Appendix E

Options for the Department of Justice

Contents

Box

Figures

Table

Appendix E
Options for the Department of Justice

GENERAL

This appendix describes and assesses several options that the Department of Justice could exercise to revise NIJ Standard 0101.03 and/or the process by which compliance with it is certified, in order to

- limit the variance in test conditions,
- provide more information on ballistic resistance of certified armor (including uncertainties and limits of ballistic resistance, dependence on wearer, etc.),
- decrease producers' financial risks as well as consumers' safety risks, and
- assure consumers that certified armor offered for sale is as good as the samples tested for certification.

Some of the options could be undertaken by the National Institute of Justice (NIJ) without additional authority or funding. Others-research and quality-assurance programs—would require substantially increased funding.

Status Quo

One option is to postpone any change to NIJ Std. 0101.03 and the current method of certifying compliance with it. The argument for this is that armor of styles certified to comply with NIJ Std. 0101.03 has saved many lives (see app. B) and is not known to have failed, in actual assaults, to stop any bullet of a type that it was certified to resist, nor to prevent lethal blunt trauma. Yet the criterion for protection from blunt trauma is not so strict that many models fail it: as of Oct. 31, 1991, of the 555 models submitted for testing for NIJ certification of compliance with the .03 standard, only 15 failed solely because of excessive backface signature (BFS), the test's index of risk of blunt trauma.

The vast majority of the failures were caused by penetration, alone (166) or in combination with excessive BFS (40). Most of the dissatisfaction of some parties with the current standard stems from these failures, or from penetrations in retests. Complaints charge that the test is "a crap shoot" (i.e., not reproducible) or too stringent. These and other arguments against the status quo were summarized in appendices A and B.

Arguments for the alternative options discussed in the remainder of this appendix are also arguments against the status quo.

Make Conventional Practices Mandatory

On several occasions since NIJ Std. 0101.03 was issued, NIJ has instructed H.P. White Laboratory, Inc., (HPWLI) in letters, telephone calls, or meetings, to perform certain test procedures in certain ways consistent with the standard. In effect, these instructions rule out other ways of performing test procedures that could reasonably be considered consistent with the printed standard. Sometimes this was done to clarify a portion of the standard; in other cases it was done with the intent of reducing variability of results that might be attributable to variability of test procedures. For example, in 1988 an official at NIST directed that the test facility use only 124-grain, FMJ 9-mm bullets made by Remington. [82]

On other occasions, HPWLI has informed NIJ that, unless instructed otherwise, it would henceforth perform certain test procedures only in certain ways but not in other ways consistent with the printed standard. Again the intent was to reduce variability. Sometimes NIJ would indicate its concurrence; sometimes NIJ would object, proposing a different procedure. For example, on March 28, 1988, HPWLI informed NTJ-in response to a modification made earlier in the month by TAPIC that the locations of shots 4 and 5 be altered slightly to ensure nonalignment with each other and the new location of shot 6-that shot 5 be raised 1 inch and shot 4 left unchanged. In May of the same year, TAPIC responded with a letter approving the new shot locations. [82]

On still other occasions, HPWLI has proposed to change certain test procedures in a way that actually departs somewhat from those specified in the printed standard, but is clearly justifiable on technical grounds. Such proposed changes are not implemented until approval is received. For example, on October 10, 1989, HPWLI proposed that the 30-degree obliquity of the fourth and fifth shots be rotated so as to be combination of horizontal and vertical obliquity, as opposed to the present situation

in which all shots lie in a horizontal plane with respect to the vertical vest. [82]

In at least one instance, NIJ has presented a major procedural change-the replacement of the flat-faced block of clay with a curved, abstractly torso-like fixture (containing a smaller flat-faced block of clay) on which the vest is mounted by its own straps as if worn by an officer-as a possible modification to the 0101.03 standard. This possible change highlights issues always present, albeit perhaps to a lesser degree, when the test procedure is changed:

1. **Does the** change make the test harder or easier to pass? Either way, vests already tested might experience a different outcome if tested again. Manufacturers of vests that failed the earlier test will want a repeat opportunity, while those whose vests passed will seek to avoid further testing.
2. Does the change confer a particular advantage on certain manufacturers?
3. What artificialities have been introduced? While it would be naive to suppose that any test or test procedure could avoid all artificialities, it is wise to consider these artificialties before they are introduced. In the case of the curvilinear test fixture, one might well ask what will happen to a vest if its straps give way during the test. Will it be picked up off the floor and reattached? If so, vest manufacturers will strive for the most tenuous possible attachment so that their vests can be picked up and smoothed out as many times as possible, reducing or even eliminating bunching and balling. If not, does the vest fail if its straps come undone? What if one strap breaks and the vest droops, obscuring the next shot's line of fire? What if an unfair shot penetrates the vest and hits the opposite panel, arguably weakening it?

The underlying point is that procedural changes have become de facto parts of the standard. NIJ should consider incorporating them into the next version of NIJ Std. 0101. Of course, some of these instructions and practices may become obsolete if the current standard is changed in other respects. It would be especially important to incorporate the applicable instructions and practices into the standard if NIJ should authorize a different laboratory to test armor for certification (or quality assurance).

Specify Backing Material

A simple but possibly helpful change would be to specify the backing material to be used. In practice, only one backing material, Roma Plastilina No. 1 modeling clay, is used by HPWLI for NIJ certification tests. However, NIJ Standard 0101.03 does not require it; it *defines* "backing material" as "a block of nonhardening, oil-base modeling clay placed in contact with the back of the test specimen during ballistic testing." This is confusing, because a variety of materials other than modeling clay are often used as backing in tests for other purposes than NIJ certification. Examples include 10-percent ballistic gelatin, 20-percent ballistic gelatin, rigid foamed polystyrene (Styrofoam), foamed polyurethane rubber, RTV silicone rubber, soap, plywood, human and animal cadavers, and live animals. Of these, only Styrofoam and soap are sufficiently inelastic for use for deformation measurement in an NIJ-like test (i.e., without high-speed cinematography or other expensive techniques).

The definition is also confusing because, although clay is *placed in* contact with the back of the test specimen at the *beginning* of ballistic testing according to NIJ Standard 0101.03, the standard prohibits ''disturbing the relationship between the armor and the backing material' to assure that the clay *remains in* contact with the back of the test specimen *during* ballistic testing (or for any other purpose). Amending the definition of backing material in section **3** (Definitions) of the standard would improve clarity, whether or not a particular backing material is specified in section 4 (Requirements) or section 5 (Test Methods).

Laboratories in England, France, and Germany have used other types of modeling clay as backing material and found that deformation is affected by choice of material. For example, researchers in England have calibrated deformation in Plastilina to deformations in Plasticize and PP2 as a function of bullet velocity. In these comparisons all three backings were conditioned so as to pass the drop test specified in NIJ Std. 0101.03. This required heating Plasticize to temperatures higher than the maximum allowed by NIJ Std. 0101.03. [28, 29, 84] As noted above, some experts consider backing temperature unimportant provided the drop test is satisfied. However, strict adherence to all provisions of NIJ Std. 0101.03, including allowable temperature, would exclude use of Plasticize and perhaps some other

backings sometimes used. This has not been an issue in NIJ certification testing; H.P. White Laboratory uses only Roma Plastilina No. 1.

Even if different backing materials can pass the drop test at temperatures within the allowed range, specifying only one of them might improve reproducibility. It is possible that the consistency (flowability) of candidate backing materials might depend strongly, but differently, on the rate of deformation.[1]

Some backing materials conditioned to produce comparable drop-test results yield different backface signatures at the much higher deformation velocities typical of a ballistic test conducted in accordance with NIJ Std. 0101.03. For example, in tests conducted by the British Police Scientific Development Branch, under otherwise similar conditions the average (viz., fitted) backface signatures produced in U.S.-made Plastilina and U.K.-made Plasticize were similar at impact velocities of **350** m/s but differed by about 4.4 mm for each 100 m/s above or below 350 m/s. **[29;** cf. 28] Thus, the drop test does not assure that backface signatures produced in different backing materials behind similar armors by similar bullets impacting at similar velocities will be the same. Some materials are known to yield different results; others, not yet tested by NIJ or NIST, could diner more dramatically. Specification of a backing material would eliminate this potential source of variation in-or operator influence on—test conditions.

Although clay composition demonstrably affects the results of the deformation test (for protection from nonpenetrating bullets), it is not certain that it affects the results of the penetration test. More research would be needed to find out whether it does.

Reduce Allowable Range of Backing Material Temperature

One way to reduce or at least limit the variability of test conditions is to reduce the range of acceptable temperatures of the backing material. Currently, the clay's temperature can be anywhere between 15 and 30 'C, i.e. 59 and 86 'F. Tightening this tolerance up, however, might make little real difference because the backing material must also pass a drop test, in which a special weight is dropped 2 meters and the resulting dent must be between 22 and 28 millimeters in depth. Some experts consider backing temperature unimportant provided the drop test is satisfied. [69, 29] The standard does not require use of Roma Plastilina No. 1, but does point out that this nonhardening modeling clay fulfills the requirements of the test.

Research by the Aerospace Corp. indicated that the volume (especially) and surface area of the crater produced in Roma Plastilina No. 1 by the drop test is very sensitive to temperature, and the Aerospace Corp. recommended that the temperature of this backing material be maintained in the range 68 to 72 OF. [8] The Aerospace Corp. calculated crater volume and surface area from depth and diameter measurements, assuming the crater to be a right circular cone. Using the same approximation, OTA has reconstructed the unrecorded depth and diameter measurements and found that crater depth is less sensitive to temperature than is crater volume (see figure E-1).[2]

The drop test, if performed at the beginning and at the end of a test, would standardize the consistency to some extent, but it is doubtful that it is an adequate substitute for temperature control. For example, if the clay block were left for many hours in an area colder than 59 'F, then brought into an area maintained at 59 'F and kept there for 3 hours, the surface of the clay block might warm enough so that the drop test could be passed, indenting the clay only about 25 mm. But in subsequent testing, a shot might push the armor into a deeper, colder, stiffer layer of the clay-e.g., to the BFS limit. Were the clay at that depth warmer, as required by the standard, the BFS test would be failed. But in practice, in testing observed by OTA, clay temperature is not measured during testing, nor is the drop test performed after the beginning of a test.[3] Thus in practice temperature may not be controlled to within specified tolerances, which would allow considerable inadvertent or operator-controlled variation in test conditions.

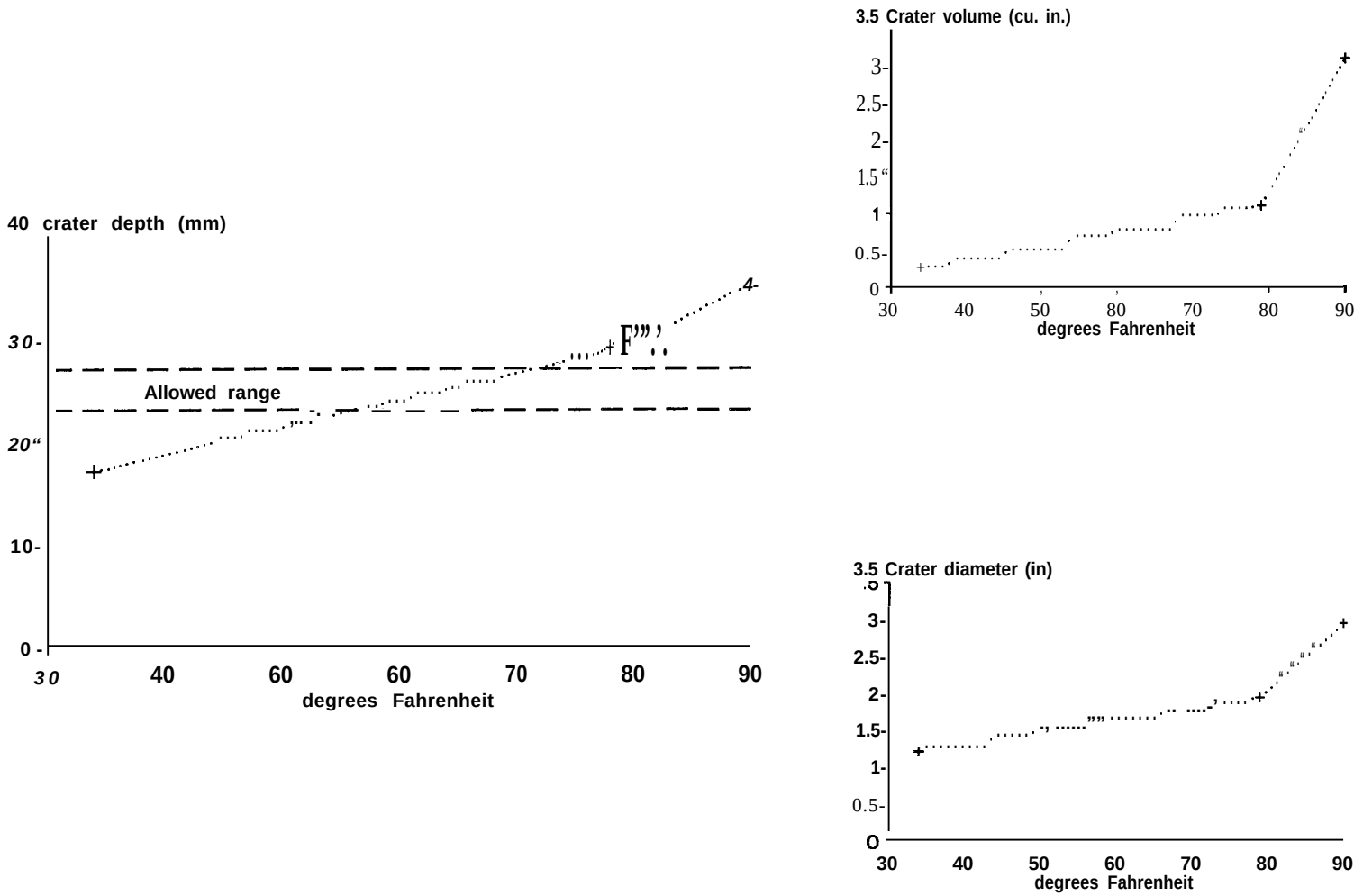
---

1 The consistency of Silly Putty®, a familiar toy item, illustrates strong strain-rate-dependence.

2 Damon the temperature sensitivity of Plasticize are shown in [281.

3 In testing observed by OTA at H.P. White Laboratory, Inc., clay was conditioned and ambient temperature was maintained within the tolerances allowed by NIJ Standard 0101.03. Moreover, clay was routinely stored at the temperature used for conditioning, even before the 3-hour conditioning period prescribed by the standard. The conditioning temperature was warmer than the ambient temperature, and the face of the clay block cooled during testing. To recondition the block as prescribed by the standard, warmer, softer clay was taken from storage and used to fill the craters made by previous shots in the test sequence.

