IEEE Access
The journal for rapid open access publishing

Received September 23, 2014, accepted November 14, 2014, date of publication December 4, 2014, date of current version December 17, 2014.

Digital Object Identifier 10.1109/ACCESS.2014.2377633

A Comparison of Ballistic Resistance Testing Techniques in the Department of Defense

THOMAS H. JOHNSON, LAURA FREEMAN, JANICE HESTER, AND JONATHAN L. BELL
Institute for Defense Analyses, Alexandria, VA 22311, USA

Corresponding author: T. H. Johnson (thjohnso@ida.org)

This work was supported by the Office of the Secretary of Defense.

ABSTRACT Ballistic resistance testing is conducted in the Department of Defense (DoD) to estimate the probability that a projectile will perforate the armor of a system under test. Ballistic resistance testing routinely employs sensitivity experiment techniques where sequential test designs are used to estimate a particular quantile of the probability of perforation. Statistical procedures used to estimate the ballistic resistance of armor in the DoD have remained relatively unchanged for decades. In the current fiscal atmosphere of sequestration and budget deficits, efficiency is critical for test and evaluation. In this paper, we review and compare sequential methods, estimators, and stopping criteria used in the DoD to those found in literature. Using Monte Carlo simulation, we find that the three-phase optimal design, a probit model, and a break separation stopping criteria are most accurate and efficient at estimating V_{50}, while the three-phase optimal design or Robbins–Monroe–Joseph method should be used to estimate V_{10}.

INDEX TERMS Department of Defense, sequential design, design of experiments, ballistic resistance testing, three-phase optimal design.

I. INTRODUCTION

A variety of fields use sensitivity experiments to characterize the probability of a binary outcome as a function of a stimulus or stress. Department of Defense (DoD) ballistic characterization tests employ sensitivity experiments to characterize the probability of perforating a material or armor as a function of a projectiles velocity. These ballistic characterization tests are essential to understanding the vulnerability and lethality of military equipment and are conducted on systems ranging from body armor to vehicle and aircraft armor. In recent years, ballistic characterization tests were conducted for the Enhanced Small Arms Protection Inserts (body armor plates), the Enhanced Combat Helmet, the new floor paneling of the CH-47F, cockpit armor for the KC-46, and armor paneling of the Joint Light Tactical Vehicle, to name a few.

Ballistic testing is destructive and DoD ballistic testing can be expensive in terms of both test and material costs. Sample sizes are generally limited and ballistic characterization tests are almost always limited to fewer than 20 shots on the armor. For example, in lot acceptance testing of body armor a ballistic characterization test might be completed in as few as 6 shots. In this case, however, previous testing provides a good expectation of the armors performance.

In addition to test size limitations, the DoD has historically preferred ballistic characterization tests that are simple to execute and that require no statistical expertise from the test team. In particular, the velocity of the next shot in a test sequence must be quickly and simply determined on the test range. Modern computing and software provide an opportunity to shift towards techniques based on maximum likelihood estimation, which can now be automated for the test team. These techniques should make more efficient use of limited test resources.

Ballistic characterization testing frequently focuses on a specific percentile. Ballistic characterization tests must therefore provide sufficient data to accurately estimate the percentile of interest. The most common percentile of interest is the velocity at which the projectile has a 50 percent probability of perforating the armor called the ballistic limit or V_{50} of the armor for the particular projectile. Historically, the V_{50} was sufficient to characterize armor. It can also be estimated more precisely and with fewer shots than other percentiles. With modern armors, however, the V_{50} might not be sufficient to characterize an armor and users might be more interested in the velocity at which the probability of perforation is 10 percent, or even lower. Estimating the V_{10} is another reason to explore the efficiency of test techniques based on maximum likelihood estimation.

The two methods most commonly used in the DoD to estimate V_{50} are the up-down method developed by Dixon and Mood (1948) [1] and the Langlie (1962) method [2]. The up-down method is specified in MIL-STD-662F and in several US Army Test and Evaluation Center test operations procedures including those for testing of body armor and combat helmets. MIL-STD-662F provides a general approach for determining the ballistic resistance of various types of armor against small arms projectiles [3]. The Langlie method has been used, for example, for the KC-46 Pegasus, OH-58 Kiowa Warrior, CH-47 Chinook, and the CV-22 Osprey.

Other sequential test methods from the statistical literature include the Delayed Robbins Monroe method [4], the Robbins Monroe Joseph method [5], Neyers method [6], and most recently Wu and Tians Three-phase Optimal Design (3POD) approach [7]. These methods are not commonly used in DoD ballistic characterization testing, but Ray, Roediger, and Neyer (2013) highlight limited success integrating Neyers method into MIL-STD-331C and into software used by the US Army Armament Research, Development and Engineering Center [8].

Numerous studies have compared the sensitivity testing procedures with fairly consistent results. In 1990, Body and Tingey [9] compared sequential methods in the context of armor testing. They compared variants of the Robbins Monroe procedures with Langlies method and the Up and Down method and concluded that the Delayed Robbins Monroe method performed best. In 1994, Young and Easterling [10] investigated methods intended to estimate extreme quantile values and recommended Neyers method over all others. Simpler methods, not considered by Young and Easterling, such as the Bias Coin Design and K-in-a-Row design were compared by Orron [11] in 2006. Orron concluded that K-in-a-Row is superior to the Bias Coin Design in all practical aspects. Most recently, Wu compared 3POD with Neyers method among others in 2012 and recommended 3POD [12].

In this paper we compare different ballistic test methods under limited sample sizes, typical of conditions in DoD testing. Three simulation studies compare test methods in terms of their efficiency and accuracy to estimate V_{50} and V_{10}. The paper is organized as follows. Section two provides an overview of the different sensitivity test procedures. We discuss Neyer's method and the 3POD method simultaneously to highlight similarities and differences. Section three provides a motivating example based on the testing of the Advanced Combat Helmet. Section four outlines the simulation study and results of the comparison. Section five draws conclusions and makes recommendations for improving the current practice of DoD ballistic testing.

II. OVERVIEW OF EXISTING BALLISTIC TEST METHODS

Numerous aspects of a ballistic resistance testing can impact the quality and consistency of the results. These include the laboratory setup, location of projectile impact, obliquity angle, temperature, projectile type, among others. In this paper, we focus on statistical aspects of the test: the stopping criteria, estimators, and sequential methods.

A. STOPPING CRITERIA

Stopping criteria specifies when testing is complete and indirectly dictates the number of shots in a test. The first stopping criteria employed in this paper is the "3 and 3" stopping criteria (33SC). Data collection under 33SC ceases when 3 non-perforations and 3 perforations are recorded within a specified velocity range. Another commonly used stopping criteria in DoD tests is the "5 and 5" stopping criteria. Use of the 33SC in this paper is motivated by widespread use of this criteria in Lot Acceptance Tests (LAT) across a variety of Personal Protective Equiptment (PPE).

The second stopping criteria we consider is the "break separation" stopping criteria (BSSC). Separation is an undesirable characteristic of a ballistic limit data set because it prevents a unique maximum likelihood estimate of the generalized linear model without applying a correction to the model. Separation is "broken" and "overlap is achieved" when the following criteria is met.

1. The highest velocity of the non-perforated shots is greater than the lowest velocity of the perforated shots.
2. The lowest velocity of the non-perforated shots is less than the lowest velocity of the perforated shots.
3. The highest velocity of the perforated shots is greater than the highest velocity of the non-perforated shots.

BSSC is typically used with maximum likelihood estimators. It is also used in the DoD, but in our experience is less common than 33SC. Both 33SC and BSSC stopping criteria, under reasonable sequential test procedures, result in small test sizes, which is why it is important to compare them in this paper focusing on DoD ballistic resistance testing.

B. ESTIMATORS

Estimators are the mathematical expressions used to calculate V_P. We address two estimators: probit maximum likelihood estimator (Probit-MLE) and the arithmetic mean estimator (AME).

In a probit model, the response of an armor target to a ballistic projectile can be characterized as perforation or non-perforation. Let $y_i = 1$ or 0 denote the binary outcome of the i_{th} shot, perforation or non-perforation, where $i = 1, 2, 3 \ldots N$ are the first, second, third, and final shot, respectively. Let $F(x_i)$ denote the probability that $y_i = 1$ for the velocity of the i_{th} shot. The location-scale probit model, which we use in our paper to characterize the ballistic resistance of armor, is $F(x, \mu, \sigma) = \Phi((x-\mu)/\sigma)$, where Φ is the standard normal cumulative distribution function [13]. We define x_P to be the P_{th} quantile of the distribution, where $F(x_P) = P$. In this formulation, μ is the estimator of V_{50}, is estimated using maximum likelihood estimation, and is referred to as the Probit-MLE estimator. Despite the plethora of alternatives, we chose a Probit model over other generalized linear models since work by NIST showed that among a probit, logit, and complementary log-log model, "...none of them distinguished itself from the others in terms of armor performance estimation [14]." Additionally, we do not consider the impact of model misspecfication in this study. Wu and Tian show that incorrectly specifying that generating data from a logistic model and fitting the probit model results in larger bias and mean square error in the estimates as expected [7].

