JOURNAL OF AIRCRAFT
Vol. 41, No. 1, January–February 2004

Statistical Analysis of Computational Fluid Dynamics Solutions from the Drag Prediction Workshop

Michael J. Hemsch*
NASA Langley Research Center, Hampton, Virginia 23188-2199

A simple, graphical framework is presented for robust statistical evaluation of results obtained from N-version testing of a series of Reynolds averaged Navier–Stokes computational fluid dynamics codes. The solutions were obtained by a variety of code developers and users for the June 2001 Drag Prediction Workshop sponsored by the AIAA Applied Aerodynamics Technical Committee. The aerodynamic configuration used for the computational tests is the DLR, German Aerospace Research Center DLR-F4 wing–body combination previously tested in several European wind tunnels and for which a previous N-version test had been conducted. The statistical framework is used to evaluate code results for 1) a single cruise design point, 2) a drag polar at a single Mach number, and 3) drag rise at three values of lift.

Nomenclature

C_D	=	drag coefficient
C_L	=	lift coefficient
C_{L_0}	=	intercept for lift coefficient fit
C_{L_α}	=	slope for lift coefficient fit
C_m	=	pitching-moment coefficient
H	=	analysis of means coverage factor
K	=	coverage factor for individual values
k	=	number of data groups
M_∞	=	Mach number
n	=	number of observations in a sample
Re	=	Reynolds number based on the mean aerodynamic chord
x	=	value of an observation
$\bar{x}$	=	sample mean of a set of observations
$\tilde{x}$	=	sample median
$\bar{\bar{x}}$	=	grand average of sample means
κ	=	slope of drag coefficient fit
μ	=	population mean
$\hat{\mu}$	=	estimate of the population mean
σ	=	population standard deviation
$\hat{\sigma}$	=	estimate of the population standard deviation

Introduction

IN June 2001, the AIAA Applied Aerodynamics Technical Committee conducted a Drag Prediction Workshop (DPW) to determine the state of the art in the use of computational fluid dynamics (CFD), in an industrial design setting, for transonic cruise drag predictions of subsonic transports. The workshop challenge was to compute the lift, drag, and pitching moment for the DLR, German Aerospace Research Center DLR-F4 wing–body configuration[1,2] for three sets of conditions (all at $Re = 3.0 \times 10^6$): 1) cruise at $M_\infty = 0.75$, $C_L = 0.5$ (required), 2) drag polar at $M_\infty = 0.75$ (required), and 3) drag rise at $C_L = 0.4$, 0.5, 0.6 (optional). The DLR-F4 wing–body was chosen because of the availability of experimental data[1,2] and because it had been used for a previous challenge.[3]

The focus of the workshop was drag prediction accuracy. For the cruise point, it was required that each submitter use one of the required grids generated before the workshop announcement. Submitters could use their own grids for additional solutions at the cruise point and for the other sets of conditions. Additional details about the workshop may be found in Refs. 4 and 5. The DLR-F4 wing–body configuration is shown in Fig. 1.

The paper is divided into five main sections. The next section presents typical customer requirements for experimental and computational simulations of performance and describes the available experimental results for the cruise point.[1,2] In the following section, the purpose of the statistical analysis and N-version testing is discussed and several statistical methods are presented for comparing the solutions from the various codes and users. This section also contains a discussion of the robustness of the methods in the face of results that differ significantly from the majority. Results of the application of the statistical framework to the DPW N-version tests are given in the following sections. The paper concludes with some final remarks.

Customer Requirements and Experimental Results

Customer requirements for wind-tunnel reproducibility can vary significantly depending on the intended use of the data, but there seems to be industrywide consensus on the needs for performance simulation of subsonic civil transports.[6,7] Typical data quality goals are given in Table 1 for $\pm 2\sigma$ coverage.

For purposes of comparison with the preceding data quality goals and with the results of the DPW, typical corrected results from wind-tunnel testing[1,2] of the DLR-F4 wing–body configuration are presented in Table 2 for the primary DPW challenge condition of cruise at $C_L = 0.5$, $M_\infty = 0.75$, and $Re = 3.0 \times 10^6$. The drag coefficient values in the data tables on the computer disks that are associated with Ref. 1 are rounded to 10 counts. For the purposes of the DPW, the roundoff error was reduced considerably by use of an enhancement technique developed by Vassberg et al.[4] The results in Table 2 were obtained by linear interpolation in the lift and pitching moment curves and in the drag polars plotted as functions of the lift squared. Note that the scatter values of the wind-tunnel drag and pitching moment results are considerably larger than the requirements listed in Table 1.

Multiplying the scatter in the angle of attack, for which $C_L = 0.5$, by the lift-curve slope (roughly 0.12 deg^{-1}) gives the likely scatter of C_L if the angle of attack (AOA) were held constant. That scatter value is ± 0.005 for $\pm 2\sigma$ coverage and is typically acceptable based on the requirements given in Table 1.

Statistical Approach for N-Version Testing

Background

For this paper, the point of view is taken that the scatter (dispersion) of the results computed by the participants for any given

Presented as Paper 2002-0842 at the AIAA 40th Aerospace Sciences Meeting and Exhibit, Reno, NV, 14 January 2002; received 8 April 2003; revision received 8 April 2003; accepted for publication 9 April 2003.

*Aerospace Engineer, Quality Assurance, Mail Stop 280. Associate Fellow AIAA.

output coefficient at a given set of conditions represents the reproducibility of the computational process just as if the individual computed realizations were obtained from a replicated measurement process. In measurement, such a process is called *N*th-order replication.[8] In computation, a similar process is called N-version testing.[9] Reproducibility is defined[10] for measurement as closeness of the agreement between the results of measurements of the same measurand carried out under changed conditions of measurement. The changed conditions of "measurement" for the DPW are, of course, the different codes, solution methods, turbulence models, grids, computing platforms, observers (people who carried out the computational process), and so on.

For this type of analysis, no individual outcome (computational realization) is considered the "right" answer or "best" result. Instead, the collective computational process is considered to consist of all of the individual processes used and the dispersion of the results to be noise in that collective computational process. (Note that, for each turbulence model, a distinct true value would be expected in the limit.) This viewpoint has been suggested by Youden[11,12] of the former National Bureau of Standards for precision measurements of physical constants at different laboratories. It is also consistent with the frequentist[13] interpretation of probability and with the new standard for reporting measurement uncertainty.[10] For reporting purposes, the new standard[10] rejects the categorization of errors into random and systematic in favor of two other classes: Those which are evaluated using conventional statistical methods (type A) and those which are not (type B). With the preceding definition of a collective computational process, a type A analysis is used herein to evaluate what would normally be thought of as systematic differences.