Figure E-l—Variation of Drop-Test Crater Dimensions with Temperature



SOURCE: Office of Technology Assessment, 1992.

Specifying that backing temperature be measured at several depths and locations and that the drop test be performed both before and after (and perhaps during) ballistic testing would insure that backing temperature is controlled to the current standard, and reducing the allowed temperature range further (e.g., to68to72‘F) would further improve control of test conditions and possibly the reproducibility of test results.

Another reason for doubting that the drop test is an adequate substitute for temperature control is the fact that deformation depends in a nonlinear way on the momentum of the dropweight or, in testing, the bullet. [8, 122, 123] These are quite different: the l-kilogram dropweight has a calculated momentum of 0.64 kg-m/son impact; but an 8-gram 9-mm bullet at 332 m/s (the nominal type II-A velocity) would have a calculated momentum of 2.7 kg-m/s. Deformation also varies nonlinearly with temperature, as shown in figure E-1, so the variation with (i.e., the sensitivity to) temperature at the momenta of bullets probably differs from that at the momentum of the dropweight on impact. However, we have no data characterizing the sensitivity to temperature at the momenta of bullets.

Although the drop test was developed to test the consistency of backing material for the purpose of standardizing the deformation test (for protection from nonpenetrating bullets), variation of consistency such as that shown in figure E-1 may also affect the results of the penetration test. Research would be needed to find out whether it does.

Certify Wet and Dry Ballistic Resistance Separately

The wet test could be mandatory or optional. The case for certifying dry ballistic resistance even if armor does not have, or is not tested for, wet ballistic resistance is that because of cost or comfort, many purchasers and wearers prefer armor with inadequate or untested wet ballistic resistance. They may suspect that the risk of its becoming dangerously wet is so low that they would accept it.

However, to learn what the risk is, they would have to weigh their armor regularly to measure and record water retention and analyze the records to calculate frequency with which retention exceeds dangerous levels. There is a risk that some may err in this, or not attempt it.

Even if it is done correctly, so that purchasers and wearers make an informed choice to accept the risk, it will be a higher risk than they would be exposed to if they bought *and wore* wet-certified armor. But in compensation, wear rate might be increased among those who find armor with inadequate wet ballistic resistance more affordable or comfortable but who also value NIJ’s certification.

Officers could weigh their armor panels at the beginning and end of each shift to measure moisture pickup, which they could record. However, this would indicate moisture *content,* which affects ballistic resistance, only if the armor were completely dry at the beginning of the shift. Some officers complain (to us) that their armor does not dry completely between shifts. Some officers may require two or more garments each in order to have a dry garment to wear while others are drying.

Even if officers measure and record the wetness of their armor, predicting the risk of future wetness and the uncertainty in the risk would be complicated, beyond the abilities of most officers and many departments. Aids in the form of worksheets or computer software would be required, along with training. The frequency with which dangerous wetness has occurred in the past is a reasonable (viz., a maximum-likelihood) estimate of the risk of dangerous wetness in the future, under similar conditions (e.g., season and duty). However, because the occurrence of dangerous wetness is apparently rare, there would be a substantial chance that the estimated risk would be inaccurate. To assess this risk, purchasers or wearers would have to calculate confidence limits on the estimated risk.

Subjecting armor only to the dry testing specified in the NIJ standard would reduce the stringency of the test, even for armor that performs as well wet as dry. For example, armor that is unaffected by moisture and has a 97-percent mean probability of stopping a bullet would have a 70 percent probability of passing a 12-shot dry test and would probably pass it; but if subjected to a wet-dry test (or a double dry test) of 24 shots, the same armor would more likely than not have failed (52 percent probability). If NIJ wished to compensate for this and maintain the stringency of the test, it could offer a choice of the current wet-dry test or a double-dry test with the same number of fair shots required.

To halve the cost of testing, one industry source has proposed testing and certifying dry ballistic

resistance *or* wet ballistic resistance, but not requiring both tests. This is based on the premise that no conceivable type of armor has less ballistic resistance when dry than when wet. This is plausible, but even if true, armor would have a higher probability of passing a wet-only test than a wet-dry test with twice as many shots.

ASSESSING RESISTANCE TO PENETRATION

Smooth Armor Between Shots

[This topic was discussed in vol. 1.]

Use a Torso-Shaped Test Fixture

Appendix C notes that one of the technical issues surrounding the .03 standard is its requirement that armor be tested by removing its ballistic panels, strapping each to a flat block of clay, and shooting. This deprives the armor, in such testing, of any benefit (e.g. against bunching) it might derive from its own strapping or the carrier garment itself. A torso-shaped test fixture, be it a mannequin or a ''curV,' ' would lessen or eliminate these problems.

Use *Resilient Backing for Penetration Test*

NILECJ-STD-0101.00 , issued in 1972, specified the use of 'a block of nonhardening modeling clay" as backing for the ballistic deformation test it described but *not* for the ballistic penetration test, which was to be air-backed. Three reasons were later given for the choice of air backing:

> First, excluding the backing material greatly simplifies the . . . projectile-fabric interaction; not only is the overall experimental scatter [variation in results] reduced, but the test results may be directly related to projectile-fabric interaction [alone].
>
> Second, exit velocities of the projectiles were desired; . . .
>
> Last, high-speed photography is much simpler without a backing material. [7]

However, the frost advantage cited was offset by the fact that there was little data relating air-backed test results to the projectile-fabric interaction on a torso, human or otherwise. Moreover, high-speed photography and measurement of exit velocities, although useful in research, are unnecessary in a test of resistance to penetration, and indeed NILECJ-STD-0101.00 did not require them. Accordingly,NILECJ-STD-O1O1.O1, which was issued in 1978, specified the use of a nonresilient backing material for testing both deformation and penetration. Like the current NIJ standard, it noted that Roma Plastilina No. 1 modeling clay was "found to be suitable" as a backing material but did not require its use, although it did specify a drop test to be performed to check the consistency of backing material.

As noted in appendix C, some critics of the current NIJ standard contend that the best technical option would be to use an inelastic backing such as clay for the blunt trauma test and an elastic backing for the penetration test.[4]

Other ballistic measurement techniques using costly apparatus might be adapted to measure deformation of resilient backing. Examples include multiflash photography, which has been used to measure deformation versus time in air backing; [39] multiflash x-radiography, which has been used to measure penetration (hence deformation) versus time in composite armor;[5] and Doppler radar, which has been used to measure velocity versus range of small projectiles impacting and penetrating media transparent to microwaves.[6] As of late 1990, the range resolution of the radar was 6.25 cm—too coarse to measure backface deformations with the accuracy needed for predicting blunt trauma. A planned improvement in signal processing was expected to improve (decrease) the range resolution tenfold, to 0.625 cm—still too coarse. Higher frequency, and costlier, millimeter-wave radar would probably be needed to provide the range resolution needed for predicting blunt trauma. Such apparatus could conceivably be afforded and used by a major ballistic testing facility such as H.P. White. However, specification of a backing that would require their use would void a major objective of the NIJ test procedure-to be reproducible at ballistic facilities typical of those used by many police departments, with no equipment more costly than a ballistic chronograph.

---

4 Dr. Martin Fackler proposed this at the NIJ Body Armor Users Workshop in Reston, Virginia, on June 6, 1990: "Maybe we need the springiness for the repeated testing, for the repeated shots; and for the backface deformation the clay. Maybe we need both of them." See transcript p. 244ll. 14-17.

5 M.S. Stephenson "A Flash X-Ray Study of the Penetration of Ceramic Faced Composite Armours," pp. 143-159 in [134].

6 J.L.M.J. van Bree and E.J.M. van Riet, "Use of a Doppler Radar for Velocity Monitoring of Small-Calibre Projectiles," pp. 261-269 in [134].

Standardize Test Bullets

The probability with which a commercially available bullet of specified mass and caliber will penetrate armor at a specified velocity depends sensitively on details of the bullet's construction and composition, which determine the hardness of the bullet and, more generally, its tendency to deform or fragment when impacting on armor. [28] A bullet that deforms may be stopped by relatively few layers of armor; many more layers may be needed to stop sharp fragments of a hard or steel-jacketed bullet.

Uncommon projectiles, ranging from the Teflon™ Thunderzap [121] to fragment-simulating projectiles, with a variety of so-called "cop-killer bullets" in between, span a greater range of penetration probabilities. This wide array of threats has led the PPAA [113] and the U.K. Home Office [28, 29] to specify test bullets more specifically than does the NIJ standard. In fact, even nominally identical bullets display considerable variation, sometimes even between different bullets in the same box of 50. Some years ago, the U.K. Home Office, noticing the variation in performance of 9-mm bullets of similar mass and velocity, purchased a large lot of one type of 9-mm round and has used it exclusively for the past 10 years, [29] even though variations in it have been noted since 1983. [28]

NIJ could follow this example and specify test bullets more strictly. This would probably increase reproducibility of test results, but it would decrease realism-it would not simulate the diversity of the threat faced by police officers.

Require a Full-Auto Test

According to a major survey, [102] police officers and chiefs are very interested in securing protection from automatic weapons, increasing numbers of which have been confiscated in recent years. However, to date they have been used in a very, very, small fraction of assaults on police officers, and in most of these no more than a very few shots hit the region covered by any one armor panel. As a risk to police officers, such assaults rank far below many others-head shots, for example. Nevertheless, assessment of ballistic resistance to automatic fire may be demanded. Providing it will require special equipment and will be costly.

One argument for the need for such a testis that in an actual assault with an automatic weapon, bunching and balling (ply separation) might occur and, if it does, might be patted down from *inside the armor* by "the dynamic, elastic human torso." However, this abdominal or thoracic undulation might not smooth the armor as completely as manual patting on clay backing would. One approach to assessing armor under such conditions would be to mount it on a resilient backing and expose it to automatic fire in a manner considered to be representative.

Before undertaking such an effort, one should critically examine the plausibility of the postulated biomechanical dynamics, an issue discussed in appendix C.

The Police Scientific Development Branch of the U.K. Home Office has developed a test fixture to expose armor to automatic fire in a predetermined pattern but has had difficulty achieving a reproducible shot pattern. [29]

Require a Ballistic Limit Test

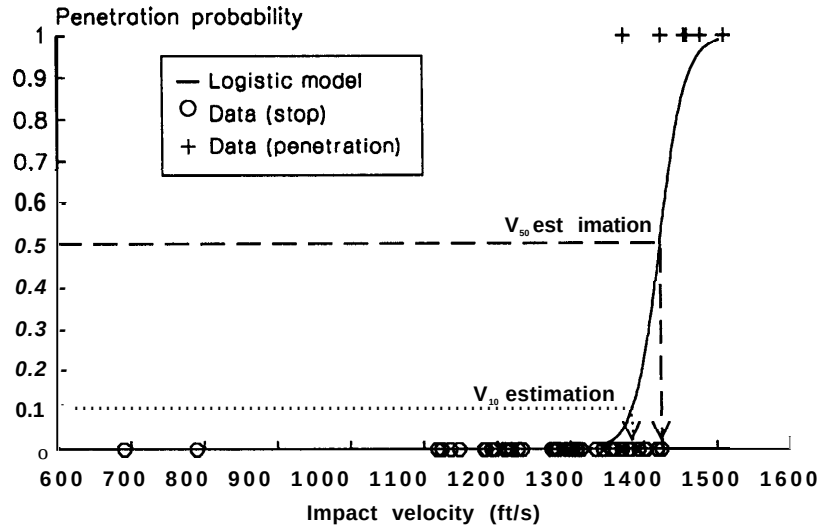
Armor could be subjected to a test to estimate its V_{50} ballistic limit-the velocity at which it has a 50-percent chance of being penetrated by the test projectile.[7] A model could be certified to have a specified type or level of ballistic resistance if the V_{50} estimated for each type of test bullet equals or exceeds a specified minimum value, and if samples also pass a test for protection from blunt trauma. But in addition, the model would be rated by the V_{50} estimate to let purchasers know the margin by which the model exceeds minimum NIJ standards.

The widely used test specified by the Department of Defense's Military Standard MIL-STD-662D [138] could be used. It uses air as the backing material (as did NILECJ 0101.00), but NIJ could specify that clay or some other backing material be used. Regardless of the material used, calibration of penetration probability in the test with penetration probability in assaults would be an issue.

An alternative score is the V_{1O}, the estimated velocity at which test bullets have a 10-percent chance of penetrating-i.e., at which the armor stops a bullet with 90-percent reliability. The V_{1O} could be estimated by logistic regression [91] using the

[7] The DoD test uses air as the backing material, but the NIJ could specify that clay or some other backing material be used. Regardless of the material used, calibration of penetration probability in the test with penetration probability in assaults would be an issue.

Figure E-2—Estimates of V_{50} and V_{IO} Obtained by Logistic Regression



.357 magnum at 30-deg incidence v. test panel on clay, smoothed between shots

SOURCE: Office of Technology Assessment, 1992.

results of a DOD-like test. (See figure E-2.) For purchasers who demand 90-percent, rather than 50-percent, reliability in stopping, the V_{IO} would be more appropriate for comparing to typical or conservative threat velocities (e.g., the minimum velocity specified for bullets in the .03 test) than would the V_{50}.. By the same token, certification could be based on the estimated V_{05} or V_{01}, but estimating these velocities, which correspond to small probabilities of penetration, would require more shots than to estimate the V_{IO} with the same accuracy, which in turn would require more shots than to estimate the V_{50}.

