four tunnels.

Figure 18 presents the same information as Fig. 17, but for each of the participating FAVOR tunnels. Clearly the existence of systematic differences between ostensibly identical polars is not confined to a single tunnel. This phenomenon is common in all four FAVOR tunnels.

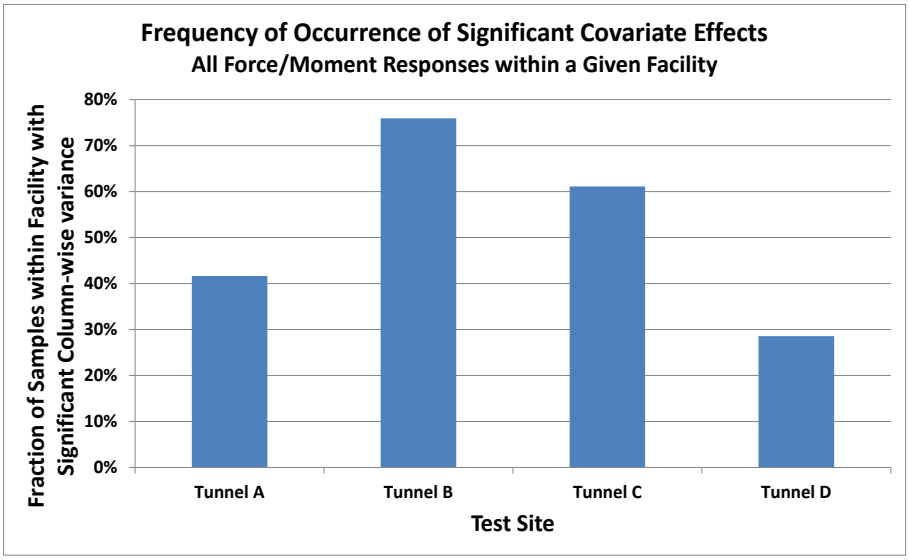

Figure 18: Frequency of occurrence of systematic unexplained variance large enough to detect with 95% confidence by tunnel, over all replicated combinations of Mach and Reynolds number and all forces and moments.

Figure 19 displays the overall frequency of occurrence of significant systematic shifts between polars assumed to be identical. Averaged over all facilities, all responses, and all sites tested in the design space, it appears as if systematic between-polar shifts occur roughly half the time.

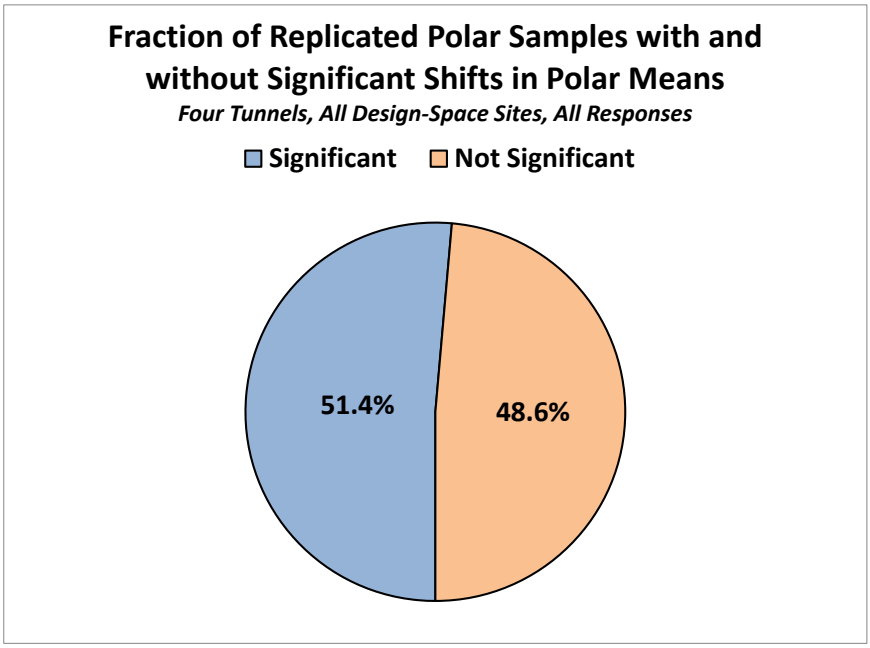

Figure 19: Frequency of occurrence of systematic unexplained variance large enough to detect with 95% confidence, over all replicated combinations of Mach and Reynolds number, all forces and moments, and all four tunnels.

