



Calhoun: The NPS Institutional Archive
DSpace Repository

Reports and Technical Reports

All Technical Reports Collection

2001-09-01

Military stochastic scheduling treated as a "multi-armed bandit" problem

Glazebrook, Kevin D.

Monterey, California. Naval Postgraduate School

<http://hdl.handle.net/10945/15377>

Downloaded from NPS Archive: Calhoun



Calhoun is a project of the Dudley Knox Library at NPS, furthering the precepts and goals of open government and government transparency. All information contained herein has been approved for release by the NPS Public Affairs Officer.

Dudley Knox Library / Naval Postgraduate School
411 Dyer Road / 1 University Circle
Monterey, California USA 93943

<http://www.nps.edu/library>

NAVAL POSTGRADUATE SCHOOL Monterey, California



Military Stochastic Scheduling Treated As a “Multi-Armed Bandit” Problem

by

Kevin D. Glazebrook
Donald P. Gaver
Patricia A. Jacobs

September 2001

Approved for public release; distribution is unlimited.

Prepared for: Director, Operational Test and Evaluation (DOT&E)
Washington, DC 20301-1700

Research also supported by the Institute of Joint Warfare Analysis (IJWA) and the Modeling, Virtual Environments and Simulation (The MOVES) Institute at the Naval Postgraduate School

NAVAL POSTGRADUATE SCHOOL
MONTEREY, CA 93943-5000

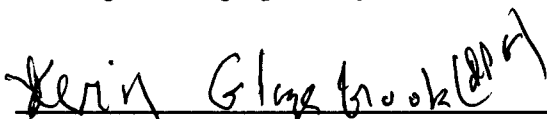
RADM David R. Ellison
Superintendent

Richard Elster
Provost

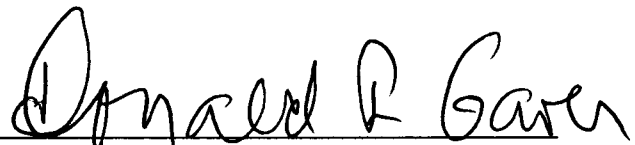
This report was prepared for and funded by the Director, Operational Test and Evaluation (DOT&E), The Pentagon (Room 3E318), Washington, DC 20301-1700. Research also supported by the Institute of Joint Warfare Analysis (IJWA) and the Modeling, Virtual Environments and Simulation (The MOVES) Institute at the Naval Postgraduate School.

Reproduction of all or part of this report is authorized.

This report was prepared by:



KEVIN D. GLAZEBROOK
Professor of Operations Research

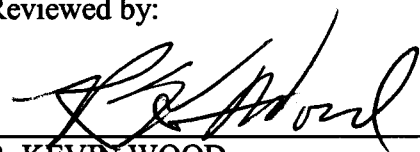


DONALD P. GAVER
Distinguished Professor of
Operations Research




PATRICIA A. JACOBS
Professor of Operations Research

Reviewed by:

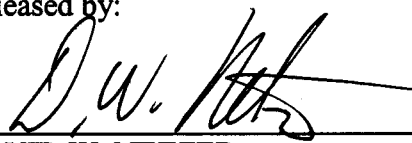


R. KEVIN WOOD
Associate Chairman for Research
Department of Operations Research

Released by:



JAMES N. EAGLE
Chairman
Department of Operations Research



DAVID W. NETZER
Associate Provost and Dean of Research

REPORT DOCUMENTATION PAGE

Form approved

OMB No 0704-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503.

1. AGENCY USE ONLY (Leave blank)**2. REPORT DATE**

September 2001

3. REPORT TYPE AND DATES COVERED

Technical Report

4. TITLE AND SUBTITLE

Military Stochastic Scheduling Treated As a "Multi-Armed Bandit" Problem

5. FUNDING

MIPR NO. DVAM10001

6. AUTHOR(S)Kevin D. Glazebrook
Donald P. Gaver
Patricia A. Jacobs**7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)**Naval Postgraduate School
Monterey, CA 93943**8. PERFORMING ORGANIZATION
REPORT NUMBER**

NPS-OR-01-010

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)Director, Operational Test and Evaluation (DOT&E)
The Pentagon (Room 3E318)
Washington, DC 20301-1700**10. SPONSORING/MONITORING
AGENCY REPORT NUMBER****11. SUPPLEMENTARY NOTES****12a. DISTRIBUTION/AVAILABILITY STATEMENT**

Approved for public release; distribution is unlimited.

12b. DISTRIBUTION CODE**13. ABSTRACT (Maximum 200 words.)**

A Blue airborne force attacks a region defended by a single Red surface-to-air missile system (SAM). Red is uncertain about the Blues he faces, but is able to learn about them during the engagement. Red's objective is to develop a policy for shooting at the Blues to maximise the value of Blues shot down before he himself is destroyed. We show that index policies are optimal for Red in a range of scenarios and yield effective heuristics more generally. The quality of such index heuristics is confirmed in a computational study.

14. SUBJECT TERMS

multi-armed bandit, index policies, air defense

**15. NUMBER OF
PAGES**

23

16. PRICE CODE**17. SECURITY CLASSIFICATION
OF REPORT**

Unclassified

**18. SECURITY CLASSIFICATION
OF THIS PAGE**

Unclassified

**19. SECURITY CLASSIFICATION
OF ABSTRACT**

Unclassified

**20. LIMITATION OF
ABSTRACT**

UL

Military Stochastic Scheduling Treated As a “Multi-Armed Bandit” Problem

Kevin D. Glazebrook
Department of Statistics
University of Newcastle
Newcastle-upon-Tyne
NE1 7RU England

Abstract

A Blue airborne force attacks a region defended by a single Red surface-to-air missile system (SAM). Red is uncertain about the Blues he faces, but is able to learn about them during the engagement. Red’s objective is to develop a policy for shooting at the Blues to maximise the value of Blues shot down before he himself is destroyed. We show that index policies are optimal for Red in a range of scenarios and yield effective heuristics more generally. The quality of such index heuristics is confirmed in a computational study.