Increase Total Shots and Allow Penetrations

If a very large number of apparently identical armors of the same model and style are subjected to apparently identical tests as specified by NIJ Std. 0101.03, some of the armors would pass and some would fail. Some of the variation in test results might be caused by subtle variations in the armors; another component of the variation might be caused by slight variations in procedure from one test to another. Some of the variation in test results would remain unexplained at any stage in the scientific understanding of the process. Some of the variation—perhaps a small fraction-would be caused by fundamentally random quantum-mechanical processes.

Revising the standard or using a different one could alter-increase or decrease-the variation in test results. However, there will always be a random influence on test outcomes. As a result, an armor of a model that had passed 99 development tests conducted in accordance with NIJ Std. 0101.03 could fail the one NIJ certification test it is allowed. It is likewise possible for an armor of a model not subjected to development tests conducted in accordance with NIJ Std. 0101.03 to pass an NIJ certification and subsequently fail 99 acceptance tests or quality-assurance tests conducted in accordance with the standard.

Clearly these possibilities pose risks-different kinds, to be sure-to manufacturers, purchasers, wearers, and standard-setting authorities. Manufacturers want assurance that good armor does not fail certification testing because of chance variation (''a crap shoot' '), and purchasers and wearers want assurance that bad armor is not certified by a fluke. Certification of bad armor poses a safety risk to wearers as well as a liability risk to manufacturers and departmental purchasers. Any indication that good armor has flunked or that bad armor has been certified, even if not statistically significant, may provoke a challenge to the credibility of the testing and certifiication procedure.

There is a way to decrease the probability of certifying bad armor while at the same time decreasing the probability of flunking good armor. Reducing the consumers' risk requires more testing-g., repetitions of the protocol specified in NIJ Std. 0101.03. Extra testing will, of course, increase cost. Reducing the producer's risk requires allowing some penetrations. The following illustration of tradeoffs between several options is modeled after an analysis Keith Eberhardt of NIST prepared for NIJ in April 1991. [60] OTA performed all calculations used here.

To simplify presentation, we will consider as options only repetitions of the test prescribed by NIJ Std. 0101.03, and we will neglect the possibility of BFS failures. In this context, the phrase "mean stopping probability" means the geometric mean of the stopping probabilities of the 48 fair shots required by the protocol; individual stopping probabilities may vary with shot location and order, test bullet, and panel-front or back, wet or dry. The fraction of fair shots stopped in a particular test or series of tests is *not the* mean stopping probability; it is the mean stopping probability plus an unknown "sampling error.'

We define "good armor" and "bad armor" in terms of mean stopping probability.[8] This is a policy choice; it should be decided by NIJ if NIJ elects to use this approach. For illustration only, we define "good armor" as armor having a mean stopping probability of at least 0.999, and "bad armor" as armor having a mean stopping probability of no greater than 0.95.

Second, we define the options for testing and certification. For illustration, we consider only two:

Option 1: Subject panels of the model to the test prescribed by NIJ Std. 0101.03 (at a specified ballistic-resistance level), and certify it if and only if no fair shots penetrate.

Option 2: Subject panels of the model to *three repetition* of the test prescribed by NIJ Std. 0101.03 (at the specified ballistic-resistance level), and certify it if and only if no more than one fair shot penetrates.

Under Option 1, the model is subjected to 48 fair shots and certified if none penetrate. Under Option 2, the model is subjected to 144 fair shots and certified if no more than one penetrates.

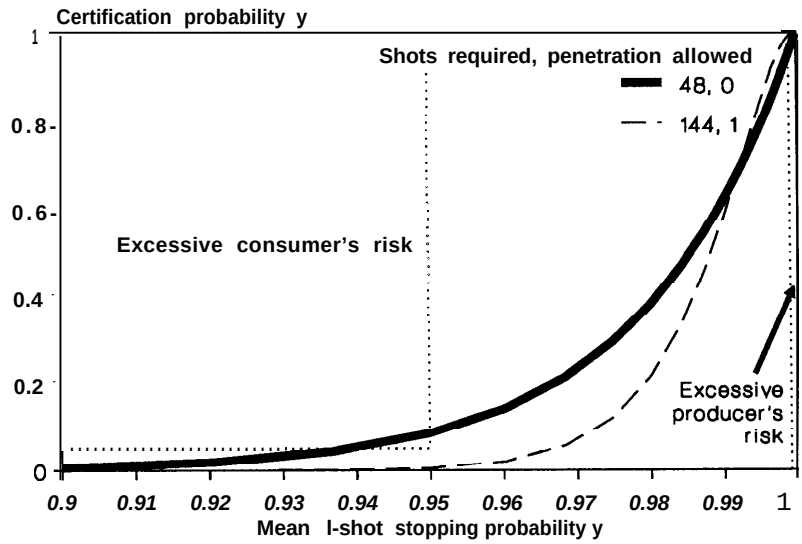
Third, we define "producer's risk" as the probability that good armor, as defined above, fails to be certified. We define ''consumers' risk' as the probability that bad armor, as defined above, is certified, recognizing that this also poses a financial risk to the producer.

Figure E-3 shows how the certification probability would vary with the mean stopping probability under each option. Note that the maximum consumers' risk of Option 2 (0.5 percent) is only about 1/17 that of Option 1 (8.5 percent), and its maximum producer's risk (0.9 percent) is also lower-only about 1/5 that of Option 1 (4.7 percent). However, Option 2 requires three times as many shots and would cost about three times as much as Option 1.

There are, of course, many other options. Even if one restricts consideration to repetitions of the .03 test sequence, one could require 1 repetition and allow O, 1,2, or up to 48 penetrations, or one could require 2 repetitions and allow O, 1, 2, or up to 96 penetrations, and so on. If upper bounds on consumers' risk and producer's risk are specified by policy, then it is a (solvable) technical problem to find the minimum number of repetitions required and the number of penetrations that must be allowed. In figure E-3, the upper left rectangular region labeled "Excessive Consumers' Risk" illustrates an upper bound of 0.05 (5 percent) on the consumers' risk, and the lower right rectangular region labeled "Excessive Producer's Risk" illustrates an upper bound of 0.05(5 percent) on the producer's risk. The graph (called an operating characteristic) for Option 1 passes through the region labeled "Excessive Consumers' Risk" and hence violates one of the bounds (hypothetically) set by policy. The operating characteristic for Option 2 avoids both prohibited regions and would be acceptable, but is not optimal, because the operating characteristic for an option not shown-two repetitions of the .03 sequence, allowing one penetration-also avoids both prohibited regions but requires fewer repetitions. However, Option 2 would be optimal if consumers' risk and

[8] More generally, **one could define "good armor" and "bad armor" in terms of mean single-shot passing probability, which we define as the probability of stopping the [fair] shot and also leaving an acceptable BFS,** if it is a shot after which **BFS** is measured.

Figure E-3—Certification Probability y Versus Mean Stopping Probability for Two Certification Criteria



SOURCE: Office of Technology Assessment, 1992.

producer's risk were both prohibited from exceeding 1 percent.

Figure E-4 plots producer's risk versus consumers' risk for several options; it helps identify the minimum-cost certification criterion meeting the constraints on consumers' risk and producer's risk. Each curve corresponds to a certain number of repetitions of the 48-shot .03 test sequence and is therefore a curve of constant cost. Each break-point on it corresponds to the maximum number of penetrations allowed; the uppermost point on each curve-the one with greatest producer's risk—corresponds to allowing O penetrations, the next lower point to allowing 1 penetration, and so on. Options outside the rectangular region at lower left have excessive producer's risk, excessive consumers' risk, or both; bounds of 5 percent on producer's risk and consumers' risk are illustrated.

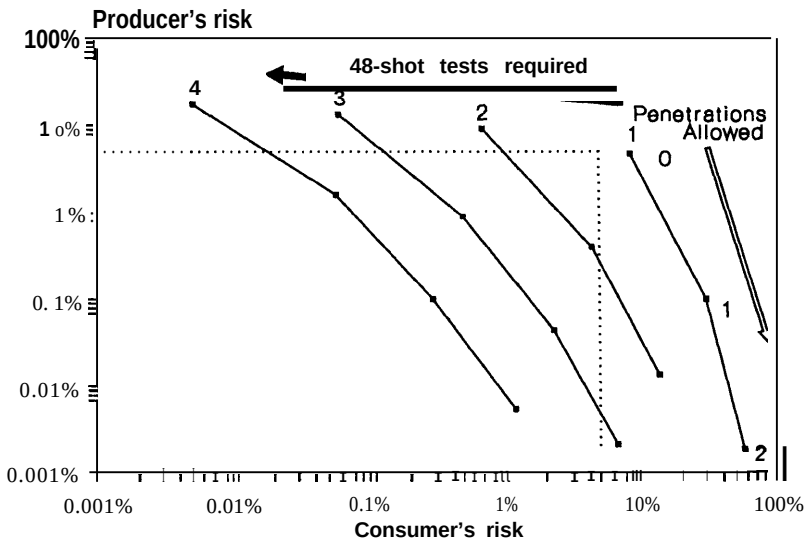
To identify acceptable minimum-cost criteria, one first examines the 1-test (48-shot) curve, and discovers that all points on it lie outside the acceptable region (only the first few points on it, including Option 1, are plotted). One next examines the 2-test (96-shot) curve, and discovers that only 1 point on it—the one corresponding to allowing 1 penetration—lies inside the acceptable region. This, then, is the unique minimum-cost criterion satisfying the constraints on producer's risk and consumers' risk.

In some cases there may be more than one minimum-cost criterion, requiring a choice between one criterion that minimizes producer's risk and another that minimizes consumers' risk. For example, if the bound on producer's risk is 10 percent and the bound on consumers' risk is 5 percent, then two repetitions would suffice, but one may allow one penetration or none. If 1 penetration were allowed, the producer's risk would be 0.4 percent and the consumers' risk 4.4 percent; if O penetrations were allowed, the producer's risk would be 9.2 percent and the consumers' risk 0.7 percent.

An alternative would be sequential testing with a stopping rule that allows testing to stop as soon as it demonstrates that both producer's and consumers' risks are acceptable. The number of tests required would not be fixed but would depend on the number of penetrations that occur as testing proceeds.

For example, a model could be certified if it withstood 96 shots with O penetrations, but if 1 penetration occurred in the first 96 shots, the armor could still be certified if it withstood 48 more shots with no more penetrations (i.e., if it withstood a total of 144 shots with no more than 1 penetration). This test would have a consumers' risk of 1 percent (slightly higher than that of the 96-shot test with no

Figure E-4-Consumers' Risk Versus Producer's Risk for Several Certification Criteria



SOURCE: Office of Technology Assessment, 1992.

penetrations allowed) and a producer's risk of 0.8 percent (much lower than that of the 96-shot test with no penetrations allowed). In some cases, it would require more testing and hence would cost more than the 96-shot test, but bad armor would have at most a 3.7-percent chance of needing more than 96 shots, and good armor would have at most a 8.7-percent chance of needing more than 96 shots.

A test *requiring 144* shots and allowing 1 penetration would have a slightly higher producer's risk (0.9 percent) but only half the consumers' risk (0.5 percent). Of course, it would cost more, on the average.

What effect would requiring more shots and allowing more penetrations have on reproducibility? It's a matter of definition. As noted in appendix C, neither these changes nor any others could provide more statistical confidence that the mean stopping probability is high enough for the model to pass a retest identical to the certification test with a specified probability. If this is the desired improvement in reproducibility, it is simply unattainable. However, if a retest is defined as, say, a 48-shot test with no penetrations allowed, then requiring several such tests for certification would reduce the probability that certified armor will fail such a retest (as distinct from the entire sequence of tests required for certification).

As noted in appendix C, the expected variance in outcomes of repeated testing is greatest when the probability of passing is one half. It approaches zero as the probability of passing approaches zero or one. Reducing the producer's risk can increase the probability that good armor will pass to as close to 1 as one desires; this will reduce the variance in outcomes of repeated testing of good armor to as small a value as maybe desired. Independently, the consumers' risk may be reduced, reducing the probability that bad armor will pass to as close to O as one desires and is willing to pay for; this will reduce the variance in outcomes of repeated testing of bad armor to as small a value as may be desired. The variance in outcomes of repeated testing of questionable armor-that having a stopping probability between that of good armor and that of bad armor-could still be high, but at least it could be argued that the variance in repeated testing of *good* armor would be low. This is qualitatively true of NIJ Std. 0101.03 and others standards such **as** PPAA STD-1989-05, but there are differences among these, and they do not define "good armor" quantitatively.

Some may object to allowing penetrations in a certification test in the belief that many, if not most, law-enforcement officers would not understand the statistical rationale and, in particular, might not trust or buy armor of a style that had been penetrated by

a round of **a type it is certified to stop, even if it** stopped 99.9 percent of such rounds. This is a valid concern for NIJ to weigh in deciding whether to allow penetrations. However, NIJ should also weigh a related danger-that allowing no penetrations allows purchasers and wearers of armor who are so inclined to believe, unscientifically, that certified armor will certainly stop, in testing and in use, all rounds for which it is rated. Although NIJ Standard 0101.03 and NIJ Guide 100-87, *Selection and Application Guide to Police Body Armor,* caution purchasers and wearers that there is no such thing as bullet-proof armor, neither specifies the statistical confidence with which the probability of stopping rated rounds can be said to be at least 90, 95, or 99 percent on the basis of certification. Purchasers and wearers should know that neither the NIJ test nor any other provides more than O percent confidence that the probability of stopping a specified round is 100 percent.

ASSESSING RISK OF TRAUMA FROM STOPPED BULLETS

Several changes could be adopted to improve the validity, accuracy, and reproducibility of the current test for acceptable risk of blunt trauma, which consists of shooting the test armor on an unspecified but calibrated inelastic backing material, measuring the. depths of craters made in the backing, and failing the model if any crater is deeper than 44 mm.

