CHANCE
Using Data to Advance Science, Education, and Society

Adaptive Testing of DoD Systems with Binary Response

Douglas M. Ray and Paul A. Roediger

The U.S. Army Armament Research, Development, and Engineering Center (ARDEC) is the Army's center for lethality, supporting the majority of armament systems for the U.S. military. ARDEC's professional statisticians, data scientists, and statistical engineers collaborate and consult with integrated product engineering teams on a wide variety of projects and programs spanning the product lifecycle, providing "cradle to grave" analytics support.

Design of Experiments (DoE) is a statistical data collection methodology that involves the systematic selection of factor-level combinations to best support a credible empirical model. This mathematical model, which translates input settings into output predictions can be used for screening (to identify factors of interest), comparison, characterization, and optimization. Some unique challenges arise when the response data being collected are binary, but those challenges can be addressed effectively using modern DoE techniques. Approaching test activities as designed experiments, rather than pass-fail events, results in the collection of richer information for decision-making and insights about the system or process under study.

Imagine we are working for a company that is developing ruggedized smartphone cases for the new "xPhone" and there are several prototype design candidates—let's say two different material densities at two thicknesses, giving us four total combinations to evaluate (let's call them A-1, A-2, B-1, and B-2). How could we evaluate the performance of the different designs—both relative to one another, and in terms of their ability to protect the phones at a specified drop height (say, 5 feet)? In this hypothetical example, the response data are binary, meaning that when we drop a phone, it either breaks or survives the fall.

One approach often employed in product verification is a "zero-failure" reliability test, which is derived from the binomial distribution. As Jovanovic and Levy say in "A Look at the Rule of Three," simplification of a special case of the zero failure reliability test is the "Rule of Three" for 95% confidence at some specified reliability level. For example, if the specified reliability we seek to demonstrate in testing is 90% (R = 0.90) with a 95% Lower Confidence Bound (LCB), then the Rule of Three approximation requires using 3/(1-R) = 3/0.10 = 30 samples dropped at 5 feet for each smartphone case design.

This will be an expensive test, since it will result in potential destruction of many of the test samples. If our reliability requirement were 99%, 99.9%, 99.99%, or greater (instead of 90%) at a 95% confidence level (n = 300, 3,000, 30,000, respectively), then the sample size quickly becomes infeasible for most applications. In addition, we would not have gained much insight about each design in terms of safety margin; in other words, we will have some limited information about performance at 5 feet, but no insight about breakage probability at other stimulus levels. Also, the test results may be ambiguous and not provide a clear path for decision-makers to select an optimal product design.

For example, what if the tests of all four designs result in zero breakages? An ideal outcome in some respects, but we will not be able to conclude whether any one design is better than another. Even if some of the four designs experienced a few breakages out of the 30 tests, we would be hard-pressed to establish a significant difference with statistical confidence. Often, all that can be learned when we design a test with only a pass/fail criterion is whether the product passed or failed the test.

An alternative to this approach is known in reliability engineering as a “Probit Test,” as described by Prairie. This would involve spreading the 30 samples (or however many samples we can acquire for testing) across multiple levels of the drop height stimulus, and then analyzing the resulting test data using Binary Logistic Regression or Probit Regression to develop a predictive model for drop height vs. Probability of Survival (non-breakage) for each of the four prototype design variants (see Figure 1).

