ANOTHER "NEW" APPROACH FOR "VALIDATING" SIMULATION MODELS

Arthur Fries, Institute for Defense Analyses
1801 N. Beauregard St., Alexandria, VA 22311

ABSTRACT

When observed test data are sparse and widely scattered across numerous experimental factors, the issue of validating any complementary simulation modeling is problematic. We focus on the extreme case — limited data within a vast highly-dimensional factor space, and only a single replicate per test — although results readily generalize. Our only assumptions are that, for each test, we can measure the associated experimental factor values and input these into the model to generate extensive simulated data. Under the null hypothesis that the model accurately portrays reality, this "distribution" of outcomes facilitates calculation of a *p*-value. Fisher's combined probability test, for synthesizing results from different experiments, is then applied to obtain an overall characterization of the degree of simulation model "validity". Other variants of this combination methodology are also discussed, including generalizations to goodness-of-fit tests for a uniform distribution. Unique aspects of the model validation problem, vice the standard statistical hypothesis testing regime, are also noted.

INTRODUCTION

Modeling and simulation (M&S) plays an ever-expanding test and evaluation role within the U.S. Department of Defense (DoD), especially when actual testing of expensive systems is limited by time, budget and resource constraints. Central to the credible application of M&S is the systematic planning for and implementation of codified verification, validation and accreditation (VV&A) procedures (Army, 1997):

- "Verification is the process of determining if the M&S accurately represents the developer's conceptual description and specifications and meets the needs stated in the requirements document."

- "Validation is the process of determining the extent to which the M&S adequately represents the real-world from the perspectives of its intended use."

- "Accreditation is the official determination that the M&S is acceptable for its intended purpose."

An essential element of these processes should be the investigation of the degree to which M&S results are consistent with observed data, from actual combat operations, training operations, of dedicated testing experiences. Statistical comparisons of this sort complement efforts to better understand and improve the M&S. They also yield quantitative evidence that can support formal declarations of whether VV&A requirements have been satisfied.

This paper considers the statistical problem of trying to establish the consistency between observed data and predicted outcomes derived from the M&S being scrutinized. When data are plentiful many standard methodological approaches are applicable (e.g., Law & Kelton, 1991). Our focus, however, is on the extreme, but not uncommon, set of circumstances in which the extent of data is quite limited, generally no replicates are available, and the factor space for the M&S input variables is highly-dimensional.

Section 2 defines our setting more precisely, and reviews and critiques the potential role of "standard" statistical approaches within this context. A "new" analysis methodology relying on Fisher's Combined Probability Test is introduced in Section 3, and extended to encompass general goodness-of-fit procedures in Section 4. Statistical characterizations of these alternative methodologies, namely Type I error rate and power, are documented in Section 5. A brief summary and discussion is given in Section 6.

OUR SETTING

Models of the performance of defense systems typically include a large number of factors — to describe operating environments (e.g., terrain, weather, light, background, clutter) and engagement factors (e.g., force types and ratios, relative geometries and velocities, tactics and countermeasures). The total number of factor combinations is exceedingly large.

Often individual data observations for DoD systems are precious commodities — rarely occurring naturally (e.g., in regularly scheduled training exercises) and extraordinarily expensive to obtain from dedicated large-scale system tests (e.g., $1 million *or more* per data outcome). This naturally provokes the fundamental question of whether it is prudent at all to even attempt to contrast M&S results to "real" data. A common argument that is raised all too often is that 10-30 data replicates would be needed to undertake such a statistical comparison for any *individual* set of factor combinations of interest. To cover any meaningful portion of the complete factor space therefore would require a prohibitively large sample size. Indeed, often resource and budget constraints limit the total number of potential testing opportunities to be of the neighborhood 10-50.

The viewpoint taken in the preceding argument essentially is that a completely self-sufficient experiment must be conducted at each design point in the factor space. Modern statistical design of experiment (DOE) principles eschew this perspective and can be utilized to more efficiently allocate meager testing resources across the M&S factor space: Sacks et al. (1989a, 1989b), Morris (1991), Currin et al. (1991), and Saltelli et al. (1993, 1999). For purposes of this paper, however, we only assume that a rationale subset of the factor design space has been prescribed for investigation (e.g., relatively homogeneous in both a uniformity and completeness sense). Such an experimental design generally can be constructed even when a comprehensive sensitivity study of M&S outputs is lacking. Ideally, statistical efficiency should not be the sole concern. A practical emphasis should be placed on those regions of the factor space that are most likely to be encountered in tactical conditions, and on those specific circumstances that stress touted performance enhancements and new capabilities.