The described approach has several benefits:

1) Credible quantitative estimates can be made of the scatter and the mean of the virtual population of possible computational outcomes, including determining if some of the codes are not performing as well as the majority.

2) Quantitative predictions of the performance of each code can be made by using all of them together as a collective.

3) It can be determined if the scatter is small enough to be useful to designers.

4) Otherwise hidden effects and opportunities for process improvement can be revealed by "seeing into" the scatter.

5) Uncertainty measures for comparison of the collective CFD results to experiment can be obtained.

By doing this, together with continual improvement of the codes and processes, the possibility is created, across the design community, of making CFD predictions with credible, enduring statements of reproducibility.

Three types of statistical graphical methods are used for analysis of the DPW N-version testing results for the cruise point and the drag polar: 1) running records of individual computational outcomes, together with centerlines and scatter bands; 2) histograms of individual outcomes; and 3) analysis of means (ANOM).

Table 1 Typical (design) customer uncertainty goals ($\pm 2\sigma$) for performance simulations[6,7]

Coefficient	Increments	Absolute
C_L	0.005	0.01
C_D	0.00005	0.0001
C_m	0.0005	0.001

The first two methods are valuable because they display all of the data in the presence of the scatter.[14] One of the biggest concerns in statistical analysis is possibly drawing incorrect conclusions from aggregated data.[14–19] The advantage of running records and histograms is that they display all of the data and that they do it in quite different ways, thus, allowing the analyst to check conclusions drawn from one or the other display. In addition, several questions are asked of the first two methods.

1) Which solutions constitute a reasonable "core" set and what does that core set look like?

2) Which solutions lie outside that core?

3) What is the central location of the core solutions?

4) What is the dispersion of the core solutions?

The third method is useful because it allows discernment of possibly significant results despite the presence of scatter.[15]

In the rest of this section, these three graphical analysis methods are reviewed, before analyzing the results of the DPW challenge for the cruise point and the drag polar in the following two main sections.

Method 1: Running Records of Individual Outcomes

In an analysis of experimental repeatability and reproducibility, it is useful to display the data values on the vertical axis and the time of the data point acquisition on the horizontal axis.[17] However, because time is irrelevant in the DPW challenge, the individual solution values are plotted herein as a function of an integer index. The order of the plotting of the solution values along the abscissa was randomly assigned.

To enhance the possibility of discerning significant results in the running record, an estimate of the population mean $\hat{\mu}$ of the plotted data points is made and shown on the graph as a centerline. In addition, scatter limits are placed about the centerline as follows:

$$\text{limits} = \hat{\mu} \pm K\hat{\sigma} \tag{1}$$

where $\hat{\sigma}$ is an estimate of the population standard deviation and K is an appropriate coverage factor.[14–18] The area between the scatter limits is considered to be a noise zone within which it is not possible to discern significant results for the individual values, at least not on the basis of statistics.

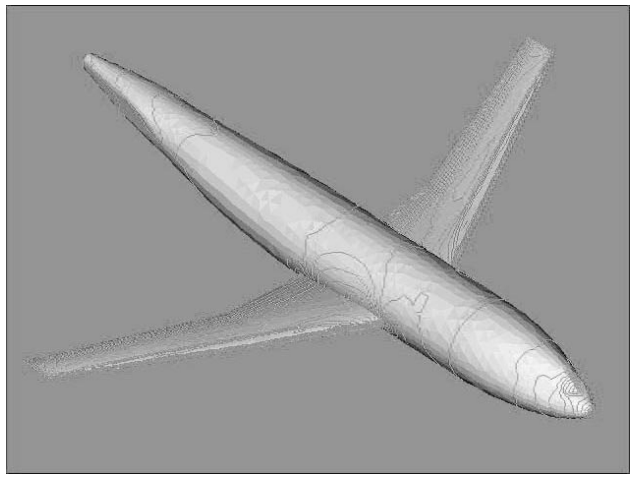


Fig. 1 Geometry of DLR-F4 model used for the N-version testing.

Table 2 Experimental results[a] from three tunnels for $C_L = 0.5$ and $M_\infty = 0.75$[1,2]

Variable	NLR[b]	ONERA	DRA	Sample mean	Sample range	Use
AOA, deg	0.153	0.192	0.181	0.175	0.039	0.18 ± 0.04
C_D	0.0288	0.0289	0.0281	0.0286	0.0008	0.0286 ± 0.0008
C_m	−0.130	−0.126	−0.137	−0.131	0.011	-0.13 ± 0.01

[a]Uncertainty is estimated at a coverage of $\pm 2\sigma$. [b]National Aerospace Laboratory in the Netherlands (NLR).

To summarize, there are six steps for the preparation of individual value running records for graphical statistical analysis:

1) Graphically display the outcomes, for supposedly identical conditions, in random order in a pseudotime series.

2) Estimate the population mean and standard deviation of the outcomes using a robust method.

3) Display the estimate of the population mean as the centerline on the graph.

4) Establish limits for the scatter of the individual outcomes based on the estimated population standard deviation and a reasonable coverage factor.

5) Display the limits obtained in step 4 as lower and upper bounds on the graph.

6) Look for significant results, that is, outcomes that are outside the noise zone delineated by the scatter limits. Such outcomes could represent poor solutions relative to the others, significant differences in solution approaches, or opportunities for process improvement.

The most common form of the histogram is obtained by splitting the range of the data into equal-sized bins (called classes). Then, for each bin, the number of points from the data set that fall into each bin are counted. That is

Vertical axis: Frequency (i.e., counts for each bin)

Horizontal axis: Response variable

The histogram can be used to answer the following questions [qualitatively]:

1. What kind of population distribution do the data come from?
2. Where are the data located?
3. How spread out are the data?
4. Are the data symmetric or skewed?
5. Are there outliers in the data?

Robust Estimation of the Population Parameters

The only two population parameters usually of interest in the estimation of uncertainty are the mean and the standard deviation.[17]

Conventionally, these parameters are estimated using the sample average

$$\hat{\mu} = \bar{x} \equiv \frac{1}{n}\sum_{i=1}^{n} x_i \tag{2}$$

and the sample standard deviation (SSD)

$$\hat{\sigma} = \text{SSD} \equiv \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{3}$$

These estimators are the most efficient for homogeneous data sets. However, they can give significantly offset estimates of the population mean and grossly inflated estimates of the population standard deviation if the sample is contaminated with values that do not belong to the population of interest.[14,16,20] For the purposes herein, the simplest way to achieve robustness is to use the sample median to estimate the population mean[16,20]

$$\begin{aligned} \hat{\mu} &= \tilde{x} \\ \tilde{x} &\equiv x_{(n+1)/2} && n \text{ odd} \\ &\equiv 0.5\left(x_{n/2} + x_{(n/2)+1}\right) && n \text{ even} \end{aligned} \tag{4}$$

and the sample median absolute deviation (MAD)[20,21] to estimate the population standard deviation

$$\hat{\sigma} = 1.483\sqrt{n/(n-1)}\,\text{median}\,(|x_i - \tilde{x}|) \tag{5}$$

Establishing the Limits

For this paper, limits for judging the significance of individual outcomes will be established at 100:1 odds, that is, observations would not be found outside the limits more than one time in 100 in the long run by chance (99% coverage). Because there will never be enough observations for this type of application to reasonably determine the population probability density function (PDF), a distribution will have to be selected for use as a guide in determining the coverage factor and the normal distribution is a convenient choice.[15]

With the normal distribution as a guide and with 99% coverage (100:1 odds) for infinite degrees of freedom, $K = 2.58$ for insertion into the scatter limits given by Eqs. (1).