Figures 20 through 25 display standard errors for the random, systematic, and combined error components for each of the six force/moment response variables. In all cases, the systematic error dominates the random error. For

practical purposes, ignoring the random error would have little effect on the total error due to unexplained variance, which is dominated by systematic between-polar shifts.

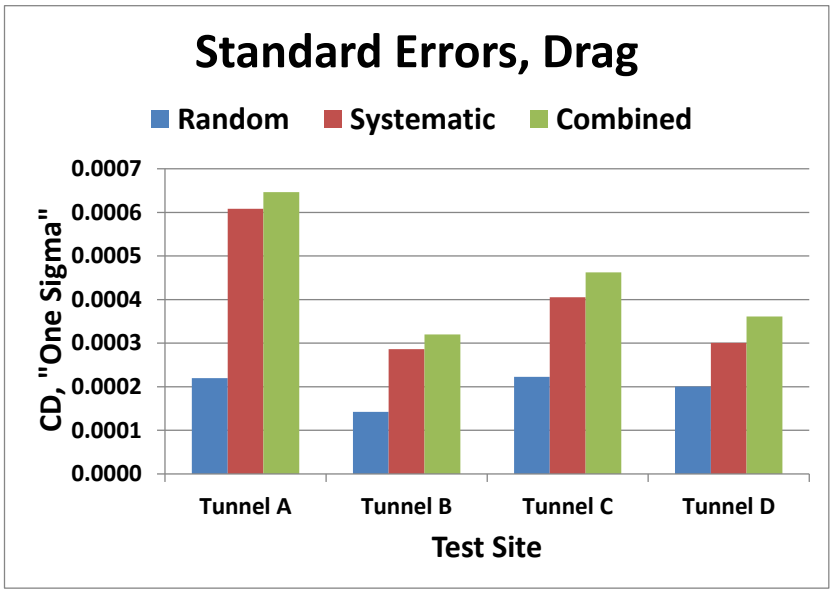

Figure 20: Relative magnitudes of random, systematic, and combined standard errors for drag, pooled over all replicated combinations of Mach and Reynolds number.

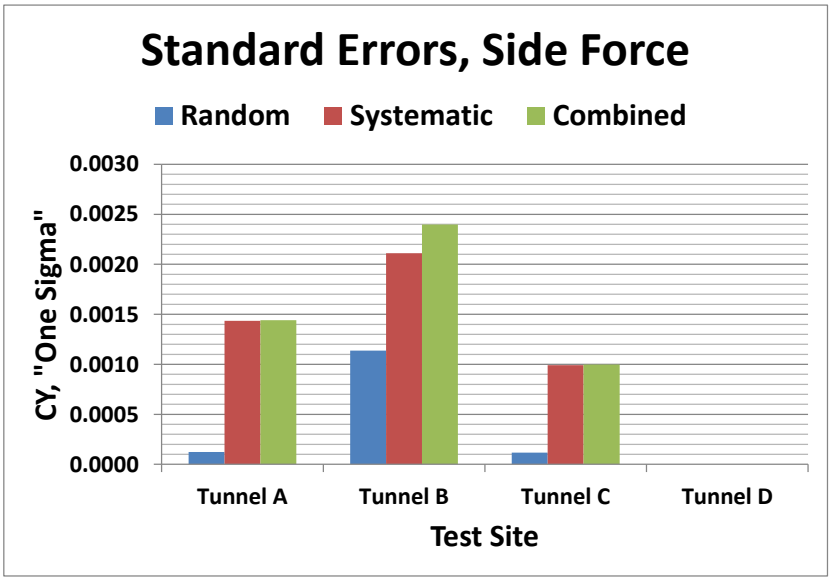

Figure 21: Relative magnitudes of random, systematic, and combined standard errors for side force, pooled over all replicated combinations of Mach and Reynolds number. No side force reported by Tunnel D.