1. Introduction and Basic Scenario

The following scenario is a simplified version of one occurring when a Blue airborne force attacks a Red region defended by a Red missile system; see Barkdoll, et al (2001).

A single Red surface-to-air missile (SAM)—hereafter, simply Red—can attack and be attacked by a collection of N Blue airborne attackers, labelled 1 through N . Blues come in B types, but Red only has imperfect information concerning the nature of the Blues he is facing. Red is able to construct N (independent) prior distributions $\Pi^1, \Pi^2, \dots, \Pi^N$ which summarise his beliefs about the type identities of the Blues before any shooting starts. Hence Π_b^j is the probability that Red assigns to the event “Blue number j is of type b , $1 \leq j \leq N$, $1 \leq b \leq B$ ” in advance of action. At each time $t = 0, 1, 2, \dots$ Red shoots at a single Blue and that Blue retaliates by firing back on Red. Red has a (constant) probability r_b of destroying a type b Blue with a single shot, and has (constant) probability θ_b of being destroyed by a retaliatory strike. Red knows when a Blue has been destroyed because no retaliatory strike follows. All shooting outcomes are assumed to be independent of each other. If Red destroys a type b Blue with his t^{th} shot then he receives

a reward $V_b \alpha^t$, where V_b is the utility associated with this occurrence and $\alpha \in [0,1]$ is a discount rate. Red's goal is to maximise the expected utility of Blues destroyed prior to his own destruction.

A crucial feature of the model concerns Red's capacity to update his beliefs about the Blues he is facing in the light of the outcomes of past engagements, by using Bayes' Theorem. In particular, if Blue target j has been involved in n engagements and he and Red have survived them all (note that this is the only event of interest for future decision-making) then the prior Π^j becomes the posterior $\Pi^{j,n}$ given by

$$\Pi_b^{j,n} = \frac{\Pi_b^j (1-r_b)^n (1-\theta_b)^n}{\sum_d \Pi_d^j (1-r_d)^n (1-\theta_d)^n}, 1 \leq b \leq B. \quad (1)$$

Hence, Red's beliefs about the Blues are evolving and this will plainly impact his shooting decisions.

2. An Index Result for a Class of Generalised Bandit Problems

The above problem will be analysed by means of a result due to Nash (1980), in a contribution that developed the classical index result of Gittins and Jones (1974). Nash envisages N "bandits", the j^{th} of which is in state $X_j(t)$ at time t . A decision-maker chooses one of the bandits to process at each time $t = 0, 1, 2, \dots$. The effect of choosing bandit j at time t is as follows:

- (i) bandit j experiences a Markovian change of state $X_j(t) \rightarrow X_j(t+1)$. Bandits not chosen remain fixed;
- (ii) a reward $\alpha^t \left\{ \prod_{i \neq j} q_i \{X_i(t)\} R_j \{X_j(t)\} \right\}$ is generated.

The novelty of this model concerned the multiplicatively separable reward structure in (ii) above. Here all bandits make a contribution to the rewards generated when j is chosen through the so-called influence functions q_i . The q 's and R 's are non-negative and bounded.

If some policy v is used for choosing bandits, with $v(t)$ used for the choice made at t , then the total return under Nash's model can be written

$$E_v \left[\sum_{t=0}^{\infty} \alpha^t \left[\prod_{i \neq v(t)} q_i \{X_i(t)\} \right] R_{v(t)} \{X_{v(t)}(t)\} \right]. \quad (2)$$

The goal is to choose v to maximise the return in (2).

Nash was able to show that, under certain conditions (which are satisfied in all of the problems discussed here), this problem has an index solution of the following character: at each time t , compute a calibrating index

$$G_1 \{X_1(t)\}, G_2 \{X_2(t)\}, \dots, G_N \{X_N(t)\}$$

for each bandit in its current state. An optimal policy will always choose that one of the bandits with the largest index. It does not matter how ties are broken.

We can deploy Nash's model to solve our problem as follows: the bandits correspond to the N Blues. The state $X_j(t)$ of Blue j at time t has three components, labelled $\Pi^j(t)$, $I_R^j(t)$ and $I_B^j(t)$. Here $\Pi^j(t)$ is the posterior distribution for Blue j describing Red's current beliefs about it—see (1). Both $I_R^j(t)$ and $I_B^j(t)$ are indicator functions as follows:

$$I_R^j(t) = \begin{cases} 0, & \text{if by time } t, j \text{ has destroyed Red,} \\ 1, & \text{otherwise,} \end{cases}$$

and

$$I_B^j(t) = \begin{cases} 0, & \text{if by time } t, j \text{ has been destroyed by Red,} \\ 1, & \text{otherwise.} \end{cases}$$

To deploy Nash's model for our problem we make the following choices for each j :

$$\begin{aligned} q_j \{X_j(t)\} &= I_R^j(t), \\ R_j \{X_j(t)\} &= 0 \text{ whenever } I_R^j(t) = 0 \text{ or } I_B^j(t) = 0. \end{aligned} \quad (3)$$

Otherwise R_j records a single return when Blue j is destroyed.

The effect of the choices in (3), when placed within Nash's reward structure, is

- (a) to wipe out any further returns following Red's destruction by any Blue j ; and

(b) to wipe out further returns from Blue j following its own destruction.

This is precisely what we want. The total return in (2) is now exactly the expected utility of Blues destroyed until Red's own destruction.