Likewise, this paper does not address the related question of how many data observations, and associated sets of M&S predictions, must be accommodated. We could speculate, based purely on personal intuition, that something like 20-40 data outcomes might constitute a universal minimally acceptable sample size. Beyond that admittedly simplistic and naive "rule", additional detailed analyses would need to be undertaken for each individual M&S program of interest. For example, simulation studies could directly examine statistical power properties across particular relevant families of alternative hypotheses (that characterize departures between M&S and real world results). It should be acknowledged, however, that in many practical circumstances, especially within DoD, a relatively meager maximum available sample size likely will be prescribed — attributable not to statistical considerations, but rather to the unyielding impact of time, budget, and resource constraints.

Since data are sparse and at a premium, replications (e.g., repeat tests under identical sets of experimental factors) are considered to be secondary to the notion of expanding the coverage of the entire factor space. We thus assume, for simplicity of exposition, that observed data contain no replications.

Figure 1 illustrates the nature of the analytical problem that confronts us. Each box corresponds to a single combination of factor settings. (Only 8 combinations out of a total sample size of N are depicted.) Each small dot within a box represents a single predicted value obtained via M&S, with the mean value being represented by the large circle. In this particular example, the predictions, as well as the associated observed data outcomes, happen to be bivariate. They could just as well be univariate or of higher dimensionality. The large star denotes the value of the single data observation obtained with the same factor settings as input to the M&S runs.

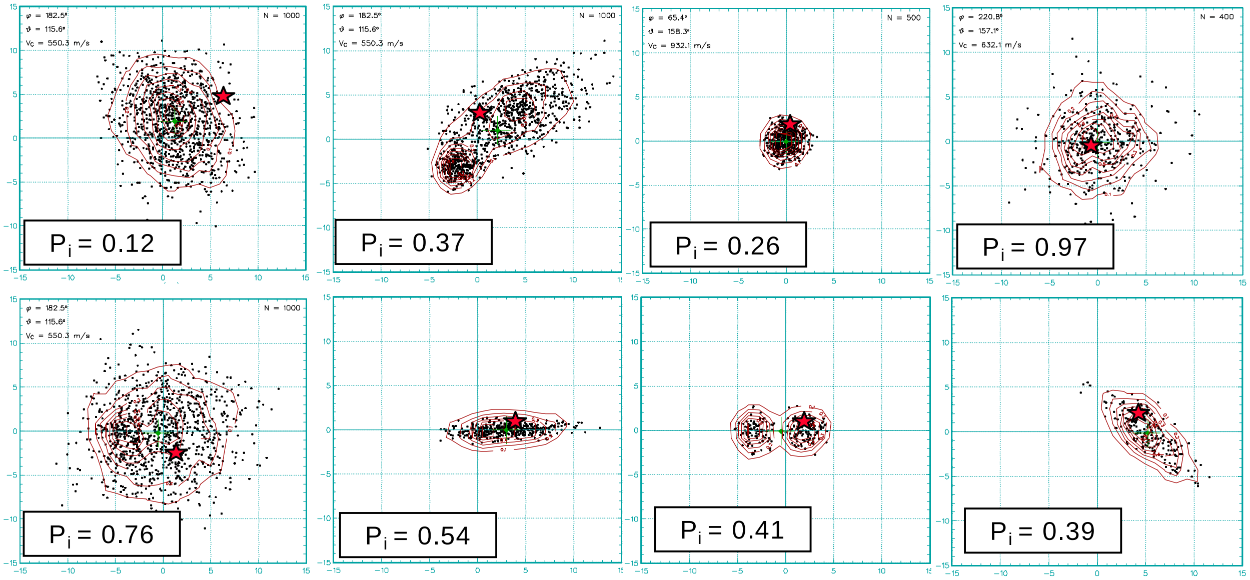

Figure 1. Ensemble Distribution of Tail Probabilities

Observe that the distributional characteristics of the individual M&S "clouds" differ dramatically. Some are tightly congested while others are more dispersed. Some are centered at the origin, while others are not. Most are roughly ellipsoidal in shape, but the orientation is not at all constant. One "cloud" is actually bimodal.