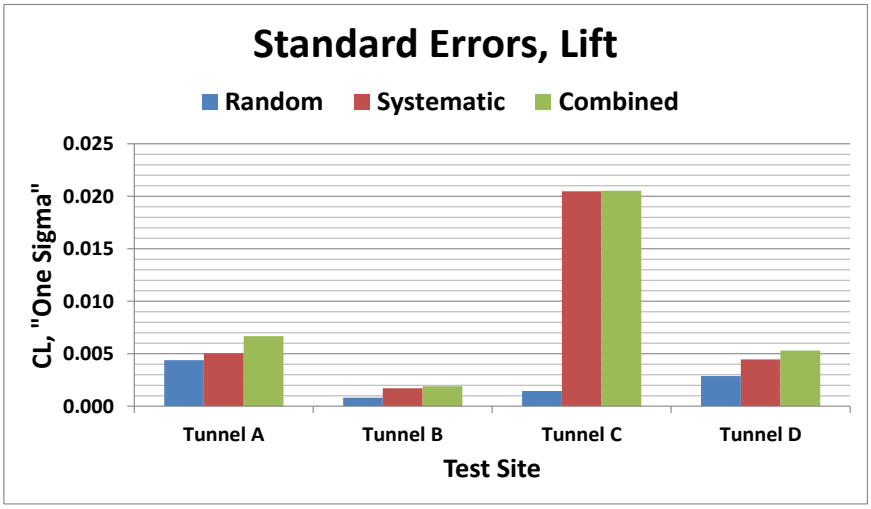

Figure 22: Relative magnitudes of random, systematic, and combined standard errors for lift, pooled over all replicated combinations of Mach and Reynolds number.

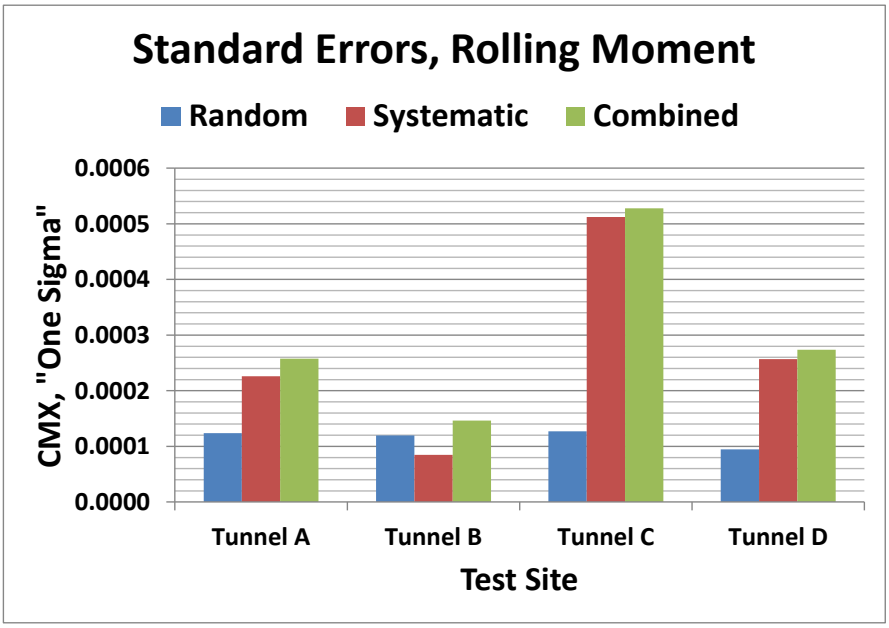

Figure 23: Relative magnitudes of random, systematic, and combined standard errors for rolling moment, pooled over all replicated combinations of Mach and Reynolds number.