For example, specifying the backing material to be used and reducing the currently allowed tolerance on its temperature might improve reproducibility, but perhaps not significantly. Reproducibility could also be improved (in the limited sense defined in the discussion of penetration resistance) by measuring more backface signatures and optionally, allowing some to exceed the specified limit [9] Reproducibility might also be improved by options for improving validity, such as those described below.

To improve the validity[10] Of the current test, NIJ could elect any of several options. If NIJ retains the current type of deformation test with a single BFS limit applicable to all bullets, velocities, types of armor, and wearers, there is evidence (see app. D, that the BFS limit corresponding to 90-percent safety exceeds 44-mm, with 95-percent confidence. NIJ could **increase the** BFS limit and still provide 90-percent safety with 90-percent confidence while reducing producers' risk.

Alternatively, NIJ could undertake to assess risk of blunt **trauma** based on the diameter(s), and perhaps also the depth, of backface signatures using **a** parametric lethality model similar **to those proposed** by Army researchers in the 1970s. A model appropriate for use does not exist today, but one could be developed, partly on the basis of reenactments and, if desired, partly on the basis of expert opinion informed by analogous animal experiments such as those performed for the NILECJ by the Army in the 1970s. Such a criterion might lead to different BFS limits for wearers of different sizes or weights and for armors of different areal density (i.e., mass per unit area); there could also be different BFS limits for portions of armor covering different parts of the body. This would increase complexity, but could be more accurate, hence more valid.

(A simpler and more conservative-hence less accurate-alternative would be to certify armor only in sizes greater than some minimum size that depends on test results or for wearers heavier than a minimum weight that depends on test results.)

There is also the option of using tests that would require additional instrumentation than that currently used (primarily, a ballistic chronograph, a thermometer, and rulers). Possibilities include measuring pressure in the backing during impact, or measuring velocity and deformation simultaneously, to use in predicting lethal trauma according to a ' 'viscous Criterion. '

The same procedures used to establish maximum allowable depths or other limits for each ballistic-resistance class could be used thereafter to revise those limits on the basis of new data on experiments with animals or assaults on humans.

Determine BFS Limits Based on Animal Experiments

As noted in appendix A, the current 44-mm BFS limit was originally derived specifically for the case of .38 Special round-nose lead bullets impacting on

9 By averaging, PPAA STD-1989.05 [113] allows more than half the measured backface signatures to exceed the specified limit of 44 mm.

10 For purposes of this discussion, we say a test is a' 'valid' test if there is scientific evidence that the test accomplishes the purpose for which it was designed-in this context, the NILECJ's safety criterion, until it is superseded by a new NIJ safety criterion.

7-ply, 400/2-denier, Kevlar-29 armor at about 800 feet per second. The **animal testing that would have** been required **to** derive BFS **limits** for other threats and armors was begun but not completed.[11] Nevertheless, NILECJ-Std.-0101.01 and its successors, including NIJ-Std.-0101.03, specify a 44-mm BFS limit for all classes ("levels") of ballistic-resistance, for all types of armor.

No rationale for this generalization was documented. It was proposed by Lester Shubin, then Director of Science and Technology at the NILECJ, who in 1991 explained the rationale as the combination of

1. his judgment **that it might be unsafe to allow** higher energy bullets to produce a deeper BFS than the maximum deemed safe for .38 Special bullets impacting 7-ply Kevlar-29 at 800 ft/s;[12]
2. the absence of data showing that the BFS limit for higher energy bullets should be less than44 mm; and
3. the urgency of the need, inasmuch as armor **was then** being certified (under NILECJ-Std.-0101 .00) and worn without any test for protection from stopped bullets.

One option for improving the validity of the current test would be to conduct

1. additional experiments on animals, similar to those performed by Goldfarb et al.; [74] and
2. corresponding ballistic **tests,** analogous to *those performed.* by Prather et al., *[114]* to determine the backface signatures (or ether ballistic measurements) on clay backing (or whatever backing may be specified) that correlate with the the various degrees Of injury observed in the animal experiments.

One set of animal and ballistic experiments would be neededl for each combination of threat bullet and velocity) and and for which a BFS limit is to be determined. Umpublished records of the NILECJ-funded Army shootings of armored goats with .357 Magnum and 9-mm bullets, which remain in Ballistics Research Laboratory files, could supply some of the data needed to determine BFS limits appropriate for these bullets impacting the particular types of armor used in those experiments.

In principle, this approach has the potential **to** predict the probability of lethality from blunt **trauma** more accurately than can approaches that rely on (simple) parametric lethality models (described below). However, there are several disadvantages to this approach:

1. It would be expensive and **time-consumin g to perform the** large number of experiments that would needed just to determine BFS limits for the threat-armor combinations already tested under PTL_LStd.-OlOl.O3.
2. There would be a delay: until such experiments are performed and their results analyzed, there would be no explicit rationale for certifying armor (other than 7-ply, 400/2-denier, Kevlar-29 armor) as reducing the risk of blunt trauma (from threats other than .38 Special round-nose lead bullets at about 800 feet per second) to an acceptable level.
3. There would likewise be a barrier to technological innovation: armor not of the generic types tested in the experiments could not be certified. Developers of novel armor material—for example, synthetic spider silk-would have to fund experiments to estimate the deformation-trauma correlation in armor made from their material, or else lobby for Federal funding for such. experiments, and convince NIJ of the validity of the results before they couldi have any hope of having their product incorporated in NIJ-certified armor.
4. Extrapolation of the experimental results from **animals to** humans would be judgmental, **as it** was in the study by Goldfarb **et al.**

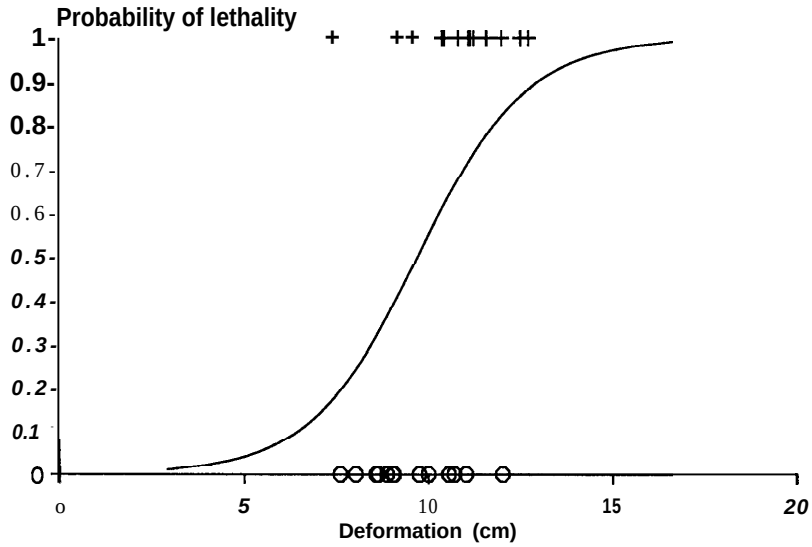
Determine BFS Limits Based on Parametric Lethality Models

Another option for improving the validity of the current test would be to base BFS limits on parametric lethality models of the type described in appendix A. An advantage of this approach, relative to the one @t described, is that extending it to additional **threats or types** of armor does not require additional biomedical **tests** (read: shooting large manmals, **and** killing some); it requires only additional ballistic tests: shooting the armor of interest

11 The NILECJ funded Army experiments in which armored goats were shot with .357 Magnum and 9-mm **bullets, as was** armor on clay backing, but the research was not **completed** or **published.**

12 Shubin worried, and some others still worry, that a BFS limit less than 44 mm might be appropriate for higher energy bullets, especially rifle bullets.

Figure E-5—Lethality Versus Prediction Based on Deformation



SOURCE: Office of Technology Assessment, 1992.

with bullets of interest at velocities of interest, using a backing such as clay.

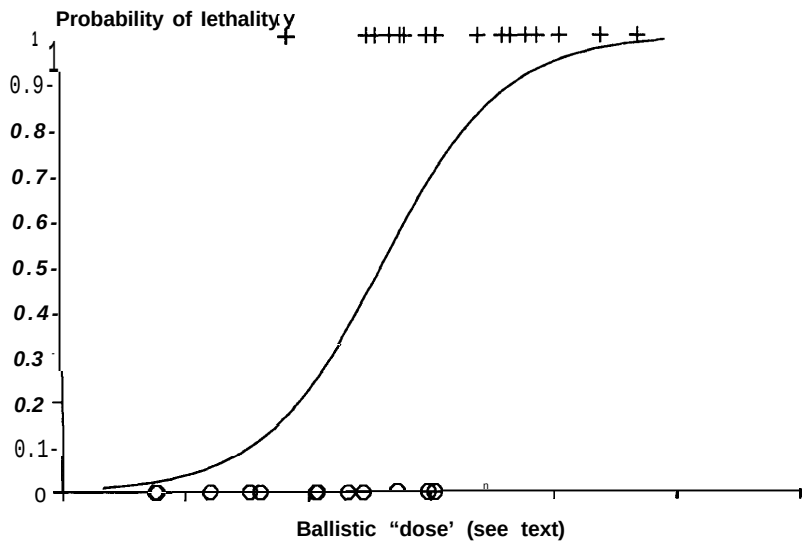
A simple parametric lethality model is a mathematical formula or graph that predicts the probability that a single shot would cause lethal blunt trauma, based on the value of a single parameter, such as BFS. Such a model could be used to derive a maximum acceptable BFS from the maximum acceptable probability of lethality specified by policy. Similarly, models of lethality or serious injury may also be developed and used (see app. D).

More complicated models, such as those proposed by Clare et al. [35] and by Sturdivan, [130] predict probability of lethality based on the values of several parameters, some describing the wearer (e.g., body mass and body-wall thickness), some describing the threat (e.g., bullet mass and velocity), some describing ballistic test results (e.g., the diameter of the crater made in flesh-simulating backing by the armor when hit by a bullet), and some describing properties of the armor (e.g., areal density: mass per unit area). In general, using more parameters provides more information and may improve the model, at the expense of the cost of making the additional measurements and the increased complexity of calculating the predicted lethality from them.

For example, figure E-5 shows the results—deaths (+) and survivals (o)-of shooting 29 goats over the liver with blunt, nonpenetrating projectiles simulating nonpenetrating ballets hitting armor, and the probability of death predicted on the basis of the maximum momentary deformation of each goat's abdomen, which is comparable to the depth of the crater the projectile would make in clay; table E-1 shows the data. Figure E-6 shows the same results, but with the probability of death predicted (by OTA) on the basis of a ballistic "dose" that depends on maximum deformation and five other parameters (the projectile's mass, diameter, and velocity, and the goat's weight and body-wall thickness). It is apparent that ordering the results by the ballistic dose as in figure E-6 separates the deaths from the survivals better than ordering them by deformation as in figure E-5. A vertical line can be drawn in figure E-6 to separate deaths from survivals with only 5 misclassifications; a similar line in figure E-5 would produce 9 misclassifications. Moreover, the model (i.e., the estimated probability of lethality) in figure E-6 predicts the results (deaths and survivals) with 67 times the likelihood predicted by the model in figure E-5.

A model (prediction) similar to that used in figure E-6 could be used for certifying acceptable protection from the impact of stopped bullets on the basis of multiple measurements.

Figure E-6-Lethality Versus Prediction Based on Multiple Measurements



SOURCE: Office of Technology Assessment, 1992.

Figure E-7 shows another example-a logistic discriminant model developed by OTA that discriminates perfectly the survivals from the fatalities of goats shot on the chest with blunt, nonpenetrating projectiles as reported by Clare et al. [35] Each shot is described in terms of a ''victim size parameter," which depends on the subject's weight, squared weight, and body-wall thickness,[13] and a "bullet-armor parameter,' which depends on the mass, speed, and diameter of the blunt projectile[14] used to simulate a combination of bullet, velocity, and armor. The model describes a straight line that separates shots that were survived from those that resulted in fatalities.

Both examples illustrate the general principle that the use of more parameters allows a model to better fit the data to which it is fitted, and may allow it to predict lethality with greater reliability. However, using more parameters may decrease the statistical confidence with which one can accept (i.e., not reject) the model. By using enough parameters, a model can be made to fit perfectly the data to which it is fitted, but this provides no confidence that the model would have been rejected had the data been different.

Appendix A described a method proposed by Prather, et al., for treating a bullet stopped by an-nor as a blunt projectile, and a multiparameter lethality model developed by Sturdivan [130] to estimate the probability that such a nonpenetrating projectile will cause lethal blunt trauma to the *thorax.* Here we will discuss how the model could be used to assess the acceptability of protection from lethal blunt trauma. Assessment of the acceptability of protection from lethal or critical trauma using a different parametric model-e.g., one based on data from reenactments—would proceed in a similar manner.

Sturdivan's model for probability of lethality, P(L), is

$$P(L) = 1/(1 + \exp(34.13 - 3.597 \ln(MW^2/W^{1/3}TD)))$$

where M denotes the projectile mass (g), V the projectile velocity (m/s), W the victim's body mass (kg), T the victim's body-wall thickness (cm), and D the projectile diameter (cm). D is estimated as the diameter of the crater made in clay backing, which is measured in a ballistic test. M and V are estimated from D, the bullet mass, M_P, and velocity, VP, and

13 The victim-size arameter is given (approximately) by the expression 86.89 W -0.9996 W^2 + 185.5 T, where W is the victim's weight (kg) and T is the victim's body-wall thickness (cm).

14 The bullet-armor parameter is given (approximately) by the expression 1.8434 M + 11.77 V -0.5788 D, where M is the mass of the blunt projectile (g), D is its diameter (mm), and V its velocity (m/s).