We return now to Nash's general model, but we shall exploit the fact that our q functions (from (3)) all have starting values 1, which remain there until a possible transition to 0. This simplifies the index structure considerably. Consider Blue j in some state x for which $q_j(x) = 1$. We shall describe the index $G_j(x)$ which is used in determining the optimal policy. Imagine Red shooting at Blue j (from initial state x at $t = 0$) until some positive-valued stopping time τ , defined with respect to Blue's evolving state. Define the reward rate $G_j(x, \tau)$ earned up to τ by

$$G_j(x, \tau) = \frac{E\left[\sum_{t=0}^{\tau-1} \alpha^t R_j\{X_j(t)\} \mid X_j(0) = x\right]}{1 - E\left[\alpha^\tau q_j\{X_j(\tau)\} \mid X_j(0) = x\right]}. \quad (4)$$

The index $G_j(x)$ is the largest such reward rate, namely

$$G_j(x) = \sup_{\tau > 0} G_j(x, \tau). \quad (5)$$

In the next section we show how to develop indices for the problem in Section 1. A general methodology for index computation for Nash's model may be found in Glazebrook and Greatrix (1995). Other discussions of Nash's model are found in Fay and Glazebrook (1987), Glazebrook and Owen (1991), Glazebrook and Greatrix (1993), and Glazebrook (1993).

3. Indices for the Blues

The problem in Section 1 may be formulated as a Bayes sequential decision problem (in which the expected reward is taken with respect to the prior distributions Π^j , $1 \leq j \leq N$, as well as over the realisations of the engagement) whose structure conforms to Nash's generalised bandit, as outlined in Section 2. Hence, all we have to do is specify what the indices G_j are which determine optimal policies for Red. In discussing this we can concentrate on individual Blues, and hence, drop the Blue identifier j .

Consider a Blue target whose associated prior is Π and which has had n engagements with Red, which have left both of them intact ($I_R = I_B = 1$). Refer to this state as (Π, n) . For the purposes of Red's decision making it is only such Blues and such states which are of interest. In Red's next engagement with this Blue, three things can happen: (1) Blue is destroyed and Red not; (2) Red is destroyed and Blue not; and (3) neither is destroyed. In the formulation as a Bayesian sequential decision problem we use the posterior in (1) to develop the probabilities of these three events as

$$(1) \quad p[\text{Blue destroyed and Red not}] = \sum_b \Pi_b r_b (1-r_b)^n (1-\theta_b)^n / D(\Pi, n),$$

$$(3) \quad p[\text{neither destroyed}] = \sum_b \Pi_b (1-r_b)^{n+1} (1-\theta_b)^{n+1} / D(\Pi, n),$$

$$(2) \quad p[\text{Red is destroyed and Blue not}] = 1 - p[\text{Blue destroyed and Red not}] - p[\text{neither destroyed}].$$

In (1) and (3) we take $D(\Pi, n) = \sum_b \Pi_b (1-r_b)^n (1-\theta_b)^n$.

Further, the expected return for Red from the next engagement is given by

$$\frac{\alpha \sum_b \Pi_b V_b r_b (1-r_b)^n (1-\theta_b)^n}{\sum_b \Pi_b (1-r_b)^n (1-\theta_b)^n}.$$

Now, in following the prescription for computing the index at the end of Section 2 we only need (for theoretical reasons) to consider certain kinds of stopping time τ in our determination of the index $G(\Pi, n)$ of the Blue under discussion. Specify positive integer $r(\geq 1)$. We write τ_r for Red's stopping time in which, from time 0 (at which point the state of the Blue is assumed to be (Π, n)), Red has r further engagements with Blue unless one or other of them is destroyed first. The random variable τ_r is the number of shots from Red that results from this, and cannot exceed r or be less than one. The expected reward up to τ_r , which is required for the numerator in (4), may be expressed as

$$\frac{\sum_b \Pi_b (1-r_b)^n (1-\theta_b)^n \left\{ \sum_{s=0}^{r-1} \alpha^{s+1} V_b r_b (1-r_b)^s (1-\theta_b)^s \right\}}{D(\Pi, n)} \quad (6)$$

and the expression in the denominator (recall that q is just I_R) is

$$1 - \frac{\sum_b \Pi_b (1-r_b)^n (1-\theta_b)^n \left\{ \sum_{s=0}^{r-1} \alpha^{s+1} r_b (1-r_b)^s (1-\theta_b)^s + \alpha^r (1-r_b)^r (1-\theta_b)^r \right\}}{D(\Pi, n)}. \quad (7)$$

From (5), (6) and (7) the index $G(\Pi, n)$ may be developed as

$$G(\Pi, n) = \max_{r \geq 1} \left[\frac{\sum_b \Pi_b (1-r_b)^n (1-\theta_b)^n \left\{ \sum_{s=0}^{r-1} \alpha^{s+1} V_b r_b (1-r_b)^s (1-\theta_b)^s \right\}}{\sum_b \Pi_b (1-r_b)^n (1-\theta_b)^n \{1 - A_{1b} - A_{2b}\}} \right] \quad (8)$$

where

$$A_{1b} = \sum_{s=0}^{r-1} \alpha^{s+1} r_b (1-r_b)^s (1-\theta_b)^s$$

and

$$A_{2b} = \alpha^r (1-r_b)^r (1-\theta_b)^r.$$

We can now implement an optimal policy. If Red is still alive, then he computes all the indices for the still live Blues and engages next whichever live Blue has the largest index.

In order to understand index structure, introduce the so-called ‘‘one-step index’’ $H(\Pi, n)$ obtained by taking $r = 1$ in (8) as

$$H(\Pi, n) = \frac{\alpha \sum_b \Pi_b (1-r_b)^n (1-\theta_b)^n V_b r_b}{\sum_b \Pi_b (1-r_b)^n (1-\theta_b)^n \{(1-\alpha) + \alpha \theta_b (1-r_b)\}}. \quad (9)$$

It is straightforward to establish the following:

- (i) If $H(\Pi, n)$ is decreasing in n , then the maximum in (8) is attained at $r = 1$ for all n and it then follows that $G(\Pi, n) = H(\Pi, n)$ for all n . If this behaviour holds good for all Blues then the index policy is quasi-myopic (a one-step look ahead rule). Here indices are always decreasing, and so in an optimal policy, which always targets the Blue with the largest index, Red will switch his targeting of the Blues frequently.
- (ii) If $H(\Pi, n)$ is increasing in n , then the maximum in (8) is attained for all n in the limit as $r \rightarrow \infty$. Here $G(\Pi, n)$ can be shown to be increasing in n . If this behaviour holds good for all Blues then Red will, in an optimal policy, persist in targeting individual Blues in turn until each is destroyed;

- (iii) If there are just two Blue types ($B = 2$), then $H(\Pi, n)$ will be either increasing or decreasing in n ;
- (iv) $H(\Pi, n)$ may be thought of (somewhat crudely) as a weighted average (with respect to the posterior distribution) of a vulnerability index

$$V_b r_b / \{(1 - \alpha) + \alpha \theta_b (1 - r_b)\}$$

for Blues of type b . This vulnerability index is high when V_b and r_b are large and when θ_b is small. It is plainly such Blues that Red would like to shoot at. In fact, $H(\Pi, n)$ takes expectations for the numerator and denominator of the vulnerability index separately. The index formula in (8) tells Red exactly how to choose.