The second estimator we use is AME. AME is widely used in military testing. It is typically used with 33SC, where V_{50} is calculated by averaging together the three highest velocity non-perforations with the three lowest velocity perforations that are within a certain velocity range. AME is only used to calculate V_{50}.

C. SEQUENTIAL METHODS

There are seven sensitivity test methods investigated in this paper. The methods are chosen based on their prevalence in military armor testing, ease of implementation, and overall effectiveness at estimating V_{50} and V_{10}. The methods compared are: the Up and Down Method (UD), the Langlie Method (LM), the Delayed Robbins Monroe Method (DRM), Wu's three-phased approach (3POD), Neyer's Method (NM), the Robbins Monroe Joseph Method (RMJ), and K-in-a-row (KR).

UD, also called the Bruceton Method, was developed in 1948 by Dixon and Mood [1]. First, the experimenter guesses (σ_G, μ_G) the true parameters (σ_T, μ_T) of the ballistic response curve. The step size, d, is set equal to σ_G. The velocity of the first shot is equal to μ_G. The velocity of the ensuing shots is equal to $x_{i+1} = x_i - d$ if x_i resulted in a perforation, and equal to $x_{i+1} = x_i + d$ if x_i resulted in a non-perforation. UD is typically used with 33SC and AME because UD is designed to converge to V_{50}; hence, three non-perforations and three perforations are likely to occur near V_{50} from which the arithmetic mean of these velocities will provide a reasonable estimate of V_{50}.

KR was introduced by Gezmu in 1996 [11]. It is not typically used in DoD testing, but was included in this study due to its simplicity, and its ability to estimate V_{10}. KR is similar to UD in that is also uses a constant step sizes (equal to σ_G). In KR, first, the experimenter sets K, which is an integer value typically equal to 2, 3, 4 or 5. On a given shot, if the projectile does perforate the armor, the velocity of the ensuing shot is decreased by d. If the projectile does not perforate the armor K times in a row, then the velocity is increased. If the projectile does not perforate, and it has not perforated K times in a row, then the velocity remains the same. KR has the effect of placing shots near the lower extreme of the response curve, which is useful for estimating values such as V_{10}. The experimenter can optimize KR to estimate a particular value of V_P by selecting the proper integer value of K prior to the beginning of test. KR has been shown to converge towards the P_{th} quantile, where $P = 1 - (1/2)^{1/k}$. For values of K equal to 2, 3, and 4, this equation yields values of P equal to 0.29, 0.21, and 0.16.

A disadvantage of the constant step size employed by UD and KR is that if σ_G is too large, either method will not be capable of converging to V_{50} with sufficient resolution. To fix this problem, the Robbins Monroe method decreases the step size as the test continues using the equation:

$$x_{i+1} = x_i - \frac{1}{N} C(y_i - P) \tag{1}$$

The step size decreases with the total number of shots taken (N). In this equation, C is a constant. Large values of C lead to large step sizes, while small values lead to small steps. An optimal value of C was found to be equal to the slope of the response curve at the P_{th} quantile value and is calculated as $C_{opt} = 1/F'(x_P)$ [15]. In cases where the first shot is taken far from the true value of V_{50}, many of the first tests will be wasted closing in on the mean. The Delayed Robbins-Monroe test attempts to fix this problem by keeping the step size constant until mixed results (at least one perforation and one non-perforation) are obtained [16].

Joseph [5] recognized that the Robbins Monroe procedure, originally developed for continuous data, is not well suited for binary data. Joseph exploited the fact that the variance of binary data is different from continuous data and used this idea to create RMJ. The method begins by taking the first shot at $x_1 = \Phi^{-1}_{\mu_G,\sigma_G}(P)$, where $\Phi^{-1}_{\mu_G,\sigma_G}$ is the inverse of the cumulative normal distribution with mean equal to μ_G and standard deviation equal to σ_G, and P is the quantile value of interest. Next, the constant parameter β is calculated, where ϕ is the normal probability density function with mean and standard deviation equal to μ_G and σ_G.

$$\beta = \frac{\phi_{\mu_G,\sigma_G}(\Phi^{-1}_{\mu_G,\sigma_G}(P))}{\phi(\Phi^{-1}(P))} \tag{2}$$

Then, the initial parameters (Equations 3 through 8) are calculated as shown below, and the velocity of the second shot, x_2, is determined.

$$\tau_1 = \sigma_G \tag{3}$$
$$\nu_1 = \beta^2 \tau_1^2 \tag{4}$$
$$c_1 = \frac{\nu_1}{(1+\nu_1)^{1/2}} \phi\left(\frac{\Phi^{-1}(P)}{(1+\nu_1)^{1/2}}\right) \tag{5}$$
$$b_1 = \Phi\left(\frac{\Phi^{-1}(P)}{(1+\nu_1)^{1/2}}\right) \tag{6}$$
$$x_2 = x_1 - \frac{c_1(y_1 - b_1)}{\beta b_1(1-b_1)} \tag{7}$$
$$\nu_2 = \nu_1 - \frac{c_1^2}{b_1(1-b_1)} \tag{8}$$

The velocity of ensuing shots are calculated using Equations 9 through 12.

$$c_{i+1} = \frac{\nu_i}{(1+\nu_i)^{1/2}} \phi\left(\frac{\Phi^{-1}(P)}{(1+\nu_i)^{1/2}}\right) \tag{9}$$
$$b_{i+1} = \Phi\left(\frac{\Phi^{-1}(P)}{(1+\nu_i)^{1/2}}\right) \tag{10}$$

$$x_{i+1} = x_i - \frac{c_i(y_i - b_i)}{\beta b_i(1-b_i)} \tag{11}$$

$$v_{i+1} = v_i - \frac{c_i^2}{b_i(1-b_i)} \tag{12}$$

An attractive feature of RMJ is that it does not use maximum likelihood techniques, making it a practical option to use by experimenters with limited statistical software or coding experience. Despite this fact, RMJ is not typically used in DoD testing.

Langlie created LM in 1962 as a reliability test method for ''one-shot'' items [2]. Langlie was tasked with developing a sequential method that requires no a priori assumption regarding the standard deviation and can be performed satisfactorily with sample sizes of around fifteen or twenty. Along with UD, LM serves as the backbone of the sequential methods used by the DoD. Since its inception, modifications have been made to repair LM's susceptibility to poor initial velocity guesses. The version of LM used in this simulation and that is currently in use by the Army is as follows [17]:

1. Select lower (μ_{min}) and upper (μ_{max}) projectile velocity limits (gates), where we assume μ_{min} will consistently result in non-perforations and μ_{max} will consistently result in perforations.
2. Fire the first shot (x_1) at the midpoint of μ_{min} and μ_{max}.
3. If the outcome of the first shot (y_1) results in a perforation ($y_1 = 1$), the second shot (x_2) equals $(x_1 + \mu_{min})/2$. If $y_1 = 0$, x_2 equals $(x_1 + \mu_{max})/2$.
4. If $y_1 = 1$ and $y_2 = 0$, or $y_1 = 0$ and $y_2 = 1$, $x_3 = (x_1 + x_2)/2$. If $y_1 = 1$ and $y_2 = 1$, x_3 equals $(x_2 + \mu_{min})/2$. If $y_1 = 0$ and $y_2 = 0$, x_3 equals $(x_2 + \mu_{max})/2$.
5. If all shots are perforations, subtract $2\sigma_G$ from μ_{min} and μ_{max}, and set the next shot (x_{i+1}) equal to $(x_i + \mu_{min})/2$. If all shots are non-perforations, add $2\sigma_G$ to μ_{min} and μ_{max}, and set the next shot (x_{i+1}) equal to $(x_i + \mu_{max})/2$. If all perforations or all non-pentrations are observed, repeat step 5.
6. The proceeding shots are calculated as follows. The rule for obtaining x_{i+1}, having completed i shots, is to work backward in the test sequence, starting at the ith shot, until a previous mth shot is found such that there are as many perforations as non-perforations in the mth through ith shots. x_{i+1} is then obtained by averaging x_i with x_m. If there exists no previous velocities satisfying the requirement stated above, then x_{i+1} is obtained by averaging x_i with μ_{min} or μ_{max} according to whether y_i was a perforation or a non-perforation.

