# $E_{50}$ testing for stab-resistant body armour evaluation

**S. Jenkin[1,2]**, B. Cazzolato[1], C.F. Jones[1] and D. Thompson[2]

*[1]School of Electrical and Mechanical Engineering, The University of Adelaide, Adelaide, 5005 South Australia, Australia,*
*scott.jenkin@adelaide.edu.au*
*[2]Armor Australia Pty Ltd, 126 Frederick Street, Welland, 5007 South Australia, Australia*

**Abstract.** Current stab-resistant body armour certification standards, such as National Institute of Justice (NIJ) Standard-0115.00 and Home Office Body Armour Standard (CAST 2017), rely on fixed-energy impact testing, providing limited insight into armour performance across a range of impact energies. In contrast, bullet-resistant armour evaluation typically employs Penetration-Backface Signature (P-BFS) testing at reference velocities and Ballistic Limit ($V_{50}$) testing, which uses statistical modelling to estimate penetration probability as a function of velocity. Despite the advantages of $V_{50}$ testing – enabling quantile-based performance assessment and greater statistical confidence – statistical penetration modelling has not been applied to stab-resistant armour. This paper introduces $E_{50}$ testing as a framework for evaluating stab-resistant body armour, adapting statistical methodologies applied in ballistic $V_{50}$ testing. Experimental depth of penetration (DOP) data was collected using a precision, instrumented, stab testing apparatus compliant with NIJ Standard-0115.00, across impact energies ranging from 25 J to 65 J. Testing was performed using modern, multi-threat armour, with 78 tests for the P1/A Edged Blade threat and 68 tests for the Spike threat. Binary regression techniques characterised penetration probability as a function of impact energy, and Monte Carlo simulations assessed the efficiency and accuracy of various sensitivity test designs – including Bruceton (up-and-down), Langlie, Neyer and 3pod2.0 methods – in estimating $E_{50}$ and $E_{05}$ quantiles. Simulations examined how sample size ($n$=12, 24, 36, 48) influenced test accuracy and efficiency within each test design, where each was simulated across different initial energy levels and energy increments. The results provide insight into the expected precision of $E_{50}$ testing, highlighting the trade-offs between test design, sample size, and statistical confidence in penetration probability estimates. By adapting statistical methods from ballistic testing, this study presents a framework for a more robust, data-driven approach to stab-resistant armour evaluation, offering greater insight into protective performance across a continuous range of impact energies.

## 1. INTRODUCTION

Current stab-resistant body armour certification standards, including the widely adopted National Institute of Justice (NIJ) Standard-0115.00 and the Home Office Body Armour Standard (CAST 2017), rely on fixed-energy impact testing at defined protection levels. These levels were derived from studies estimating the distribution of impact energies deliverable by a male population during an attack, based on impact force and acceleration measurements from human volunteers [1]. Protection Levels 1, 2 and 3 correspond to the 85$^{th}$, 90$^{th}$, and 96$^{th}$ percentile of delivered energies, respectively [2]. Each protection level involves testing at two fixed energies: a primary test energy (E1), which assesses protective performance at the defined attack energy percentile, and a higher 'overtest' energy (E2). The maximum allowable depth of penetration (DOP) - defined as the maximum length of the threat protruding from the rear surface of the armour panel - at E1 is based on anatomical studies evaluating minimum skin-to-organ distances [3, 4]. At E2, the allowable DOP is larger, as the objective is to ensure the armour does not fail catastrophically at 50% higher energy, rather than to assess injury risk.

NIJ Standard-0115.00 requires 12 impacts per threat type, distributed across three test conditions defined by energy level and impact angle of incidence: four at E1 (0°), four at E2 (0°), and four at E1 (45°). Certification requires all impacts to not exceed the maximum allowable DOP. Since armour designs are often limited by a single test condition, performance in this one condition may determine pass or fail. The small sample size per condition reduces statistical confidence, resulting in a wide confidence interval. If all four impacts pass, the estimated 90% confidence interval (Wilson score [5]) for the true probability of protection is between 0.6 and 1. CAST 2017 and Draft NIJ Standard-0115.01 employ larger sample sizes, improving statistical confidence, yet they remain restricted to testing only at E1 and E2. Consequently, they provide limited insight into protection probabilities across a broad range of impact energies, making it difficult to assess performance beyond these two energy levels.

In contrast, bullet-resistant armour certification combines Penetration-Backface Signature (P-BFS) testing at a reference velocity with Ballistic Limit ($V_{50}$) testing, which applies statistical evaluation methodologies. $V_{50}$ represents the velocity at which there is a 50% probability of complete penetration,


https://doi.org/10.52202/081621-0042

and is estimated using generalised linear models (GLMs) with link functions such as probit and logit regression. This statistical approach enables performance assessment across a continuous range of velocities, improving confidence in protective performance evaluation and allowing estimation of critical quantiles such as $V_{05}$. While stab-resistant armour testing follows a fixed-energy approach to demonstrate protective performance, similar to P-BFS testing, statistical methods such as those used in $V_{50}$ methodology have not been applied.

This paper introduces $E_{50}$ testing as a framework for evaluating stab-resistant body armour, adapting statistical methodologies applied in ballistic $V_{50}$ testing. It has two main objectives: first, to determine whether binary regression models can effectively characterise penetration probability as a function of impact energy; and second, to evaluate the accuracy and efficiency of various sensitivity test designs in generating data for regression-based estimation of $E_{50}$ and $E_{05}$ quantiles using Monte Carlo simulations. This framework provides a more statistically robust approach to assessing armour performance, enabling probabilistic evaluation across a continuous range of impact energies and improving confidence in protection levels beyond current fixed-energy certification methods.