Method 2: Histograms of Individual Values

This statistical graphics method has been used to great effect for roughly 200 years.[19] The histogram presents a very different pattern of the data values to the eye and, hence, may reveal significant results that would otherwise be obscured in tables or a running record.[14,19] Detailed discussions on the construction and use of histograms are given in Ref. 16 and 18. The following descriptive comments are taken directly from Ref. 16:

For the analysis of computational (or experimental) outcomes, there are rarely enough observations to determine skewness or kurtosis or the form of the population distribution itself. Of course, an attempt must be made to determine possible outliers and multiple modes. For single-event data sets such as the DPW results, outliers suggest mistakes or solution methods significantly different from the core group, whereas multiple modes may suggest significant effects buried in the scatter. (For time-series data obtained for an ongoing process, it would be more appropriate to interpret outliers, that is, points outside the limits, as indications of changes in the process that generates the data.) The methods of the next two subsections are designed to address the latter.

Method 3: Analysis of Means (ANOM)

The ANOM[15,22] is a graphical statistical technique used for comparison of the averages of sets of supposedly similar results in the presence of noise. For comparisons of just two averages, it is identical, in numerical result, to a hypothesis test on means.[22] The ANOM method is more general, however, because it can be used for comparison of more than two averages and, of course, it is graphical.

The ANOM decision limits are given by Wheeler[15,22] as

$$\text{limits} = \bar{\bar{x}} \pm H\hat{\sigma}\Big/\sqrt{n} \tag{6}$$

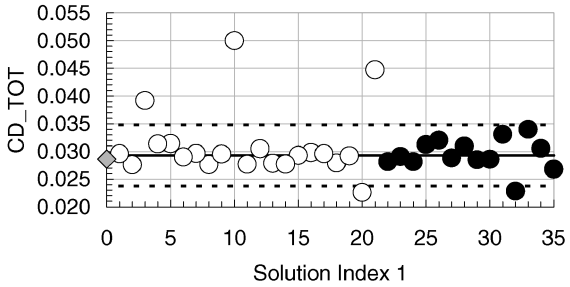
where $\bar{\bar{x}}$ is the grand average of the sample sets and H is a function of the coverage desired, the degrees of freedom for the estimate of the population standard deviation, and the number of sets of averages being compared, k. Note that, if the number of observations in the sets of sample averages are different, the limits will be different because of the $\sqrt{n}$ in the denominator of Eqs. (6). Note also that, once again, it is necessary to choose a PDF for a guide and again that the normal distribution is selected.

As with the interpretation for a running record of individual values, the distance between the ANOM decision limits represents the noise of the process within which the averages cannot be distinguished, at least not statistically. For the ANOM analyses in this paper, an exploratory level of coverage, 90%, will be used because the intention is to look for possible opportunities for improvement.

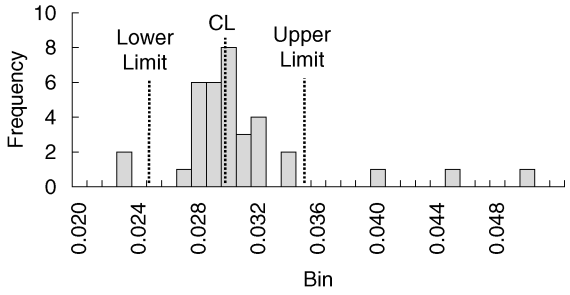
Table 3 Breakdown of location and scale for the cruise point drag solutions[a]

Parameter	CD_PR	CD_SF	CD_TOT	Experiment
$\hat{\mu}$	0.0166	0.0134	0.0293	0.0286
$\hat{\sigma}$	0.0014	0.0015	0.0021	0.0004

[a]Median and MAD are used to estimate the population parameters for the computations.



a) Running record



b) Histogram, each bin covers 10 counts

Fig. 2 All total drag solutions at $C_L = 0.5$ and $M_\infty = 0.75$; median plus MAD plus 100:1 limits: ○, provided grids; ●, other grids; ——, median; - - - -, lower limit; - - - -, upper limit; and ◇, experimental data.

Replicating the process will determine if the effects suggested by the statistical technique are real.[22]

Cruise Point at $C_L = 0.5$ and $M_\infty = 0.75$

For this paper, only the solutions provided to the DPW committee before the workshop were used. For the cruise point condition, the solution statistics are as follows:

1) There were 14 codes used: 7 structured grid, 6 unstructured-grid, and 1 Cartesian grid (Euler plus integral boundary layer).

2) There were 35 solutions computed: 24 structured, 10 unstructured, and 1 Cartesian.

3) There were 3 types of turbulence models used: a) 17 Spalart–Allmaras, 8 structured grid and 9 unstructured grid; b) 17 two-equation, 16 structured grid and 1 unstructured grid; and c) 1 integral boundary layer.

4) There were 21 solutions that used the provided grids.

5) There were 14 solutions that used other grids.

This breakdown will be used to attempt to determine possible sources of variation in the solutions.

Analysis of All Drag Results

Total Drag

The cruise point solutions for the total drag are shown in the running record and histogram of Fig. 2 plotted at random along the abscissa, together with the centerline and limits obtained as described earlier (median and MAD) and with the experimental results from Table 2. The number shown centered just below each histogram bin is the upper value for that bin interval.

For both the running record and the histogram, it seems clear that there is a central core with a wide spread on the order of 50–80 counts, together with 5 solutions that are definitely outside that core. The population parameters as estimated by the median and MAD for all of the solutions are given in Table 3. Note that the first 21 solutions shown in Fig. 2 were obtained on the provided grids (open symbols). An additional 14 solutions were obtained on other grids (filled symbols).