NM [6] and 3POD [12] have many similar features and are best described at the same time. Although Wu describes 3POD as having three distinct phases, and Neyer may argue that his method contains two phases (modified binary search, and everything else), for the sake of comparison between these methods we have chosen to describe them in three sequential parts:

1. Bound the response curve with perforations and non-perforations.
2. Break separation.
3. Refine estimate of quantile of interest.

Part 1 is accomplished using an initial design. Neyer was the first to propose a systemic method for generating a good initial design. He referred to the initial design as a modified binary search. An initial design is used to quickly generate perforations and non-perforations. The purpose of this is to provide a coarse understanding of location and slope of the response curve by bounding the response curve with a non-perforation at a low velocity and a perforation at a high velocity. This is an important step because it sets the stage for part 2. Wu, recognizing the importance of an initial design, included a similar approach in 3POD. Both 3POD and NM require the experimenter to specify μ_G and σ_G prior to the beginning of the test. μ_{min} and μ_{max} are defined as $\mu_G - 4\sigma_G$ and $\mu_G + 4\sigma_G$. The first shot of NM is fired at the midpoint of μ_{min} and μ_{max}, while the first two shots of 3Pod are fired at $x_1 = 3/4\mu_{min} + 1/4\mu_{max}$ and $x_2 = 1/4\mu_{min} + 3/4\mu_{max}$. NM and 3POD continue with the next few shots, as shown in Figures 1 and 2, depending on whether the initial gate is too far left, right, or narrow. Typically, 3-6 shots are required to move to the second part of the procedure.

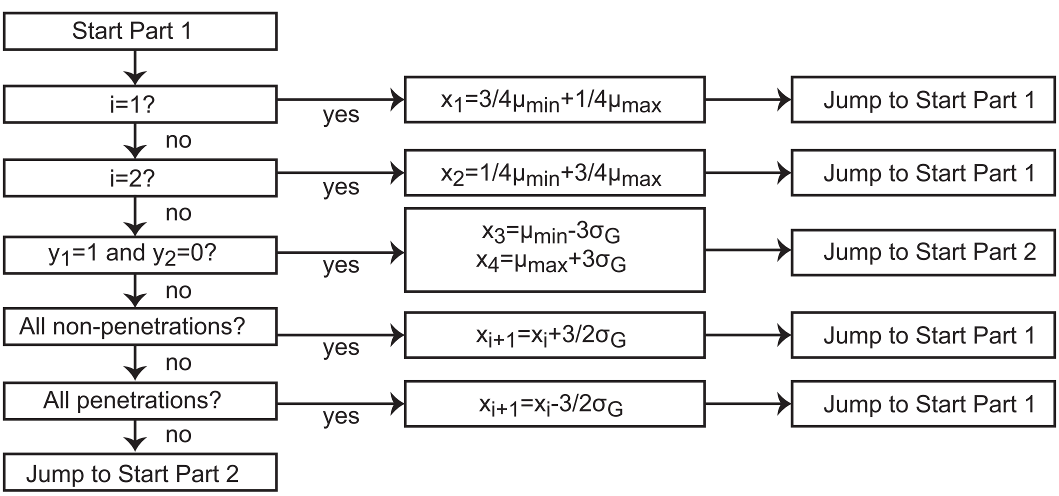

FIGURE 1. **Sequential Method Rules for 3Pod Part 1.**

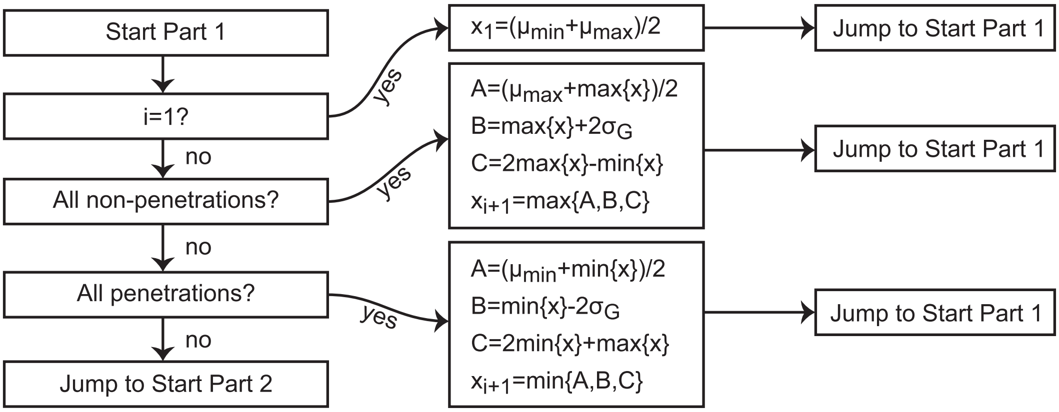

FIGURE 2. Sequential Method Rules for NM Part 1.

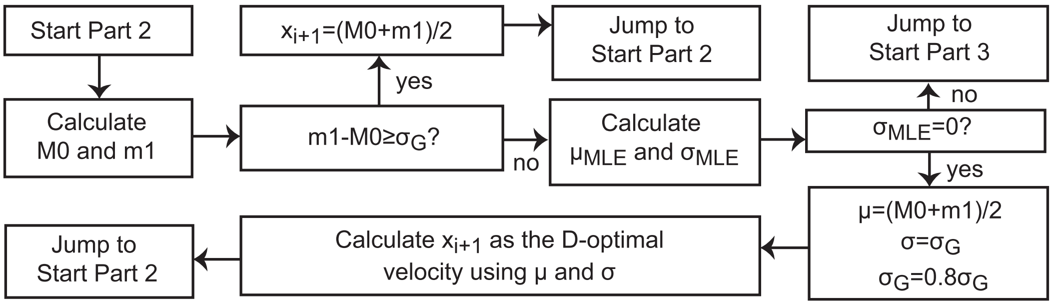

FIGURE 3. Sequential Method Rules for NM Part 2.

Part 2 of 3POD and NM is used to break separation. To break separation, the lowest velocity of the perforated shots (m1) has to be less than the highest velocity of the non-perforated shots (M0). A break in separation indicates that some degree of resolution of the response curve has been achieved, and allows for estimation of unique coefficients of the response curve that can be used to estimate any quantile value of the response curve (x_P).

Part 2 of NM, as shown in Figure 3, begins by determining the difference (δ) between the highest velocity that did not perforate (M0) and the lowest velocity that did perforate (m1). If this difference is greater than σ_G then the next shot is placed at the midpoint of M0 and m1. This process continues until δ is less than σ_G, at which point shots are placed at a D-optimal location. The D-optimal location is based on σ_G and μ_{MLE}, where μ_{MLE} is estimated from all previously recorded shots. A D-optimal velocity is estimated by finding the velocity that maximizes the determinant of the observed information matrix. This process places shots in the vicinity of V_{20} and V_{80}. After each shot, the magnitude of σ_G is decreased by multiplying σ_G by 0.8 and a new D-optimal velocity is calculated. This effectively pushes the D-optimal velocities nearer to the midpoint of the response curve. Eventually, it becomes likely that as the D-optimal velocities get closer to V_{50}, a region of overlap will be discovered. At that point, the second part of Neyers method is complete.

Part 2 of 3POD does not use a D-optimal routine, and instead relies more heavily on conditional logic. In the beginning of 3POD's part 2, shots are placed at μ_{MLE} until δ is less than or equal to $1.5\sigma_G$, where μ_{MLE} is estimated from the data that is calculated by holding σ constant and equal to σ_G. Once δ is sufficiently small and overlap has not been achieved, points are placed in the range of M0 to m1, but not at the midpoint like in Neyers method. Wu places shots slightly away from the midpoint at velocities that are $0.3\sigma_G$ away from M0 or m1. If at this point overlap is not yet achieved, the value of σ_G is decreased and part two restarts from the beginning. This process continues until separation is broken. Wu added a novel step at the end of part 2, meant to strengthen the overlap. This step is not shown in Figure 4.

In DoD testing, care should be taken to make sure that the velocity set point error is not greater than what is allowed in the conditional logic in 3POD. For example, a potential shortcoming in part 2 of 3POD occurs after the conditional step that asks if $n0 > n1$. The proceeding block prescribes the next velocity as $x_{i+1} = m1 + 0.3\sigma_G$, and determines if separation is broken if x_{i+1} is greater than m1 and $y_{i+1} = 0$. But in armor testing, x_{i+1} includes a component of set point error due to the tester not being able to perfectly execute a planned velocity. As a result, if the magnitude of error is greater than 0.3σ, there is a possibility that the proceeding step will falsely conclude that separation has been broken. In our simulation study, we purposely set the set point error smaller than 0.3σ to avoid this issue. In practice, however, a stopgap solution could be to evaluate the data set as a whole to determine if separation exists, instead of using $y_{i+1} = 0$ or $y_{i+1} = 1$ as an indicator.

