PASS (Personal Armour Systems Symposium) 2008

A statistical approach to proof testing

Ian Horsfall, Trevor Ringrose, Celia Watson
Cranfield University, Defence Academy of the United Kingdom, Shrivenham, Wilts. UK.

1. Abstract

Police body armour has undergone rapid evolution, and this is due in part to a relatively simple type approval process which matches relatively small numbers purchased by individual police forces. However, as body armour technology and usage has increased there has been a gradual change of emphasis from solving the immediate protection problems of highly specialised systems towards quality assurance towards of a standard item of equipment. In addition, as the amount and age of armour in service increases, concerns have been raised about methods for ensuring the continued performance over long periods of use or over long production runs.

These factors have drawn attention to the statistical significance of existing proof tests, in which armour systems are subjected to small numbers of stabs or ballistic tests. A number of approaches have been suggested including the addition of a V_{50} ballistic limit test to provide a fully quantitative measure of performance [1]. However this approach also lacks statistical rigour and further enhancements such as regression analysis [2,3] have been suggested to remedy this.

In the current work a different statistical approach is suggested in which conventional proof tests can be used to produce statistically robust data of known significance. Initial trials on current police body armour showed that ballistic penetration and knife penetration were similar as the data was highly random and it was difficult to statistically predict individual test results. Ballistic blunt trauma followed a more predictable pattern with simple and easily predicted test-to-test variation allowing good predictions to be made.

For the knife and ballistic penetration tests two approaches have been investigated. One method is a point estimate approach that determines failure probability as a simple ratio of pass or fail. Therefore to achieve a failure probability of lower than 0.1 (10%) no more than 1 failure in 10 would be allowed. The second option would be determine how many successful tests were needed to be sure (for instance to 95% probability) that the failure rate was no more than 0.1. This second approach is more severe and it has been shown that at least 28 successful tests are required in order to be reasonably sure that the failure rate is less than 0.1.

This paper will demonstrate the development of the statistical model which has been used within 2007 HOSDB body armour standards and shows how it is applied in both type approval and batch testing.

2. Background

Typical body armour test regimes in the UK [4,5] and USA [1,6] rely on external test houses performing body armour type approval tests against the relevant test standards. The test requires manufacturers of body armour to submit samples for type approval tests which if successful then allow that armour to be sold to police forces. This system has generally worked well but it is recognised that the increased and widespread use of body armour raises some challenges.

One problem is that manufacturers typically provide warranties on the armour for only five years. Beyond this time there is currently no means to re-test armour in order to determine whether it may have been damaged or deteriorated. Secondly the pace of armour development and deployment has now slowed so that armour designs of more than five years ago are still competitive and therefore manufacturers wish to continue to sell them. However there is no means to ensure that changes in materials or manufacture have not altered performance relative to the system which was originally tested. Even when armour designs have short lifetimes there may be cases where the production runs are very large and various factors might cause variation on the performance. Both raw materials manufacturers and armour manufacturers do usually carry out quality control checks but until 2007 there was no requirement for this to be done and more importantly no defined re-test method.

What was needed was a means to re-test armour in order to establish whether its performance has changed, and related to this a method to batch test armour as a function of production period or amount. In principle this could be achieved by simply re-testing using the type approval method however this has serious problems as any armour must have a finite failure probability and therefore will fail if tested enough times.

Ideally armour systems would be designed to never fail, but with current technology this would result in armour too thick and heavy for extended use. Even with more advanced technology it is unlikely that it would be desirable to design armour to never fail as it is also necessary to meet other criteria such as weight, bulk, and cost. Designing against any failure with no regard to the other factors would result in armour less comfortable, and therefore less likely to be worn. Other factors such as increased physical exertion and poorer mobility might even increase the overall risks to the wearer. Therefore it is necessary to design for a finite failure probability and to provide a suitable test regime. This regime must discriminate between failures due to acceptable statistical variation and failures due to poor performance.

2.1 The purpose of armour testing

The first question is to determine the purpose of the test regime, both for initial certification and for re-test. If the purpose is quality control then a thorough test regime with a statistically large sample would have to be enforced both for initial certification and subsequent re-testing. This is probably not desirable as the large cost and increased timescale would almost certainly slow down technological development by reducing the throughput of new designs and making it difficult for small or new companies to gain access to the market.