A variety of extensions to the above are available from standard index theory. Two are, perhaps, worthy of mention:

- (a) When new Blues arrive for engagement in a Poisson fashion, an index policy is still optimal. The index in (8) is not always quite the right one, but will do very well in practice. See Fay and Glazebrook (1992);
- (b) If there are several identical Red shooters operating in parallel, instead of just one, and the Red objective is to maximise the utility from destroying Blues until all Reds are destroyed, then the above index policy (operated in the obvious way) will do very well, but will not, in general, be strictly optimal. See Glazebrook and Garbe (1998).

4. Some Major Extensions

We elaborate the scenario in Section 1 by supposing that Red could be one of several (R) Red types and each Blue has at his disposal several weapons, some of which may be designed for use against particular Red types. In this situation, each Blue will seek to learn about what kind of Red type he faces as well as vice-versa. We shall assume that the individual Blues can only learn about Red independently of each other—they cannot

pool information. We shall consider a range of approaches, in increasing levels of complexity. Note that there are minor variants of most of the following proposals and most of the objects described can be j (i.e., target) dependent.

(a) **Blue's strategy known to Red**

The simplest option is to suppose that each Blue type b has a strategy for choosing successive weapons in the face of inconclusive engagements and that these strategies are known to Red. Hence, for each blue type b , there is a sequence $\{W_b(n), n \geq 1\}$ of weapons to be used. Note that we do not actually require that all Blues of type b have the same strategy—that is just here for simplicity. An index policy is still optimal and the indices concerned involve minor adjustments to (8). We write

$$\bar{\theta}(b, n) = \prod_{m=1}^n \{1 - \theta_{W_b}(m)\} \quad (10)$$

where θ_W is the kill probability for weapon W . The index for this situation may be shown to be

$$G(\Pi, n) = \max_{r \geq 1} \left[\frac{\sum_b \Pi_b (1 - r_b)^n \left\{ \sum_{s=0}^{r-1} \alpha^{s+1} V_b r_b (1 - r_b)^s \bar{\theta}(b, n + s) \right\}}{\sum_b \Pi_b (1 - r_b)^n \{1 - B_{1b} - B_{2b}\}} \right] \quad (11)$$

where

$$B_{1b} = \sum_{s=0}^{r-1} \alpha^{s+1} r_b (1 - r_b)^s \bar{\theta}(b, n + s)$$

and

$$B_{2b} = \alpha^r (1 - r_b)^r \bar{\theta}(b, n + r).$$

(b) **Blue's beliefs known to Red**

This is, in fact, a simple example of (a) in which Blue's strategies $W_b \equiv \{W_b(n), n \geq 1\}$ are developed as Blue type b updates his prior beliefs P^b about the Red type he faces. This notation presupposes that all Blues of the same type will have

the same priors, but this is **not** an essential feature. If Red has access **both** to the P^b 's and also to how Blue is using his posterior beliefs to choose successive weapons, then he has access to Blue's strategy and a suitable form of the index in (11) can be used.

(c) **Blue's beliefs not known to Red**

The approaches in (b) will yield an optimal index policy whose return $R(P^1, P^2, \dots, P^B)$ will depend upon the priors P^b describing Blue's initial beliefs about Red. How do we proceed if we drop the assumption that Red knows the P^b 's? The two classical decision-theoretic approaches are:

(1) **Suppose Red is minimax**

Here Red acts conservatively and chooses the best (i.e., index) policy for the "least favourable" priors. For most reasonable models, this will amount to Red supposing that all Blues **know** what kind of Red type he is and calculating indices accordingly.

(2) **Suppose Red is Bayes**

Here Red expresses his beliefs about the unknown P^b 's via appropriate prior distributions $\phi_b(\mathbf{p})$. We are putting priors on priors, each of the latter being an R -dimensional probability vector \mathbf{p} . Indices can now be developed as follows:

For each b we have

$$\mathbf{p} \rightarrow \text{weapon sequence } W_b(\mathbf{p}) \rightarrow \bar{\theta}(b, n, \mathbf{p}), n \geq 1,$$

extending (10). The index (11) now is developed to become

$$G(\Pi, n) = \max_{r \geq 1} \left[\frac{\sum_b \int \Pi_b (1-r_b)^n \left\{ \sum_{s=0}^{r-1} \alpha^{s+1} V_b r_b (1-r_b)^s \bar{\theta}(b, n+s, \mathbf{p}) \right\} \phi_b(\mathbf{p}) d\mathbf{p}}{\sum_b \int \Pi_b (1-r_b)^n \{1 - C_{1b}(\mathbf{p}) - C_{2b}(\mathbf{p})\} \phi_b(\mathbf{p}) d\mathbf{p}} \right] \quad (12)$$

where

$$C_{1b}(\mathbf{p}) = \sum_{s=0}^{r-1} \alpha^{s+1} r_b (1-r_b)^s \bar{\theta}(b, n+s, \mathbf{p})$$

and

$$C_{2b}(\mathbf{p}) = \alpha^r (1-r_b)^r \bar{\theta}(b, n+r, \mathbf{p})$$

and such indices determine the optimal policy for the Bayesian Red. In this formulation, Red can make inferences about Blue's evolving beliefs about what kind of Red he is. For example, if $R = 5$ and a Blue type possesses 5 weapons, each one potent against one of the 5 different Red types and ineffectual against the others, then after 4 inconclusive engagements, Red will understand that such a Blue type now almost certainly has a clear view of what kind of Red he is and that such a Blue's next retaliation could well be fatal for him. The index in (12) will reflect these developing beliefs.