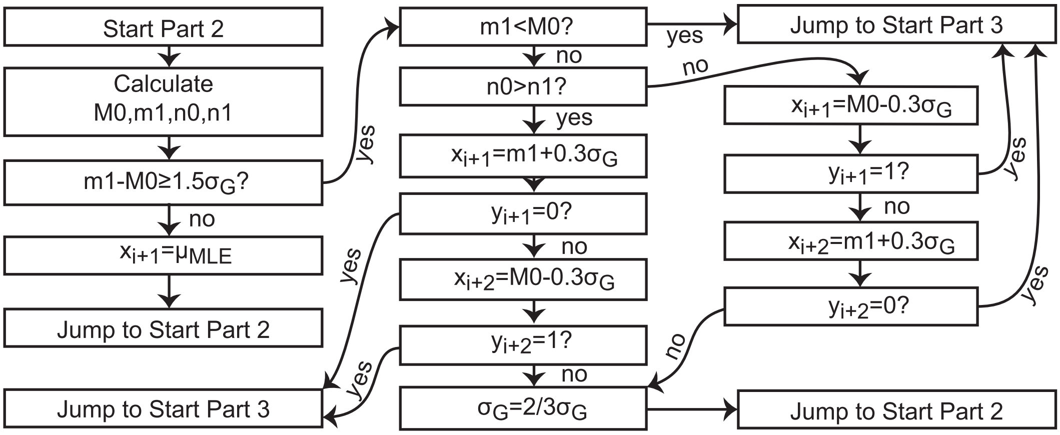

FIGURE 4. **Sequential Method Rules for 3Pod Part 2.**

Part 3 of 3POD and NM are also quite similar and are both used to refine the estimate of the quantile of interest. In NM, part 3 places shots at D-optimal velocities. A user-defined, predetermined number of shots are placed at velocities that satisfy a D-optimality criterion. Placing shots at the D-optimal location minimizes the confidence region on the coefficients, which in turn improves the estimate of the quantile of interest. The D-optimal velocity is recalculated after each shot and is defined as the velocity that maximizes the determinant of the observed information matrix. There are typically two D-optimal locations: one near V_{20} and the other near V_{80}. This effectively spreads the design space in a manner that tends not to bias any one direction. In our case, this is advantageous because it results in a less biased estimate of V_{50} when also estimating V_{10}. Part 3 of 3POD is broken into two stages and is also used to refine the estimate of the quantile of interest. The resources used in each stage are predetermined by the tester prior to the beginning of the test. For our study, we split the resources in half. The first stage follows the same D-optimal approach used by Neyer. The second stage employs the Robbins Monroe Joseph method, which places shots near the quantile value of interest.

III. MOTIVATING DoD EXAMPLE

The Advanced Combat Helmet (ACH) has been widely used by the Army and Marine Core for protection in combat since 2003. The Director, Operational Test and Evaluation (DOT&E) publishes a standard test protocol for conducting both first article testing (FAT) and lot acceptance testing (LAT) for all aramid helmets including the ACH [18]. We will focus on the protocols for ballistic penetration testing during LAT in this example. LAT testing is conducted on every lot of helmets that the DoD procures. Therefore, it is essential that LAT ensure quality and also be affordable in terms of the number of helmets expended during testing. The ballistic penetration testing is only one component of the LAT testing and therefore, it is desirable to keep the sample size small.

The LAT uses an up-and-down test method with the 33SC, with modifications to keep sample sizes small. More specifically, in the LAT, the initial round is fired approximately 100 ft/s above minimum requirement for the designated threat. If the first round results in a complete penetration (CP), second round is fired at an estimated charge velocity approximately 50 ft/s to 100 ft/s below velocity of initial round. If the first round results in a partial penetration (PP), second round is fired at an estimated charge velocity approximately 50 ft/s above velocity of initial round. For subsequent rounds, if the previous round resulted in a CP, the next round is fired approximately 50 ft/s below the previous firing velocity; if the previous round resulted in a PP, the next round is fired approximately 50 ft/s above the previous firing velocity.

After the stopping criteria of 3 partial penetrations and 3 complete perforations is met the arithmetic mean is the average of the 3 highest partial penetration velocities and the 3 lowest complete penetration velocities within a 125 ft/s spread. At least 6 rounds are made but more rounds may be needed to reach these criteria. If this cannot be met, the arithmetic mean is the average of the 5 highest PP velocities and the 5 lowest CP velocities. At this point, at least 10 shots have been taken, again it could be more. If neither of the previous conditions can be met and 5 PPs and 5 CPs have been made and the 5 highest PP velocities and 5 lowest CP velocities are above the minimum requirement then a V50 value is not calculated and the test is declared as "inconclusive."

The LAT acceptance data shown in Figure 5 was used to guide the choices of the parameters in the following simulation study. The 18 run experiment resulted in a V_{50} estimate of 2,413 ft/s with a standard deviation of 76 ft/s using a Probit model fit. This example also guides the selection of the velocity range for calculating the AME, which we set at a range of 125 ft/s.

IV. SIMULATION STUDY

In this paper we compare sequential methods, estimators, and stopping criteria using Monte Carlo simulation. We focus on the Probit model to represent the true relationship between probability of perforation and projectile velocity. We consider two sets of true parameters that are reflective of the combat

TABLE 1. Factors and levels for the first simulation study.

Factors	Levels
Sequential Method	UD,LM,DRM,3POD,NM,RMJ
σ_T	75 ft/s,150 ft/s
σ_G/σ_T	1/3,1/2,2,3
μ_G	$\mu_T - 2\sigma_T$, μ_T, $\mu_T + 2\sigma_T$
Stopping Criteria	33SC (with AME), BSSC (with Probit-MLE)

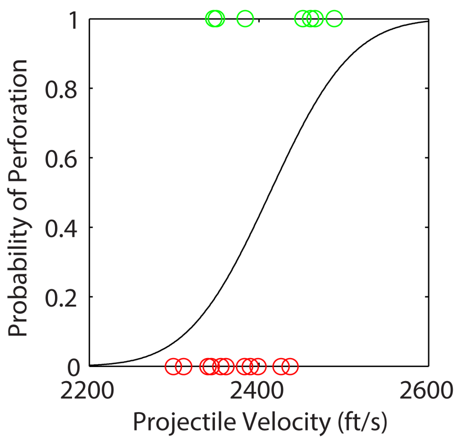

FIGURE 5. Example LAT Test.

helmet example: (1) $\mu_T = 2400$ ft/s, $\sigma_T = 75$ ft/s, and (2) $\mu_T = 2400$ ft/s, $\sigma_T = 150$ ft/s.

A simulated test is carried out in a similar manner as a physical one except that no projectiles are fired, and the outcome of whether the projectile perforated the armor is determined using a random Bernoulli draw from the probability of perforation estimated from the true model. For example, if a given simulated shot is fired at 2300 ft/s, according to the first set of true parameters, the probability that that projectile perforates the armor is $G((2300 - 2400)/100) = 0.16$. Then, a random Bernoulli number is generated that has a sixteen percent chance of being a perforation. To instill more realism into the simulation, we include a velocity set point error. For each calculated velocity, we add a random error drawn from a uniform distribution between plus or minus 10 ft/s.

Three simulation studies each employ a full factorial experiment to address three objectives. The first focuses on the number of runs required to reach the stopping criteria and the impact on V_{50} error. The second compares the V_{50} estimators. The third concentrates on the sequential methods for estimating V_{10}. As explained below, the first, second, and third full factorial experiments consist of 288, 288, and 336 trials, respectively. 1000 simulations are executed per trial. A simulation is representative of a single live fire test, consisting of anywhere from 5 to 50 sequentially fired projectiles, which we will refer to as runs. The V_{50} or V_{10} bias is calculated once the stopping criteria is satisfied. The outputs of a simulation include the runs required to reach the stopping criteria, V_{50} bias, or V_{10} bias. Both the median and interquartile range of these outputs are the response variables for each factorial trial.

We present results using two types of figures. The first shows the response variables (the median and interquartile range of the V_{50} or V_{10} bias). The second illustrates an effect screening plot. Effect screening is an efficient way to summarize and compare the results of highly dimensional factorial experiments [19]. The effects show the impacts of the factors, and interactions between factors, on the response variables. The effects are calculated by regressing each response variable on the factors of the factorial experiment, and are the coefficient estimates of the resulting regression model. Coefficient estimates are shown for all main effects and two-factor interactions. The intercept of the regression model is the grand mean of the response variable and is shown in the bottom left of the effects plot. The coefficient of a particular level of a factor describes the difference between the grand mean and the average response at that level.