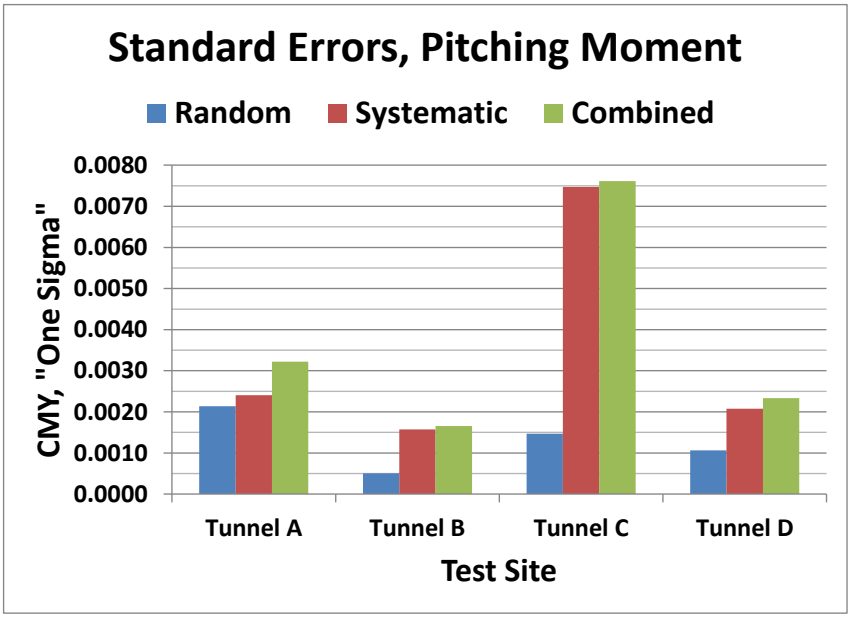

Figure 24: Relative magnitudes of random, systematic, and combined standard errors for pitching moment, pooled over all replicated combinations of Mach and Reynolds number.

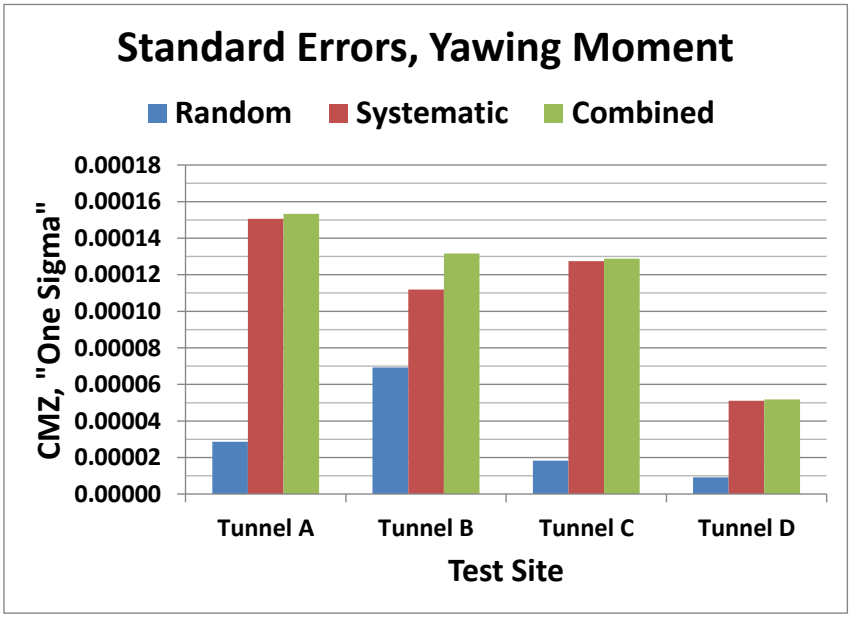

Figure 25: Relative magnitudes of random, systematic, and combined standard errors for yawing moment, pooled over all replicated combinations of Mach and Reynolds number.

Figure 26 displays the standard error for systematic unexplained variance as a multiple of the random standard error. For essentially all responses and all facilities, the systematic standard error exceeds the random standard error; often by factors of two or more and by as much as an order of magnitude in a couple cases.

Figure 27 reveals that most of the unexplained variance is systematic, a fact that has emerged in other figures as well.