*1.1 Sensitivity test designs*

Sensitivity test designs provide a structured approach to dynamically selecting stress levels, which in this context refers to the applied test variable – such as velocity in ballistic testing or impact energy in stab testing – based on observed binary outcomes. Several established test designs were considered for stab-resistant body armour evaluation, including those commonly used in ballistic $V_{50}$ testing such as Bruceton (up-and-down) (UD) [6], modified up-and-down (MUD), and Langlie (LM) [7] methods, as well as advanced designs such as Neyer (NM) [9] and Three-Phase Optimal Design (3pod) [10] methods.

UD is widely used due to its simplicity, following a staircase approach where stress levels are sequentially adjusted based on previous outcomes. Testing begins at an initial estimate of mean response, $\mu_g$, where 50% of tests are expected to result in a pass, and a constant step size, $d$, is set to an estimated standard deviation, $\sigma_g$. Subsequent stress levels are determined by $x_{i+1} = x_i - d(2y_i - 1)$, where $y_i$ represents the binary outcome of the $i^{\text{th}}$ test. While UD effectively converges on the true mean, it performs poorly when $\mu_g$ or $\sigma_g$ are inaccurate. Additionally, since it concentrates test points around the mean, it is not suitable for estimating extreme quantiles. A modified version, MUD, improves efficiency by employing a fixed number of variable step sizes that decrease after pass-fail reversals. This refinement has been widely adopted in ballistic $V_{50}$ testing, including MIL-STD-662F, NIJ Standard-0101.06, and NATO AEP-2920. Further modifications are seen in NIJ Standard-0101.06, where the initial test velocity is not set at the expected mean but at a lower, reference velocity. This adjustment increases the amount of data collected at lower stress levels, improving estimation in the lower quantile range.

LM is employed in NIJ Standard-0101.04 for $V_{50}$ estimation and in MIL-STD-331D for evaluating weapon initiation sensitivity. It begins with two initial estimates: a stress that will always result in a failure, $\mu_{max}$, and one that will always result in a pass, $\mu_{min}$. The first stress level is conducted at the midpoint, $x_1 = (\mu_{min} + \mu_{max})/2$. If it fails, the next test is placed midway between $x_1$ and $\mu_{min}$; otherwise, it is set midway between $x_1$ and $\mu_{max}$. This process continues until a pass-fail reversal is observed. Subsequent levels are determined by averaging the previous stress with the earliest prior stress where the number of passes and failures were balanced. If no such point exists, the next level is set by averaging the previous stress with $\mu_{max}$ if the last test passed, or $\mu_{min}$ if it failed. LM converges on the mean and is generally more efficient than UD, but it can perform poorly if the initial $\mu_{max}$ and $\mu_{min}$ do not bound the true transition region between pass and fail results. To improve this, a US Department of Defense modification [8] continually adjusts $\mu_{min}$ and $\mu_{max}$ by $2\sigma_g$ after the third test until a pass-fail reversal is observed.

NM, also employed in MIL-STD-331D, follows a structured three-phase approach. Phase I begins with initial estimates for $\mu_{min}$, $\mu_{max}$ and $\sigma_g$, using a binary search algorithm [9] to efficiently locate the pass-fail transition region. This search adapts dynamically, expanding its range if initial estimates are inaccurate. Phase II starts once the difference between the lowest failure and highest success is within $\sigma_g$. At this stage, temporary parameter estimates ($\hat{\mu}_T$,$\hat{\sigma}_T$) are computed, and stress levels are selected using D-optimal design, which maximises the determinant of the Fisher Information matrix. This ensures that stress levels are placed to maximise information gain, which allows for relatively efficient determination of all quantiles. Phase III is triggered once overlap occurs, meaning the lowest failure

stress is below the highest passing stress. Here, maximum likelihood estimates (MLE) of $\hat{\mu}$ and $\hat{\sigma}$ replace the temporary ones from phase II, and D-optimal stress placement continues to refine quantile estimates. 3pod also follows a three-phase approach but differs from NM in phase I and phase III. Instead of the binary search used in NM, 3pod applies an alternative algorithm in phase I to locate the transition region, continuing until overlap is achieved. Phase II is identical to NM, applying D-optimal placement using MLE estimates to refine estimation of all quantiles. In phase III, 3pod introduces the Robbins-Monro-Joseph (RMJ) method [11], which focusses test points near a specific quantile of interest, improving the accuracy of extreme quantile estimation.

Monte Carlo simulations showed that 3pod estimated extreme quantiles more accurately than NM due to the targeted quantile placement method in phase III [10]. However, Neyer [12] found that NM achieved better accuracy for extreme quantiles when $c$-optimal design was used in phase II and III, as it placed stress levels around the estimated quantile of interest. NM also outperformed 3pod in phase I when initial estimates were inaccurate, as NM expanded its search range exponentially, while 3pod expanded linearly. In response, Wang [13] revised 3pod to 3pod2.0, modifying the phase I search logic to expand more rapidly if repeated pass (or fail) sequences were observed. This study used 3pod2.0 due to its improved phase I performance.

*1.2 Binary regression models*

The binary outcomes from sensitivity tests are commonly analysed using binary regression models to estimate parameters $\hat{\mu}$ and $\hat{\sigma}$, which describe the probability of response as a function of stress. The probability of response at stress $x$ is modelled as $\pi(x) = G(z)$, where $G(z)$ is a known cumulative probability distribution function, and $z$ is the linear predictor given by $z = (x - \hat{\mu})/\hat{\sigma}$. The choice of $G(z)$ determines how the probability of response transitions between 0 and 1 as stress increases. The corresponding link function, which maps probabilities back to the linear predictor, is the inverse of $G(z)$, given by $G^{-1}(\pi(x)) = (x - \hat{\mu})/\hat{\sigma}$.