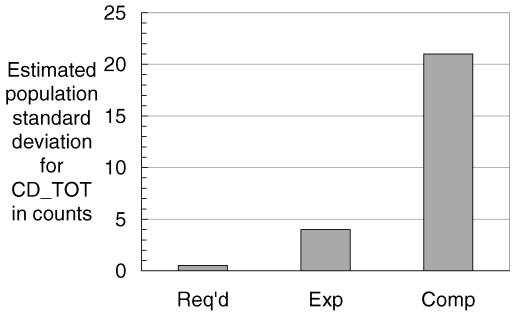
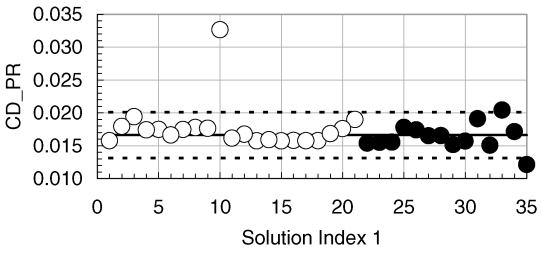
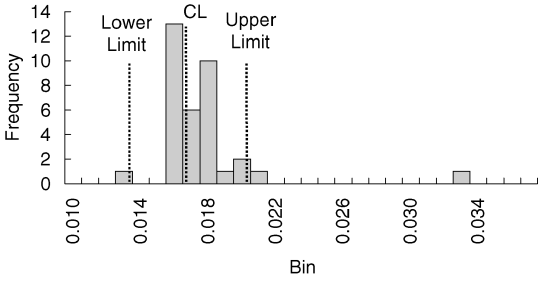


Fig. 3 Comparison of required, experimental and computational estimated population standard deviations.



a) Running record



b) Histogram, each bin covers 10 counts

Fig. 4 All pressure drag solutions at $C_L = 0.5$ and $M_\infty = 0.75$; median plus MAD plus 100:1 limits: ○, provided grids; ●, other grids; ——, median; - - - -, lower limit; and - - - -, upper limit.

A comparison of customer requirements, experimental results and the scatter of the core solutions for CD TOT is shown in Fig. 3.

Pressure and Skin-Friction Drag

The cruise point solutions for the pressure and skin-friction drag are shown similarly in Figs. 4 and 5 plotted with the same index as Fig. 2. Note that the histograms of Figs. 4 and 5 show that it is not as easy to decide which drag component solutions are in the core group or outside it when they lie close to the limits. The population parameters as estimated by the median and MAD for all of the solutions are given in Table 3.

The breakdown on the solutions/codes that had the total drag or one of its components outside the limits is as follows: 1) 7 out of 35 solutions (20%), 2) 6 out of 14 codes (43%), 3) 4 out of the 21 solutions on the provided grids (19%), and 4) 3 out of the 14 solutions on other grids (21%). Overall, there appears to be no significant difference in either location or scale between the drag solutions carried out on the provided grids and solutions carried out on grids developed or improved by the participants. However, discussion during the workshop revealed that the provided grids were not well suited to some of the codes.

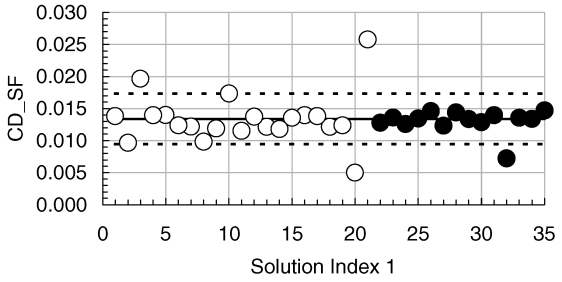
Analysis of All Angle-of Attack (AOA) Results

The cruise point solutions for the AOA values are shown in Fig. 6. The breakdown of solutions/codes that had values outside the limits

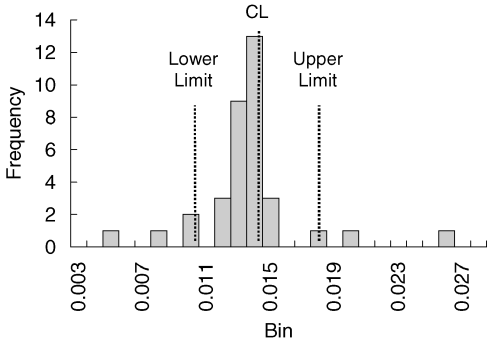
Table 4 Breakdown of location and scale for the cruise point AOA and pitching-moment solutions[a]

	AOA, deg		C_m	
Parameter	CFD	Experiment	CFD	Experiment
$\hat{\mu}$	−0.26	0.18	−0.160	−0.13
$\hat{\sigma}$	0.13	0.02	0.0084	0.005

[a]Median and MAD are used to estimate the population parameters for the computations.



a) Running record



b) Histogram, each bin covers 10 counts

Fig. 5 All skin-friction drag solutions at C_L = 0.5 and M_∞ = 0.75; median plus MAD plus 100:1 limits: ○, provided grids; ●, other grids; ——, median; - - - -, lower limit; and - - - -, upper limit.

is as follows: 1) 7 out of 35 solutions (20%), 2) 7 out of 14 codes (50%), 3) 4 out of the 21 solutions on the provided grids (19%), and 4) 3 out of the 14 solutions on other grids (21%).

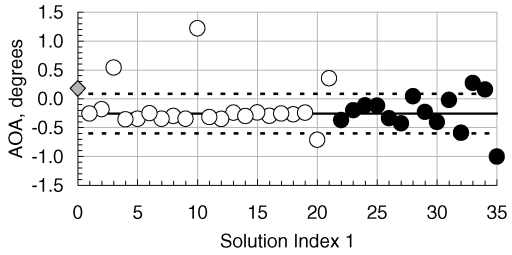
Again, there does not seem to be any significant effect of provided vs participant-generated grids on the medians of the two types of solutions. A summary of the location and scale values for the cruise point AOA solutions is given in Table 4.

Note that the difference between the estimated computed AOA location and the experimental location is 0.44 deg, which corresponds to an overprediction of the lift coefficient of about 10%. The estimated standard deviation of 0.13 deg corresponds to an estimated standard deviation for the lift coefficient of roughly 0.02, which is comparable to the estimated value for the available experimental data.

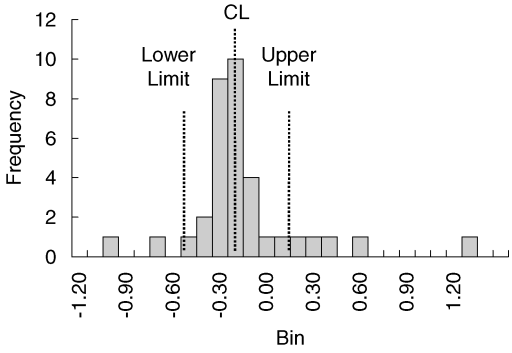
Analysis of All Pitching-Moment Results

The cruise point solutions for the pitching-moment values are shown in Fig. 7. The breakdown of solutions/codes that had pitching-moment values outside the limits is as follows: 1) 8 out of 35 solutions (23%), 2) 7 out of 14 codes (50%), 3) 4 out of the 21 solutions on the provided grids (19%), and 4) 4 out of the 14 solutions on other grids (29%). As with the drag and AOA results, there does not seem to be any significant effect of provided vs participant-generated grids on the medians of the two types of solutions. A summary of the location and scale values for the cruise point pitching-moment solutions is given in Table 4.