The current test regime can alternatively be viewed as an approval of the armour design, with an implication that quality control is the responsibility of the manufacturer and/or the purchaser. Under this regime the manufacturer takes on the risk of armour passing or failing and has to judge the appropriate margin required in order to ensure that it is successful in certification.

If this latter concept of the type approval and re-test is used then at the type approval stage we make the assumption that the armour does not work and the test is designed to prove beyond reasonable doubt that the armour does in fact work. At the re-test stage the assumption is made that the armour does in fact work and the re-test then simply has to show that this assumption is reasonable. This is similar to regimes of normal and tightened inspection in British and International standards [7].

3. Analysis of pre-2007 tests

An initial study of the statistics of armour performance focused on the stab resistance tests conducted against the pre 2007 HOSDB stab resistance test standard [8] for which there was a large database of test results. The test procedure involves a single armour sample which is stabbed four times at a specific test energy (denoted E1). This is the design requirement of the armour system and the maximum allowable penetration is 7mm. Four more tests are then conducted at 150% of this initial energy (denoted E2). This is to determine that the armour does not fail catastrophically when overmatched and maximum allowable penetration is 20mm. The certification test also requires an additional four angled stab tests and in some cases other additional tests but these have been omitted for simplicity.

The initial approach was to examine what statistical information could be obtained from the existing data and in addition some repetitive test were commissioned to provide large data sets for some types of armour. The analysis focused on the E1 and E2 tests as these test produced all the recorded failures. In these tests the level of penetration is recorded and a pass is achieved if all the measured penetrations are below the required level. If any one test exceeds the requirement then the armour fails. Because some armour designs easily meet the E1 criteria but fail E2 whilst other armour designs easily meet E2 but fail E1 it is necessary to treat these as two completely separate tests. But at each energy level the result will be the outcome of only four pass/fail tests and will consequently have a very poor statistical basis. For example if an armour has a true pass probability at E1 of 50% then the chance of it being certificated is the same as that of tossing a coin and getting heads four times

$$0.5^4 = 0.06 \text{ (6\% probability of passing)}$$

With only four pass/fails tests to work from, not only is the chance of passing a poor armour not particularly low but the chances of passing a good armour are also not particularly high. So armour which actually has a 98% probability of passing will have a probability of passing the test of

$$0.98^4 = 0.92 \text{ (8\% chance of failure)}$$

Figure 1 shows a plot of the probability of an armour passing as a function of the pass probability for E1 and E2 test criteria. It can be seen that for the displayed probability range (0.7-1) the overall pass probability range is spread out and the probability of passing changes only slowly with the actual chance of achieving the test criteria.

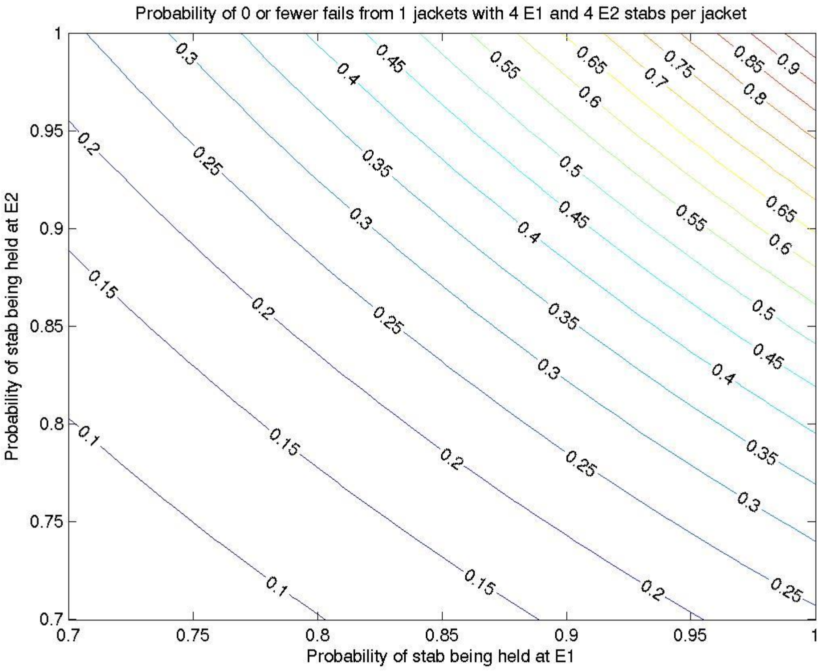

Figure 1 The probability of armour passing a four E1 plus four E2 stab test as a function of the actual probability of meeting the two test criteria.