Common link functions include the logit, probit, complementary log-log (cloglog) and log-log functions (Table 1). The logit function, based on the logistic cumulative distribution, is recommended in NIJ Standard-0101.06 for ballistic $V_{50}$ analysis. The probit function, based on the normal cumulative distribution $\Phi(z)$, is used in NATO Standard AEP-2920 for estimating the standard deviation in $V_{50}$ testing. Both assume a symmetric probability transition between pass and failure. The cloglog function, based on the Gumbel cumulative distribution, models asymmetric transitions, where failure probability increases gradually at low stress levels but sharply at higher levels, making it useful in survival analysis and reliability engineering. Conversely, the log-log function models the opposite asymmetry, where failure probability increases rapidly at low stress levels but transitions more gradually at higher levels. This makes it suitable for scenarios where failures are frequent at low stress but stabilise at higher stress.

**Table 1.** Link functions and their corresponding cumulative distributions.

| Link name | Link function, $G^{-1}(\pi)$ | Distribution function, $\pi(x)$ |
|---|---|---|
| Logit | $\ln\left(\frac{\pi}{1-\pi}\right)$ | $\left(1 + e^{-\frac{x-\hat{\mu}}{\hat{\sigma}}}\right)^{-1}$ |
| Probit | $\Phi^{-1}(\pi)$ | $\Phi\left(\frac{x-\hat{\mu}}{\hat{\sigma}}\right)$ |
| Complementary loglog | $\ln(-\ln(1-\pi))$ | $1 - e^{-e^{\frac{x-\hat{\mu}}{\hat{\sigma}}}}$ |
| Log-log | $-\ln(-\ln(\pi))$ | $e^{-e^{-\frac{x-\hat{\mu}}{\hat{\sigma}}}}$ |

## 2. BINARY REGRESSION FOR MODELLING STAB PENETRATION PROBABILITY

The suitability of binary regression methods was evaluated for modelling penetration probability as a function of impact energy in stab-resistant body armour assessment. Experimental DOP data was used in this study, which was converted to binary outcomes using thresholds of 0 mm (onset of penetration) and 7 mm (maximum allowable penetration based on an injury criterion). The experimental dataset included P1/A Edged Blade and Spike threats, capturing differences in penetration mechanics between the two.


https://doi.org/10.52202/081621-0042

### 2.1 Experimental depth of penetration testing

Testing followed NIJ Standard-0115.00 but included intermediate impact energies ranging from 25 J to 65 J using a rail-guided stab testing apparatus (Figure 1a). The soft armour panels featured a hybrid construction of para-aramid and UHMWPE fabric layers, measuring 400 x 400 mm with a total areal density of 6.7 kg/m$^2$. The stab-resistant materials were sealed within a thermoplastic polyurethane-backed 70D nylon fabric cover. This construction meets NIJ Standard-0115.00 requirements for Edged-Blade Protection Class at Protection Level 2 and Spike Protection Class at Protection Level 1. Furthermore, the panels comply with NIJ Standard-0101.06 Level II for ballistic protection, making them representative of modern, flexible, multi-threat armour solutions. Impacts were positioned with a minimum spacing of 94 mm from the armour panel edge and 60 mm from the backing material edge. Additional impacts with a 60 mm impact-to-edge spacing from the armour panel edge were also conducted during initial testing but were excluded due to observed differences in penetration behaviour, further discussed in Section 5.1.

For the P1/A Edged Blade threat, DOP was measured using a digital microscope with image processing software (ImageJ) and a calibrated scale (Figure 1b), in accordance with Home Office Body Armour Standard (2017) Section 7.7.4.1. The cut length of the rear 70D nylon fabric cover was converted to DOP using NIJ Standard-0115.00 Appendix C. For the Spike threat, DOP was recorded using a digital depth calliper (Mitutoyo, JPN) by measuring the spike length protruding from the armour rear surface. For impacts that did not fully penetrate the rear of the armour, the panel was disassembled, and an LED backlight source was used to identify the last ply with cut or displaced fibres. The number of intact plies from the rear was counted, and a negative DOP was estimated using measured ply thickness.

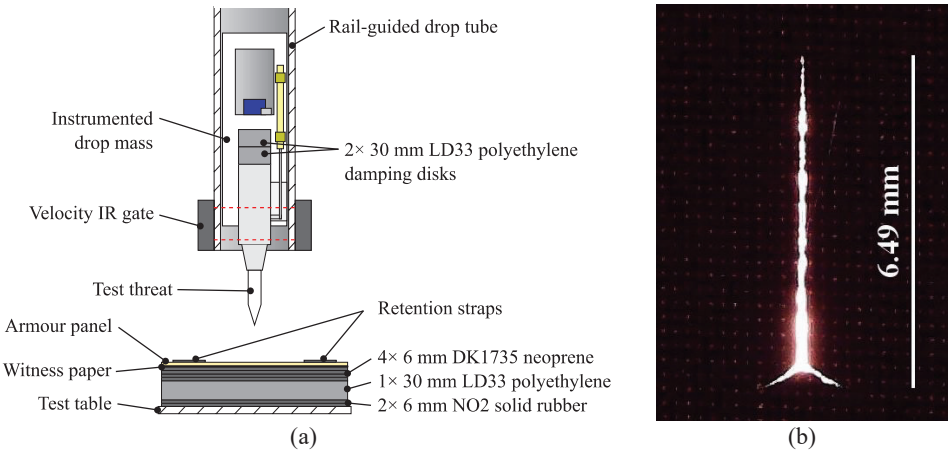



**Figure 1.** (a) illustration of the stab testing setup, and (b) digital measurement of cut length on 70D nylon fabric cover for P1/A Edged Blade threat testing, corresponding to a DOP of 15.1 mm.