Note that the difference between the estimated computed location and the experimental location is 0.03, which corresponds to



a) Running record



b) Histogram, each bin covers 0.1 deg

Fig. 6 All AOA solutions at C_L = 0.5 and M_∞ = 0.75; median plus MAD plus 100:1 limits: ○, provided grids; ●, other grids; ——, median; - - - -, lower limit; - - - -, upper limit, and ◇, experimental data.

an overprediction of the static margin of roughly 6% of the mean aerodynamic chord.

Effect of Turbulence Model and Grid Type on Drag

If the five total-drag outlier solutions, the Cartesian integral-boundary-layer solution, and the two-equation unstructured grid solution are deleted, 28 cruise-point solutions are left that break down as follows: 1) 7 solutions using the Spalart-Allmaras turbulence model on unstructured grids (3 codes with 4 observers), 2) 7 solutions using the Spalart-Allmaras model on structured grids (2 codes with 4 observers), and 3) 14 solutions using a two-equation model on structured grids (6 models and 5 codes with 7 observers). To compare these solutions and attempt to determine if there are significant effects due to turbulence model and grid type, the ANOM method described earlier will be used. First, the sample averages [Eq. (2)] and standard deviations [Eq. (3)] for each of the three distinct sets of values are calculated. (The robust, but less efficient, estimators of the earlier analyses are not needed because the major outliers have been deleted.) Second, those values are pooled to obtain $\bar{\bar{x}}$ and $\hat{\sigma}$ (Refs. 14 and 17). The results are given in Table 5 for the total drag and the two drag components. Results for AOA and pitching moment are given in Ref. 23.

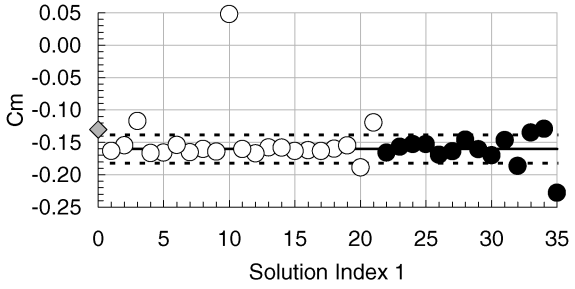
To complete the ANOM statistical graph, note that there are three sets of averages to compare for each type of outcome and 25 degrees of freedom (28 observations divided into 3 sets). Furthermore, it is desired to compare the averages with a coverage of 90%. This gives $H = 1.76$ (Refs. 15 and 22). The ANOM results are shown in Figs. 8–10, together with the corresponding running records (at 99% coverage). The randomly assigned abscissa index for the running records is different from the index used in Figs. 2, 4–7.

Effect of Turbulence Model on Drag

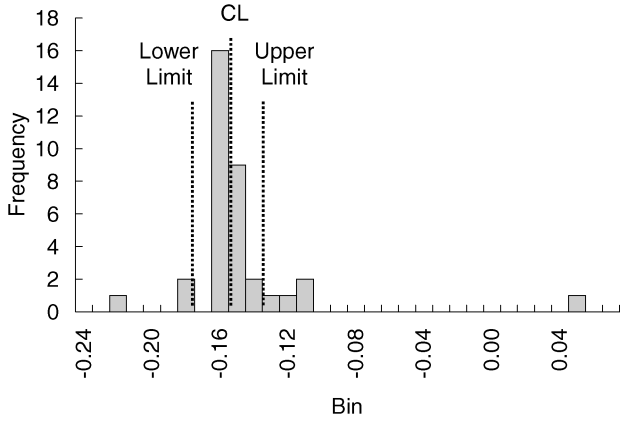
When judged from both the running record and ANOM charts of Fig. 8, there does not seem to be a significant effect of the two types of turbulence model (groups 2 and 3) on the total drag. However, the pressure and skin-friction component comparisons, Figs. 9 and 10, show a different story. The results from the two model types seem to be (barely) significantly offset by about 10–12 drag counts but with opposite sign. This result suggests that it would be worthwhile to investigate further.

Table 5 Values used for ANOM analysis of cruise point data

			CD_TOT		CD_PR		CD_SF	
Set	Turbulence model	Grid	$\bar{x}$	S	$\bar{x}$	S	$\bar{x}$	S
1	Spalart–Allmaras	Unstructured	0.0284	0.00061	0.0162	0.00053	0.0123	0.00048
2	Spalart–Allmaras	Structured	0.0299	0.00108	0.0162	0.00090	0.0138	0.00024
3	Two equation	Structured	0.0298	0.00201	0.0172	0.00141	0.0126	0.00144
	Pooled value		0.0295	0.00157	0.0167	0.00114	0.0128	0.00107



a) Running record



b) Histogram, each bin covers a delta C_m of 0.01

Fig. 7 All pitching-moment solutions at $C_L = 0.5$ and $M_\infty = 0.75$; median plus MAD plus 100:1 limits: ○, provided grids; ●, other grids; ——, median; - - - -, lower limit; - - - -, upper limit, and ◇, experimental data.

Effect of Grid Type on Drag

When judged from both the running record and ANOM charts of Fig. 8, there does seem to be a significant effect on the total drag of the two grid types (groups 1 and 2), using the Spalart–Allmaras turbulence model. Comparing the drag component results in Figs. 9 and 10; it is seen that the effect is due to a significant difference in the skin-friction drag of 15 counts.

Drag Polar for $M_\infty = 0.75$

The required drag polars were fit in the linear range ($C_L = 0.15 - 0.40$) as follows:

$$C_L = C_{L_0} + C_{L_\alpha}\alpha, \qquad C_D = C_D@C_{L_0} + \kappa C_L^2 \tag{7}$$

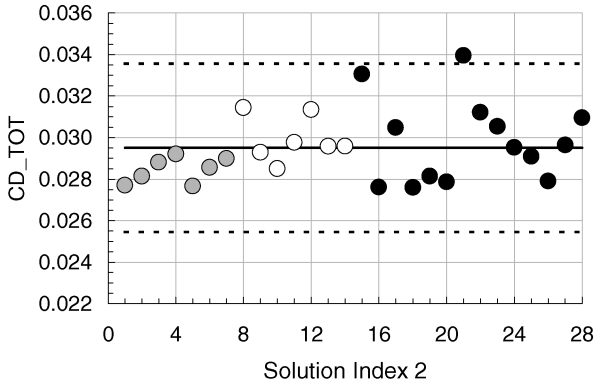
The available experimental results[1,2] were fit as well. The running records and histograms of the fit slope parameters, C_{L_α} and κ, for all of the available solutions are given in Figs. 11 and 12. The estimated values of the population means and standard deviations are given in Table 6 for all of the fit parameters, together with the estimates from the three wind-tunnel tests. The running records and histograms for the fit intercept parameters can be found in Ref. 23. Note that the solution abscissa index is different from the two used earlier.