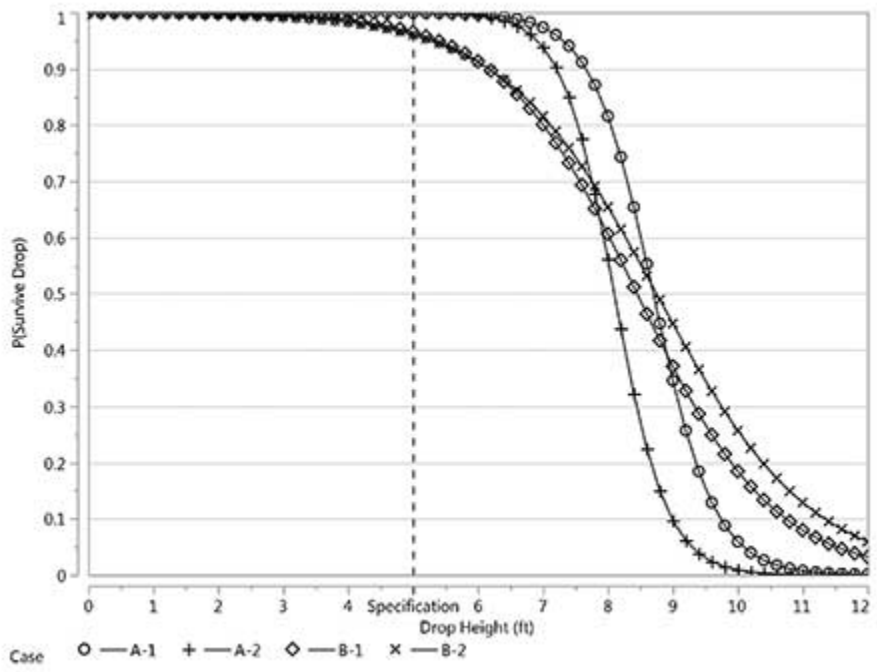

Figure 1. Drop height vs. P(survive drop) model for four candidate designs.

Assuming we tested in the correct stimulus region and that we collected high-quality data, this will provide much-more-useful information than the previous approach.

However, this approach assumes that we have some prior knowledge of a model—where the center of the curve lies, and how wide it is in terms of stimulus values corresponding to the all-break and all-survive thresholds. If we don’t already know where this region lies, then it is

possible to select stimulus levels that provide little to no useful information—a difficult position to defend when reporting test results.

This situation can arise if the analyst selects: (1) stimulus levels that are spread too far apart from one another; (2) stimulus levels that are too close together; or (3) a stimulus range that is too far above or too far below the stimulus region where the "true" curves occur. For example, if we had decided to spread our 30 samples across five drop heights, with six samples at each drop height centered at our spec of 5 feet, in 1 foot increments (3 ft, 4 ft, 5 ft, 6 ft, and 7 ft), we may end up with a highly unbalanced data set. Based on where our curves lie in Figure 1, we might see two design variants (B-1 and B-2) with only one or two failures out of 30 samples (occurring at 7ft), and the other two (A-1 and A-2) with zero failures.

Faced with this situation, regression analysis would not be able to provide us with a useful predictive model. The Probit Test approach is much more desirable when compared to the first if we have some prior knowledge of how the product will behave, but the generation of useless test data using this approach is a common occurrence.

A third approach would be an adaptive sensitivity test method. The Department of Defense has a rich history with adaptive sensitivity testing, going back to the "Bruceton Up-Down" method developed in 1948 by Dixon and Mood. Adaptive sensitivity testing is similar to the Probit Test in that it seeks to generate a set of data that supports a predictive regression model for the probability or quantile associated with various stimulus levels. Where it differs is that it adapts to the results as each data point is generated and analyzed, thereby significantly reducing the risk of generating useless or unbalanced data sets.

Modern sensitivity test methods (such as Neyer's D-optimal test or Wu and Tian's 3pod) require the use of a computer to execute real-time calculations to select stimulus levels for each sample based on previous response data.

As Kiefer describes, D-optimality is a computational design generation method commonly used in DoE as an alternative to factorial-based classical designs. The D-optimality criterion seeks to maximize the determinant of the Fisher information matrix through optimal placement of points within the design space. However, the distinction in adaptive sensitivity test approaches is that this algorithm is executed after each individual sample has been tested.

Adaptive sensitivity tests begin with an initial guess for the mean (where the curve will be centered, or the stimulus value corresponding to the response 50th percentile) and the standard deviation, or sometimes the stimulus values associated with the upper and lower tails (0.95 response probabilities).