### 2.2 Binary regression evaluation

The penetration probability (DOP > threshold) for both the P1/A Edged Blade and Spike threat DOP datasets was modelled using MLE with the four link functions in Table 1. Model fit was assessed using Akaike Information Criterion (AIC) [14] and leave-one-out cross-validation (LOO-CV) [15]. AIC, computed as $AIC = -2\ln(L_{max}) + 2d$, provides a relative measure of model fit, where $L_{max}$ is the maximised likelihood, and $d$ is the number of estimated parameters. Since all four models had the same number of parameters, rankings were determined solely by likelihood, with lower AIC values indicating better fit. LOO-CV estimated model misclassification error by systematically removing each observation, fitting the model to the remaining data, and predicting the excluded case. The cross-validation error rate, calculated as the average proportion of incorrect classifications, quantified model generalisation, with lower values indicating better predictive accuracy. Estimated $\hat{E}_{50}$ and $\hat{E}_{05}$ quantiles, AIC scores, and CV error rates for each link function and DOP threshold are summarised in Table 2.

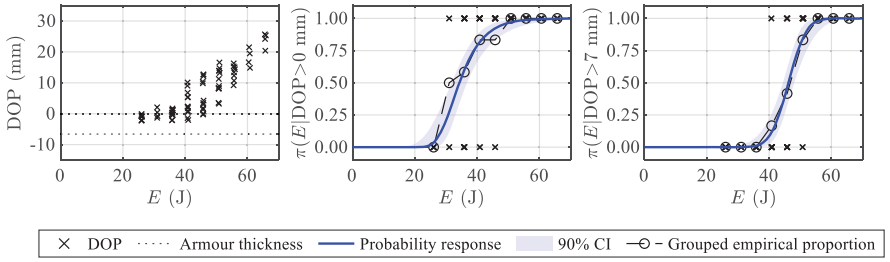
For the P1/A Edged Blade at the 0 mm threshold, the log-log function (Figure 2b) provided the best fit, capturing the asymmetric response with a sharp increase in penetration probability at 25 J. While above the threshold, penetrations at this energy were shallow (≤2 mm), indicating a sudden transition

from non-penetration to minor penetration (Figure 2a). At the 7 mm threshold, the response was more symmetric, making the probit function the best fit (Figure 2c).

For the Spike threat, the cloglog function provided the best fit for both 0 mm (Figure 3b) and 7 mm (Figure 3c) thresholds, indicating penetration probability increased gradually at lower energies but transitioned sharply to near-certain penetration above 45 J. This behaviour differed from the P1/A Edged Blade, which progressively cuts fibres, whereas the Spike primarily displaced them, allowing penetration to continue with minimal additional resistance once initiated.

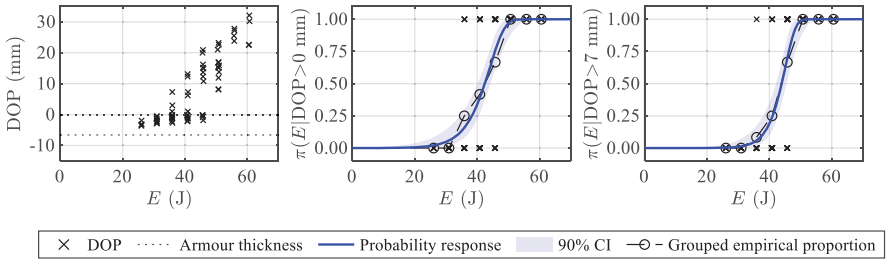
**Table 2.** Estimated $\hat{E}_{50}$ and $\hat{E}_{05}$ quantiles, AIC scores, and cross-validation error rates for four link functions applied to P1/A Edged Blade and Spike data, with DOP thresholds of 0 mm and 7 mm. The best-fitting model for each model and DOP threshold in highlighted in bold.

| Threat | Description | DOP ≤ 0 mm | | | | DOP ≤ 7 mm | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Logit | Probit | Cloglog | Loglog | Logit | Probit | Cloglog | Loglog |
| P1/A | $\hat{E}_{50}$ (J) | 35.0 | 35.2 | 35.6 | 34.3 | 46.3 | 46.2 | 47.0 | 45.4 |
| | $\hat{E}_{05}$ (J) | 22.7 | 23.2 | 15.2 | 26.6 | 38.6 | 38.8 | 37.4 | 39.4 |
| | AIC | 52.1 | 51.9 | 53.3 | **51.0** | 43.3 | **42.9** | 43.1 | 44.0 |
| | CV Error (%) | **14.5** | **14.5** | 21.1 | **14.5** | **11.8** | **11.8** | **11.8** | 14.5 |
| Spike | $\hat{E}_{50}$ (J) | 41.6 | 41.6 | 42.4 | 40.7 | 43.3 | 43.2 | 43.9 | 42.4 |
| | $\hat{E}_{05}$ (J) | 31.8 | 32.3 | 30.2 | 33.3 | 36.3 | 36.2 | 35.3 | 36.3 |
| | AIC | 52.6 | 52.0 | **51.4** | 53.3 | 41.2 | 41.0 | **39.8** | 43.4 |
| | CV Error (%) | **17.6** | **17.6** | **17.6** | 22.1 | **11.8** | **11.8** | **11.8** | **11.8** |




(a) (b) (c)

**Figure 2.** P1/A Edged Blade threat ($n$=76) depth of penetration results and best-fit probability response curves: (a) depth of penetration vs impact energy, (b) penetration probability response curves using loglog link for 0 mm threshold, and (c) penetration probability response curves using probit link for 7 mm threshold.




(a) (b) (c)

**Figure 3.** Spike threat ($n$=68) depth of penetration results and best-fit probability response curves: (a) depth of penetration vs impact energy, (b) penetration probability response curves using cloglog link for 0 mm threshold, and (c) penetration probability response curves using cloglog link for 7 mm threshold.

The difference in $\hat{E}_{50}$ estimates between 0 mm and 7 mm thresholds further highlighted distinct penetration mechanics between the threat types. For P1/A Edged Blade threat, the 11 J difference indicated that deeper penetration required additional energy, likely due to progressive fibre cutting. In


https://doi.org/10.52202/081621-0042

contrast, for the Spike threat, this difference was only 2 J, highlighting that once material fibre displacement occurred, further penetration required minimal added energy.