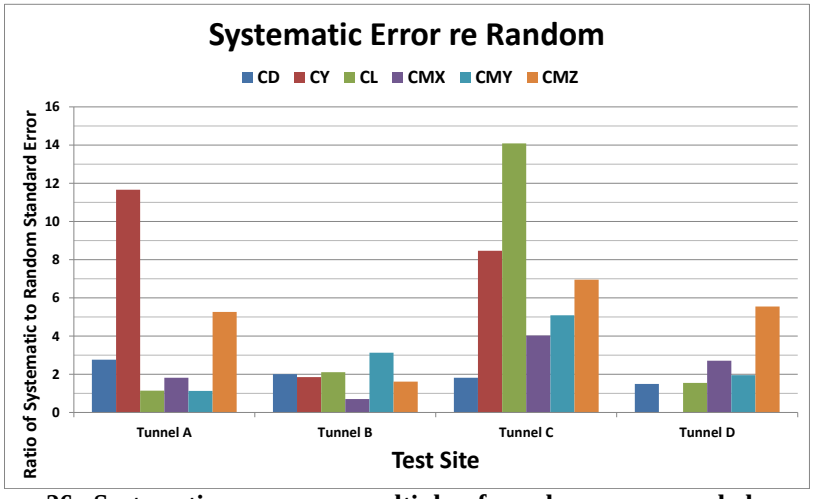

Figure 26: Systematic error as a multiple of random error, pooled over all replicated combinations of Mach and Reynolds number.

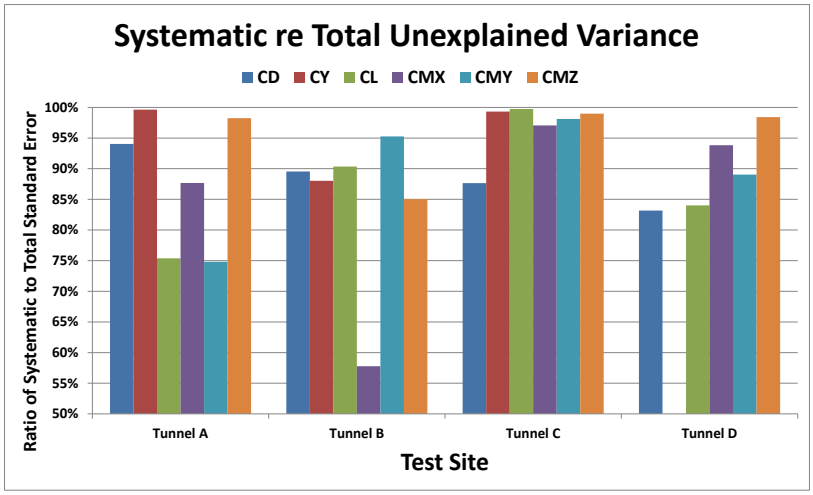

Figure 27: Systematic error as a fraction of total uncertainty due to unexplained variance, pooled over all replicated combinations of Mach and Reynolds number.

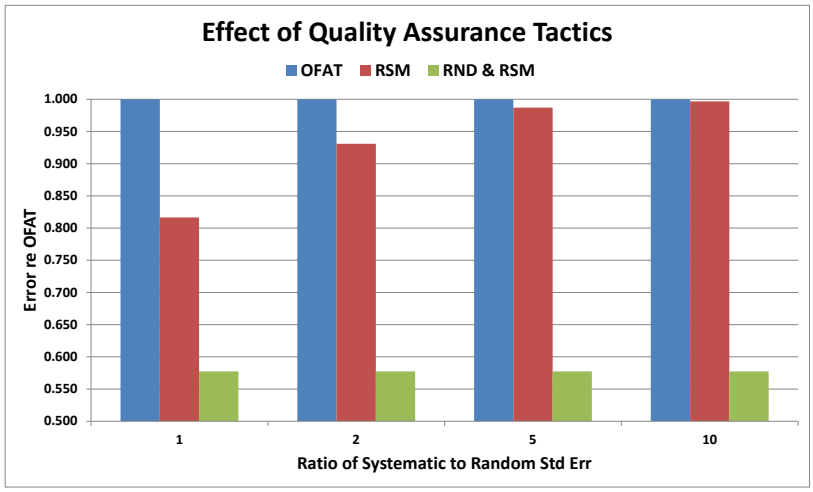

Figure 28: Error reduction due to various quality assurance tactics, for four different degrees of systematic unexplained variance.