Table E-l—Lethality of Blunt Trauma to Liver v. Characteristics of Projectile and Victim

M (g)	V (m/s)	D (mm)	W (kg)	T (cm)	Depth(cm)	Survival?
300	49.2	74	47.31	2.7	7.60	yes
300	38.6	74	55.88	2.1	11.68	no
300	41.2	74	48.46	2.4	9.75	yes
300	46.1	74	52.60	2.3	10.51	no
300	43.0	74	55.64	2.6	8.87	yes
300	39.4	74	54.67	1.9	8.58	yes
300	38.6	74	52.20	2.3	9.24	no
300	56.2	74	57.86	4.1	8.63	yes
300	49.1	74	58.19	2.9	7.46	no
300	45.0	74	56.75	2.6	9.24	no
300	36.0	74	56.02	2.1	9.86	no
300	48.5	74	55.65	3.3	9.03	yes
300	31.2	74	43.63	1.6	11.34	no
300	40.9	74	55.35	2.7	9.06	yes
300	41.2	74	50.23	2.6	10.54	no
430	42.0	100	53.15	1.8	11.25	no
430	38.2	100	44.02	2.4	10.70	yes
430	37.0	100	37.21	2.5	8.02	yes
430	32.5	100	40.98	1.5	10.00	yes
430	33.4	100	57.90	1.8	12.81	no
430	58.3	100	62.06	4.2	11.02	yes
430	34.3	100	48.96	1.6	12.01	yes
430	38.2	100	59.30	3.6	11.20	no
430	32.4	100	56.06	2.1	12.09	no
430	29.9	100	41.80	1.8	9.69	no
430	30.4	100	46.04	1.4	12.60	no
430	33.5	100	48.62	2.0	10.49	no
430	27.0	100	43.21	1.7	10.92	no
430	58.4	100	63.31	3.2	10.55	yes

Legend: M: projectile mass.
V: projectile speed.
D: projectile diameter.
W: weight of victim (goat).
T:thickness of victim's bodywall (skin, fat, muscle) at impact point.
Depth: maximum momentary depth of depression of victim's (goat's) skin by projectile.

SOURCE: U.S. Army Chemical Research, Development, and Engineering Center, April 4,1991 [131].

the areal density of the armor, a_d(g/cm*), using the formulas

$$M = \quad + 3.14\ (D/2)^2 a_d$$
$$V=(Mp/M) \qquad V_p.$$

To assess risk of blunt trauma to a particular wearer, the wearer's body mass W and body-wall thickness Tare measured and used, along with D, M, and V in the formula for P(L). This procedure would be reversed in a certification test: P(L) would be set equal to the maximum acceptable probability of lethality, P(L)_, and the equation

$$P(L)_{max} = 1/(1 + \exp(34.13 - 3.597 \ln(MV^2/W^{1/3}TD)))$$

would be solved for $W^{1/3}T$. The value obtained would be the minimum allowable value:

$$(W^{1/3}T)_{min} = [\exp(-34.13)\ (1 - P(L)_{max})/P(L)_{max}]^{1.3.597}\ MV^2/D$$

That is, the armor could be certified to provide acceptable protection from lethal blunt trauma to wearers having a body mass W and body-wall thickness T large enough so that $W^{1/3}T$ equals or exceeds a value, $(W^{1/3}T)$min, derived from the specified threat Mp, Vp), the ballistic test result (D), and the areal density of the armor (a_d). Figure E-8 illustrates the process.

For a maximum acceptable probability of lethality of 10 percent $(P(L)^{max} = O. 1)$, $W^{1/3}T$ must equal or exceed 0.0001395 MV^2/D.

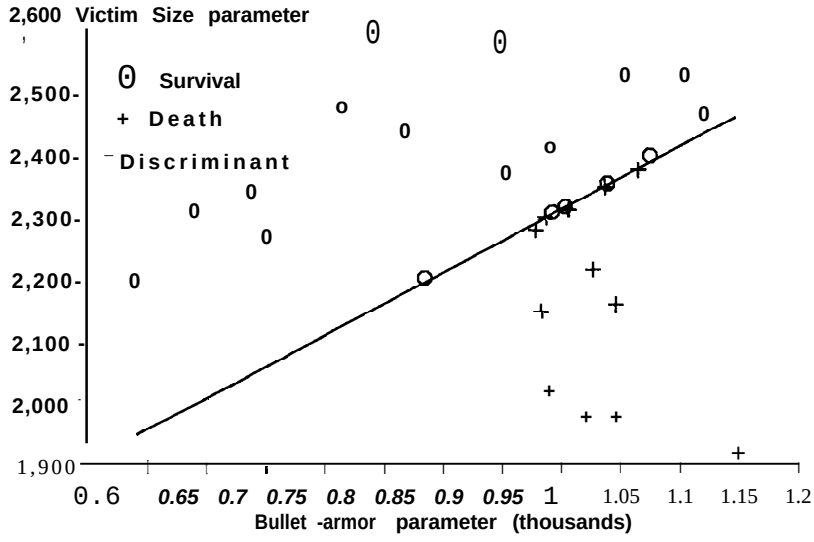
If it is not desired to certify armor for wearers having at least a specified value of $W^{1/3}T$, a conservative alternative would be to certify armor unconditionally if its calculated value of $(W^{1/3}T)$min exceeds the the value corresponding to a small fractile of officers, perhaps $(W^{1/3}T)$min = 7.6, for W= 55 kg and T = 2 cm. Of course, conservatism has its risks-of decreasing wear rate and increasing producer's risk unnecessarily. The option of certifying armor only for wearers having at least a specified value of $W^{1/3}T$, and variations of this, are discussed in greater detail below.

We will illustrate the calculation of $(W^{1/3}T)_{ti}$, assuming $P(L)_{max}$ =0.1, for a test in which a .38-cal., 158-grain (10.2-gram) lead round-nose bullet was fired at an armor panel made from 7-ply, 1,000-denier Kevlar 29. The impact velocity was 833 fps (V_p = 254 m/s), and the BFS was a crater 3.4 cm deep, with a roughly elliptical base measuring 6.2 cm x 5.5 cm. [114] The geometric mean of these major and minor axes (5.8 cm, the square root of 6.2 cm x 5.5 cm) should be used as the diameter D in calculating M. The nominal areal density of 1,000-denier, 31x31 Kevlar 29 fabric is 8.3 ounces per square yard (0.028 g/cm^2) per ply, so the areal density a_d of the 7-ply panel would be about 0.20 g/cm^2. Hence

$$M = Mp + 3.14\ (D/2)^2 a_d$$
$$= 10.2 + 3.14\ (6.2/2)^2\ 0.20$$
$$= 16\ g$$
$$V = (Mp/M)\ V_p$$
$$= (10.2/16)\ 254$$
$$= 162\ m/S$$
$$\sim/D = 67726, \text{ and}$$
$$(W^{1/3}T)_{min} = 9.448\ kg^{1/3}\text{-cm}$$

Measurement of the armor's areal density over the crater presents a problem: Should the portion of armor over the crater be excised, cleaned of bullet, fragments, and backing, and weighed? This may degrade the value of the armor as an archival

Figure E-7—Discriminant Model for Assessing Protection From Lethal Trauma by a Stopped Bullet



SOURCE: Office of Technology Assessment, 1992.

standard for quality-assurance. For some armor, there is an alternative: the areal density of armor made from 1000-denier, 31x31 Kevlar 29 fabric could be inferred from bullet momentum and crater depth and diameter, using a clay-cavity model published by the Aerospace Corp. [7] This procedure is illustrated in figure E-9. One could attempt to develop similar models and procedures for other armor materials, but this may be costly (although less costly than animal experiments) and may pose a barrier to innovation.

Before putting these procedures into practice, it would be advisable to adjust the lethality predicted by Sturdivan's models, or others fitted to data obtained by targeting vulnerable organs, to account for the less accurate marksmanship typical of assaults. The adjustment process would weigh the blunt-trauma lethality predicted for each vulnerable organ by an organ-specific model according to the probability that a shot on armor (or on the upper torso) would impact over that organ, as was done in the medical assessment by Goldfarb et al.

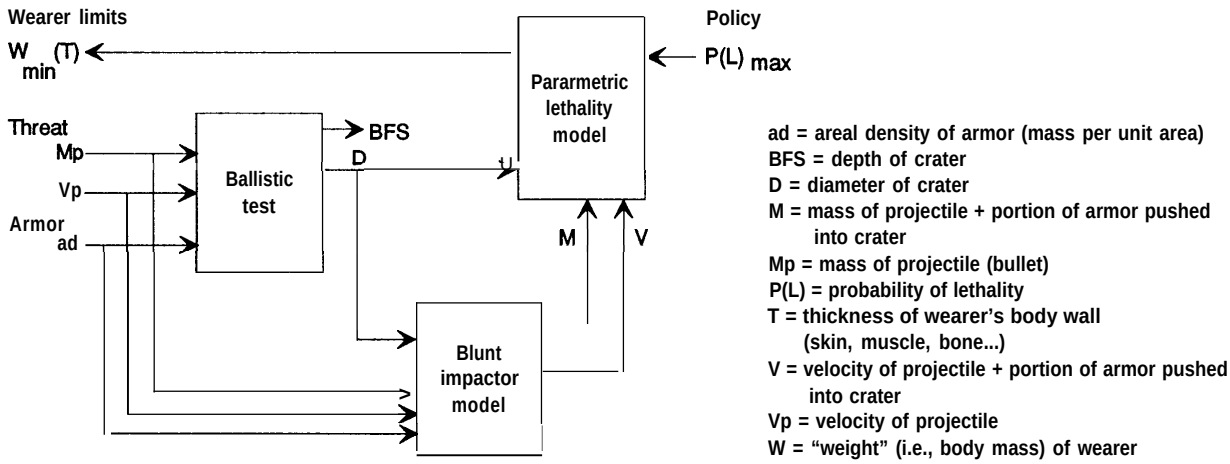
The extrapolation of predictions based on animal data to humans would be necessarily judgmental, as it was in the original body armor medical assessment sponsored by the NILECJ. Different experts, considering the animal data, might estimate different probabilities of death or trauma in humans under the same conditions. There is a procedure for combining these estimates, [95] and if this is done for c conditions, a c-parameter logistic model (counting the "dummy regressor") could be fit to the c combined estimates. Advantages of a logistic model include its great generality and the ability to update it easily on the basis of additional data [164] from reenactments of assaults.

Specify Size-Dependent BFS Limits

As noted in appendix A, the body armor medical assessment team that recommended the current 44-mm BFS limit did so to guarantee protection to light, female wearers with a thin body wall; they expected that heavier male wearers with a thicker body wall would face a lower probability of surgery or death if shot by a round that would cause a 44-mm BFS behind their armor. The parametric lethality models discussed above also support this expectation. These considerations provide a rationale for allowing a deeper BFS behind armor sized for large males, or certified only for male or female wearers heavier than specified minimum (perhaps sex-dependent) weights.

As examples of how this could be done, consider the 0.20 g/cm^2 vest mentioned above that stopped a 10.2-gram bullet that impacted at 254 m/s and made a crater measuring 6.2 cm x 5.5 cm in diameter. The calculated value of $(W^{1/3}T)_{min}$ was 9.448 $kg_{1/3}$-cm.

Figure E-8-Assessing Acceptability of Protection From Lethal Blunt Trauma Using a Parametric Lethality Model



SOURCE: Office of Technology Assessment, 1992.

The vest could be certified to provide acceptable (viz., "90-percent") protection from lethal blunt trauma to wearers having $W_{1/3}T = 9.448$ $kg_{1/3}$-cm or greater. A certification of compliance could state the restriction in this way, or it could portray the restriction in graphical or tabular form, for example:

> This armor complies with NIJ-Std.0101.xx and provides 90-percent protection from lethal trauma from a stopped bullet to wearers weighing at least
>
> 54 kg and having a body-wall thickness of at least 2.5 cm, or
> 61 kg and having a body-wall thickness of at least 2.4 cm, or
> 70 kg and having a body-wall thickness of at least 2.3 cm, or
> 80 kg and having a body-wall thickness of at least 2.2 cm, or
> 92 kg and having a body-wall thickness of at least 2.1 cm, or
> 106 kg and having a body-wall thickness of at least 2.0 em.

This is more complicated and cumbersome than the current procedure. On the other hand, it could provide a rationale for certification of protection against blunt trauma caused by other than Type I bullets hitting Kevlar armor. It also would allow qualified certification of armor that would fail if required to provide the smallest wearers with acceptable protection from lethal blunt trauma.[15]

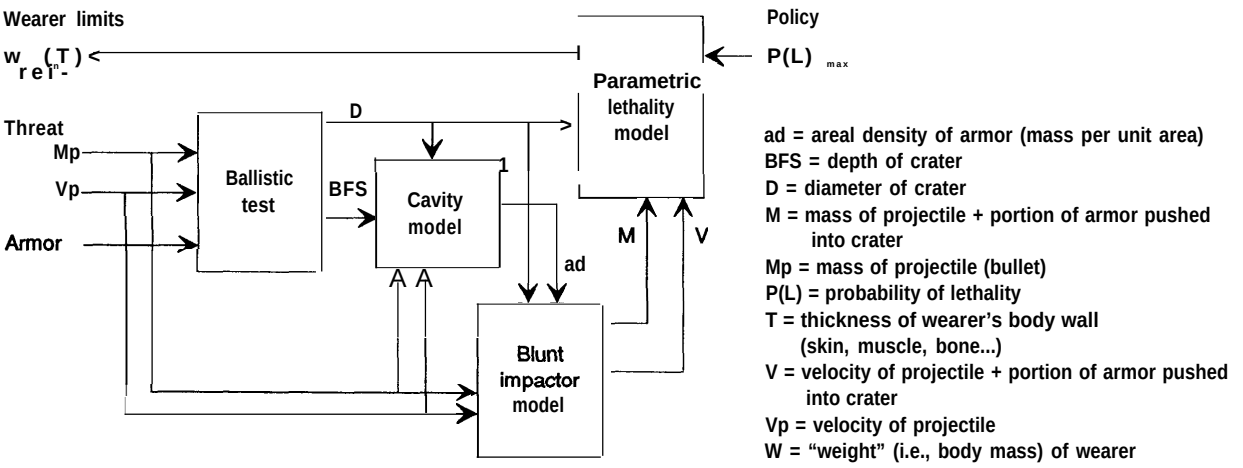
However, another drawback of the procedure must be addressed: it requires knowing the wearer's body-wall thickness, which is not readily measured. It might require a computed axial tomography (CAT) scan. This could be avoided, perhaps with some loss of reliability, by using a parametric lethality model that does not depend on T. For example, Clare et al. [35] developed a model of blunt-trauma lethality as a function of MV^2/WD. A related approach is to use, in a model that depends on T, an estimate of T in terms of other variables. For example, Sturdivan [132] has found that T is roughly proportional to $W_{1/3}$ in both goats and man. If a is the constant of proportionality, then one could use a $W_{1/3}$ in place of T in $MV^2/DW_{1/3}T$, resulting in a model that depends on $aM^2/DW^{2/3}$. OTA has determined that this procedure results in negligible reduction in goodness-of-fit to some data[16] but reduces goodness-of-fit to other data substantially.[17] Other such models could be developed; however, other things being equal, requiring a model not to depend on T may reduce the reliability with which it correctly predicts lethality.