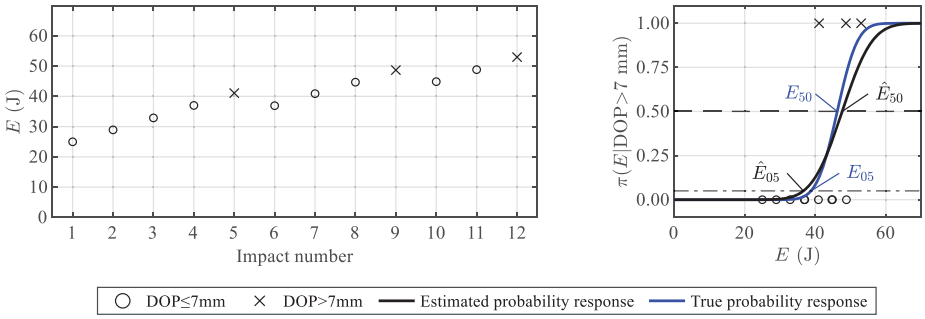
For the remainder of this study, 0 mm was selected as the DOP threshold for the Spike threat, given the small energy difference between shallow and deep penetration, it aligned with current CAST (2017) acceptance criteria for the Spike threat, and the potential biohazard risk associated with any penetration in correctional environments [16]. For the P1/A Edged Blade threat, the 7 mm threshold was used as a practical threshold, as the cutting penetration mechanism ensured shallow penetrations remain stable, aligning with NIJ Standard-0115.00 acceptance criteria.

## 3. MONTE CARLO SIMULATIONS

Monte Carlo simulations were conducted to evaluate the accuracy and efficiency of nine sensitivity test designs for $E_{50}$ and $E_{05}$ estimation. These included UD, LM, MUD, and three variations each of NM and 3pod2.0. NM variants shared the same phase I design but differed in subsequent test placement: NMD used D-optimal design, NMC applied $c$-optimal design targeting $\hat{E}_{05}$, and NMDC employed a hybrid approach, switching from D-optimal to $c$-optimal after the first 24 impacts. For the 3pod2.0 variants, 3pod2A used only phases I and II, while 3pod2B and 3pod2C targeted $\hat{E}_{50}$ and $\hat{E}_{05}$, respectively, in phase III, which initiated after 24 impacts.

The simulations were repeated for two underlying true probability distributions, based on the best-fitting experimental models, to provide penetration probabilities at each simulated test energy: a normal (probit) model for the P1/A Edged Blade ($\mu$ = 46.2 J, $\sigma$ = 4.52 J, 7 mm threshold) and a Gumbel (cloglog) model for the Spike threat ($\mu$ = 44.2 J, $\sigma$ = 4.71 J, 0 mm threshold). To reflect real-world uncertainty in $\mu$ and $\sigma$, each test design was evaluated across 81 initial parameter combinations ($\mu_g$ = 20-60 J, $\sigma_g$ = 1-9 J). The MUD method followed an adapted NIJ Standard-0101.06 ballistic $V_{50}$ testing approach, starting at 24 J and adjusting step sizes ($2\sigma_g$, $1.5\sigma_g$, or $\sigma_g$) based on pass-fail reversals, resetting every 12 impacts. For LM, NM and 3pod2.0 test designs, initial search bounds were set to $\mu_G \pm 4\sigma_G$, as recommended by Langlie [7] and previously validated as suitable for NM and 3pod2.0 methods [9, 10].

Each test design, parameter combination and assumed true distribution were simulated for 1000 test series with a sample size of $n$ = 48 impacts. To assess sample size effects, results were also analysed after 12, 24, and 36 impacts. Simulated impact energy was sampled from a normal distribution centred at the target energy, $E_{target}$, with heteroscedastic error $\sigma = aE_{target}{}^{b}$ (where $a$ = 7.12e-4 and $b$ = 1.35, based on experimental variability), and was constrained between 5 J and 70 J to maintain realistic test conditions. A single simulated UD test series with sample size $n$ = 12 is illustrated in Figure 4a. At the conclusion of each test series, parameter estimates $\hat{\mu}$ and $\hat{\sigma}$ were obtained using MLE, and $\hat{E}_{50}$ and $\hat{E}_{05}$ estimates were compared against their true values, $E_{50}$ and $E_{05}$, respectively (Figure 4b).




(a) (b)

**Figure 4.** Simulated UD test series with $\mu_g$=25 J and $\sigma_g$=4 J: (a) binary outcomes and energies showing test series progression, and (b) estimated probability response from the simulated data (blue) and the underlying true probability response (black).

https://doi.org/10.52202/081621-0042

## 4. RESULTS OF SIMULATED TEST DESIGNS

The accuracy of each test design was evaluated by comparing $\hat{E}_{50}$ and $\hat{E}_{05}$ estimates to their true values using root-mean-square error (RMSE), as given in Equation (1). To ensure meaningful estimates, RMSE was calculated only for simulations that met validity criteria. MLE required overlap between pass and fail results, but additional criteria were imposed: the lowest simulated energy had to result in a pass, the highest had to result in a failure, the slope of the probability response curve had to be positive, and $\hat{E}_{05}$ had to be greater than zero. The proportion of simulations that met validity criteria, $P_{valid}$, represents the efficiency of a test design in generating data suitable for parameter estimation (Figure 5a), where white shading indicates greater efficiency.