The estimated standard deviations for the computed (and experimental) fit intercepts (Table 6) are not surprising given the results discussed earlier for the cruise point condition. However, the estimates of the computational and experimental population standard

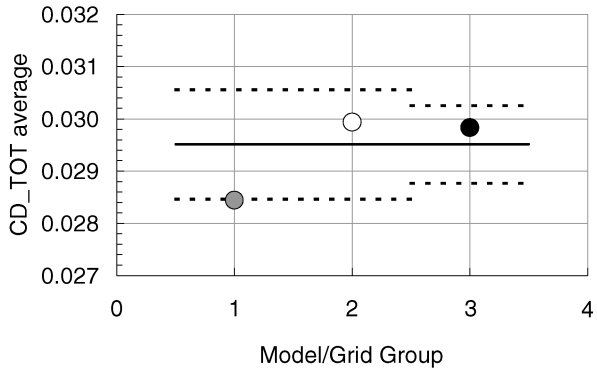
Table 6 Results from fits to computational and experimental[a] drag polars

	$\hat{\mu}$		$\hat{\sigma}$	
Parameter	CFD	Experiment	CFD	Experiment
C_{L_0}	0.531	0.473	0.021	0.003
C_{L_α}	0.120	0.114	0.0026	0.0019
C_D at $C_L = 0$	0.0198	0.0188	0.0016	0.0005
κ	0.0365	0.0374	0.0018	0.0040

[a]Experimental values were obtained using the sample averages and standard deviations.



a) Running record



b) ANOM

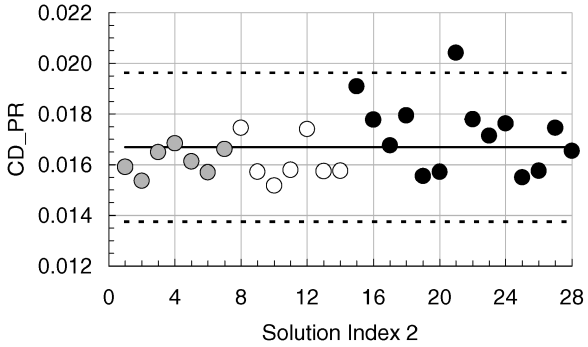
Fig. 8 Total drag; analysis of turbulence model and grid-type effects at $C_L = 0.5$ and $M_\infty = 0.75$: ◎, Spalart–Allmaras unstructured; ○, Spalart–Allmaras structured; ●, two-equation structured; ——, grand average; - - - -, lower limit; and - - - -, upper limit.

deviations for the fit slopes, C_{L_α}, and κ, seem to be considerably larger than would be expected from conventional wisdom.

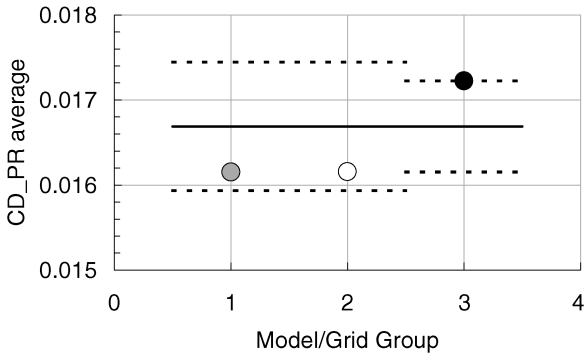
The breakdown on the fit parameters that are outside the limits given by the median and MAD for the 27 polars is as follows[23]: 1) for C_{L_0}, four (15%); 2) for C_{L_α}, two (7%); 3) for C_D at C_{L_0}, four (15%); and 4) for κ, two (7%).

Drag Rise at $C_L = 0.4$, 0.5, and 0.6

Eight drag rise solutions for all three lift-coefficient values were provided to the workshop committee: seven observers using five codes, two turbulence models (Spalart–Allmaras and Wilcox $k-\varpi$), and both grid types. The solutions, together with experimental

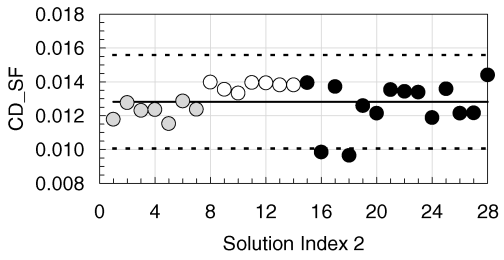


a) Running record

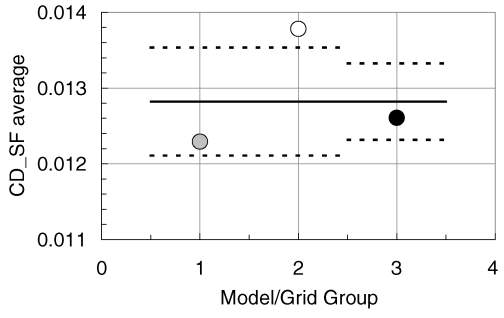


b) ANOM

Fig. 9 Pressure drag; analysis of turbulence model and grid-type effects at $C_L = 0.5$ and $M_\infty = 0.75$: ◍, Spalart–Allmaras unstructured; ○, Spalart–Allmaras structured; ●, two-equation structured; ——, grand average; - - - -, lower limit; and - - - -, upper limit.

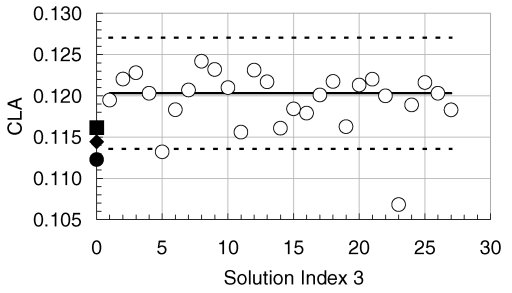


a) Running record

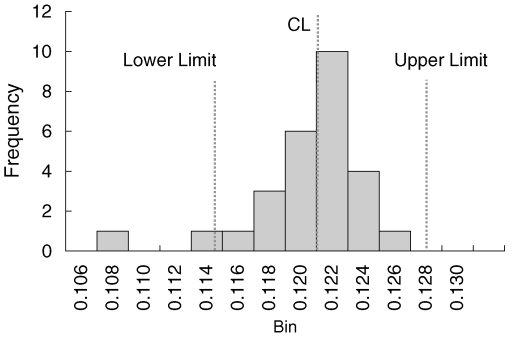


b) ANOM

Fig. 10 Skin-friction drag; analysis of turbulence model and grid-type effects at $C_L = 0.5$ and $M_\infty = 0.75$: ◍, Spalart–Allmaras unstructured; ○, Spalart–Allmaras structured; ●, two-equation structured; ——, grand average; - - - -, lower limit; and - - - -, upper limit.