A. SETUP OF SIMULATION STUDY 1

The first simulation study determines the number of runs required to reach each stopping criteria and the impact of those stopping criteria on V_{50} bias. Since 33SC does not readily accommodate Probit-MLE and, likewise, BSSC does not accommodate AME, this simulation study paired AME with 33SC and Probit-MLE with BSSC. Resources are critically important in live fire testing so it is important we consider the tradeoff between runs required and V_{50} bias.

Table 1 shows the factors and levels for this experiment. To provide a fair comparison between sequential methods, we set the following parameters for each method. The step size for UD is equal to μ_G. The quantile value of interest for DRM is set to 0.5. The upper and lower limits for LM, NM, and 3Pod are $\mu_G \pm 4\sigma_G$. Only the first two phases of 3POD are considered, and the methods are stopped after the stopping criteria is satisfied.

The response variables for the effect screening are the median number of runs required to satisfy the stopping criteria, the median V_{50} bias (referred to as V_{50} bias), and the interquartile range of the V_{50} bias (referred to as V_{50} IQR).

We limit the maximum number of runs per simulation to 50. Typically, the stopping criteria is satisfied before that maximum. For this factorial experiment, 99.0 percent of all simulations required fewer than 50 runs to reach the stopping criteria.

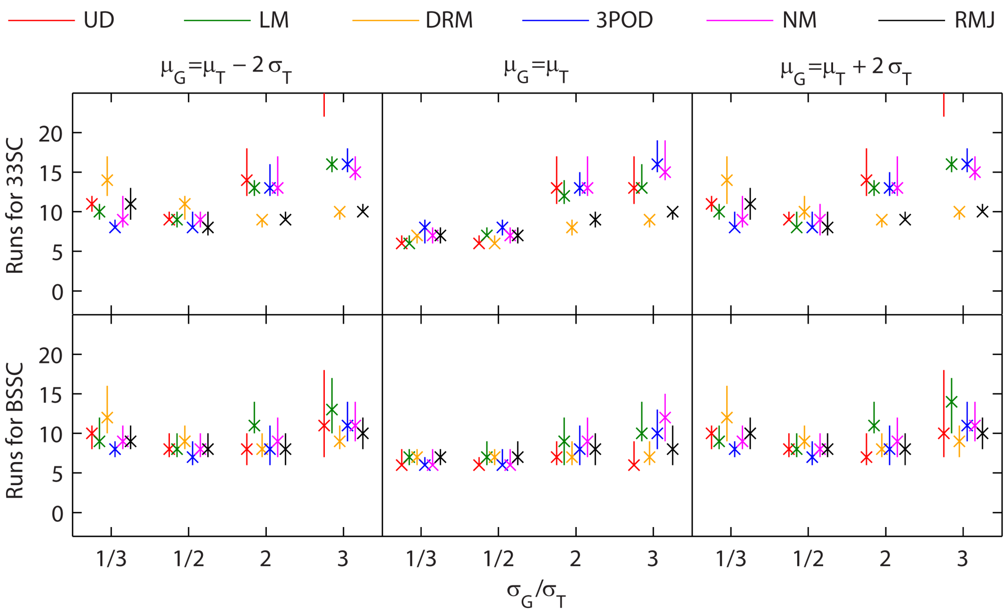

FIGURE 6. Simulation Study 1: Runs to Satisfy Stopping Criteria.

B. RESULTS OF SIMULATION STUDY 1

1) RUNS TO SATISFY STOPPING CRITERIA

Figure 6 shows the median and interquartile range of the number of runs required to satisfy each stopping criteria. Overall, it appears BSSC requires fewer runs than 33SC. The majority of simulations require less than 20 runs to satisfy both stopping criteria, except for UD where $\sigma_G/\sigma_T = 3$ and $\mu_G = \mu_T - 2\sigma_T$ or $\mu_G = \mu_T - 2\sigma_T$. In that case, the step size for UD is too large, hence it lacks the resolution to converge to V_{50}, resulting in a poor V_{50} estimate using 33SC. BSSC on the other hand did not exhibit the same problem.

The effect screening analysis in Figure 7 shows the coefficients of the model that regresses the median number of runs required to reach the stopping criteria on each factor in the experiment. The intercept indicates that on average across the entire design space approximately 10.5 runs are needed to satisfy the stopping criteria. The main effects that have the largest contribution are σ_G/σ_T and stopping criteria. 33SC requires approximately 1.8 more runs than the grand mean on average, while BSSC needs 1.8 fewer. The largest increase to the number of runs occurs when $\sigma_G/\sigma_T = 3$; however, DRM and RMJ perform well under that condition. The largest decrease occurs when $\sigma_G/\sigma_T = 1/2$.

Numerous two-factor interactions influence the runs required to satisfy the stopping criteria. The interaction that has the largest increase in runs occurs between $\sigma_G/\sigma_T = 3$ and UD, which is consistent with our observations from the previous figure. The interactions also show that 3POD requires fewer runs for BSSC than 3SSC, which is as expected since 3POD employs BSSC by design. Surprisingly, UD also requires fewer runs for BSSC than 33SC. This was due to the large number of runs required for the 33SC case where $\sigma_G/\sigma_T = 3$. 33SC requires fewer runs for small values of σ_G/σ_T, while BSSC requires fewer runs with larger values.

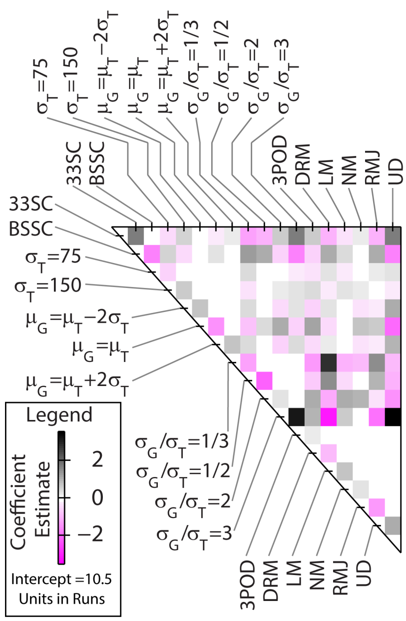

FIGURE 7. Simulation Study 1: Effects Plot of Median Runs to Satisfy Stopping Criteria.

2) MAGNITUDE OF THE MEDIAN V_{50} ERROR

Figure 8 shows the median and interquartile range of the V_{50} bias for the sample sizes presented in Figure 6. Not surprisingly, UD provides an inaccurate estimate of V_{50} when $\sigma_G/\sigma_T = 3$. It also appears that the bias tends to be smaller for $\mu_G = \mu_T$, which is an obvious favorable condition because it implies the true value of V_{50} is known prior to the test. As for the non ideal situation where $\mu_T - 2\sigma_T$ or $\mu_T + 2\sigma_T$, DRM appears to perform poorly when $\sigma_G/\sigma_T = 1/3$. This is because the step size used by DRM in this case is too small, which prevents DRM from placing shots near the true value of V_{50} when the initial guess of V_{50} is offset by plus or minus $2\sigma_T$. Essentially, the stepping algorithm never reaches the true V_{50}.

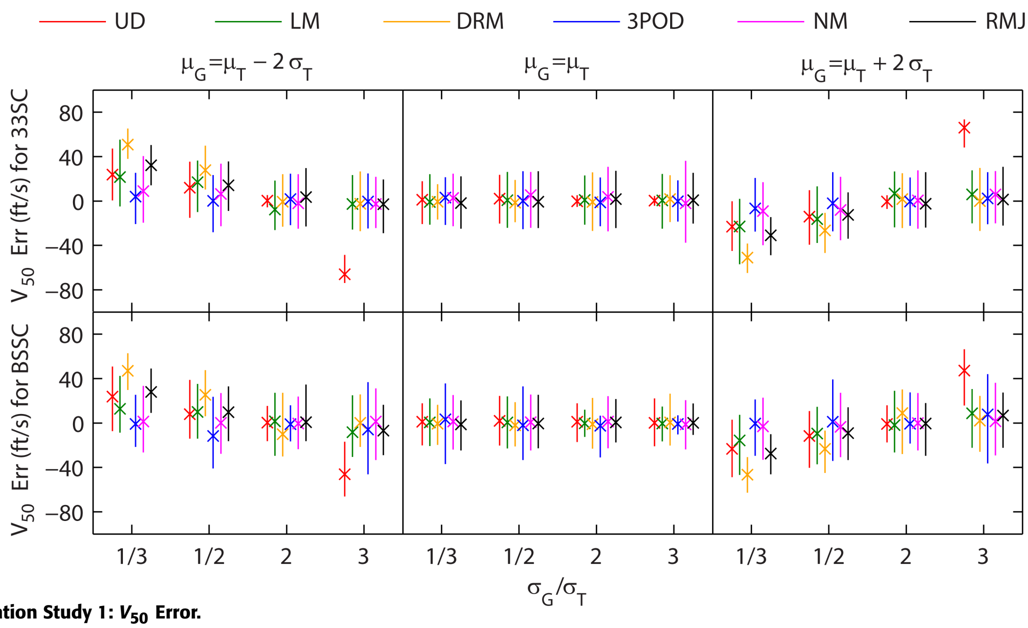

FIGURE 8. Simulation Study 1: V_{50} Error.