To compare the accuracy across test designs, RMSE values were normalised by their respective true quantiles to account for differences in energy scale. The mean normalised RMSE, $\overline{NRMSE}$, was computed across both quantiles, as given in Equation (2), and is presented in Figure 5b as an overall accuracy score, where white shading indicates greater accuracy. Although $\overline{NRMSE}$ equally weighted both quantiles, rankings were more influenced by $\hat{E}_{05}$ accuracy due to its typically larger error. This was intentional, as precise estimation of $E_{05}$ was more critical from a protective equipment perspective, whereas $E_{50}$ accuracy primarily reflected overall model fit.

$$RMSE_p = \sqrt{\sum_{j=1}^{n_{valid}} \frac{1}{n_{valid}} \left(\hat{E}_{p,j} - E_p\right)^2} \tag{1}$$

$$\overline{NRMSE} = \sum_{p\in\{5,50\}} \frac{1}{2}\left(\frac{RMSE_{p,P1/A}}{E_{p,P1/A}} + \frac{RMSE_{p,SPIKE}}{E_{p,SPIKE}}\right) \tag{2}$$

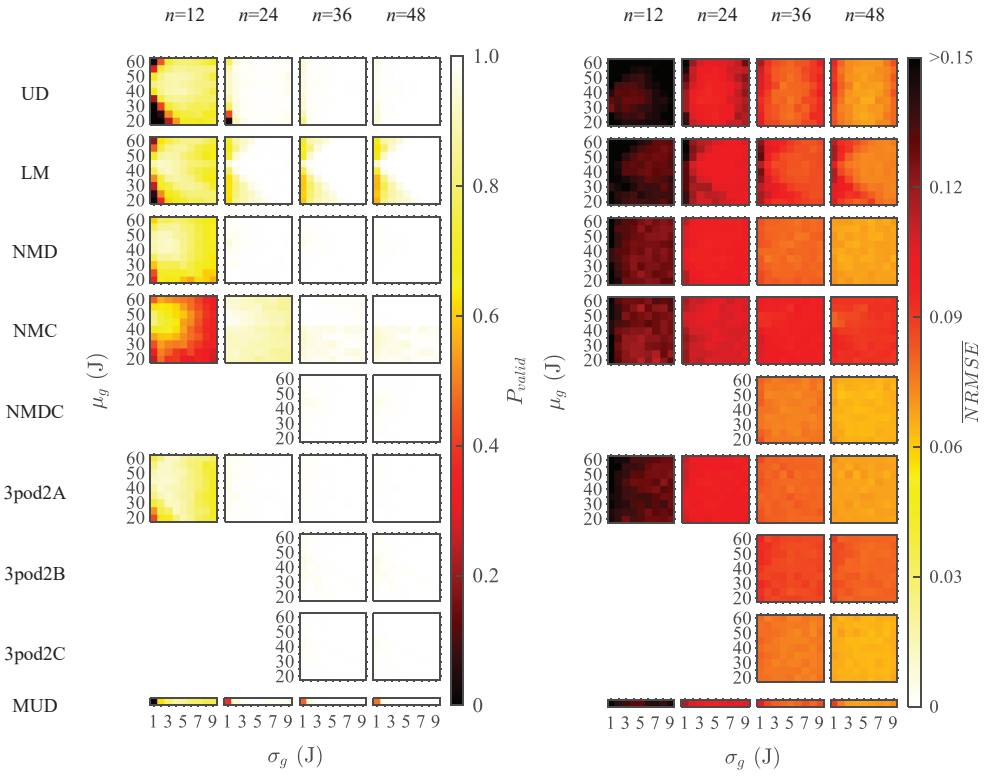



**Figure 5.** Efficiency and accuracy of simulated test designs for *n*=12, 24, 36, and 48: (a) proportion of simulated tests that achieved validity criteria, where white shading indicates greater efficiency, and (b) mean normalised root-mean-square error, where white shading indicates greater accuracy.


https://doi.org/10.52202/081621-0042

## 5. DISCUSSION OF SIMULATED TEST DESIGNS

The accuracy of each test design aligned with its intended purpose – methods focused on $E_{50}$ estimation performed best for $E_{50}$, while those designed for $E_{05}$ estimation excelled in that metric. UD and LM yielded the lowest $RMSE_{50}$, with typical errors of 1.8 J (4.4% from true values), 1.4 J (3.0%), 1.0 J (2.5%), and 0.9 J (2.1%) for $n$ = 12, 24, 36 and 48, respectively, when initial parameters of $\mu_g$ (40-50 J) and $\sigma_g$ (1-5 J) were close to true values. However, outside this range, $P_{valid}$ was low at smaller sample sizes (Figure 5a). UD improved efficiency with additional sample size, whereas LM remained constrained by its poor initial bounding and was slow to adapt. Despite low $RMSE_{50}$, these methods had higher $\overline{NRMSE}$ (Figure 5b), as improved $E_{50}$ accuracy came at the expense of $E_{05}$ estimation, with errors typically 6.2 J (18%), 4.8 J (14%), 3.8 J (11%), and 3.5 J (10%) for each sample size. MUD improved $\hat{E}_{05}$ estimates compared to UD by incorporating additional data in the lower energy range. MUD performed best with initial parameter $\sigma_g$ = 5 J, with $RMSE_{50}$ values of 2.2 J (5%), 1.6 J (3.5%), 1.2 J (2.8%) and 2.5%, while $RMSE_{05}$ was 5.8 J (17%), 4.5 J (13%), 3.5 J (10%), and 3.1 J (9%). Further work could investigate the effect of balancing step size increments and the number of impacts before resetting the energy to 24 J.