To address the issue of statistical consistency between the M&S results and the observed data, what "standard" statistical methodologies might be applicable? Based on limited personal experience, M&S practitioners seem to rely on simplistic hypothesis testing procedures. For example, a given M&S "cloud" (or "distribution") can be reduced to a single summary value (either the mean or median), with the entire collection of *N* M&S summary points being contrasted to their corresponding *N* data observations via *t*-tests. (We will use " *t*-test" to denote either the standard univariate rendition, or the Hotelling T^2 extension to multiple dimensions.)

Two-sample *t*-tests have been used for this purpose, but they do not account for the wide variability of predicted values across the factor space. The estimated intrinsic variance tends to be inflated, making it easier to accept the null hypothesis (i.e., "validate" the M&S) and more difficult to reject the null hypothesis (i.e., "invalidate" the M&S). One-sample paired *t*-tests, conducted on the *N* differences (e.g., observed outcome minus associated M&S mean) are clearly more preferable, but they still suffer from some formidable shortcomings.

First, they focus entirely on difference in means while ignoring other distributional aspects such as variability. Thus, as long as the individual "cloud" means and corresponding data outcomes are fixed, *t*-test conclusions are invariant to the size of the M&S "clouds". For example, if all of the M&S "clouds" are centered at the origin, then uniformly doubling/halving their sizes leaves the computed value of the *t*-statistic unaffected. But in the limit as the "cloud" size is uniformly increased/shrunk to some value considerably larger/less than the maximum/minimum of the *N* observed data outcomes, one certainly would expect that conclusions about M&S validity should be affected!

Other shortcomings of *t*-test approaches are that they treat distinct combinations of factor settings as being statistically equivalent (e.g., all true differences between M&S means and observation means are assumed to be normally distributed with constant moments), and they make no use of any information related to the particular factor setting values themselves. Similar comments apply to nonparametric approaches (excluding, of course, the standard normal distribution assumption). Regression-based approaches could be pursued to attempt to incorporate factor value information, but they still must contend with the other obstacles reported above. In addition, it may be difficult to find linear regression models or other simple regression formulations that accurately depict the influence of the factors on the M&S results.

FISHER'S COMBINED PROBABILITY TEST

For each of the boxes in Figure 1 (i.e., M&S predictions and solitary test outcome for a common specific combination of factor values) one can ask the question "How rare is the observed data outcome relative to the empirical reference distribution established by the complete set of M&S results?" The most direct quantification is in terms of a *p*-value derived from the empirical reference distribution itself — either in terms of the proportion of M&S values that is further away from the center of the "cloud" than the observed data point is, or a similar probability obtained by generating contour plots for the M&S values.

Whatever the mechanism, one can construct and associate a single *p*-value with each box depicted in Figure 1. The perspective taken here is the most common one in which the major validation issue is whether M&S results are conservative relative to observed outcomes (Figure 2). Expressed, admittedly loosely, in the language of hypothesis testing, we have

$$H_0\text{: M\&S distribution = "test" distribution,}$$

$$H_1\text{: M\&S distribution < "test" distribution.}$$

One could instead focus on whether the M&S predictions tend to be pessimistic, e.g., the "clouds" exhibit an unwarranted excessive variability. All that would be required is to reverse the roles of p and $1\text{-}p$, and to rephrase H_1 accordingly. But such an emphasis would be extremely unusual. A more plausible concern would be to simultaneously guard against the M&S being *either* optimistic *or* pessimistic, i.e., to take a two-sided approach. Again the generalization is immediate:

$$p \rightarrow 2p \text{ when } p < 0.5,\ p \rightarrow 2(1\text{-}p) \text{ otherwise.}$$

An alternative two-sided procedure, typically more powerful, is to conduct two distinct one-sided tests each utilizing a significance level one-half of the nominal prescribed level.

Under any formulation of the null hypothesis, the *p*'s are uniformly distributed on the unit interval [0,1] and the transformed variables $-\ln(p)$ adhere to the exponential distribution. Summing over the N observations, the test statistic $X = -2 \sum \ln(p)$ follows a chi-square distribution with $2N$ degrees of freedom. The null hypothesis is rejected for sufficiently large X.

This is precisely Fisher's combined probability test (Fisher 1932), originally introduced to assimilate the statistical results from multiple related, but not identical, experiments sharing a common null hypothesis (especially when the sample size and/or statistical power for each experiment is small). In our context, each comparison between M&S distribution and observed test outcome (i.e., each box in Figure 1) corresponds to a single "experiment". The same H_0 applies across all our experiments, which, since they involve completely different combinations of factor settings clearly are not identical. Finally, are test sample sizes, all equal to 1, obviously are small.