Increasing the number of tests has the effect of reducing the spread of probabilities so that there is a much faster change from low pass probability to high pass probability. However for larger number of tests there is an increasing tendency to pick up the tail of the distribution curve so although it becomes more likely for a bad armour to fail is also becomes less likely for even a very good armour to pass. Therefore as the number of test increases it is desirable to condone some level of excess penetration.

Figure 2 shows the probability of an armour passing if the number of test is increased to 24 and one over-penetration is condoned. If this is compared to the earlier case (figure 1), then it can be seen that the effective pass probability of a reasonable jacket can remain the same whilst the chance of failing better jackets decreases slightly and the chance of passing poor jackets decreases markedly.

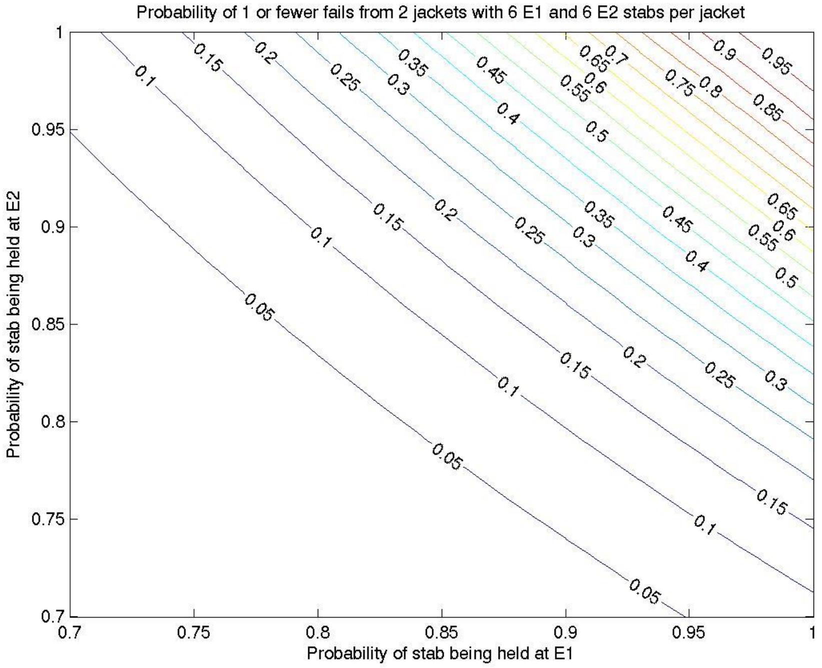

Figure 2 The probability of an armour passing, based on a total of 24 tests with one failure allowed.

The result of this analysis is to show that the number of tests must be increased to provide a reliable method of certification, but it does not indicate what this new number should be.

4. Re-testing

When a multiple test regime is introduced some additional effects need to be accounted for. If a single over-penetration is allowed in every batch of tests then it follows that if the armour is tested 10 times then up to 10 fails will have to be condoned. More importantly if an armour is expected to average 1 fail in every batch then in some batches no over-penetrations will occur whilst in others two or more will occur and the armour will fail. This is unlikely to be a problem for a design sold in small numbers as it might be expected to be retested only once or twice. Providing its probability of passing is relatively high (95%) then its chance of passing twice more will be

$$0.95^3 = 0.86 \text{ (86\%)}$$

However if the same armour were sold in large numbers, for instance 5000 sets and re-test was set as one in every 500 then the probability of the armour passing on all occasions becomes

$$0.95^{10} = 0.59 \quad (59\%)$$

Therefore for a re testing procedure a different approach may be required in order to provide a method which allows good armours to continues to pass but which spots a deteriorating system.