Typically, the algorithm's determination of the first several stimulus levels in testing is logic-based and seeks to identify a stimulus region that results in a mix of responses, often referred to as the zone of mixed results. At this point, both Neyer and 3pod implement a D-optimal procedure, recalculating the D-optimal point after every stimulus value, with each response recorded.

The interesting thing about this test approach is that even if the initial guess for the center or range (or both) of stimulus values is wrong, the test will adjust and adapt based on the previous responses, and will usually obtain a data set with overlap in five to 15 samples, which can then be analyzed using Binary Logistic Regression to develop a tentative predictive model.

Using D-optimality after achieving overlap means that the algorithm's placement of the rest of the points is optimized to provide the greatest improvement in predictive precision. Usually this means that the points will be alternately placed near the 17th and 83rd quantiles of the curve.

The 3pod procedure stands for three-phase optimal design of sensitivity experiments. It is similar in some ways to Neyer's D-optimal test procedure, mainly in the execution of phase 2, but contains a third phase, the Robbins-Monro-Joseph (RMJ) procedure, with what Wang, Tian, and Wu described as the more-recently developed skewed RMJ option, which is a nonparametric quantile estimation method. A unique aspect of 3pod is flexibility and modularity: The three phases of 3pod can be configured to meet specific sensitivity test goals.

Using a modern adaptive sensitivity test method minimizes the risk of generating useless test data, making for more defensible testing. With 20–30 samples per smartphone case design, we can generate a rich data set, providing the ability to develop relatively precise models for each prototype design, which enables us to predict the probability of smartphone survival (reliability) at any stimulus level of interest at a specified confidence level.

Figure 1 shows that the means of design A-1 and B-2 are the largest in terms of drop height (i.e. resistance to breakage from greater heights). In fact, they cross very close to their 50th percentiles, and B-2 is slightly better, based on the mean alone. However, the slope of the curve is determined by the standard deviation, where the smaller standard deviation (less system variability) results in a steeper curve. As the standard deviation approaches zero, the curve would approach a step function, which is often just as important as the mean.

With this in mind, design A-1 is the clear winner. The mean is nearly as high as B-2, but the standard deviation is smaller, leading to better overall performance up to approximately 9 feet. This means that at our specification of 5 feet, we can see that A-1 has much better reliability than B-2. Figure 2 shows the 90% two-tailed confidence intervals (or 95% lower confidence bound) for the prediction model for probability of survival, or reliability. At the 5 feet drop height specification, the model prediction point estimate is 0.9996, while the 95% LCB for reliability is 0.9831.

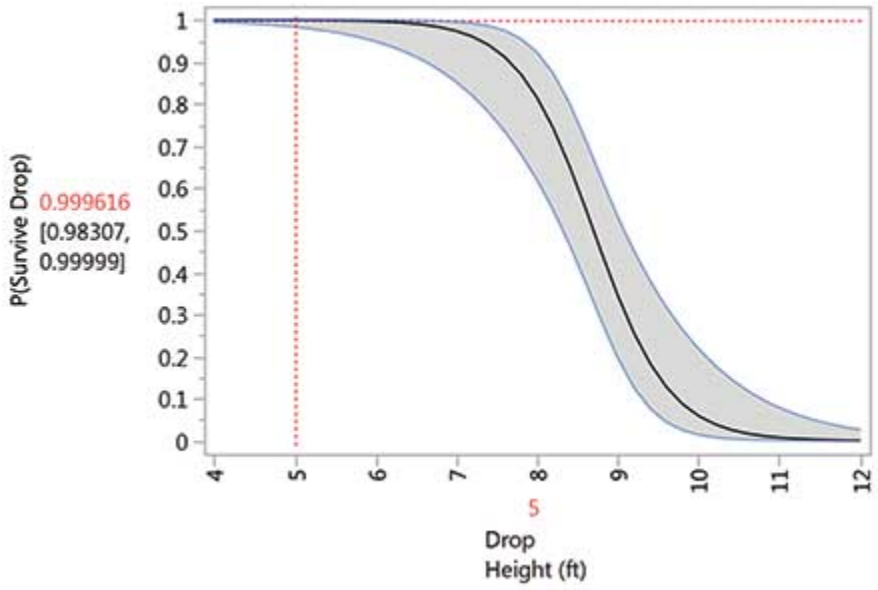

Figure 2. Drop height vs. reliability for a 90% two-tailed confidence interval (95% LCB).