One well-known property of Fisher's methodology is that a solitary small *p*-value can by itself lead to rejection of the null hypothesis — even when all N-1 of the remaining *p*-values are nominally large. Thus one "outlier" value can dominate completely. This has motivated the construction of a number of alternative more tempered meta-analysis procedures for *p*-value amalgamation: Folks (1984), Rice (1990) and Olkin (1995). Nonetheless, we continue to endorse application of the traditional Fisher procedure as it forces us to confront directly the issue of whether a single test data outcome should "invalidate" the M&S under scrutiny. In addition, it should provoke the discussion of how and why "outlier" *p*-values resulted, whether and with what fidelity the underlying physical causes are represented within the M&S, whether other tested factor combinations logically could generate similar "outliers", etc.

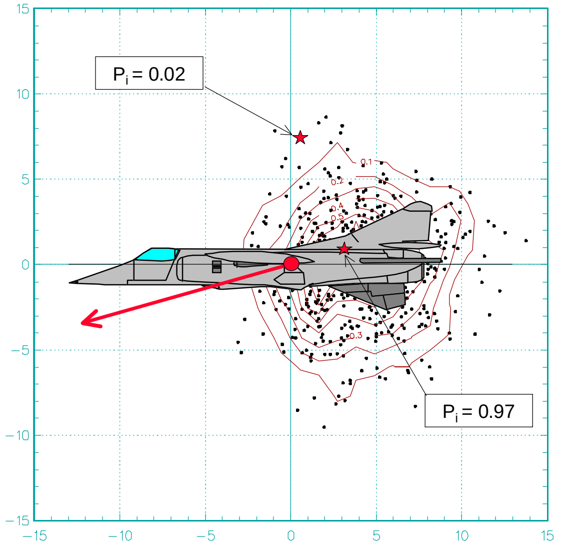

Figure 2. One-Sided "Tail" Probability

GOODNESS-OF-FIT PROCEDURES

Fisher's combined probability test can be interpreted as a goodness-of-fit (GOF) procedure checking observed *p*-values for consistency with a uniform distribution, one that attaches extra weight to skewed departures away from H_0. Taking this perspective, there are a number of other standard GOF methodologies that could be utilized to test H_0. For instance, the Kolmogorov-Smirnov test is based on the maximum difference between the empirical and theoretical cumulative distribution functions (cdf's) for the *p*-values. Other tests, such as the Anderson-Darling, Cramér von Mises, Kuiper, Watson, and "C", rely on various integrated differences of the two cdf's.

One naturally would reason that these classical GOF procedures are superior to Fisher's test for some classes of alternative non-uniform hypotheses that do not focus on a preponderance of small *p*-values. Imagine, for example, a clustered concentration of *p*-values all in close proximity to 0.5. Clearly this is a non-uniform pattern that should be readily detectable by omnibus goodness-of-fit tests. Simulation studies, summarized in Section 5 below, support this expectation. Fisher's procedure, however, is completely insensitive to such circumstances, as, by construct $X/2N = \ln(2) < 1$ when $p \equiv 0.50$.

STATISTICAL PROPERTIES

When the M&S is "good", i.e., accurately represents reality, we desire to accept H_o with high probability. The merit of any statistical procedure is thus measured by the extent to which its Type I errors reflect prescribed nominal significance levels. Likewise, when the M&S is "bad" we seek to reject H_o with high probability. Higher rejection probabilities correspond to more powerful procedures.

Here we report results from a simulation study aimed at characterizing the statistical properties, i.e., Type I errors and powers, of Fisher's combined probability test, the various GOF tests listed in Section 4, and the one-sample paired *t*-test described in Section 2. The simulation study focused on a univariate setting, with N = 25 test observations each associated with a known M&S distribution of predictions. For simplicity, and to facilitate comparisons of the Fisher test to the *t*-test when the latter is known to be optimally powerful, we took each M&S distribution to be a standardized n(0,1) normal distribution (with mean 0 and standard deviation 1). Note that we could begin with the more realistic situation in which these distributions initially are dissimilar (recall Figure 1). But, since in any simulation study we have the advantage of knowing what the true distributions are, we can always invoke appropriate transformations (e.g., inverse cdf functions) to force the reference distributions to the prescribed n(0,1) form. Within this construct, we represented various alternative hypotheses, i.e., "test data" distributions, by generalized normal distributions (with varying means and variances, and even raised to different arithmetic powers).