The advanced methods NMD and 3pod2.0, using D-optimal design, generally performed well across all sample sizes, offering a better balance between $E_{50}$ and $E_{05}$ estimation. Both yielded similar $RMSE_{50}$, typically 2.1 J (4.8%) and 1.5 J (3.4%) for $n$ = 12 and 24, respectively, while $RMSE_{05}$ was typically 5.7 J (16%) and 4.6 J (13%). Both methods were robust across a range of initial parameter estimates, except for very small $\sigma_g$, highlighted by the shading uniformity across the combinations of initial parameters (Figure 5b). NMC provided the most accurate $E_{05}$ estimates, however at the expense of $E_{50}$ estimation. NMC also struggled to achieve the overlap condition at low sample sizes, potentially due to its early transition to $c$-optimal design. A modification where $c$-optimal placement is triggered immediately after overlap, rather than when the highest pass and lowest failure differ by less than $\sigma_g$, may improve its performance.

After 24 impacts, NMDC and 3pod2C were the best-performing designs, with similar results that were robust across all initial parameters. Typical $RMSE_{50}$ values were 1.4 J (3.1%) and 1.3 J (3.0%) for $n$ = 36 and 48, respectively, while $RMSE_{05}$ was typically 3.3 J (9.7%) and 2.7 J (7.9%). These methods highlighted the effectiveness of first refining parameter estimates using D-optimal design before switching to $c$-optimal, concentrating energies near $\hat{E}_{05}$. Further work may assess the effect of switching the design approach earlier for smaller sample sizes. The distribution of the estimated response curves for the best-performing test designs at each sample size is illustrated in Figure 6.

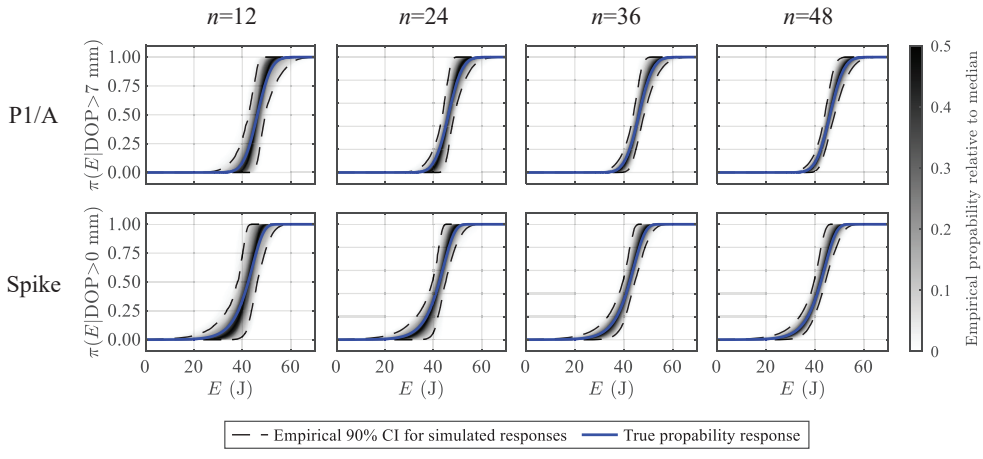



**Figure 6.** Distribution of probability response curves from 1000 simulations of the best-performing test designs for each sample size: (a) NMD, $\mu_g$=55 J, $\sigma_g$=5 J, (b) MUD, $\sigma_g$=5 J, (c) NMDC, $\mu_g$=40 J, $\sigma_g$=4 J, and (d) NMDC, $\mu_g$=50 J, $\sigma_g$=2 J. Simulations not meeting validity criteria were excluded.

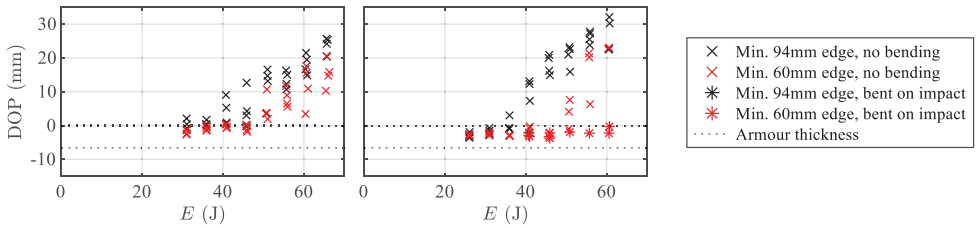
The findings in this study were based on a single armour construction, which may limit their applicability to other material systems. Armour types such as ring mesh, metal plate, or rigid epoxy-fibre

https://doi.org/10.52202/081621-0042

composites, may exhibit different penetration mechanics that were not captured in this study. Further investigation is required to determine whether these materials follow similar probability distributions. However, provided penetration behaviour can be described by a monotonically increasing cumulative distribution function, then $E_{50}$ methodology should remain applicable. Additionally, while 1000 simulations were conducted per test design, some statistical variation in the results is expected.

### 5.1 Effect of impact location on depth of penetration

Impacts placed 60 mm from the armour edge, while valid under NIJ Standard-0115.00, exhibited significantly lower DOP than equivalent tests performed closer to the centre. Figure 7 shows DOP results at an intermediate stage in testing, before edge impacts were discontinued. This discrepancy was attributed to lower effective stiffness of the armour and backing material due to lack of lateral constraint. As a result, greater material displacement occurred during impact, reducing peak impact forces and decreasing DOP. For Spike threats, asymmetric deformation of the armour and backing near the edge promoted plastic buckling of the slender spike (Figure 7b), particularly at higher energies. While edge effects can influence all armour types, flexible, loose-ply, fabric systems may be particularly susceptible.




(a) (b)

**Figure 7.** Depth of penetration results for minimum 94 mm and 60 mm impact-to-edge distance on armour panel, (a) P1/A Edged Blade threat, and (b) Spike threat.

If edge impact locations had been retained in the experimental dataset, the estimated standard deviation of the penetration probability response would have been artificially inflated. In practical terms, this would increase the likelihood of achieving a passing result when testing near an edge, leading to an overestimation of $\hat{E}_{50}$ and other quantiles such as $\hat{E}_{05}$, making the assessment less conservative.