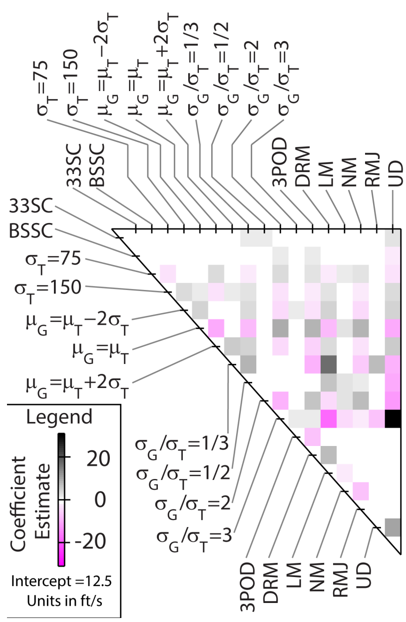

FIGURE 9. Simulation Study 1: Effects Plot of Magnitude of Median V_{50} Error.

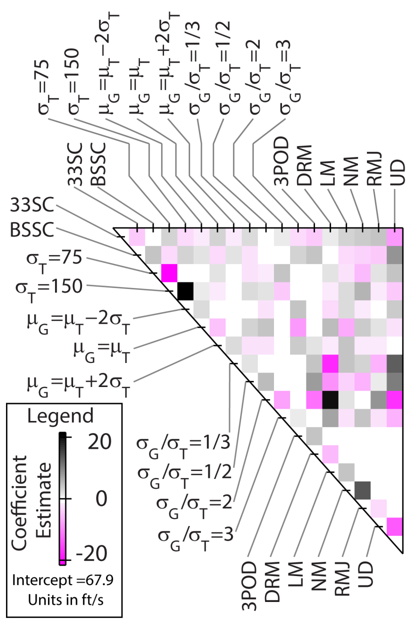

FIGURE 10. Simulation Study 1: Effects Plot of Interquartile Range of V_{50} Error.

Figure 9 presents the coefficients of the model that regresses V_{50} bias on each factor in the experiment. The stopping criteria has no affect on the V_{50} bias. The average V_{50} bias for each stopping criteria is approximately equal to the grand mean, which is equal to 12.5 ft/s. Among sequential methods, 3POD and NM provide the greatest reduction in bias as shown by the main effects. Comparing the interactions between sequential methods and all other factors, 3POD shows nearly all zeroes, which implies it is the most robust method in this case.

3) INTERQUARTILE RANGE OF V_{50} ERROR

Figure 10 presents the coefficients of the model that regresses V_{50} IQR on each factor in the experiment. The figure shows that among main effects the V_{50} IQR is most influenced by σ_T. V_{50} IQR increases as σ_T increases. This result is expected since an increase in the precision error in the true ballistic response curve should result in an increase in the precision error in the estimate of the response curve. The difference in the average of V_{50} IQR between $\sigma_T = 75$ and $\sigma_T = 150$ is 42.4 feet per second. To a lesser degree, the stopping criteria also impacts V_{50} IQR. The average V_{50} IQR is 4.5 feet per second less than the grand mean while the average for BSSC is 4.5 feet per second more than the grand mean. The additional runs afforded to 33SC contribute to the decrease in V_{50} IQR.

UD and DRM reduce the interquartile range of the V_{50} bias more than the other sequential methods according to Figure 10. UD and DRM perform particulary well when

$\mu_G = \mu_T$. In spite of that, we saw in Figure 9 that UD and DRM demonstrate the worst median V_{50} bias. Therefore, despite the fact that they provide superior precision over other sequential methods, their V_{50} estimates cannot be trusted with poor guess of σ and μ.

C. SETUP OF SIMULATION STUDY 2

The purpose of the second simulation study is to compare the estimators in terms of their V_{50} bias. To provide a level playing field to compare estimators, the stopping criteria is set as the minimum number of runs that satisfies both the 33SC and BSSC. This is important because we would like to compare estimators with respect to the smallest possible sample sizes that would be encountered in ballistic resistance testing.

TABLE 2. Factors and levels for the second simulation study.

Factors	Levels
Sequential Method	UD,LM,DRM,3POD,NM,RMJ
σ_T	75 ft/s,150 ft/s
σ_G/σ_T	1/3,1/2,2,3
μ_G	$\mu_T - 2\sigma_T$, μ_T, $\mu_T + 2\sigma_T$
Estimator	AME, Probit-MLE
Stopping Criteria	Max(33SC, BSSC)

Table 2 shows the factors and levels for this experiment. To provide a fair comparison between sequential methods, we set the following parameters for each method. The step size for UD is equal to μ_G. The quantile value of interest for DRM is set to 0.5. The upper and lower limits for LM, NM, and 3Pod are $\mu_G \pm 4\sigma_G$. Only the first two phases of 3POD are considered.

The response variables for the effect screening analysis are V_{50} bias and V_{50} IQR.

D. RESULTS FOR SIMULATION STUDY 2

Figure 11 presents the coefficients of the model that regresses the V_{50} bias on each factor in the experiment. Figure 12 shows the coefficients of the model that regresses the V_{50} IQR on each factor in the experiment. In this study we only present the effects plot since the results are quite similar to the previous study. The median number of runs to satisfy both stopping criteria is recorded for each factorial trial; the average of which was 12.9 runs.

The estimator has no affect on V_{50} bias and IQR as shown in Figure 11 and Figure 12. Furthermore, there are no substantial interactions between estimator and any other factor. This suggests that under the conditions tested the estimator has no consequence on the V_{50} estimate.

While most of the results from this simulation study are consistent with the first simulation study, there is one result that stands out. 3POD performs better using the new stopping criteria. This result makes sense since 3POD is designed for BSSC, and the stopping criteria in this study ensures that 3POD has at least enough runs to satisfy BSSC, whereas in the previous study 33SC was used in half of the trials. Compared to the previous study, 3POD retains the best performance in terms of V_{50} bias and improves in terms of V_{50} IQR. 3POD also exhibits a reduction in interactions with other factors compared to the previous study, making it the most robust method in terms of V_{50} IQR.

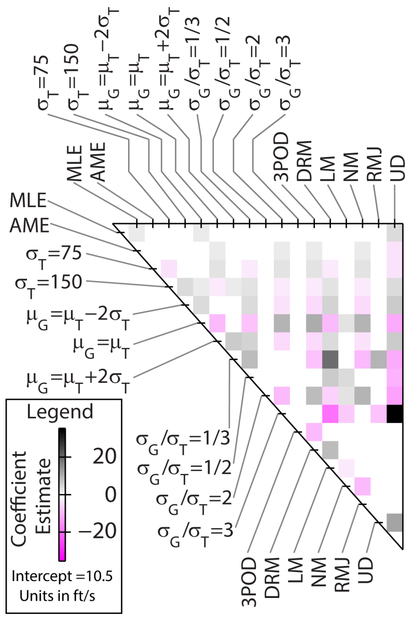

FIGURE 11. Simulation Study 2: Effects Plot of Magnitude of Median V_{50} Error.

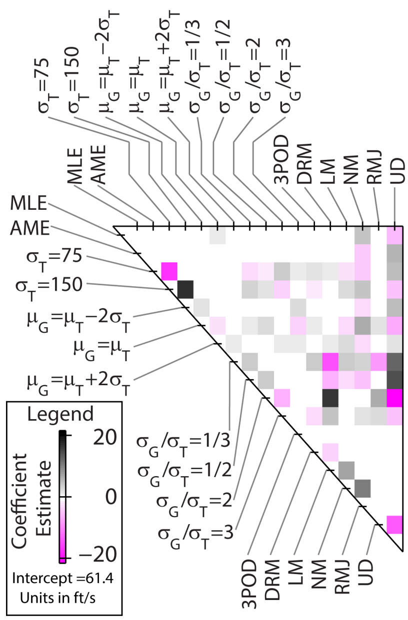

FIGURE 12. Simulation Study 2: Effects Plot of Interquartile Range of V_{50} Error.