15 For example, in a similar test also using a .38-cal., 10.2-g LRN bullet fired at a 7-ply, 1,000-denier, Kevlar-29 panel, the impact velocity was 787 fps (V_p = 240 rids), and the BFS was a crater 4.6 cm deep, with a circular base 6.0 cm in diameter. [114] This result would have failed the armor under NIJ-Std.- 0101.03, but the procedure discussed here would allow the armor to be certified for wearers having $W^{1/3}T = 8.786$ $kg^{1/3}$-cm or greater. For example, the armor could be certified for wearers weighing 75 kg with body walls at least 2.1 cm thick.

16 For example, the data on lethality of blunt impacts to goat abdomen over the liver in table E-1.

17 For example, the data on lethality of blunt impacts to goat thorax in table 1 of [35].

Figure E-9-Alternative Procedure for Estimating Probability of Blunt-Trauma Lethality From Backface Signature and Parametric Lethality Model



SOURCE: Office of Technology Assessment, 1992.

Revise BFS Limit(s) Based on Field Experience

The Army's initial medical assessment of body armor and the parametric lethality models described above are based on animal experiments performed before data were available on shootings of humans wearing such armor.

Now more than 20 assaults (but only 2 that resulted in death or critical injury) have been reenacted, several times each. OTA's analysis of the results (see app. D) concludes that the 44-mm BFS limit in NIJ Standard 0101.03 is smaller than necessary to limit the risk of death or life-threatening injury from a bullet that impacts at the maximum velocity for which protection is certified and is stopped by the armor to 10-percent, a goal specified by the NILECJ in 1976. However, the analysis does not show that the test reliably discriminates unsafe armor from safe armor; if it does, more reenactments will be needed to prove it.

If NIJ decides that a 10-percent risk is still acceptable (this is a policy choice implying a value judgment), the BFS limit could be increased. This might increase the risk to wearers of armor (perhaps only slightly) but might increase the frequency with which officers wear their armor. It would decrease the risk, to manufacturers, that armor that actually limits risk as required would fail the test.

To increase the confidence with which conclusions may be inferred (as in app. D), more reenactments of more assaults-especially assaults in which officers were killed or critically injured by stopped bullets-are needed. This will require monitoring assaults and collecting detailed data on those suitable and most important for reenactment.

If and when such reenactments have been performed, the Walker-Duncan procedure [164] could be used to revise any of the logistic models described in app. Din light of the new data. A new model with more parameters would have to be fitted, using separate-sample logistic regression, [9] to the cumulated data in order to estimate BFS limits for different cases—i.e., threat-, armor-, and wearer-dependent limits.

Specify Tests Other Than BFS

Someday, certification of acceptable protection from blunt trauma could be based in whole or in part on tests other than BFS measurements. Proposals include measuring pressure in the backing during impact, or measuring velocity and deformation simultaneously, to use in predicting lethal trauma according to a "viscous criterion. " These tests would require more sophisticated, expensive instrumentation than that currently used—primarily, a ballistic chronograph, a thermometer, and special rulers-and it is not yet known whether such tests

would be more accurate than the current one or the other tests, discussed above, based on BFS measurements.[18]

Pressure Criteria

Some experts expect that the peak pressure measured in backing would be a better predictor of specific types of blunt trauma than would any test based on BFS. One such type is the laceration or rupture of arteries or other organs compressed suddenly by the intense pressure wave generated by the impact of a nonpenetrating bullet on armor. Such trauma has caused the death of one police officer, whose armor, in stopping a rifle bullet, penetrated his chest.[19]

Research has also demonstrated that a brief, intense pressure pulse, similar to the early portion of the pressure pulse generated by a nonpenetrating ballistic impact, may block conduction by cardiac nerves. [122, 123] Some deaths caused by automobile accidents and baseball impacts might be attributable to this mechanism or to apnea (cessation of breathing) or other effects. [154, 155] It might also be responsible for deaths caused by single blows of other types-e. g., the widely publicized classroom death caused by a blow delivered to the chest in a hitting game called a‘ ‘cuss game.’ Although deaths attributable to these mechanisms are apparently rare, tests based on BFS may not be a good predictor of them, because research has demonstrated that BFS is more strongly correlated with the later, longer, less intense portion of the pressure pulse than with the early, brief, intense portion. [84] However, correlation of peak pressure in backing with lethality in humans has not yet been established.

Viscous Criteria

Empirical research suggests that blunt trauma caused by automobile accidents, baseball impacts, and other causes may be classified as lethal or nonlethal based on the maximum value of the velocity of deformation times the factional compression of the body.[20] A blow is predicted to be lethal if the velocity of deformation times the fractional compression ever exceeds a certain threshold; this is called the “viscous criterion. ” [156] [21] Using it in armor certification would require

1. using a backing that simulates the deformation-versus-time history of the human torso or can be calibrated with it, and
2. measuring velocity and deformation simultaneously.

Another hypothesis holds that lethality of blunt trauma may be predicted on the basis of maximum velocity and maximum deformation (or compression), [38, 84] which occur at different times. Such a model would be easy to use for certification, because the maximum velocity can be approximated as the impact velocity, which is already estimated in NIJ certification testing, and the maximum deformation of impacted tissue could be calibrated to crater depth in the inelastic backing, which is already recorded. Variants of the general hypothesis maybe tested for consistency with animal blunt-trauma data already collected.

For example, OTA fit the logistic model

$$P(L) = 1/(1 + \exp(-a - b \ln(V) - c \ln(\text{compression})))$$

to data on survival of 29 goats shot over the liver by blunt, nonpenetrating projectiles (see table E-l). V is projectile velocity at impact in m/s, and “compression’ is maximum depth of abdominal indentation, in cm, divided by the cube root of the animal’s body mass Win kg. The cube root of W was used as a proxy, or substitute, for the thickness of the body in the direction of indentation, which was not recorded. The best fit (maximum likelihood) was obtained with a = 13.0, b = -4.15, and c = 2.58, so the fitted model is

$$P(L) = 1/(1 + e^{-1^3 V^{4.15}} / \text{compression}^{2.58})$$

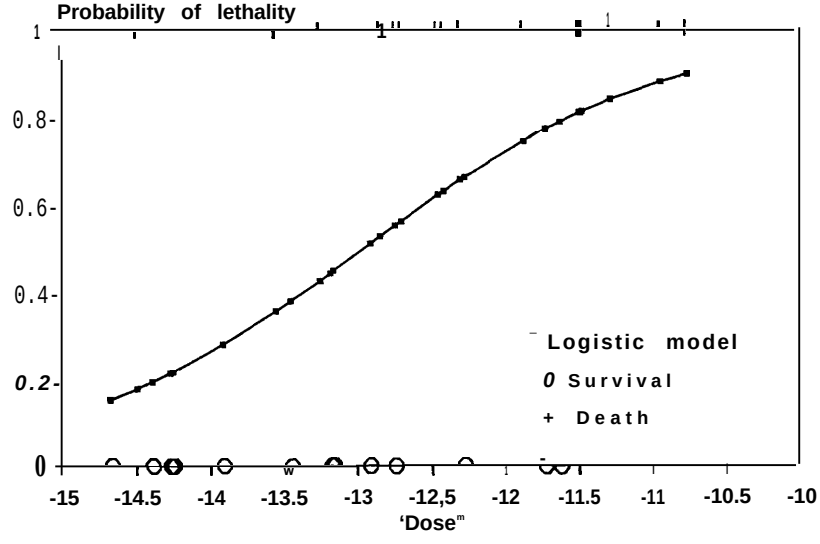
18 Iremonger and Bell [84] cited experiments that found viscous criterion values for impacts to be correlated with BFS depth, which in turn was found to be “not a sensitive measure of injury severity.’ (However, their definition of the viscous criterion differed from that of Viano and Lau [156], whom they cited.) They speculated that pressure measurement, perhaps in combination with other measurements, might be a better predictor of injury, but noted that ‘further work is required in order to quantify the damaging effect of stress wave transmission.”

19 The medical examiner attributed the cause of death not t. the penetration, per se, but to ‘‘The shockwave created by the missile, ’ which “lacerated the aorta, the pulmonary artery, and the vena cava immediately adjacent to the heart, resulting in death by insanguination into the thoracic cavities.” [133] Although the speed of sound maybe so low in lung tissue that the pressure wav may have been supersonic (hence a shock wave) there, [38, 166] the pressure wave was probably subsonic (not a shock wave) in tha orta, the pulmonary artery, and the ven cava. However, even a subsonic pressure wave, if sufficiently strong, could cause the damage noted.

20 The fractional compression is defined as the depth of deformation divided by the thickness of the body in the direction of deformation.

21 See also [94, 153, 158, 159]; cf. [84].

Figure E-10-A Logistic Model for Blunt-Trauma Lethality in Terms of Compression Times Velocity of Deformation (a "VIscous Criterion")



"dose" = 2.58 In(compression) -4.15 In(v) where
V = impact velocity in m/s
compression = (maximum depth of abdominal indentation in cm)/ $W^{1/3}$
W = body mass in kg

SOURCE: Office of Technology Assessment, 1992.

Figure E-10 shows the predicted lethality as a function of a ballistic "dose" defined by

"dose" = 2.58 In(compression) -4.15 in(V).

It may seem paradoxical that the model predicts that, of animals suffering comparable compression, those hit by higher velocity projectiles would be less likely to die.[22] Nevertheless, predictions of the model may be sensible if based on real data, because one would expect that, of similar animals, those hit by higher velocity projectiles would be more likely to suffer greater compression. What is surprising in this case is that those animals hit by higher velocity projectiles suffered less compression, on the average. Thus, although the apparently paradoxical form of the model is not surprising, the reason for it is. For whatever reason (perhaps mere chance), the data in table E-1 are peculiar, and one should doubt the validity of the OTA model based on them unless the peculiarity is explained or the model validated by other data.

Nevertheless, the model predicts the deaths and survivals in table E-1 with a likelihood (6.2×10^{-8}) more than three times that (1.9×10^{-8}) with which Sturdivan's model

$$P(L) = 1/(1 + e^{-29.0}(MV^2/W^{1/3}TD)^{4.34}),$$

predicts them. This shows that the simple viscous criterion considered here predicts lethality better than a logistic model in terms of $MV^2/DW^{1/3}T$. More complicated viscous criteria considered by OTA fit slightly better, but not as well as a nonviscous logistic model,

$$P(L) = 1/(1 + e^{14.2} M^{-32.1} V^{10.9} D^{42.6} Depth^{-5.06} W^{-11.0} T^{-0.249}),$$

which predicts the outcomes with a likelihood of 2.3×10^{-6}, which is 37 times the likelihood with which OTA's viscous model predicts the outcomes and more than a hundred times the likelihood with which Sturdivan's model predicts the outcomes. It is possible that a logistic model predicting lethality or injury in terms of the viscous criterion proposed by Viano, Lau, and colleagues could predict outcomes of other experiments (in which the required measurements are recorded) better than OTA's viscous model, or other logistic models, would. However,

[22] Similar results are common in logistic models that depend on correlated variables, such as velocity and compression in this case. The predictions of such an apparently paradoxical model are usually reasonable if the variables do not have values outside the range of values of the data to which the model was fitted.

more research would be needed to find out whether this is true.

To summarize, it is plausible that pressure criteria could predict blunt-trauma lethality from some, possibly rare, causes better than other criteria discussed here. However, there is as yet no basis for expecting that criteria based on pressure measurements in backing would significantly improve predictions; future research may, or may not, provide such a basis. Measurement of backing pressure for certification or acceptance tests based on pressure criteria would require instrumentation costing hundreds or thousands of dollars. Viscous criteria may predict lethality of ballistic blunt trauma as well as or better than parametric models developed by the Army for the NILECJ. However, it is reasonable to expect that more general parametric models including but not restricted to viscous criteria maybe better predictors of blunt-trauma lethality. Some, but not all, viscous criteria would require expensive instruments for measuring and recording backing indentation and velocity histories.

ASSURING QUALITY AT POINT-OF-SALE AND IN SERVICE

Revise NIJ Std. 0101.03 to Apply to Lot-Acceptance Testing

Some of the issues of enforcement and quality control discussed in appendix C would be solved if NIJ revised its armor certification process to be a lot-certification process rather than a model-certification process, with a separate style-certification process.

To execute this option, NIJ would have to

1. Revise the current standard to apply to lot testing, as NILECJ-0101.00 [141] did.[23]
2. Define "lot" precisely. (Must a lot be homogeneous? Why?)
3. Specify the number of samples from each lot to be tested, or a way to calculate the number from statistical criteria such as maximum probability of accepting a lot more than 1 percent of which is defective.
4. Ensure that the samples to be tested are selected randomly from each lot.

Definition of Lot

The definition of lots is usually guided by the following principles [60, 107]:

- *Lots should be natural units in commerce.*
- *Lots should be homogeneous—all* units in a lot should be made in the same time period by the same workers using the same equipment and materials, which in turn should be from the same lot, etc.

In addition, a lot should have at least enough units to provide the samples required for quality assurance (see item 3 above). For economy, the lot size should be many times the sample size, so that the cost of testing, including the cost of the samples, could be amortized over the units remaining after sampling for testing.[24]

Units of Commerce

The natural unit of commerce in armor varies widely; a large order may consist of tens of thousands of units,[25] while for custom armor it is often 1 unit. If the current test procedure is retained, shipping 1 unit of certified custom armor would require producing 7 units from which 6 could be sampled at random for testing. Even more samples would be required if high statistical confidence in high reliability[26] were demanded.