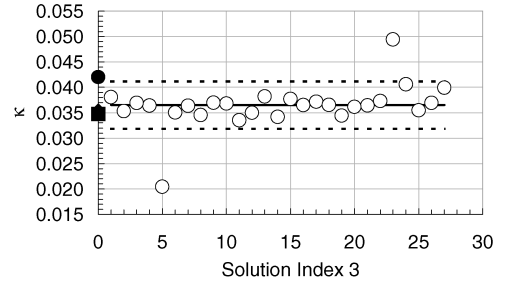


a) Running record

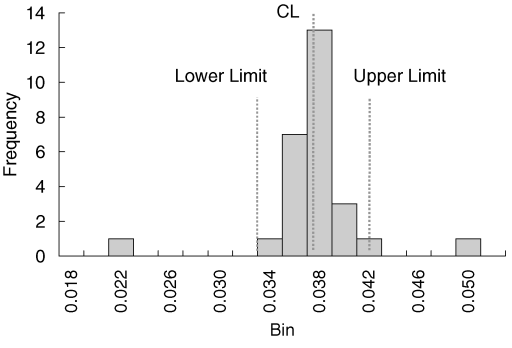


b) Histogram, each bin covers a delta slope for lift coefficient fit of 0.002

Fig. 11 Lift-curve slope from drag polars at $M_\infty = 0.75$; median plus MAD plus 100:1 limits: ○, fit values; ——, median; - - - -, lower limit; - - - -, upper limit; ●, NLR-HST; ■, ONERA-S2MA; and ♦, DRA 8 × 8.



a) Running record



b) Histogram, Each bin covers a delta κ of 0.002

Fig. 12 Drag-curve slope from drag polars at $M_\infty = 0.75$; median plus MAD plus 100:1 limits: ○, fit values; ——, median; - - - -, lower limit; - - - -, upper limit; ●, NLR-HST; ■, ONERA-S2MA; and ♦, DRA 8 × 8.

data inferred from the drag polar fits (filled symbols), are shown in Fig. 13. Note that one of the solution drag-rise curves is considerably outside the scatter of the other solutions. Also note that the total drag appears to be underpredicted for the highest M_∞, C_L combination (0.8, 0.6), but otherwise seems to scatter reasonably about the experimental data.

For the type of flows encountered in this challenge, it seems reasonable to ask if there is any effect of the Mach number and lift coefficient on the scatter. To that end, the SSDs of the solutions at each M_∞, C_L combination (not including the "structured Wilcox 2" solution) are shown in Fig. 14.

There are several points worth discussing.

1) The computed SSDs lie within a band of three counts except for the two highest Mach numbers for $C_L = 0.6$.

2) The estimated population standard deviation for the drag rise points is about half that of the estimates obtained for the cruise point and drag polar analyses.

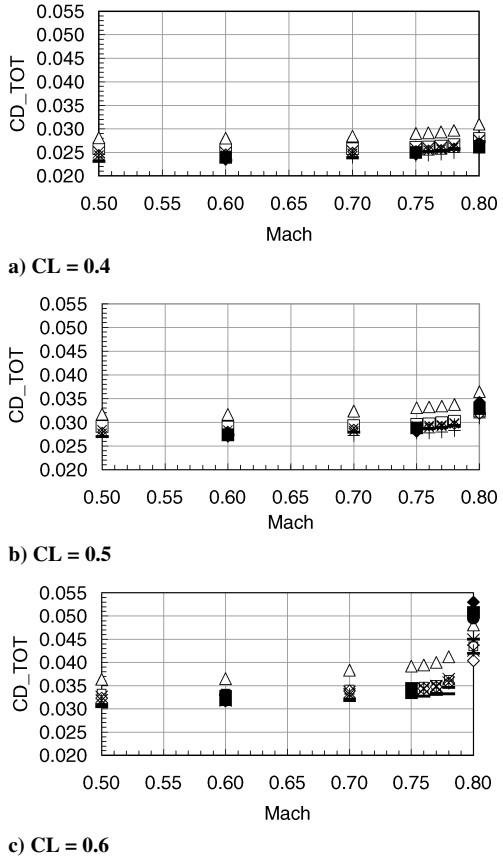


Fig. 13 Comparison of drag rise solutions: ○, structured Spalart–Allmaras 1; □, structured Spalart–Allmaras 2; ◇, structural Wilcox 1; △, structural Wilcox 2; ✳, unstructured Spalart–Allmaras 1; ×, unstructured Spalart–Allmaras 2; +, unstructured Spalart–Allmaras 3; ——, unstructured Spalart–Allmaras 4; ●, NLR-HST; ■, ONERA-S2MA; and ♦, DRA 8×8.

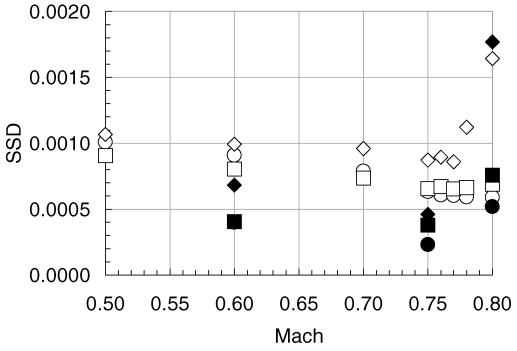


Fig. 14 Running record of estimated standard deviations from drag rise studies: ○, computed C_L= 0.4; □, computed C_L, = 0.5, ◇ computed C_L = 0.6; ●, experimental C_L = 0.4; ■, experimental C_L = 0.5; and ♦, experimental C_L = 0.6.

3) The computational $\hat{\sigma}$ is roughly a factor of two larger than the experimental $\hat{\sigma}$.

The first observation suggests the possibility that some additional source of variation manifests itself for the highest M_∞, C_L combinations. Perhaps the shock has become strong enough to create significant flow separation. This possibility could be checked by comparing the solutions for the lower and higher SSDs. The second observation is probably the result of fewer opportunities for variation to be expressed. There were 7 drag rise solutions used for this analysis vs 35 and 27 for the cruise point and drag polar analyses, respectively. The third observation suggests that more work is needed to reduce the computational scatter before validation is possible to the accuracy of the experimental data.