Figure 28 shows how various quality assurance tactics can reduce the combined standard error due to random and systematic unexplained variance. The blue bars labeled "OFAT" at a value of "1" serve as a reference. This corresponds to the "one sigma" level of uncertainty if no quality assurance tactics are invoked. The bars labeled "RSM" show the impact of hidden replication on the random component of unexplained variance when response surface modeling methods are used to estimate responses from fitted data. It can be shown[11] that the standard prediction error for a polynomial response surface model of any order in any number of factors, averaged over all the points used to fit the model, is:

$$\sigma_{RSM} = \sigma\sqrt{\frac{p}{n}} \tag{7}$$

where p is the number of terms in the model (including the intercept) and n is the number of points used to fit the model. Since there must be at least one point for every term in the fitted model, the square root term will never be more than 1. If three points are acquired for every term in the model as required to achieve levels of precision that are widely deemed adequate, then $n = 3p$ and the square root term has a numerical value of 0.577. In that case, the standard error for model predictions is just over half the standard error for ordinary single-point measurements. The reason for the reduction in uncertainty is the "hidden replication" that occurs because residuals of a well-fitted model tend to fall with equal likelihood above or below the fitted response model predictions, thus partially canceling. This reduction has been applied to the random component of the total unexplained variance in all the "RSM" cases of Fig. 28. It cannot be applied to the systematic component of the total error and so the net benefit tends to be smaller for cases dominated by systematic error.

The purple bars show the effect of randomizing and also fitting the data. Both uncertainty components are then reduced by the 0.577 factor and so the total uncertainty is as well. Randomizing and fitting combined has the same percentage effect relative to the OFAT case no matter how much of the unexplained variance is systematic, because randomizing converts all error due to unexplained variance to random error. There is, however, a progressively greater benefit of both randomizing and fitting the data (essentially randomizing and replicating, through hidden replication) relative to fitting the data only, the more severe the systematic unexplained variance is. This is seen in the widening gap between the "RSM" and "RSM & RND" bars.

V. Concluding Remarks

The existence of a dominant systematic component of the unexplained variance is apparently common, and does not reflect negatively on a given facility. On the contrary, none of these tunnels seems likely to remain in a state in which polar means are time-independent indefinitely. For example, the total variance associated with a commercial jet transport drag coefficient polar is typically on the order of 10^{-4}. The standard error variance associated with a "two-sigma" error budget of a half drag count is on the order of 10^{-10}. The error budget for unexplained variance would therefore be parts per million of the total variance in this example, a level that is difficult to maintain consistently in a facility as complex and energetic as a modern wind tunnel for a typical tunnel entry durations of several weeks.

The FAVOR results suggest that is no more an indictment of a wind tunnel to exhibit a systematic component of the unexplained variance than it is to exhibit a random component, since both are apparently so common. On the contrary, any report of the complete absence of systematic unexplained variance should be regarded as at best a short-term result obtained under ideal circumstances that are unlikely to be maintained indefinitely, and viewed the same way as a report that all *random* error is also absent.

The existence of systematic error in wind tunnel facilities is not so much a problem as the traditional failure to understand and effectively cope with the impact that the corresponding lack of statistical independence among data points has on reproducibility. The true imperative is not to identify and eliminate all causes of systematic error, which is impractical, but rather to ensure that each measurement is statistically independent in spite of the inevitable systematic error that may be reliably assumed to exist. Methods to do so through the design of the test matrix and by such quality assurance tactics as randomization and replication, including hidden replication by response surface modeling, are well known and are incorporated as a standard element of the MDOE process. They are relatively easy to implement in a wide range of cases, and have been practiced in other industries for nearly one hundred years. As long as the aerospace industry invokes unwarranted assumptions of stability as a rationale for ignoring systematic error variance instead of adopting proactive measures to cope with it, reliably consistent reproducibility in wind tunnel test results is likely to remain an elusive goal.

VI. Acknowledgements

This work was supported by the NASA Aeronautics Test Program Office.

VII. References

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

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

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

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

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

[6]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.

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

[8]International Organization for Standardization, *Guide to the Expression of Uncertainty in Measurement*, ISBN 92-67-10188-9, ISO, Geneva, 1993 [Corrected and reprinted, 1995]

[9]Satterthwaite, F. E. (1946), "An Approximate Distribution of Estimates of Variance Components.", *Biometrics Bulletin* **2**: 110–114,

[10]Welch, B. L. (1947), "The generalization of "student's" problem when several different population variances are involved.", *Biometrika* **34**: 28–35

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