To obtain the analog of a single box in Figure 1, no random draws were necessitated to establish the reference M&S distribution and only one random draw was taken from the "test" distribution. In essence we assumed that the modelers could run the M&S infinitely many times and exactly duplicate the theoretical standardized normal distribution. This process was then repeated N = 25 times to generate a complete real-world situation, i.e., all of the paired M&S and "test" data available to support an analytical investigation of M&S validity. Each such "situation" was then replicated a large number of times (typically 100,000) to produce estimates of Type I error and power.

Type I error turns out to be not much of an issue. All of the statistical procedures recovered nominally prescribed significance levels, nearly exactly or, for some of the GOF tests, within acceptable errors due to small sample size effects. Such an outcome would by expected for the Fisher and GOF procedures regardless of how the M&S reference distributions are defined. For the *t*-test, we basically contrived this outcome by prescribing normal distributions.

First we consider two-sided comparisons between the Fisher test and the *t*-test, all for a nominal significancle level of α = 0.05. (Similar results hold for other choices of α and for one-sided tests.) Let $S_i \sim$ n(0,1) and $T_i \sim [n(\mu_i,\sigma_i^2)]^{\rho}i$ respectively denote the M&S (H_o) and "test" (H_1) distributions, i = 1, 2, ... , 25. For those circumstances that the *t*-test was <u>not</u> designed to address, it does not do well. For instance, when the variance is allowed to be *any* value other than the null hypothesis specification of 1, its statistical power (i.e., probability of rejecting H_0) is miniscule — equal to α = 0.05. This holds regrettably even for relatively large σ, e.g., set to 2 or 5. In each case, however, the Fisher test has a power essentially equal to 1.00, i.e., P(Fisher) = 1.00. Similar results hold when the "test" distribution distorts H_o via a simple power transformation. When $T_i \sim [n(0,1)]^3$, P(Fisher) = 0.87 >> P(t) = 0.03. But how does the Fisher procedure fare relative to the *t*-test when H_1 is just a shift in the mean away from H_o, i.e., for those situations for which the *t*-test is by theory optimal? (The *Z*-test actually is optimal since σ = 1 is prescribed, but since N is as large as 25 there is only an insignificant reduction in power for the *t*-test.) It turns out the Fisher procedure is extremely efficient within this class of alternative hypotheses. For the three examples μ_i = 0.1, 0.5, [(i-1)/24], the corresponding (P(Fisher),P(t)) values are (0.07,0.08), (0.63,0.67), and (0.60,0.63), respectively. Taken together, these results suggest the Fisher methodology should always be preferred to the "standard" *t*-test.

How do the GOF procedures compare to the Fisher approach? Results vary with the specific H_1 under examination and the particular GOF test. In some cases one can find a GOF test whose power is similar to that of the Fisher methodology. For instance, when $T_i \sim$ n(0.1,1) we have already observed that P(Fisher) = 0.07. In contrast, P(GOF) ranges from 0.04 to 0.05 for the different choices of the test procedure. If we increase the mean somewhat, say to μ_i = 0.8, both sets of powers increase, but much more so for the Fisher approach: P(Fisher) = 0.94 and 0.18 < P(GOF) < 0.41. To illustrate the effect of changing variances, we considered $T_i \sim$ n(0,1.8^2) resulting in P(Fisher) = 0.94 and 0.63 < P(GOF) < 0.95. These results again suggest the superiority of the Fisher test.

But we should not forget the hypothetical example presented in Section 4, which clearly establishes that the Fisher test is not uniformly more powerful than GOF tests (across all possible H_1 designations). The fundamental characteristic of the H_1 considered in that example was a "compressed" distribution — with "test" observations tending to occur near the middle of the reference H_o distribution established by the M&S. To explore this notion further, we considered two H_1 distributions: $T_i \sim n(0,0.6^2)$ and $T_i \sim n(0.4,0.6^2)$. GOF procedures were markedly superior for the former:

$$\text{P(Fisher)} = 0 < \text{P}(t) = 0.05 << 0.55 < \text{P(GOF)} < 0.83.$$

For the latter, the *t*-test was actually best, although some of the GOF tests were relatively efficient:

$\text{P(Fisher)} = 0.04 < 0.21 < \text{P(GOF)} < 0.38 < \text{P}(t) = 0.41$.