An assumption that Blues can pool their information about Red will induce stochastic dependence among the Blues. Appropriately developed indices can do well but will not be strictly optimal. See, for example, Boys, Glazebrook and McCrone (1996).

References

- D. Bertsimas and J. Niño-Mora, "Conservation laws, extended polymatroids and multi-armed bandit problems: a polyhedral approach to indexable systems," *Mathematics of Operations Research*, **21**, 257-306, 1996.
- R.J. Boys, K.D. Glazebrook and C.M. McCrone, "A Bayesian model for the optimal ordering of a collection of screens," *Biometrika*, **83**, 472-476, 1996.
- N.A. Fay and K.D. Glazebrook, "On the scheduling of alternative stochastic jobs on a single machine," *Advances in Applied Probability*, **19**, 955-973, 1987.
- N.A. Fay and K.D. Glazebrook, "On a 'no arrivals' heuristic for single machine stochastic scheduling," *Operations Research*, **40**, 168-177, 1992.
- T.C. Barkdoll, D.P. Gaver, K.D. Glazebrook, P.A. Jacobs, and S. Posadas, "Enemy Air Defense (EAD) Suppression as an Information Duel," Naval Postgraduate School Working Paper 2001.
- J.C. Gittins and D.M. Jones, "A dynamic allocation index for the sequential design of experiments," in *Progress in Statistics*, ed. J. Gani and I. Vince, 241-266, North-Holland, Amsterdam.
- K.D. Glazebrook, "Indices for families of competing Markov decision processes with influence," *Annals of Applied Probability*, **3**, 1013-1032, 1993.
- K.D. Glazebrook and R. Garbe, "Almost optimal policies for stochastic systems which almost satisfy conservation laws," *Annals of Operations Research*, (to appear), 1998.
- K.D. Glazebrook and S. Greatrix, "On scheduling influential stochastic tasks on a single machine," *European Journal of Operations Research*, **70**, 405-424, 1993.
- K.D. Glazebrook and S. Greatrix, "On transforming an index for generalised bandit problems," *Journal of Applied Probability*, **32**, 168-182, 1995.
- K.D. Glazebrook and R.W. Owen, "New results for generalised bandit processes," *International Journal of Systems Science*, **22**, 479-494, 1991.
- P. Nash, "A generalised bandit problem," *Journal of the Royal Statistical Society*, **B42**, 165-169, 1980.

Appendix A

Issues for the Blue force

The scenario is as in Section 1 of the main report where the primary focus is on Red's decision-making. However, the controller of the Blue force also faces some issues. A natural first question for Blue concerns what force he needs to deploy in order to destroy an optimally shooting Red with a given large probability, 0.95, say. This, in fact, turns out to be straightforward to assess. Suppose that N_b type b Blues are deployed, $1 \leq b \leq B$. The probability of Red's ultimate survival (having destroyed all Blues) does not depend upon his strategy for engaging them. Hence, we may as well suppose that Red engages each Blue in a continuous fight until one or the other is destroyed. In such an engagement it is easy to show that

$$p[\text{Blue of type } b \text{ is destroyed}] = r_b / \theta_b (1 - r_b) = \psi_b, \text{ say, } 1 \leq b \leq B.$$

Hence, the probability that Red survives the battle with N_b type b Blues, $1 \leq b \leq B$, is given by

$$\psi_1^{N_1} \psi_2^{N_2} \dots \psi_B^{N_B}$$

and the controller of Blue requires this to be less than or equal to 0.05, say. If there is only one Blue type, then the choice is of the smallest number N to deploy such that

$$\psi^N \leq 0.05.$$

If we now ask how the Blue force should accomplish the destruction of Red with given probability at least cost to itself, then the strategy for Red does come into play since, for example, Red may tend to engage "expensive" Blues first. Hence, we shall suppose that Red shoots optimally, and will consider a simple situation for Blue in which $B = 2$ and the loss of each type b costs him C_b , $b = 1, 2$. We note from the main report that when $B = 2$, all indices are either increasing or decreasing in n . We shall suppose that the

former is the case for all Blues and so Red's optimal shooting policy engages each Blue non-preemptively until one or other is destroyed.

Let the expected cost to the Blue force of the deployment of N_b type b 's, $b = 1, 2$ against an optimally shooting Red be denoted $C(N_1, N_2)$. Blue's optimisation problem is

$$\underset{N_1, N_2}{\text{minimise}} C(N_1, N_2)$$

such that (13)

$$\psi_1^{N_1} \psi_2^{N_2} \leq \varepsilon$$

where $1 \geq \varepsilon \geq 0$ and $1 - \varepsilon$ is the desired probability of killing Red.

We describe a scenario in which $C(N_1, N_2)$ may be computed easily. Red's priors for the Blue types he faces are obtained by moderating his ignorance about them (as initially expressed by $p(\text{Blue is of type } b) = 0.5, b = 1, 2$) by means of information obtained from a sensor. This sensor can only judge Blue type with error. We have

$$p[\text{Blue judged to be of type } b_1 | \text{Blue is of type } b_2] = \phi_{b_1 b_2}.$$

Hence, Red allocates to each Blue one of two possible priors $\Pi^b, b = 1, 2$ according to the judgement of the sensor. We have

$$\Pi_1^1 = p[\text{Blue is of type 1} | \text{Blue judged to be of type 1}] = \phi_{11} / (\phi_{11} + \phi_{12})$$

and similarly for the other probabilities. Let X_1 be a $\text{Bin}(N_1, \phi_{11})$ random variable representing the number of the N_1 type 1 Blue types judged by the sensor to be of type 1 and hence, given prior Π^1 by Red. Similarly, $X_2 \sim \text{Bin}(N_2, \phi_{22})$. Red faces $X_1 + N_2 - X_2$ Blue types to which he allocates prior Π^1 and initial index G_1 and $N_1 - X_1 + X_2$ Blue types to which he allocates prior Π^2 and initial index G_2 . Suppose $G_1 > G_2$ and so Red first engages all those Blues judged to be of type 1, followed by those judged to be of type 2. We assume that if Red faces two or more Blues with the same index then he chooses between them at random.