Using a modern adaptive test approach translates to improved decisions in terms of system reliability and performance, increased fidelity to the true design margins relative to product specifications, and minimized risk of tests which result in generation of low-utility data sets, while potentially reducing test quantities by an order of magnitude or more, thereby saving substantial test cost, schedule, and hardware.

One drawback of adaptive sensitivity test methods is that currently, procedures are only available for experiments with a single factor. There might be the opportunity for an extension to the adaptive sensitivity test that can also treat the material thickness and material density of the smartphone case as two design factors, while the stimulus factor—drop height—is the "noise" factor. This becomes a three-factor, robust-parameter DOE problem, executed adaptively.

As Ray, Roediger, and Neyer suggest, adaptive test approaches that can handle multiple design factors, multiple stimulus variables, or some combination of the two would be a welcome tool in the analyst's toolkit.

The U.S. Army ARDEC statisticians have developed a customized sensitivity testing implementation in R called Gonogo. With it, experimenters are equipped to perform four adaptive protocols in real time: the Neyer test; Wang, Tian, and Wu's newly revised 3pod2.0; and the historically important Bruceton and Langlie tests as shown in the DoD's MIL-STD-331D. The powerful and flexible 3pod2.0 test approach has been adapted to dozens of armament-related systems over the past several years. By including the Neyer test, this comparably powerful procedure is now readily accessible without having to acquire a commercial license for SenTest.™

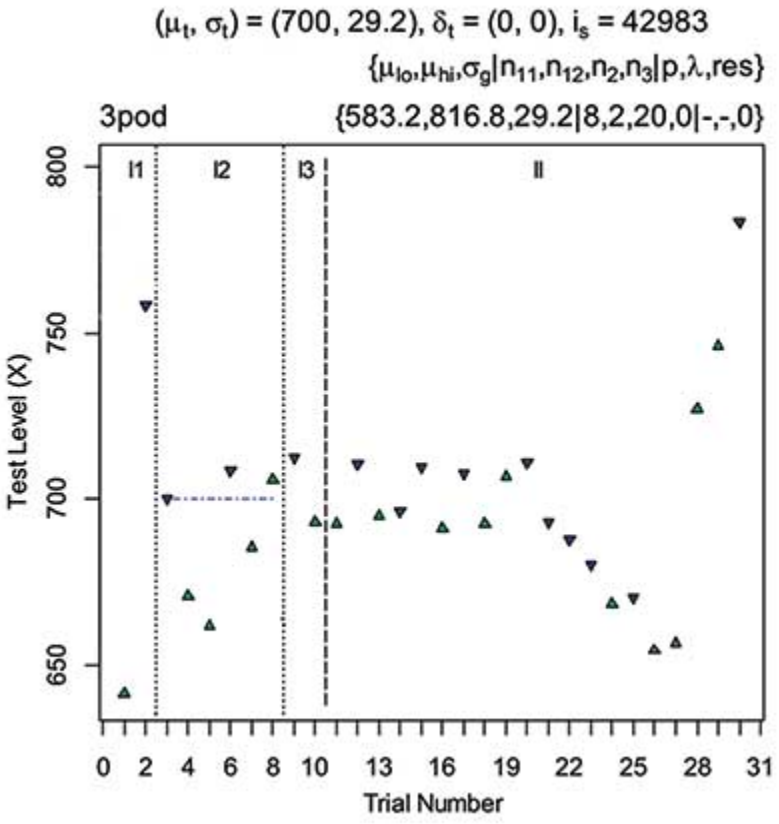

Figure 3. Test 42983 (a randomly generated 3pod test).