5. Confidence Intervals

In order to solve the problem posed above it is necessary to determine how our confidence in the test results varies as a function of the number of tests or test failures. It is proposed that this could be done by calculating the likely failure probability based on the test data, using confidence intervals (CIs)

There are several approaches for the construction of confidence intervals (CIs) for binary probabilities, in this case the probability of a failure of a given armour in a given test. The CIs given here are the 'exact' Clopper-Pearson intervals, which are generally regarded as the best. However in this case we really only care about the upper limit, i.e. how high could the failure probability be, so that we consider one-sided intervals. The lower limit of all the intervals quoted is implicitly zero. The upper confidence limit is given by that value of the failure probability θ, such that we are 95% confident that the true value of θ is less than or equal to this.

Suppose we have 16 stabs with no failure. Our best guess of the failure probability is clearly zero, but this is not very useful. The upper limit to the possible value of θ is taken to be a value such that the observed result of no failures has some specific low probability. Typically this might be 5%, giving 95% CIs.

For example, if the chance of the armour successfully holding the stab is 0.8, then the failure probability θ is 1 - 0.8 = 0.2 and the chance of no failures, which is what we observe, is just 0.02815 or 2.8%.

$$0.8^{16} = 0.02815$$

Similarly if θ is 1 - 0.9 = 0.1 then the chance of getting no failures is 0.18530 or 18.5%.

$$0.9^{16} = 0.18530$$

A failure probability of 1-0.829250=0.170750 gives a probability of getting no failures of 0.05000 or 5%.

$$0.829250^{16} = 0.05000$$

Hence if the failure probability exceeds 0.170750 then the chance of getting no failures is less than 5%, whilst if the failure probability is less than 0.170750 then the chance of getting no failures is greater then 5%. In this sense, we are therefore ‘95% confident’ that the failure probability θ is less than or equal to 0.170750. Any value in the interval (0, 0.1707450) is a credible value for the true failure probability θ in the sense that it gives a non-trivial chance of observing as few failures as we actually did. When one or more failures are observed then the idea is similar except that we now consider the probability of observing the number of failures which we actually did or fewer.

6 An approach to testing and re-testing

We start with a strict test on new armour, so that if it passes then we are pretty sure that it is good. The burden of proof is placed on the manufacturer to convince us beyond reasonable doubt that the armour is adequate.

In the re-test, the burden of proof is relaxed, similarly to conventional sampling procedures [7] where manufacturers with a good past record are given the benefit of the doubt. Hence the re-tests are much smaller and easier to pass.

This strict-then-relaxed approach reduces the problem of multiple tests increasing the chance that armour fails at some point, but does not remove it. There is always the chance that good armour will fail at some point or that bad armour will pass. The result is that if an armour fails a re-test then it loses the benefit of the doubt and must pass a full test again.

6.1 Initial test

Table 1 shows the upper confidence limits for the failure probability for various choices of number of stabs (10-50) and number of failures allowed (zero or one). Hence if we choose 30 stabs then if there are no failures we are 95% confident that the true failure probability is less than or equal to 0.0950. This is of course an upper limit on the likely failure probability, so the true value is very probably lower. Similarly, if there is one failure then we are 95% confident that the true failure probability is less than or equal to 0.1486. If we want to be 95% sure that the true failure probability is least than 0.1 then we need to test (at least) 29 stabs with no failures allowed. Alternatively, it could be (at least) 46 stabs with one failure allowed.

6.2 Re-test

One possibility for this is to say that the re-test is passed if the point estimate of the failure probability in the re-test is within the confidence interval from the original test. However, this is of course a semi arbitrary criterion and other criteria could be used. Hence if the original test is 29 stabs with no failures, so that the upper confidence limit is 0.0981, then a re test could allow 1 fail in 10.