Now the cost of engaging b_1 (fixed) type 1 Blues and b_2 (fixed) type 2 Blues in random order can be computed recursively by

$$c(b_1, b_2) = \frac{b_1 \phi_1 \{C_1 + c(b_1 - 1, b_2)\}}{(b_1 + b_2)} + \frac{b_2 \phi_2 \{C_2 + c(b_1, b_2 - 1)\}}{(b_1 + b_2)}$$

$$c(0, 0) = 0.$$

Hence, the desired expected cost to Blue of the chosen deployment is given by

$$C(N_1, N_2) = E\{c(X_1, N_2 - X_2) + \psi_1^{X_1} \psi_2^{N_2 - X_2} c(N_1 - X_1, X_2)\}.$$

This can now be used in (13).

In more complicated situations, Blue's expected cost may be computed via suitable development of the methodologies described by Bertsimas and Niño-Mora (1996) for multi-armed bandits.

Appendix B

Shoot-Look-Shoot for Red

We elaborate the scenario described in Section 1 of the main report in two ways:

- (i) after every shot by Red, the targetted Blue is inspected and categorised (with error) according to type and alive/dead. Write

$$p[\text{Blue judged to be of type } b_1 | \text{Blue is alive of type } b_2] = \phi_{b_1 b_2}$$

$$p[\text{Blue judged to be of type } b_1 | \text{Blue is dead of type } b_2] = \phi_{b_1 \bar{b}_2},$$

where $1 \leq b_1, b_2 \leq B$.

- (ii) the Blue targetted by Red may or may not retaliate. We now have δ_b for the probability of retaliation to a single shot for live Blues of type b . Dead Blues do not fire back.

Inter alia, (ii) enables us to consider the deployment of decoys by Blue.

Red now gathers information about the Blues he is facing in a much more complicated way than previously. Index policies are still optimal, but the index structure is more complex and simple closed forms as in (8) above must not be expected. Consider a Blue target with assigned prior Π . Sufficient statistics gleaned from the history of Red's past engagements with this Blue, which will determine Red's posterior distribution for this target, are:

- (a) the number of Red shots faced by this target (n);
- (b) the outcome of the subsequent inspections ($\underline{b} = \{b_1, b_2, \dots, b_n\}$);
- (c) the number of retaliations by Blue (m);
- (d) the shot by Red to which Blue last retaliated (k).

Note that $m \leq k \leq n$. The posterior probability that, given n, \underline{b}, m, k , Blue is of type b and is still alive is proportional to

$$\Pi_b (1 - r_b)^n \left(\prod_{i=1}^n \phi_{b_i, b} \right) (1 - \theta_b)^m \binom{k-1}{m-1} \delta_b^m (1 - \delta_b)^{n-m} \equiv \Pi_b P_b(n, \underline{b}, m, k) \equiv \Pi_b P_b(H) \quad (14)$$

where H is used as a shorthand for the history (n, \underline{b}, m, k) . The posterior probability that, given n, \underline{b}, m, k , Blue is of type b but is now dead is proportional to

$$\begin{aligned} & \Pi_b (1-r_b)^k (1-\theta_b)^m \binom{k-1}{m-1} \delta_b^m (1-\delta_b)^{k-m} r_b \\ & \times \sum_{t=0}^{n-k-1} (1-r_b)^t \left(\prod_{i=1}^{k+t} \phi_{b,b} \right) \left(\prod_{i=k+t+1}^n \phi_{b,\bar{b}} \right) (1-\delta_b)^t \\ & \equiv \Pi_b \bar{P}_b(n, \underline{b}, m, k) \equiv \Pi_b \bar{P}_b(H) \end{aligned} \quad (15)$$

as before. Hence, given history H , the posterior probabilities are given by

$$\begin{aligned} p[\text{Blue alive of type } b|H] &= \Pi_b P_b(H) / \sum_b \Pi_b \{P_b(H) + \bar{P}_b(H)\} \\ p[\text{Blue dead of type } b|H] &= \Pi_b \bar{P}_b(H) / \sum_b \Pi_b \{P_b(H) + \bar{P}_b(H)\}. \end{aligned}$$

The corresponding *one-step index* for a Blue with prior Π and history H is given by

$$\frac{\alpha \sum_b \Pi_b P_b(H) V_b r_b}{\sum_b \Pi_b \{P_b(H) [1 - \alpha \{r_b + (1-r_b)(1-\delta_b \theta_b)\}] + \bar{P}_b(H)(1-\alpha)\}},$$

and will frequently yield good shooting policies.

In order to develop the index $G(\Pi, H)$ for a Blue with prior Π and history H that can be used to determine optimal policies for Red, we require an iterative procedure due to Glazebrook and Greatrix (1995). Denote by $\Omega(H)$ the set of histories reachable (in the obvious sense) from history H and $B\{\Omega(H)\}$ the set of bounded functions on $\Omega(H)$.