Lot Homogeneity

In some approaches to quality control, it is important that a lot be homogeneous, i.e., that all units in the lot be alike. In the approach to acceptance sampling described above in *Increase Total Shots and Allow Penetrations*,[27] lot homogeneity is important because it provides a rationale for *assuming* that all units in a lot have the same reliability, so that the reliabilities of the units not

---

23 NIJ Guide 100-87, *Selection and Application Guide to Police Body Armor [145],* might also need to be revisal.

24 An alternative for attaining hi@ confidence with small sample sires is to use Bayesian methods of risk assessment, which are explicitly subjective and hence controversial. However, they have been used to assess the safety of nuclear power plants and space launch vehicles. [11]

25 A large order may consist of tens of thousands of units of various sizes. We argue that size may affect ballistic resistance both in tests and in service, so otherwise similar armor of various sizes should not be considered a single lot, according to the usual definition of a lot.

26 Viz., probability of passing.

27 The approach is a form of 'acceptance sampling on the basis of parameters.''

tested may be inferred from the results of the tests of the units selected from the lot to be tested. This assumption may be wrong, and it may be unnecessary.

- It may be *wrong* because subtle, unnoticed variations in manufacturing processes could cause the reliabilities of apparently identical units to differ. Ballistic test results could be subjected to a statistical test to decide whether they are.[28] But,
- It may be *unnecessary,* depending on type of reliability one is interested in. Two distinctly different concepts of reliability that should be distinguished are (1) the reliability of an individual unit of armor, and (2) the (''average") reliability of a lot, which is, by definition, homogeneous in the lot. In either case, a lot could be any set of armor labeled as such by the manufacturer-not necessarily homogeneous in ballistic resistance nor in any other respect, such as size-provided it passes statistical tests, based on the results of ballistic tests, to limit the risk of accepting bad armor as well as the risk of rejecting good armor.

  Concept (l)-of the reliability of an individual unit-is problematical in the classical, frequentist interpretation of probability, which holds that reliability (i.e., probability of success) is a meaningful concept only if it is possible to conduct identical, repeated trials.[29] However, if the individual units of a lot may differ, perhaps invisibly, and especially if the purpose of testing is to determine whether they do differ, then tests of samples from the lot cannot be *assumed* to be *identical* repeated trials.[30] [31]

  Concept (2), the reliability of a lot (which an adherent to concept (1) could call the average reliability of a lot), is an admissible concept in the classical paradigm of statistical inference. Sampling and testing (e.g., as described above in *Increase Total Shots and Allow Penetrations)* provides information directly about the reliability of a lot, which may be all that some consumers care about. But, together with information about lot size and sample size, it also provides information about the distribution of the individual reliabilities in a lot.

Sample Size

In fact, if one is concerned about individual reliabilities in a lot, the minimum sample size will be determined by the lot size, themmaximum acceptable risk of accepting unreliable armor, and the maximum acceptable risk of rejecting reliable armor. If one is concerned only about the reliability of a lot, the minimum sample size will not depend on the lot size, but only on the maximurn acceptable consumers' and producers' risks.

It is simpler to illustrate this by focusing on the number of tests required (rather than the number of shots required), the number of test-failures allowed (rather than number of penetrations allowed), and the probability that a unit will pass the test (rather than the reliability of stopping each shot) .32 Also, for purposes of this discussion, we consider a' 'unit" of armor to be a set of however many identical garments are required for a test-e.g., 4 garments for a 2-caliber wet/dry NIJ test of standard-type ballistic resistance, or 1 garment for a l-caliber wet-only or dry-only test of special-type ballistic resistance. An 8.53-percent probability of passing a 48-shot test corresponds to a 95-percent geometric-mean single-shot probability of passing (the boundary between "bad" and "marginal" armor in the example above[33]), and a 95.3-percent probability of passing a 48-shot test corresponds to a 99.9-percent geometric-

28 For example, a 2-sided, 1-sample Kolmogorov-Smirnov test [45] could be used to test goodness of fit to a binomial distribution, which the number of passes would have if all units had the same probability of passing. It gives an upper bound on the statistical significance—i.e., a significance level—at which a discrete distribution, such as a binomial distribution, may be rejected.

29 If so, the reliability is the limit that the relative frequency (i.e., fraction) of successes is almost certain to approach as then-r of trials increases without bound.

30 One can nevertheless contrive scenarios in which the reliability of an individual unit of an inhomogeneous lot would make sense in the classical paradigm. For example, even though lot 1 may contain only 1 unit of size-38 model A armor, one could argue that it is meaningful to speak of its reliability, because one could, if one wanted, make additional units of size-38 model A armor and test them. This still assumes, however, that their properties-including the invisible ones being tested—would be identical.

31 The reliability of an individual unit is a meaningful concept in the *Bayesian* paradigm of statistical inference [11, 80, 81].

32 Otherwise, it would be necessary to introduce such arcane concepts as the arithmetic mean (i.e., the average) of the geometric-mean single-shot probabilities of passing.

33 See *Increase Total Shots and Allow Penetrations,* above.

mean single-shot probability of passing (the boundary between "marginal" and ''good' armor in the example above) .34

Suppose now, for example, that a lot consists of 10 units, that 2 of the units are selected randomly and tested, and that both pass. Exact l-sided binomial confidence limits on the average passing probability are easily calculated for this case;[35] the average passing probability is at least 0.0853 with 99.3-percent statistical confidence. If the average passing probability were no greater than 0.0853, there would be no more than a 0.7-percent chance that the results would have been as good as those obtained. Thus the consumers' risk is only 0.7-percent.[36 37]

There is, however, a greater risk that one or more or the units in the lot has a passing probability lower than 0.0853. The probability of a pass (the reliability of the lot) is the sum (over all units) of the probability that the unit will be selected times the probability that it will pass if tested. Each unit has the same probability of being selected: the reciprocal of the lot size. Thus probability of a pass is the average of the individual probabilities of passing. In the present example, the 2 units tested could each have a passing probability of 0.4265 while the 8 units not tested could have a passing probability of O, and the average passing probability would be 0.0853. By such calculations one may deduce lower confidence limits on individual passing probabilities from the lower confidence limits on the average passing probability. In general, individual passing probabilities may be much lower than the average passing probability, at the same confidence level, especially if the lot size is much larger than the sample size. In contrast, confidence limits on the average passing probability are insensitive to lot size, but sensitive to sample size.

If a maximum acceptable consumers' risk and a maximum acceptable producer's risk are specified, one may prepare a control chart, such as the example shown in figure E-11, to indicate whether a lot must be rejected to limit the consumers' risk or accepted to limit the producers' risk. The chart is for l-percent maximum consumers' risk of accepting a lot with a passing probability worse than $0.95^{48} = 0.0853$ and l-percent maximum producers' risk of rejecting a lot with a passing probability better than 0.999* = 0.9531. These illustrative values are arbitrary; similar charts could be prepared for other choices. Figure E-12 shows how the control limits (the boundaries of the must-accept and must-reject regions) change as the maximum acceptable consumers' and producers' risks are increased to 5 or 10 percent.

What should be done if the test results lie in the discretionary region between the lower and upper control limits? In the interest of reproducibility, such a decision should not be made arbitrarily on a case-by-case basis; a policy (even if arbitrary) governing such cases should be established. One option would be to require testing to continue; this might well consume all the armor in a lot, but it would not violate either the maximum acceptable consumers' risk or the maximum acceptable producer's risk. Another option would be reject the lot; this would be consistent with a desire to minimize consumers' risk without exceeding the maximum acceptable producer's risk. The opposite extreme would be reject the lot; this would be consistent with a desire to minimize producer's risk without exceeding the maximum acceptable consumers' risk. Many other policies are conceivable; the choice would be a value judgment for NIJ.

To recapitulate, specification of sample sizes implies a judgment about the risk NIJ will accept of accepting a lot with more than a maximum allowable percentage of "defective" units. (See box E-l.) A clearer alternative would be to specify the maximum acceptable risks explicitly and a means of calculating the sample sizes they require in specific cases (e.g., for sequential testing).

34A better definition of ''bad'' would include a trauma-survivability criterion, for example: For purposes of this standard, "bad armor'' is armor having a (geometric) mean stopping probability of no greater than 0.95 or a probability per shot of exceeding the backface signature limit of greater than 0.05.

35 The 1-sample Kolmogorov-Smirnov test [19, 45] also provides 1-sided confidence limits on the average passing probability, but they are conservative, not exact, for discrete distributions such as the binomial distribution.

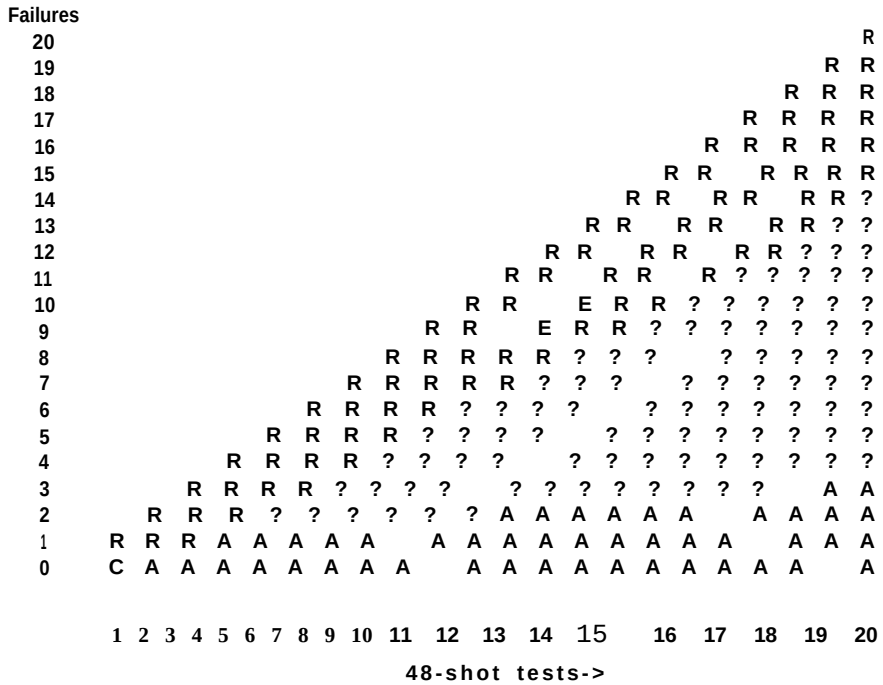
36 This is also the significance level—i.e., probability of error—at which on can reject the hypothesis that the lot is bad—i.e., has a probability of passing lower than 8.53 percent.

37 If the ballistic test were a V_{50} test (or some other test that results in a' 'score" rather than a pass or failure), a 1-sided, l-sample Kolmogorov-Smirnov test [19, 45] could be used to calculate a kind of consumers' risk or significance level-the probability that the actual distribution of V_{50}s in the lot exceeds the empirical distribution of measured V_{50}s (i.e., is worse) by some specified margin.

Figure E-n-Example of Control Chart for Acceptance Testing

1% Consumer's Risk, $p_B = .950^{48} = 0.0853$

1%. Producers' Risk, $p_G = .999^{80} = 0.953$



Legend:
R= REJECT—Consumers' Risk too great if accepted
A= ACCEPT—Produoers' Risk too great if rejected
?=Could ACCEPT or REJECT
C= Conflict must ACCEPT and REJECT (so require more tests)
p^B=maximum probility that bad armor will pass(definfiionof bad armor).
p_G=minimum probability that good armor will pass (definition of good armor).
SOURCE: Office of Technology Assessment 1992.

Sample Selection

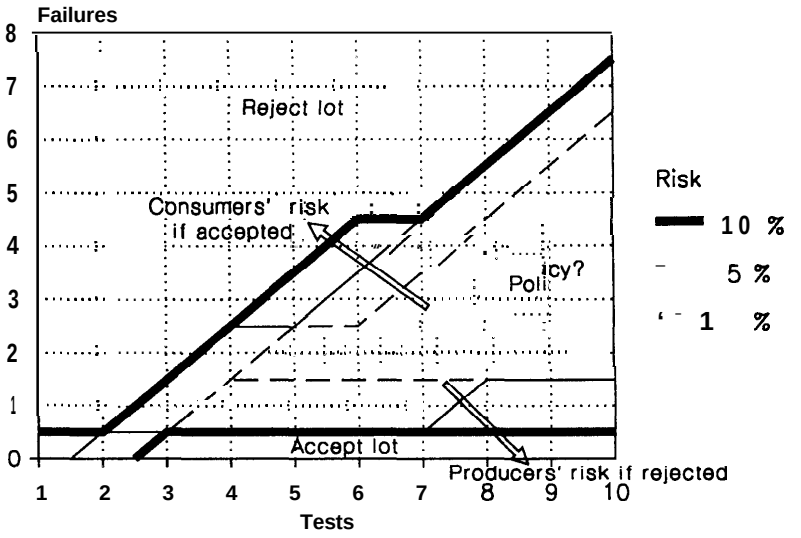
A lot-certification process could require a lot submitted for sampling and testing to be inventoried, tagged, and sampled by (or as prescribed by) NIJ, and the samples to pass a sequential test such as that described above. The armor need not all be shipped to NIJ; it could be inventoried and sampled on the manufacturer's premises by an agent of NIJ. The samples would be sealed and shipped for testing, while the balance of the armor would remain sealed on the manufacturer's premises until the samples are certified to have, or found not to have, the specified level of ballistic resistance,

All armor labeled as belonging to the lot would have to be inventoried. Marketing a unit of armor labeled as belonging to a lot that has been certified when in fact the unit was kept aside from, or produced after, the NIJ inventory and sampling would be false and deceptive labeling, an offense punishable under existing statutes enforced by the FTC. However, detecting such a practice would require a government surveillance program, which could be run by NIJ. It might require undercover purchases on the open market, which might require substantial funding, unless sellers agree to reimburse the costs of obtaining the samples randomly.