Conclusions

One aim of the analysis in this paper has been to determine if the replicated outcomes appear to have been drawn at random from a single (virtual) population. If so, it is meaningful to talk about the parameters of the population of the collective, including the mean and the standard deviation. It is also then meaningful to talk about quantifiable predictability. However, note that the analysis herein was for a single data set. Hence, any conclusions must be considered tentative, that is, exploratory. Confirmatory results can only be obtained by determining that the CFD results for the population parameters are predictable, that is, stable, and that requires established processes and repeating the processes over time.

One can check the stability of any process, including CFD, by making the running record of individual values into a true time series of repeated outcomes and tracking it as a process behavior chart.[14–18] With these caveats, it does seem that using the median to estimate the location and MAD to estimate the scale did allow discernment of outlier solutions in the running records and histograms without losing the meaning of the "core" solutions. Also, there do seem to be credible values of the CFD population parameters of interest for the outcomes studied. Again, whether these values are durable can only be seen by repeating this exercise.

The analysis also suggests that some set of best practices and quantitative sanity checks is needed to avoid outliers, especially those that might show up when the luxury of multiple solutions from diverse codes and models, or even grid convergence, is not available. The continued existence of such outliers would force acceptance of much bigger numbers for the scatter. For example, for $\hat{\sigma}$ for the cruise point condition, the changes would be 1) CD_TOT, 51 counts vs 21 counts; 2) AOA, 0.37 deg vs 0.13; and 3) CM, 0.040 vs 0.008. Unfortunately, even the scatter of the core solutions is unacceptable compared to the oft-stated requirements of the user community[6,7] and for reasonable validation as well, as can be seen in the comparison of Fig. 3.

Acknowledgments

The author is grateful to Richard A. Wahls and the other members of the Drag Prediction Workshop Committee for suggesting this analysis and to Ollie J. Rose, James M. Luckring, and Eric L. Walker for many helpful suggestions.

References

[1]Redeker, G., "DLR-F4 Wing Body Configuration," *A Selection of Experimental Test Cases for the Validation of CFD Codes*, AR-303, Vol. 2, AGARD, Aug. 1994, pp. B4.1–B4.21.

[2]Redeker, G., Muller, R., Ashill, P. R., Elsenaar, A., and Schmitt, V., "Experiments on the DLR-F4 Wing Body Configuration in Several European Wind Tunnels," *Aerodynamic Data Accuracy and Quality: Requirements and Capabilities in Wind Tunnel Testing*, CP-429, AGARD, July 1988, pp. 2.1–2.15.

[3]Elsholz, E., "The DLR-F4 Wing/Body Configuration," *ECARP—European Computational Aerodynamics Research Project: Validation of Turbulence Models*, Notes on Numerical Fluid Mechanics, Springer Verlag, Heidelberg, Germany, Vol. 58, 1997, pp. 429–450.

[4]Vassberg, J., Buning, P. G., and Rumsey, C. L., "Drag Prediction for the DLR-F4 Wing/Body Using OVERFLOW and CFL-3D on an Over-set Mesh," AIAA Paper 2002-0840, Jan. 2002.

[5]Levy, D. W., Zickuhr, T., Vassberg, J. C., Agrawal, S., Wahls, R. A., Pirzadeh, S., and Hemsch, M. J., "Summary of Data from the AIAA CFD Drag Prediction Workshop," AIAA Paper 2002-0841, Jan. 2002.

[6]Steinle, F., and Stanewsky, E., "Wind Tunnel Flow Quality and Data Accuracy Requirements," AR-184, AGARD, Nov. 1982.

[7]Carter, E. C., and Pallister, K. C., "Development of Testing Techniques in a Large Transonic Wind Tunnel to Achieve a Required Drag Accuracy and Flow Standards for Modern Civil Transports," *Aerodynamic Data Accuracy and Quality: Requirements and Capabilities in Wind Tunnel Testing*, CP-429, AGARD, July 1988, pp. 11.1–11.20.

[8]Moffat, R. J., "Contributions to the Theory of Single-Sample Uncertainty Analysis," *Journal of Fluids Engineering*, Vol. 104, No. 2, June 1982, pp. 250–260.

[9]Hatton, L., "The T Experiments: Errors in Scientific Software," *IEEE Computational Science and Engineering*, Vol. 4, No. 2, 1997, pp. 27–38.

[10]*U.S. Guide to the Expression of Uncertainty in Measurement*, American National Standards Institute/National Conference of Standards Labs., Rept. ANSI/NCSL Z540-2-1997, Boulder, CO, Oct. 1997.

[11]Youden, W. J., "Enduring Values," *Technometrics*, Vol. 14, No. 1, 1972, pp. 1–11.

[12]Youden, W. J., "Systematic Errors in Physical Constants," *Physics Today*, Vol. 14, No. 9, 1961, pp. 32–43.

[13]Von Mises, R., *Probability, Statistics and Truth*, Dover, New York, 1981.

[14]Wheeler, D. J., and Chambers, D. S., *Understanding Statistical Process Control*, 2nd ed., SPC Press, Knoxville, TN, 1992.

[15]Wheeler, D. J., *Advanced Topics in Statistical Process Control: The Power of Shewhart's Charts*, SPC Press, Knoxville, TN, 1995.

[16]Croarkin, C., and Tobias, P. (Eds.), *NIST/Sematech Engineering Statistics Handbook* (online) http://www.itl.nist.gov/div898/handbook/ [Accessed 28 Sept. 2001].

[17]Mandel, J., *The Statistical Analysis of Experimental Data*, Dover, New York, 1984.

[18]*Manual on Presentation of Data and Control Chart Analysis*, 6th ed., ASTM Manual Series: MNL 7, American Society for Testing and Materials, Philadelphia, PA, 1995.

[19]Tufte, E. R., *The Visual Display of Quantitative Information*, 2nd ed., Graphics Press, Cheshire, CT, 2001.

[20]Mosteller, F., and Tukey, J. W., *Data Analysis and Regression: A Second Course in Statistics*, Addison-Wesley, Reading, MA, 1977.

[21]Muller, J. W., "Possible Advantages of a Robust Evaluation of Comparisons," *Journal of Research of the National Institute of Standards and Technology*, Vol. 105, No. 4, 2000, pp. 551–555.

[22]Wheeler, D. J., *Understanding Industrial Experimentation*, 2nd ed., SPC Press, Knoxville, TN, 1990.

[23]Hemsch, M. J., "Statistical Analysis of CFD Solutions from the Drag Prediction Workshop," AIAA Paper 2002-0842, Jan. 2002.