If $H = (n, \underline{b}, m, k)$ there are two distinct ways in which the history can evolve immediately from H , depending upon whether Blue retaliates or not during the next engagement with Red. If Blue does retaliate we have an evolution of the form

$$H \rightarrow H(b, \text{ret}) = \{n+1, (\underline{b}, b), m+1, n+1\} \quad (16)$$

on the assumption that neither party to the engagement is destroyed. To achieve the transition in (16), Blue needs to be judged by the sensor to be of type b and also to retaliate. If Blue does not retaliate, we have an evolution of the form

$$H \rightarrow H(b, \text{nonret}) = \{n+1, (\underline{b}, b), m, k\}. \quad (17)$$

Let $u \in B\{\Omega(H)\}$ and $H' \in \Omega(H)$. From Glazebrook and Greatrix (1995) we need to consider the transform $T_H: B\{\Omega(H)\} \rightarrow B\{\Omega(H)\}$ defined by:

$$\begin{aligned}
\{T_H(u)\}(H') = \max & \left[\left(\frac{\alpha \sum_b \Pi_b P_b(H') V_b r_b}{D(H')} + \frac{\alpha \sum_b \Pi_b P_b(H')}{D(H')} \left[r_b \sum_d \phi_{d\bar{b}} u\{H'(d, \text{nonret})\} \right. \right. \right. \\
& + (1-r_b)(1-\delta_b) \sum_d \phi_{db} u\{H'(d, \text{nonret})\} \\
& \left. \left. + (1-r_b)\delta_b(1-\theta_b) \sum_d \phi_{db} u\{H'(d, \text{ret})\} \right] \right. \\
& \left. + \frac{\alpha \sum_b \Pi_b \bar{P}_b(H')}{D(H')} \sum_d \phi_{d\bar{b}} u\{H'(d, \text{nonret})\} \right]; \\
& \left(\frac{\alpha \sum_b \Pi_b P_b(H) V_b r_b}{D(H)} + \frac{\alpha \sum_b \Pi_b P_b(H)}{D(H)} \left[r_b \sum_d \phi_{d\bar{b}} u\{H(d, \text{nonret})\} \right. \right. \\
& + (1-r_b)(1-\delta_b) \sum_d \phi_{db} u\{H(d, \text{nonret})\} \\
& \left. \left. + (1-r_b)\delta_b(1-\theta_b) \sum_d \phi_{db} u\{H(d, \text{ret})\} \right] \right. \\
& \left. + \frac{\alpha \sum_b \Pi_b \bar{P}_b(H)}{D(H)} \sum_d \phi_{d\bar{b}} u\{H(d, \text{nonret})\} \right) \Bigg]. \tag{18}
\end{aligned}$$

In (18) we use the notations established in (16) and (17), together with

$$D(H) = \sum_b \Pi_b \{P_b(H) + \bar{P}_b(H)\}$$

with similar usage for $H' \in \Omega(H)$. We compute the index $G(\Pi, H)$ by noting that

$$\lim_{n \rightarrow \infty} \{T_H^n(u)\}(H) = G(\Pi, H) \tag{19}$$

for any $u \in B\{\Omega(H)\}$. Observe that in (17), T_H^n denotes an n -fold application of T_H —i.e.,

$$T_H^n = T_H(T_H^{n-1}) = T_H(T_H(T_H^{n-2})) = \dots$$

Appendix C

Some Other Extensions

The main report mentions some developments of the simple scenario of Section 1. In (i) – (iii) below, we identify some further elaborations for which index policies remain optimal. In (iv) we identify other possible extensions for which index policies will perform well, while not always being strictly optimal.

- (i) Each Blue type has a finite number of bullets (known to Red). This requires a modest elaboration to the index structure and index policies remain optimal.
- (ii) Red has a finite number of bullets. Here we have a “finite horizon” version of the (potentially infinite) battle depicted in Section 1. The index policy based on $H(\Pi, n)$ remains optimal for the case that these are all decreasing in n .
- (iii) Here we elaborate the simple scenario in Section 1 by allowing all Blues that are still alive to take a shot at Red (after each of Red’s shots), and not simply that Blue which was targetted. Suppose that Blue number j has a probability η_j of killing Red (irrespective of which Blue type he is) when he is not the Blue targetted. Typically the η ’s will be much smaller than the θ ’s. Under certain plausible additional conditions, the index in (8) will be replaced by the following for Blue number j :

$$G_j(\Pi, n) = \max_{r \geq 1} \left[\frac{\sum_b \Pi_b (1-r_b)^n (1-\theta_b)^n \left\{ \sum_{s=0}^{r-1} \alpha^{s+1} V_b r_b (1-r_b)^s (1-\theta_b)^s \right\}}{\sum_b \Pi_b (1-r_b)^n (1-\theta_b)^n \left\{ (1-\psi_j) - A_{1b} - (1-\psi_j) A_{2b} \right\}} \right].$$

- (iv) Other versions of the “finite horizon” problem in (ii) for which the indices are *not* all decreasing are not strictly indexable, but index policies will usually continue to do well. The same holds for a suggested development in which each Blue would remain in the targetting zone for Red for just a finite amount of time before leaving (having, for example, run out of fuel).

Appendix D

Simulation Study

This appendix reports on results from a simulation model implemented by P.A. Jacobs. The scenario is as in Section 1 of the main report with Blue targets being of two types. There are b_1 type 1 Blue targets and b_2 type 2 Blue targets. Red uses a sensor to initially estimate the type of each Blue target. The probability that Red classifies a type i target as type i is ϕ_{ii} ; otherwise it is classified as the other type. Natural priors for Red to use in this context are (see Appendix A):

- (a) for those Blues judged to be of type 1:

$$\Pi_1^1 = \phi_{11}/(\phi_{11} + \phi_{12}) = 1 - \Pi_2^1;$$

- (b) for those Blues judged to be of type 2:

$$\Pi_1^2 = \phi_{21}/(\phi_{21} + \phi_{22}) = 1 - \Pi_2^2.$$

The simulation model implements two shooting policies for Red: (i) an index policy (as in Section 3 of the main report) with assigned values of $V_1 = V_2 = 1$, $\alpha = 1$; and (ii) random shooting in which, at each decision epoch, Red chooses to engage one of the remaining Blues chosen at random (with equal probabilities). Some results are presented in Tables 1 and 2. In each cell of both tables we report the estimated mean number of Blues killed prior to Red's destruction, with the corresponding standard error in brackets. The upper figures in each cell correspond to the index policy and the lower to the random shooting policy. In all runs we take $\phi_{11} = \phi_{22} = \phi$. All entries are based on 100 replications.