Gonogo provides graphics and tabular outputs, including sequential plots of the data and, when appropriate, confidence intervals computed via Fisher Matrix, GLM, or Likelihood Ratio methodologies. It includes a fair amount of documentation to give the user fast access to the procedures described.

Gonogo also includes a simulation suite of R functions and graphics. With it, users can study the performance of any of Gonogo's four adaptive procedures, under various conditions of interest. The following is one such example.

Example: Suppose a study is to be conducted on the electrical sensitivity of an initiator being developed by the U.S. Army. The purpose of the study is to reduce the size and cost of the item's componentry by reducing the nominal mean (V_{nom}) and variance (σ^2_{nom}) of initiation voltage. Qualification testing of this next generation initiator will be accomplished via 30 shot sensitivity tests, either Neyer or 3pod, having initial starting values of $\mu_{min} = V_{nom} - 4\sigma_{nom}$, $\mu_{max} = V_{nom} + 4\sigma_{nom}$ and $\sigma_{guess} = \sigma_{nom}$. The evaluation of all tests will depend on the estimation of two quantities [12]:

1. Maximum No Fire Voltage (*MNFV*), with the voltage having at most a 0.001 probability of firing (at 95% confidence), and

2. Maximum Allowable Safe Stimulus (*MASS*), with the voltage having a 1 in 1 million probability of firing (point estimate)

The proposed qualification criterion will be: *Max* (*MNFV*, *MASS*) > 500 volts.

The design team would like to know:

A. If V_{nom} can be reduced to 700 volts, how small does σ_{nom} have to be to ensure a 95% probability of qualifying the new initiator?

B. Which test, Neyer or 3pod, is better suited for the job?

While a Gonogo simulation generates a single test at a time, generating many at a time will require writing your own customized, and usually, brief R script. Table 1 is one test we came across in our first simulation study—of 2,000 3pod tests of size 30 for $\sigma_{nom} = 29.2$.

Gonogo includes a handy function to compute confidence intervals—one bounding probability for a given stress level, and the other bounding stress for a given probability. For the 3pod test above, it returns MNFV and MASS estimates in the following format (see Table 1).

A Confidence Interval Calculation for Test 42983					
Stress (q)			Probability (p)		
q_{lo}	q	q_{hi}	P_{lo}	P	P_{hi}
249.7810	**455.2194 (MASS)**	660.6579	0.000000	**0.000001**	0.213193
408.6442 (MNFV)	541.5460	674.4479	0.000000	**0.001000**	0.298157

Table 1—90% 2-sided Confidence Interval Estimates (Computed via the GLM Methodology)

The 2,000 MNFV and MASS estimates are plotted in Figure 4.