The fact that the Probit-MLE did not out perform the AME is counterintuitive. The small sample sizes contributed to this result. Additionally, the simulation study was based on PPE testing, where reasonable ranges for the AME calculation are known. In other scenarios, either consisting of

TABLE 3. Factors and levels for the third simulation study.

Factors	Levels
Sequential Method	UD,LM,DRM,3POD,NM,RMJ,KR
σ_T	100 ft/s,400 ft/s
σ_G/σ_T	1/3,1/2,2,3
μ_G	$\mu_T - 2\sigma_T$, μ_T, $\mu_T + 2\sigma_T$
Estimator	Probit-MLE
Stopping Criteria	N=20,N=40

larger sample sizes or unknown armor performance we would expect MLE to outperform AME. Additionally, the Probit-MLE has the advantage that the full curve is estimated, and therefore all of the quanitles can be estimated.

E. SETUP OF SIMULATION STUDY 3

The third simulation characterizes the impact of the factors on the median and interquartile range of the V_{10} bias, while also considering the impact on V_{50} estimation. The factors and levels for this experiment are shown in Table 3. AME was not used in this experiment because it cannot be used to estimate V_{10}. To make a fair comparison between sequential methods, we set the following parameters for each method with the primary goal of estimating V_{10} and the secondary goal of estimating V_{50}.

We set the quantile value of interest for RMJ, DRM and 3POD equal to 0.1. For RMJ, τ was set equal to σ_G. We set K equal to five for KR, which yields a quantile value of interest equal to 0.13. We set the step size for KR and UD equal to σ_G. NM, 3Pod and LM use an upper and lower limit of $\mu_G \pm 4\sigma_G$. For RMJ, we set τ equal to σ_G. For 3Pod, we allocate half of the remaining runs for phase 2 and the other half for phase 3.

The response variables for the effect screening analysis are the magnitude of the median V_{10} bias (referred to as V_{10} bias) and V_{50} bias .

F. RESULTS FOR SIMULATION STUDY 3

1) V_{10} MEDIAN BIAS

RMJ and DRM reduce the V_{10} median bias more than other sequential methods as demonstrated in Figure 14. RMJ and DRM consistently provide decreased V_{10} bias for all values of μ_G and σ_G. Figure 13 shows that 3POD is next best, followed by LM, NM, and lastly UD.

Despite the fact that DRM is the second most accurate sequential method in terms of V_{10} bias, DRM has difficulty breaking separation. For each simulation, if separation is broken, we use the Probit-MLE estimator to estimate V_{10}. If separation persists, we discard the run. It turns out that DRM fails to break separation in 49.9 percent of the experiment (61.2 percent for N = 20 runs, and 38.5 percent for N = 40 runs). This is by far the worst performance among sequential methods which makes DRM look far better than it really is since many of its runs are discarded. The second

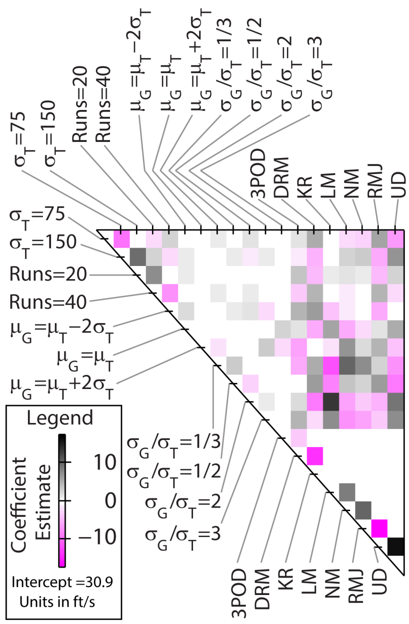

FIGURE 13. Simulation Study 3: Effects Plot of Magnitude of Median V_{10} Error.

worst is RMJ which fails to break separation 15.2 percent of the time in this experiment. 3POD on the other hand performs best, breaking spearation 99.4 percent of the time in this simulation study.

DRM's failure is caused by its step size that decreases as the test continues (decreases with N). At a certain point, the step size becomes smaller than the velocity set point error. Recall from earlier that we include a uniform random error between plus or minus five feet per second to each calculated velocity. When DRM's step size become smaller than the uniform random error, the sequence no longer moves in a coherent direction. The problem is pronounced when estimating V_{10} because the step size is multiplied by $1-P$ or 0.1 (see Equation 1). This causes DRM's step size to become smaller than the uniform random error within the first few shots, which causes separation in the data set and erratic estimates of V_{50}.

Figure 14 shows that the V_{10} median error is bias in one direction for all methods except DRM and RMJ. This is because the other sequential methods place runs closer to V_{50}, thereby biasing the V_{10} estimate closer to V_{50}. This in effect

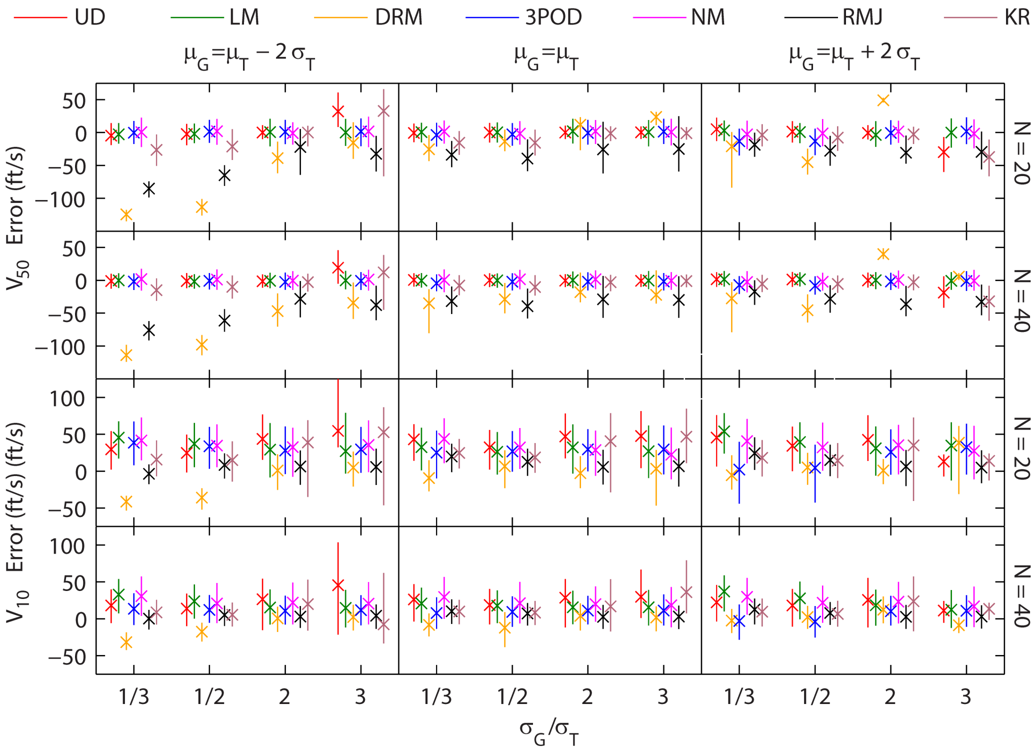

FIGURE 14. Simulation Study 3: V_{10} and V_{50} Error.

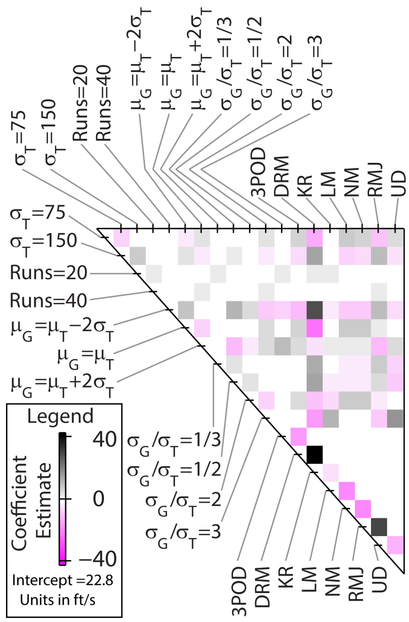

FIGURE 15. Simulation Study 3: Effects Plot of Magnitude of Median V_{50} Error.

is magnified for LM, NM, and UD, since these sequential methods place runs closer to V_{50} by design.