Quality-Control Options

Some manufacturers have extensive in-house quality-control programs; here we consider how purchasers and wearers could be assured of product

Figure E-12—Testing More Samples Can Reduce Both Consumers' and Producers' Risks



This figure shows the boundaries between the rejection, indeterminate, and acceptance regions of figure E-1 1, as well as boundaries for 5-percent consumers' and producers' risks and for 10-percent consumers' and producers' risks. (For all cases, $p_B = 0.0853$ and $p_G = 0.953$.)

SOURCE: Office of Technology Assessment, 1992.

quality by an independent third party, such as NIJ, with expertise and a vested interest in quality assurance, and none in armor sales.

In general, the testing and certification could be done by the government or by the private sector (e.g., UL or HPWLI), with or without government (NIJ or OSHA) supervision, and could be voluntary or compulsory. However, a compulsory program, such as would be authorized by enactment of H.R. 322, might be limited to inspection and ballistic testing of products (e.g., lot certification). The alternatives described in this section would require intimate access to the manufacturing process and the cooperation with the manufacturer; they are probably only feasible if voluntary.

An alternative to certifying lots is to certify models (as is now done) and also test samples of units of certified models produced after certification to decide whether they differ significantly from the samples tested for model certification. If they do, certification of the model would be suspended until the production process is corrected. If the decision is made by statistical inference, this is called statistical process control (SPC). Other options rely more on inspection-of samples of armor as well as the production process—and less on ballistic testing, to attain a desired level of confidence in product quality.

In one option for SPC, NIJ would require V_{50} measurements [38] as part of the certification test, to provide a baseline against which V_{50}s of future samples of the same model could be compared to check consistency of physical properties. However, certification of a model would not depend on the measured V_{50}s; it would continue to depend on a test of ballistic resistance, such as those specified by NIJ Standard 0101.03.

At least two V_{50}s would have to be measured in certification testing to establish upper and lower control limits-values within which V_{50}s of later samples must lie if they are to be considered consistent with the samples tested for certification. The upper and lower control limits would also depend on certain assumption-e.g., that V_{50}s of baseline samples are normally distributed-and on how many standard deviations from the mean the

38 As specified by MIL-STD-662D [138]; see also app. C.

Box E-l—Lot Sampling and Acceptance Testing in NILECJ-Std.-0101.00

NILECJ-Std.-0101.00, unlike later versions of the standard, contained a section (4.1) on quality assurance and an appendix (A) on sampling. [141] The apparent purpose of these sections was to provide guidance to manufacturers, retailers, and, especially, purchasers, who might want to specify quality-assurance provisions in a purchase agreement. The text of the standard specified ballistic tests, suggested procedures and sample sizes for lot testing, but did not describe the certification process, Apparently the NILECJ considered certification of lots, but left the definition of "lot" so vague that a rnanufacturer could call his entire production of a given model a "lot," The standard recommended that a sample of more than one unit should be tested if the lot size was larger than 8 units. However, the de facto certification process required a sample of only one unit from a lot of arbitrary size. This violated the only explicit quality-assurance requirement of NILECJ-Std.-0101 .00:

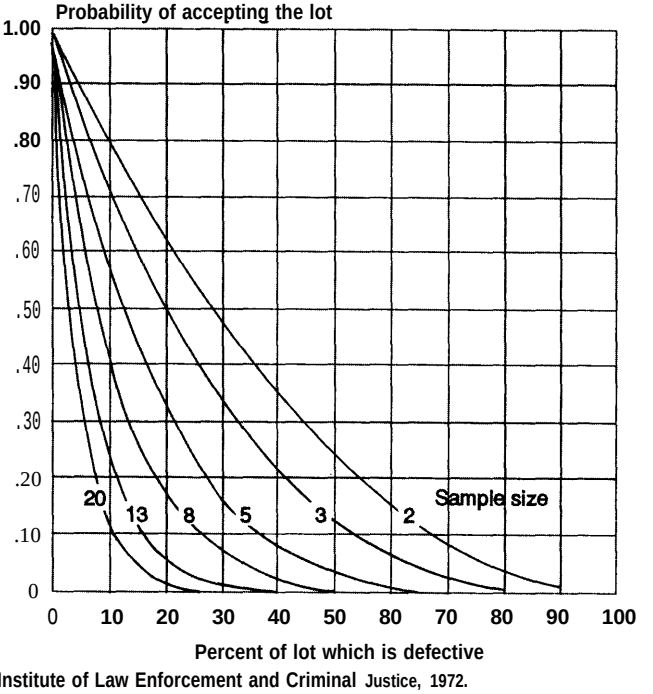
> A ssmple of each lot shall be taken for test at random, using a table of random numbers or an equivalent procedure.

If the entire production (including future production) of a given model is considered to be a lot, then one cannot, in the present, select a sample from it at random for testing. In effect, this "random sampling" requirement, the essence of which survives in the current standard, precludes considering the entire production of a model to be a lot Hence we consider certification of compliance with NILECJ-Std.-0101 .00 or its successors to be a design certification rather than any sort of lot certification--that is, it attests to the potential ballistic resistance of units of a certain design but provides no information on the actual ballistic resistance of production units. Section 4.1.1 of NILECL-Std.-0101.00 provided the following advice on sample size:

> The number of complete armors selected for test from each lot may be in accordance with the table below. This table is considered to be a reasonable compromise between an acceptable level of quality and the cost of testing. However, any desired sample size maybe selected by the purchaser, and should be specified in the purchase document. For a discussion of statistical considerations, see appendix A.

The standard recommended a sample size of 1 unit for a lot size of 1 to 8 units, and a sample size of 20 units for a lot size of 151 or more units. The recommendations imply judgments about the acceptability of risk as indicated in figure 4 of appendix A to the standard reproduced here.

Effect of Sample Size on the Probability of Accepting A Lot, As a Function of the Percent of the Lot That ь Defective



SOURCE: National Institute of Law Enforcement and Criminal Justice, 1972.

control knits should be, which can be deduced from the maximum probability of error allowed in deciding that the production process is ''out of control" when a sample's V_{50} falls outside the control limits. A typical but arbitrary choice is to choose upper and lower control limits 3 standard deviations above and below the mean; these are called '3-sigma' control limits. [31] Only 0.3 percent of the V_{50}s of samples produced by a process "in control" would lie outside 3-sigma control limits, if the V_{50}s of baseline samples were indeed normally distributed.

Once the control limits are established based on certification test results, samples of units of the model produced thereafter would be selected randomly (e.g., each unit produced having a l-percent chance of being selected) and their V_{50}s would be measured. If the V_{50} of any sample is outside the control limits, the production process would be judged to be out of control, and certification of the model would be suspended until the production process is corrected (so that sample V_{50}s again fall within the control limits).

Control limits based on certification test results could be used for other purposes, even if NIJ did not want to use them for SPC. For example, purchasers could use them as benchmarks for acceptance tests: A purchaser could make acceptance of a lot contingent on samples having V_{50}s within the control limits, or above the lower control limit. They could also be used to investigate the possibility of false or deceptive labeling: For example, if armor of a certified model failed to perform as rated in service, its V_{50} could be measured and compared to the control limits. If outside, it would indicate that the production process was out of control when the unit was produced, even if inspection revealed the failed armor to be identical in appearance to the units submitted for certification testing.

Advocates of V_{50} tests for quality testing propose that nondeformable fragment-simulating projectiles (FSPs) [139] be used, instead of bullets, for the V_{50} tests, because, being machined from steel instead of cast from lead, they are more uniform (and more penetrating) than any bullet,[39] and FSP V_{50}s of similar samples generally have less variance, than do ballistic V_{50}s of similar samples. However, they also cost more (a .22-caliber FSP costs about $1.50), and the 3-sigma control limits for ballistic V_{50}s are no more likely to be exceeded than are 3-sigma control limits for FSP V_{50}s of similar samples, although the former would be farther apart.

An advantage of using V_{50} tests, instead of pass/fail tests, for SPC is that many fewer tests (or shots) are required to establish control limits or thereafter discern an anomaly in quality at a specified level of statistical significance. One could, for example, calculate 3-sigma control limits for the number of passes (O or 1) of one .03 test, but this test statistic would not be normally distributed.[40,41] The number of passes in 30 or more .03 tests would be approximately normally distributed, but obtaining such a statistic would require submission of 180 samples of armor, and shooting at least 120 of them!

Thus FSP V_{50} tests are an economical means of detecting a *statistically significant change in* armor and are used for this purpose by the military and by some manufacturers of police armor. However, a statistically significant change in FSP V_{50} may or may not denote an unacceptable change in the type of ballistic resistance in which confidence is sought. A statistically significant change in FSP V_{50} would be grounds for subjecting additional samples to inspection and ballistic-resistance testing, but not necessarily for concluding that ballistic resistance has become unacceptable. The converse should also be considered: an unacceptable change in the type of ballistic resistance in which confidence is sought may not be reflected in a statistically significant change in FSP V_{50}. Experts believe that it would, but it would be difficult to prove that it would, for all types of bullets and armors.

FSP V_{50} tests may be more acceptable to some purchasers and wearers for SQC than certification-type tests or ballistic V_{50} tests, for psychological reasons:

1. Because the tests are different from the certification test, manufacturers might approach periodic retesting without the trepidation some

[39] See, e.g., T.A. Abbott, "The Variation of the Geometry of Fragment Simulators," pp. 205-218 in [134].

[40] Binomial confidence limits could be used in this case, if the probability of passing were assumed to be constant when the process is in control, Or a Kohnogorov-Smirnov test in any case.

[41] Another issue is that, for the process to be "in control," the probability of passing would have to be 99.7 percent—much higher than is necessary for armor to have better than even odds of being certified.

feel when contemplating repeated testing with the NIJ .03 test.
2. Purchasers and wearers who might be wary of armor certified to have been penetrated by bullets (as in a ballistic V_{50} test) might accept armor certified to have been penetrated by FSPs, which are laboratory instruments (not bullets like those used by criminals).

Other options rely more on inspection and lesson ballistic testing to attain a desired level of confidence in product quality. Some options rely on inspection of the production process as well as inspection of samples of armor. A voluntary program resembling the Classification program of Underwriters Laboratories (UL) would be based on the following principles:[42]

1. Testing to a nationally recognized standard.
2. Publication of the test results in a report that includes a comprehensive description including photos and drawings of the products.
3. Publication of a list of manufacturers and specific products that have demonstrated by tests compliance with the requirements.
4. Factory follow-up inspections at least four times a year using the report described in item 2 to assure that production units are identical to the unit which was submitted for and passed the testing.
5. Annual sample retest—this involves selection of a representative sample during one of the inspection visits and returning it to the test laboratory for retest to assure continued compliance.
6. Products produced under such a program would carry the mark of the third-party certification laboratory. This would facilitate user identification of those products that have been deemed to be in compliance with the standard.
7. The test laboratory shall maintain tight control of its mark. Compliance failure at either the factory follow-up inspection, item number 4, or annual retest, item number 5, would require corrective action, removal of the certification mark, or holding of shipment of the affected units. Additionally, certification marks could easily include lot traceability identifiers which could facilitate a recall as a last resort.

A manufacturer seeking to have a product Listed or Classified by UL pays UL to inspect and test initial samples of the product to determine whether the product meets UL standards for safety from fire and electrical shock (e.g., in the case of Listing) or some other standard (in the case of Classification). If so, and if the manufacturer agrees to allow (and pay) UL to conduct a limited number of surprise inspections of the manufacturer's production and quality-control processes (including some tests of randomly-selected production items), then UL Lists or Classifies the product, and permits the manufacturer to affix a seal ("mark") indicating that the product is Listed or Classified by UL[43]

The cost of UL or UL-like procedures for assuring the quality of body armor would depend on the standard to which they should comply, which in turn might specify how samples are to be selected, inspected, and tested, and the confidence (if any) with which the tests are to assure that the samples are identical to the original test articles or, in any case, provide the ballistic resistance required.

One option would be to test intitial samples for model certification in accordance with NIJ Standard 0101.03 or a similar standard, and thereafter to base certification of product quality (viz., similarity to the initial samples) on audits of the manufacturer's production and quality-control processes and on selection, inspection, and ballistic testing of production samples.

The feasibility of intitial testing by UL was demonstrated in June 1988, when UL conducted a series of tests of body armor for TAPIC in accordance with NIJ Standard 0101.03. The testing was overseen by a staff member of the NIST Law Enforcement Standards Laboratory to verify that the work was in conformance to the .03 standard and consistent with its interpretation at LESL. UL now estimates that such initial testing of a model could be performed for about $3,000 and about $1,500 for each additional model from the same manufacturer) tested at the same time.

An ongoing followup inspection program typically involves a basic annual charge of $435 plus an inspection fee of $72 per hour spent by the UL inspector at the manufacturing facility. UL estimates

42 Isaac I. Papier, Managing Engineer, Burglary Detection and Signalling Department, Underwriters Laboratories, Inc., personal communication, Aug. 5, 1991.

43 Today, UL Lists no armor garments but does test and certify a broad range of products that provide ballistic protection.

that a basic followup service for NIJ-like armor Classification would require 4 annual visits, each about 1 or 2 hours long, if the manufacturer's quality-control program is in good order. On one of the visits, the UL inspector would select random samples (not necessarily including samples of all models) for testing, the cost of which would be extra but much less than that for initial testing, because not all models would be tested and no report would be generated. [112] Hence the recurring annual cost to a manufacturer could be little more than about $700 to $1,000.

This option would provide neither quantitative estimates of the confidence in the program nor (the other side of the coin) of the probability of failure-i.e., the probability that a unit of production armor Classified by UL as complying with the standard of ballistic resistance actually does not (or fails a ballistic test, which is not quite the same thing). Some manufacturers might hesitate to participate in it, because they would perceive the unannounced factory inspections as intolerably intrusive.

Although this option for UL Classification would not provide purchasers of UL-Classified armor with quantitative estimates of risks, other options could. For example, lot-acceptance testing and certification, as described above, could be done in the context of UL Classification if the NIJ standard were revised to apply to lots instead of models.

If NIJ reconsiders UL Classification or an analogous option and solicits bids for such a program, several independent test laboratories might respond by proposing programs.