SUMMARY & DISCUSSION

Taken in total, our simulation results suggest that some combination of Fisher and/or various GOF procedures generally outperforms the "standard" *t*-test, with relatively little penalty associated with those circumstances for which the *t*-test is slightly more powerful. For the most common alternative hypothesis, concerned with optimistic M&S predictions, the Fisher procedure is recommended.

From an M&S programmatic perspective, it is important to note that the Fisher and GOF tests accommodate formal statistical comparisons, even when data are sparsely distributed across a highly-dimensional factor space. Thus, the M&S VV&A process can be supported rigorously by statistical quantifications without demanding that a large sample of "test" events be dedicated to any individual point in the factor space.

The degree to which such an evaluation provides meaningful VV&A information depends greatly on the representativeness and extent of the factor combinations that yield "test" data. Ideally, the M&S offers the opportunity to study the sensitivity and importance of various factors, individually and in combination. Such insights should play a central role in judiciously allocating the locations within the factor space that are actually tested. But statistical efficiency should not be the sole concern. A practical emphasis should be placed on those regions of the factor space that are most likely to be encountered in tactical conditions, and on those specific circumstances that stress touted performance enhancements and new capabilities.

As always, the statistical methodologies discussed and endorsed in this paper should not be applied in any automatic fashion, and due attention must be given to the important distinction between "statistical significance" and "practical significance". Statistical conclusions of "inconsistencies" between M&S and "test" results should not in and of themselves lead immediately to a declaration of an "invalid" model. Considerable thought should be given to what "valid" and "invalid" mean within the intended sphere of M&S application. For instance, should a single "outlier" data observation by itself imply that the M&S is "invalid"? The focus should be on ways to better understand and improve the M&S, vice on formal pronouncements of "valid" or "invalid" M&S.

ACKNOWLEDGEMENTS & DISCLAIMERS

Portions of Arthur Fries' research were undertaken at the Institute for Defense Analyses (IDA) under IDA Central Research Project C9016 and under various tasks sponsored by the Office of the Director of Operational Test and Evaluation (ODOT&E) in the Office of the Secretary of Defense (OSD) within the U.S. Department of Defense (DoD). The author gratefully acknowledges appreciate helpful comments and inputs provided by A. Rex Rivolo, IDA, and the numerous supporting analyses conducted by LeAnna Guerin, IDA summer intern.

The views expressed in this paper are solely those of the author. No official endorsement by IDA, ODOT&E, OSD or DoD is intended or should be inferred.

REFERENCES

Currin, C., Mitchell, T., Morris, M., and Ylvisaker, D. (1991), "Bayesian Prediction of Deterministic Functions, with Applications to the Design and Analysis of Computer Experiments," *Journal of the American Statistical Association*, 86, 953-963.U.S.

Army (1997), AR 5-11, *Management of Army Models and Simulations*, Washington, DC.

Fisher, R.A. (1932), *Statistical Methods for Research Workers*, Oliver and Boyd, Edingburgh.

Folks, J.L. (1984), "Combination of Independent Tests," in *Handbook of Statistics, 4, Nonparametric Methods*, P.R. Krishnaiah and P.K. Sen (eds.), North-Holland, New York.

Law, A.M. and Kelton, W.D. (1991), *Simulation Modeling and Analysis*, Mc-Graw Hill, New York.

Morris, M.D. (1991), "Factorial Sampling Plans for Preliminary Computational Experiments," *Technometrics*, 33, 161-174.

Olkin, I. (1995), "Meta-Analysis: Reconciling the Results of Independent Studies," *Statistics in Medicine*, 14, 457-472.

Rice, W.R. (1990), "A Consensus Combined *P*-Value Test and the Family-wide Significance of Component Tests, *Biometrics*, 46, 303-308.

Sacks, J., Schiller, S.B., and Welch, W.J. (1989a), "Designs for Computer Experiments," *Technometrics*, 31, 41-47.

Sacks, J., Welch, W.J., Mitchell, T.J., and Wynn, H.P. (1989b), "Design and Analysis of Computer Experiments," *Statistical Science*, 4, 409-435.

Saltelli, A., Andres, T.H., and Homma, T. (1993), "Sensitivity Analysis of Model Output: An Investigation of New Techniques," *Computational Statistics and Data Analysis*, 15, 211-238.

Saltelli, A., Tarantola, S., and Chan, K.P.-S. (1999), "A Quantitative Model-Independent Method for Global Sensitivity Analysis of Model Output," *Technometrics*, 41, 39-56.