Table 1 95% confidence limits as a function of number of stabs

	Upper 95% Confidence Limit	
Stabs	if No Failures	if One Failure
10	0.2589	0.3942
11	0.2384	0.3644
12	0.2209	0.3387
13	0.2058	0.3163
14	0.1926	0.2967
15	0.1810	0.2794
16	0.1707	0.2640
17	0.1616	0.2501
18	0.1533	0.2377
19	0.1459	0.2264
20	0.1391	0.2161
21	0.1329	0.2067
22	0.1273	0.1981
23	0.1221	0.1902
24	0.1173	0.1829
25	0.1129	0.1761
26	0.1088	0.1698
27	0.1050	0.1640
28	0.1015	0.1585
29	0.0981	0.1534
30	0.0950	0.1486
31	0.0921	0.1441
32	0.0894	0.1398
33	0.0868	0.1359
34	0.0843	0.1321
35	0.0820	0.1285
36	0.0798	0.1251
37	0.0778	0.1219
38	0.0758	0.1189
39	0.0739	0.1160
40	0.0722	0.1132
41	0.0705	0.1106
42	0.0688	0.1080
43	0.0673	0.1056
44	0.0658	0.1033
45	0.0644	0.1011
46	0.0630	0.0990
47	0.0617	0.0970
48	0.0605	0.0951
49	0.0593	0.0932
50	0.0582	0.0914

6.3 Chances of passing the test and re-test

The most important feature of any test/re-test setup is that of how likely any given armour is to pass or fail it. This, of course, can only be calculated if we use some value for the true probability of failure against a single test. To illustrate the consequences of a particular decision about what test should be used table 2 shows the probability of an armour with a given true failure probability passing tests and re-tests.

The tests illustrated are those where the test allows no failures while the re-test allows one failure. The number of stabs allowed in the re-test is chosen so that, with one failure in the retest, the point estimate from the re-test will be just inside the confidence interval from the original test. The probabilities of passing tests and re-tests are then given for true failure probabilities of 0.0001 to 0.2. These are given first for a test of 28 stabs with a re-test of 10 stabs, and for comparison these are followed by much smaller tests of 16 and 6 and much larger tests of 46 and 16 stabs respectively.

Table 2 Chances of passing a given test and re-tests (n stabs, *r* fails allowed)

True	Test	Retest	Test	Retest	Test	Retest
failure	n = 28	n = 10	n = 16	n=6	n = 46	n = 16
probability	r=0	r = 1	r=0	r = 1	r=0	r = 1
0.0001	0.9972	1.0000	0.9984	1.0000	0.9954	1.0000
0.0002	0.9944	1.0000	0.9968	1.0000	0.9908	1.0000
0.0003	0.9916	1.0000	0.9952	1.0000	0.9863	1.0000
0.0004	0.9889	1.0000	0.9936	1.0000	0.9818	1.0000
0.0005	0.9861	1.0000	0.9920	1.0000	0.9773	1.0000
0.0006	0.9833	1.0000	0.9904	1.0000	0.9728	1.0000
0.0007	0.9806	1.0000	0.9889	1.0000	0.9683	0.9999
0.0008	0.9778	1.0000	0.9873	1.0000	0.9639	0.9999
0.0009	0.9751	1.0000	0.9857	1.0000	0.9594	0.9999
0.001	0.9724	1.0000	0.9841	1.0000	0.9550	0.9999
0.002	0.9455	0.9998	0.9685	0.9999	0.9120	0.9995
0.003	0.9193	0.9996	0.9531	0.9999	0.8709	0.9989
0.004	0.8938	0.9993	0.9379	0.9998	0.8316	0.9982
0.005	0.8691	0.9989	0.9229	0.9996	0.7941	0.9971
0.006	0.8449	0.9984	0.9082	0.9995	0.7582	0.9959
0.007	0.8214	0.9979	0.8937	0.9993	0.7239	0.9945
0.008	0.7986	0.9972	0.8794	0.9991	0.6911	0.9929
0.009	0.7764	0.9965	0.8653	0.9988	0.6598	0.9911
0.01	0.7547	0.9957	0.8515	0.9985	0.6298	0.9891
0.02	0.5680	0.9838	0.7238	0.9943	0.3948	0.9601
0.03	0.4262	0.9655	0.6143	0.9875	0.2463	0.9182
0.04	0.3189	0.9418	0.5204	0.9784	0.1529	0.8673
0.05	0.2378	0.9139	0.4401	0.9672	0.0945	0.8108
0.06	0.1768	0.8824	0.3716	0.9541	0.0581	0.7511
0.07	0.1311	0.8483	0.3131	0.9392	0.0355	0.6902
0.08	0.0968	0.8121	0.2634	0.9227	0.0216	0.6299
0.09	0.0713	0.7746	0.2211	0.9048	0.0131	0.5711
0.10	0.0523	0.7361	0.1853	0.8857	0.0079	0.5147
0.11	0.0383	0.6972	0.1550	0.8655	0.0047	0.4614
0.12	0.0279	0.6583	0.1293	0.8444	0.0028	0.4115
0.13	0.0203	0.6196	0.1077	0.8224	0.0017	0.3653
0.14	0.0147	0.5816	0.0895	0.7997	0.0010	0.3227
0.15	0.0106	0.5443	0.0743	0.7765	0.0006	0.2839
0.16	0.0076	0.5080	0.0614	0.7528	0.0003	0.2487
0.17	0.0054	0.4 730	0.0507	0.7287	0.0002	0.2170
0.18	0.0039	0.4392	0.0418	0.7044	0.0001	0.1885