## 6. CONCLUSIONS

This study demonstrated that binary regression models effectively described penetration probability as a function of energy in stab-resistant body armour evaluation. From the experimental DOP data, the best-fitting models captured the distinct mechanics of each threat, with the probit model providing the best fit for the P1/A Edged Blade and the cloglog model for the Spike threat. These models were applied using DOP thresholds of 7 mm for the P1/A Edged Blade and 0 mm for the Spike, providing practical performance evaluation limits while aligning with current acceptance criteria in accordance with existing standards. Among the simulated test designs, NMDC, NMD, and 3podC provided the most accurate $\hat{E}_{50}$ and $\hat{E}_{05}$ estimates, even when initial parameters deviated significantly from true values. These methods not only provided superior precision in quantile estimation but also demonstrated robustness across a range of conditions, making them strong candidates for implementation in stab-resistant armour evaluation.

$E_{50}$ testing has the potential to enhance stab-resistant body armour evaluation by enabling probabilistic performance characterisation across a continuous energy spectrum. If adopted within standardisation and conformity assessment models, $E_{50}$ could complement existing fixed-energy testing, similar to the way ballistic standards combine P-BFS and $V_{50}$ methods. Within such a framework, quantile-based acceptance criteria, such as a minimum allowable $E_{05}$ value for a given Protection Level, could be analogous to the $V_{05}$ criterion used in NIJ Standard-0101.06 for ballistic certification. Any decision to adopt $E_{50}$ testing rests with the appropriate standardisation bodies, which would need to define suitable DOP thresholds, acceptable quantile estimation error, and sample size requirements based on regulatory and operational considerations. The framework developed in this study may assist in informing those decisions by illustrating how such parameters could be evaluated and applied.


https://doi.org/10.52202/081621-0042

If implemented, $E_{50}$ testing would likely require increased sample sizes compared to current certification methods, leading to a modest increase in certification cost. However, by aligning more closely with ballistic testing practices, it offers significant value to armour material developers, manufacturers, program managers, and end-users. It enables direct, threat-specific, performance comparison across armour systems, supports more informed procurement decisions, and improves confidence in meeting minimum performance requirements during batch, lot acceptance, and follow-up inspection testing. It also facilitates targeted design optimisation by identifying design-limiting threats within protection classes, reducing overdesign and supporting more efficient material use.

Beyond development and initial certification, $E_{50}$ testing could be applied to Surveillance Testing Programs, commonly adopted for bullet-resistant armour but not yet leveraged for stab-resistant armour. Establishing an initial $E_{50}$ dataset enables the calculation of an allowable degradation margin, as defined in NIJ STD 0101.06 Explanatory Material for Section 7.9.5. Periodic penetration probability assessments can then monitor any performance degradation as the armour ages throughout its life-of-type, critical for forecasting armour service life. This alignment with ballistic armour evaluation represents a significant step forward in the standardisation and long-term assessment of stab-resistant body armour performance.

**Acknowledgments**

This research was supported by the Australian Government through the Department of Education's National Industry PhD Program (project 36179). The views expressed herein are those of the authors and are not necessarily those of the Australian Government or the Department of Education.

**References**

[1] Horsfall I., Prosser P., Watson C. and Champion S., An assessment of human performance in stabbing', Forensic Science International, vol. 102(2-3) 1999, 79-89.
[2] NIJ Standard 0115.00, Stab Resistance of Personal Body Armor, National Institute of Justice, 2000.
[3] Connor S.E.J., Bleetman A. and Duddy M.J., Safety standards for stab-resistant body armour: a computer tomographic assessment of organ to skin distances, Injury, vol. 29(4) 1998, 297-299.
[4] Bleetman A. and Dyer J., Ultrasound assessment of the vulnerability of the internal organs to stabbing: determining safety standards for stab-resistant body armour, Injury, vol. 31(8) 2000, 609-612.
[5] Wilson E., Probable inference, the law of succession, and statistical inference, Journal of the American Statistical Association, vol. 22(158) 1927, 209-212.
[6] Dixon J.W. and Mood A., A method for obtaining and analyzing sensitivity data, Journal of the American Statistical Association, vol. 43(241) 1948, 109-126.
[7] Langlie H.J., A reliability test method for 'one-shot' items, Technical report U-1792, Aeronutronic Division of Ford Motor Company, Newport Beach, California, 1965.
[8] Johnson T.H., Freeman L., Hester J. and Bell J.L., A comparison of ballistic resistant testing techniques in the Department of Defense, IEEE Access, vol. 2 2014, 1442-1455.
[9] Neyer B., A D-optimality-based sensitivity test, Technometrics, vol. 36(1) 1994, 61-70.
[10] Wu C.F.J. and Tian, Y., Three-phase optimal design of sensitivity experiments, Journal of Statistical Planning and Inference, vol. 149 2014, 1-15.
[11] Joseph V.R., Efficient Robbins-Monro procedure for binary data, Biometrika, vol. 91(2) 2004, 461-470.
[12] Neyer B., Comparison of the Neyer D-optimal and 3pod sensitivity test designs, JSM 2014, Boston, Massachusetts, US.
[13] Wang D., Tian Y. and Wu C.F.J., Comprehensive comparisons of major sequential design procedures for sensitivity testing, Journal of Quality Technology, vol 52(2) 2020, 155-167
[14] Akaike H., Information theory and an extension of maximum likelihood principle, Selected papers of Hirotugu Akaike, Springer New York, 1998, 199-213
[15] Stone M., Cross-validatory choice and assessment of statistical predictions, Journal of the Royal Statistical Society: Series B (Methodological), vol. 36(2) 1974, 111-113.
[16] Greene M.E., Horlick J., Longhurst D.A., Miller L.A., Robinson C. and Sundstrom R.A., NIJ standards for ballistic resistance of body armor and stab resistance of body armor: new developments, Personal Armour Symposium Proceedings 2023, Dresden GER.