2) V_{50} MEDIAN BIAS

The advantage of reduced V_{10} bias by RMJ comes at the expense of V_{50} bias. Figures 15 shows that the three best sequential methods for reducing V_{50} bias are LM, NM, and 3POD. RMJ and DRM yield the worse V_{50} bias. This result is not surprising because DRM and RMJ forgo intial designs, and do not place points near V_{50}. Meanwhile, 3Pod and NM employ initial designs, and D-optimal selection criteria that balances the design space.

V. CONCLUSIONS AND RECOMMENDATIONS

The DoD uses sensitivity experiments to assess the ballistic resistance of various types of armor. This paper shows that the majority of methods employed by the DoD test community can be improved by employing more recent sensitivity test design methods. However, a change must must occur, as many of the newer sensitivity test design methods are based on generalized linear models. The first and most change the test community can make is to start to universally employ maximum likelihood estimation and generalized linear models in the analysis of ballistic limit testing. This transition provides the gateway to employ more complex and efficient sensitivity designs. Additionally, maximum likelihood estimation techniques generate the full perforation response curve providing more information for the same test resources.

In general we found that the methods compared in this study perform commensurate with the goal of the test design. The top three sequential methods that reduce V_{10} bias in descending order are RMJ, DRM, and 3POD. However, 3POD is more robust to values of μ_G and σ_G/σ_T than DRM. We also noted that DRM performs erratically for tests with large samples sizes (20 or greater) because its step size becomes smaller than the velocity set point error. On the other hand, for the same simulation studies designed around estimating V_{10}, UD, LM, 3POD, and NM resulted in the lowest bias on V_{50}. The 3POD method appears to be the most robust method of estimating multiple quantiles.

The first two simulation studies that used DoD stopping criteria showed that every method provided lower bias than the UD method. This suggests that improvements can be made to all of the current protocols employing the UD methodology. Further research might investigate using portions of Neyer's method and 3POD to determine optimal stopping criteria to meet the limited sample size needs of the DoD.

NOMENCLATURE

μ	Location parameter of probit model (equal to V_{50}).
μ_G	Guessed value of μ.
μ_T	True value of μ.
σ	Scale parameter of probit model.
σ_G	Guessed value of σ.
σ_T	True value of σ.
d	Step size.
$M0$	Highest velocity of non-perforated shots.
$m1$	Lowest velocity of perforated shots.
N	Total number of shots.
$n0$	Number of non-perforated shots.
$n1$	Number of perforated shots.
V_P	The velocity at which the projectile has a P percent chance of perforating the armor.
y_i	The binary outcome of the ith shot.
33SC	Three and three stopping criteria.
3POD	Wu's Three-phase Method.
AME	Arithmetic Mean Estimator.
BSSC	Break separation stopping criteria.
DRM	Delayed Robbins Monroe Method.
FAT	First Article Test.
KR	K-in-a-row Method.
LAT	Lot Acceptance Test.
LM	Langlie Method.
NM	Neyer's Method.
PPE	Personal Protective Equiptment.
Probit-MLE	Maximum Likelihood Estimator using Probit Model.
RMJ	Robbins Monroe Joseph Method.
UD	Up and Down Method.

REFERENCES

[1] W. J. Dixon and A. M. Mood, "A method for obtaining and analyzing sensitivity data," *J. Amer. Statist. Assoc.*, vol. 43, no. 241, pp. 109–126, Mar. 1948.
[2] H. J. Langlie, *A Reliability Test Method for 'One-Shot' Items*, Aeronautics Division of Ford Motor Company, document U-1792, Aug. 1962.
[3] *V50 Ballistic Test for Armor*, Department of Defense Test Method Standard MIL-STD-662F, 1997.
[4] D. Anbar, "A stochastic Newton-Raphson method," *J. Statist. Planning Inference*, vol. 2, no. 2, pp. 153–163, 1978.
[5] V. R. Joseph, "Efficient Robbins–Monro procedure for binary data," *Biometrika*, vol. 91, no. 2, pp. 461–470, 2004.
[6] B. T. Neyer, "A D-optimality-based sensitivity test," *Technometrics*, vol. 36, no. 1, pp. 61–70, Feb. 1994.
[7] C. F. J. Wu and Y. Tian, "Three-phase optimal design of sensitivity experiments," *J. Statist. Planning Inference*, vol. 149, pp. 1–15, Jun. 2014.
[8] R. M. Ray, P. A. Roediger, and B. T. Neyer, "Commentary: Three-phase optimal design for sensitivity experiments," *J. Statist. Planning Inference*, vol. 149, pp. 20–25. June. 2014.
[9] B. A. Bodt and H. B. Tingey, "Design and estimation in small sample quantal response problems: A Monte Carlo study," US Army Res. Lab., Adelphi, MD, USA, Tech. Rep. BRL-TR-3002, 1989.
[10] L. J. Young and R. G. Easterling, "Estimation of extreme quantiles based on sensitivity tests: A comparative study," *Technometrics*, vol. 36, no. 1, pp. 48–60, Feb. 1994.
[11] M. Gezmu, "The geometric up-and-down design for allocating dosage levels," Ph.D. dissertation, Dept. Math. Statist., American Univ., Washington, DC, USA, 1996.
[12] C. F. J. Wu and Y. Tian, "Three-phase sequential design for sensitivity experiments," in *Proc. Spring Res. Conf., Enabling Inter. Statist. Eng.*, 2012.
[13] J. C. Collins, "Quantal response: Practical sensitivity testing," US Army Res. Lab., Adelphi, MD, USA, Tech. Rep. ARL-TR-6022, 2012.
[14] D. Mauchant, "Analysis of three different regression models to estimate the ballistic performance of new and environmentally conditioned body armor," Dept. Commerce, National Inst. Standards Technol., Gaithersburg, MD, USA, Tech. Rep. NISTIR 7760, 2011.
[15] J. L. Hodges and E. L. Lehmann, "Two approximations to the Robbins–Monro process," in *Proc. 3rd Berkeley Symp. Math. Statist. Probab.*, 1955, pp. 95–104.
[16] W. G. Cochran and M. Davis, "The Robbins–Monro method for estimating the median lethal dose," *J. Roy. Statist. Soc.*, vol. 27, no. 1, pp. 28–44, 1965.
[17] J. C. Collins and L. L. C. Moss, "Langmod users manual," US Army Res. Lab., Adelphi, MD, USA, Tech. Rep. ARL-TN-437, 2011.
[18] *Standardization of Hard Body Armor Testing*, Office of the Secretary Of Defense, Washington, DC, USA, Apr. 2010.
[19] R. H. Myers, D. C. Montgomery, and C. M. Anderson-Cook, *Response Surface Methodology: Process and Product Optimization Using Designed Experiments*. New York, NY, USA: Wiley, 2009.

THOMAS H. JOHNSON was born in Algonquin, IL, USA, in 1984. He received the B.S. degree from Boston University, Boston, MA, USA, in 2002, and the M.S. and Ph.D. degrees from Old Dominion University, Norfolk, VA, USA, in 2006 and 2011, respectively, all in aerospace engineering. In 2011, he joined the Operational Evaluation Division, Institute for Defense Analyses, Alexandria, VA, USA, as a Research Staff Member. He has been supporting the test and evaluation of the apache helicopter and other Army acquisition programs since 2011.

LAURA FREEMAN is currently a Research Staff Member with the Institute for Defense Analyses, Alexandria, VA, USA, where she leads the Test Science Task providing support to the Director, Operational Test and Evaluation on the use of statistics in test and evaluation. Her areas of statistical expertise include designed experiments, reliability analysis, and industrial statistics. In addition, she has a background in aerospace engineering. She received the B.S. degree in aerospace Engineering, and the M.S. and Ph.D. degrees in statistics from the Virginia Polytechnic Institute and State University, Blacksburg, VA, USA. Her Ph.D. research was on design and analysis of experiments for reliability data.

JANICE HESTER received the B.S. degree in physics from Arizona State University, Phoenix, AZ, USA, in 2002, and the Ph.D. degree in physics from Princeton University, Princeton, NJ, USA, in 2008. She joined the Institute for Defense Analysis, Alexandria, VA, USA, as a Research Staff Member, in 2010. She currently works on the live fire test and evaluation of personal protective equipment.

JONATHAN L. BELL is currently a Research Staff Member with the Institute for Defense Analyses (IDA), Alexandria, VA, USA. He received the Ph.D. degree in materials science and engineering from the University of Illinois at Urbana-Champaign, Champaign, IL, USA, in 2008, and the bachelor's degree in materials science from Carnegie Mellon University, Pittsburgh, PA, USA. His work at IDA is focused on operational testing of ground vehicle systems with a specific emphasis on reliability.