0.19	0.0027	0.4068	0.0343	0.6799	0.0001	0.1632
0.20	0.0019	0.3758	0.0281	0.6554	0.0000	0.1407

7. Conclusions

Using table 1 it is possible to determine the number of tests required in order to establish the required minimum failure probability. It would be desirable to work towards the lowest possible failure probability in order to maximise the wearer's safety. However the number of tests required will increase to unmanageable levels if the failure probability is set too low. The main disadvantages will be costly and lengthy certification procedures, which will in turn prevent new entrants to the body armour market and reduce the frequency of new designs being certificated. One of the major triumphs of the current system has been the dramatic reduction in armour weight and stiffness over the last 15 years which has lead to widespread acceptance of modern armour for continuous and universal use by patrolling officers.

A failure probability of 0.1 appears to be a relatively high failure rate if taken at face value. However if this is assessed by an upper CI approach it ensures that the actual failure probability is less, and probably much less than this value. From table 2 it can be seen that an armour with a true failure probability of 0.1 will only stand a 5% (0.0523) chance of passing a 28 stab test with no failures. In order to stand even a 50% chance of passing, an armour will have to have a true failure probability of less than 0.03.

If the test regime were to be designed around a true failure probability of 0.1 with no failures then this would require a minimum of 29 stabs (or shots). This should be compared to the older ballistic test [9] which uses 6 shots of each calibre (even if these are combined the upper CI is only 0.22), or the stab test in which only four E1 stabs are conducted. Therefore in the 2007 HOSDB test standards [4.5] the certification test for ballistic resistance against handguns requires 30 shots for each of the two ammunitions, whilst for stab resistance 30 tests at E1 are required. All other tests including testing carried out for batch or re test purposes are based on 10 stabs or shots with 1 failure being allowed.

8. Acknowledgements

The authors would like to acknowledge HOSDB for supporting this work and particularly John Croft and Daniel Longhurst for their support and guidance.

8. References

1. NIJ 0101.04, Ballistic resistance of personal body armor, US department of Justice, 2000.
2. Kneubűhl, B.P, Improved test procedures for body armour, Proc. PASS 1996.
3. Gotts, P.L, Fenne, P.M, Leeming, D.W, The application of critical performance analysis to UK military and police personal armour, Proc. PASS 2004.

4, HOSDB Body Armour Standard for UK Police (2007), Part 2. Ballistic Resistance Publication No.39/07/B.

5. HOSDB Body Armour Standard for UK Police (2007), Part 3. Knife & Spike Resistance Publication No.39/07/C.

6. NIJ 0115.00, Stab Resistance of Personal Body Armor: US department of Justice, 2000.

7. (BS 6001-0 (ISO 2895-0), Sampling procedures for inspection by attributes.

8. PSDB Body Armour Standards for UK Police (2003), Part 3 Knife and Spike Resistance Police Scientific Development Branch, Home Office Publication No. 7/03/C.

9. PSDB Body Armour Standards for UK Police (2003), Part 2 Ballistic Resistance, Publication No 7/03/B. Home Office Police Scientific Development Branch.