TABLE 1
(Blue types very different: $r_1 = 0.9$, $\theta_1 = 0.1$; $r_2 = 0.1$, $\theta_2 = 0.9$)

ϕ	(b_1, b_2)	(2, 8)	(4, 6)	(6, 4)	(8, 2)
1	(i)	2.04 (0.04)	3.98 (0.05)	5.65 (0.15)	7.82 (0.15)
	(ii)	0.39 (0.05)	0.83 (0.11)	1.41 (0.14)	3.42 (0.27)
0.95	(i)	1.66 (0.08)	3.36 (0.13)	5.25 (0.17)	7.28 (0.18)
	(ii)	0.34 (0.06)	0.72 (0.11)	1.33 (0.17)	2.74 (0.26)
0.9	(i)	1.35 (0.10)	2.82 (0.15)	4.60 (0.21)	6.46 (0.23)
	(ii)	0.33 (0.06)	0.59 (0.09)	1.59 (0.16)	3.03 (0.26)
0.85	(i)	1.19 (0.09)	2.29 (0.16)	4.20 (0.21)	6.13 (0.24)
	(ii)	0.39 (0.08)	0.81 (0.10)	1.32 (0.18)	3.27 (0.27)
0.8	(i)	0.89 (0.09)	1.96 (0.16)	3.38 (0.21)	5.86 (0.27)
	(ii)	0.24 (0.05)	0.82 (0.11)	1.49 (0.16)	3.41 (0.29)
0.7	(i)	0.69 (0.10)	1.39 (0.14)	2.88 (0.20)	4.28 (0.27)
	(ii)	0.44 (0.08)	0.74 (0.11)	1.28 (0.14)	2.70 (0.24)
0.6	(i)	0.41 (0.06)	1.06 (0.12)	2.07 (0.17)	3.72 (0.27)
	(ii)	0.30 (0.06)	0.69 (0.10)	1.58 (0.16)	3.21 (0.25)
0.5	(i)	0.30 (0.06)	0.86 (0.11)	1.49 (0.16)	2.80 (0.25)
	(ii)	0.30 (0.05)	0.81 (0.11)	1.52 (0.16)	2.56 (0.25)

The mean number of Blues killed by Red prior to Red's own destruction under (i) an index policy and (ii) a random shooting policy.

TABLE 2
(Blue types more alike: $r_1 = 0.7$, $\theta_1 = 0.3$; $r_2 = 0.3$, $\theta_2 = 0.7$)

ϕ	(b_1, b_2)	(2, 8)	(4, 6)	(6, 4)	(8, 2)
1	(i)	2.13 (0.11)	3.40 (0.19)	4.19 (0.27)	5.44 (0.33)
	(ii)	0.80 (0.11)	1.38 (0.19)	2.24 (0.23)	3.08 (0.33)
0.95	(i)	2.07 (0.15)	3.04 (0.20)	4.08 (0.26)	4.38 (0.32)
	(ii)	0.96 (0.10)	1.30 (0.16)	2.50 (0.28)	3.29 (0.29)
0.9	(i)	1.55 (0.14)	3.19 (0.22)	3.55 (0.25)	4.74 (0.33)
	(ii)	0.85 (0.11)	1.48 (0.19)	2.41 (0.25)	3.38 (0.30)
0.85	(i)	1.61 (0.14)	2.47 (0.21)	3.99 (0.28)	4.75 (0.32)
	(ii)	0.95 (0.16)	1.41 (0.15)	1.66 (0.19)	2.78 (0.26)
0.8	(i)	1.22 (0.14)	2.11 (0.18)	3.32 (0.25)	4.74 (0.32)
	(ii)	0.87 (0.11)	1.27 (0.17)	2.09 (0.21)	2.58 (0.28)
0.7	(i)	1.30 (0.14)	1.70 (0.18)	2.95 (0.26)	4.09 (0.32)
	(ii)	0.74 (0.12)	1.71 (0.21)	2.38 (0.24)	3.12 (0.30)
0.6	(i)	1.19 (0.15)	1.83 (0.20)	2.43 (0.21)	3.91 (0.31)
	(ii)	0.82 (0.11)	1.34 (0.16)	2.20 (0.24)	3.20 (0.29)
0.5	(i)	0.98 (0.13)	1.57 (0.20)	1.67 (0.20)	3.09 (0.29)
	(ii)	0.80 (0.13)	1.33 (0.17)	2.05 (0.21)	2.57 (0.29)

The mean number of Blues killed by Red prior to Red's own destruction under (i) an index policy and (ii) a random shooting policy.

Although plainly a more extensive simulation study (with more replication) is desirable, certain major features are already transparent from Tables 1 and 2. As we might expect, the index policy outperforms the random shooting policy other than at $\phi = 0.5$, where the sensor does no better than the flip of a fair coin and the two policies are virtually identical. The level of excess number of Blues killed achieved by the index policy is remarkably high when Red receives high quality information from the sensor assets (i.e., ϕ is high). However, even rather mediocre information ($\phi = 0.6$, say) can be put to very good use by Red. The value of the information to Red is unsurprisingly greater when the Blue types are more distinct.

Initial Distribution List

1. Defense Technical Information Center 2
8725 John J. Kingman Road, STE 0944
Ft. Belvoir, VA 22060-6218
2. Research Office (Code 09) 1
Naval Postgraduate School
Monterey, CA 93943-5000
3. Dudley Knox Library (Code 013) 2
Naval Postgraduate School
Monterey, CA 93943-5002
4. Richard Mastowski (Editorial Assistant) 2
Dept. of Operations Research
Naval Postgraduate School
Monterey, CA 93943-5000
5. Prof. Donald P. Gaver (Code OR/Gv) 1
Dept. of Operations Research
Naval Postgraduate School
Monterey, CA 93943-5000
6. Prof. Patricia A. Jacobs (Code OR/Jc) 1
Dept. of Operations Research
Naval Postgraduate School
Monterey, CA 93943-5000
7. Prof. Kevin D. Glazebrook 6
Dept. of Statistics
Newcastle University
Newcastle-upon-Tyne
NER1 7RU England