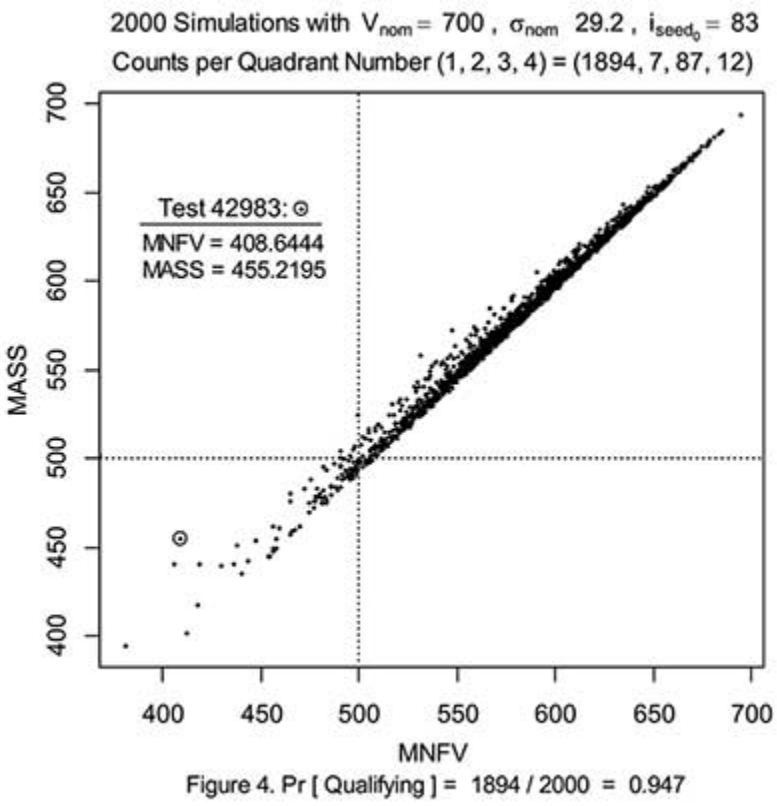

Figure 4. Pr [Qualifying] = 1894 / 2000 = 0.947

Repetition of the process (that led to the .947 estimate for σ_{nom} = 29.2) was done 11 more times, yielding estimates having an aggregate mean of **.9542**. The entire 3pod case was completed after repeating this procedure for other σ_{nom}.

A Neyer case counterpart subsequently completed in a similar manner allowed eventual assembly of Table 2.

	$P = Pr[Qualifying \mid V_{nom} = 700, \sigma_{nom}]$									
σ_{nom}	28.4	28.6	28.8	29.0	29.2	29.4	29.6	29.8	30.0	30.2
P_{3pod}	.9609	.9614	.9582	.9572	**.9542**	.9490	.9501	.9458	.9420	.9395
P_{Neyer}	.9611	.9581	.9566	.9545	.9532	.9497	.9475	.9455	.9407	.9388

Table 2—Each P is an Average Obtained from 12 Simulations (2,000 Tests of Size 30 for Each)

This simple summary abated the team's biggest concern: that limitations imposed on σ_{nom} by our analysis could far exceed the new item's process control capability. Table 2 also reveals that the two test protocols are comparable for this application.

Sensitivity testing has a wide variety of applications in the DoD, including testing energetic materiel to determine impact sensitivity thresholds, precision guided munition fuze component

reliability where the voltage vs. P[initiation] predictions can then be rolled up to system-level reliability models via block diagrams, ballistic testing of soldier protective armor (by hand-loading cartridges to varying velocities), and ammunition penetration testing (often referred to as V50 testing).

The need for this method of testing in the DoD is driven by the types of systems in use, and the sometimes-destructive nature of the testing. However, sensitivity testing is still underused in private industry. One reason may be that the applications are less obvious, and none of the techniques have been adapted in DoE textbooks or commonly available software implementations.

A few application areas beyond the DoD use variations of some of the approaches discussed here, including psychoacoustics (e.g., hearing tests with different volume "beeps"), and dose-ranging in pharmaceutical research. Recently, the U.S. Army ARDEC consulted with Sartorius Stedim Biotech (an international supplier of equipment and services to the biopharmaceutical industry) on implementing modern sensitivity testing procedures, and they have already applied 3pod to a variety of products using Gonogo.

In one example, the adaptive process enabled their engineers to assess the fragility of containers protecting high-value product quickly while consuming a reduced number of prototypes. Other applications now being investigated include estimating lifetimes for parts that cannot be inspected during use, and improved methods for determining the low temperature characteristics of plastic film materials.

Although classical Design of Experiments has origins in agricultural applications, modern DoE is a powerful and flexible family of statistical tools with specialized techniques that can be adapted to a wide variety of applications: Mixture DoE for chemical formulations; Space-Filling Designs for computational modeling and simulation experiments, and Uncertainty Quantification (UQ) studies (see Sacks, Welch, Mitchell, and Wynn); optimal designs for complex real-world problems, such as test constraints or hard-to-change factors; and covering arrays for software testing (see Dalal and Mallows), to name a few.

DoE practitioners understand that it is generally preferred to capture continuous response measurements rather than binary response data, whenever possible. Adaptive sensitivity testing is a powerful tool that belongs in the toolkit of any DoE practitioner in industry, as well as the DoD. When continuous response measurement is not possible, these test methods are the best available tools for dealing with the challenges and risks inherent to binary response data collection when faced with real-world resource constraints.

Currently, statisticians from the U.S. Army ARDEC and Air Force Institute of Technology (AFIT) are collaborating on developing a universally available R package to execute the U.S. Army ARDEC's customized sensitivity testing capability. To obtain the latest documentation and version of Gonogo, visit *https://www2.isye.gatech.edu/~jeffwu*.

Further Reading

Jovanovic, B.D., and Levy, P.S. 1997. A Look at the Rule of Three. *American Statistician 51*(2):137–139.

Prairie, R.R. 1967. Probit Analysis as a Technique for Estimating the Reliability of a Simple System. *Technometrics 9*(2):197–203.

Dixon, J.W., and Mood, A.M. 1948. A Method for Obtaining and Analyzing Sensitivity Data. *Journal of the American Statistical Association 43*:109–126.

Neyer, B.T. 1994. A D-Optimality-Based Sensitivity Test. *Technometrics 36*:61–70.

Wu, C.F.J., and Tian, Y.B. 2014. Three-phase optimal design of sensitivity experiments. *Journal of Statistical Planning and Inference 149*:1–15.

Kiefer, J. 1959. Optimum Experimental Designs. *Journal of the Royal Statistical Society, Series B, 21*:272–319.

Wang, D., Tian, Y.B., and Wu, C.F.J. 2015. A Skewed Version of the Robbins-Monro-Joseph Procedure for Binary Response (PDF download). *Statistica Sinica 25*(4):1,679–1,689.

Ray, D.M., Roediger, P.A., and Neyer, B.T. 2014. Discussion of Three-phase optimal design for sensitivity experiments. *Journal of Statistical Planning and Inference 149*:20–25.

Wang, D., Tian, Y.B., and Wu, C.F.J. Comprehensive Comparisons of Major Design Procedures for Sensitivity Testing. *Journal of Quality Technology* (to appear).

MIL-STD-331D, Appendix G. 2017. Statistical Methods to Determine the Initiation Probability of One-Shot Devices. Washington, DC: Department of Defense.

Neyer Software LLC, SenTest,™ Version 1.0.

MIL-DTL-23659 F. 2010. Initiators, Electric, General Design Specification for. Washington, DC: Department of Defense.

Sacks, J., Welch, W.J., Mitchell, T.J., and Wynn, H.P. 1989. Design and Analysis of Computer Experiments. *Statistical Science 4*(4):409–423.

Dalal, S.R., and Mallows, C.L. 1998. Factor-Covering Designs for Testing Software. *Technometrics 40*(3):234–243.

About the Authors

Douglas Ray, PStat® is the lead statistician at the U.S. Army ARDEC in Picatinny Arsenal, NJ. His work focuses on application of industrial statistics and analytics to armament systems spanning the engineering lifecycle; he is an experienced Design of Experiments (DOE) practitioner, and focuses on statistical quality control and process improvement, reliability data analysis, data mining analytics, and Uncertainty Quantification (UQ)/Probabilistic Optimization of computational models and simulations. Ray holds a BS in applied mathematics and an MS in statistical engineering, and is currently working toward his PhD in systems engineering analytics at Stevens Institute of Technology. Ray is an ASA Accredited Professional Statistician,™ Lean Six Sigma Black Belt, and ASQ Certified Reliability Engineer. He is also a combat veteran of the U.S. Army.

Paul Roediger is a subject matter expert (SME) in mathematics and statistics with UTRS, Inc. He retired from federal service in 2012 as lead statistician at the U.S. Army ARDEC, Picatinny Arsenal, NJ. He holds a BS degree in mathematics and MS degrees in both mathematics and statistics. While with UTRS, Inc. and supporting the U.S. Army ARDEC, he developed a suite of R programs to conduct, analyze, and simulate two of the most-modern sensitivity experiments available—Jeff Wu, et. al.'s 3pod and Barry Neyer's SenTest™ experiments.