Equivalence tuning of *Strike Limit Algorithms*

Geof H. Givens[*], André E. Punt[+] and Tiffany A.O. Bernstein[‡]

Contact e-mail: geof@stat.colostate.edu

ABSTRACT

Equivalence tuning involves adjusting a candidate aboriginal whaling management *Strike Limit Algorithm* (*SLA*) to enable fair comparison with respect to its ability to satisfy the objectives for aboriginal subsistence whaling. Two methods for equivalence tuning (*'depletion tuning'* and *'H-tuning'*) are reviewed and compared. Conceptually, *H-tuning* is appealing because it accounts for aboriginal subsistence need as well as risk explicitly, whereas *depletion tuning* is based only on risk. However, *H-tuning* is only approximate, whereas *depletion tuning* is exact. Whale dynamics are slow so the choice among alternative *SLA*s is likely to be one related to a simple catch/risk trade-off. Hence, it is reasonable to favour the simpler *depletion tuning* approach if it can be implemented in a manner that facilitates fair and reasonable comparison. However, in one example shown, *H-tuning* was more successful at finding a comparison level that reflected an appropriate catch/risk balance.

KEYWORDS: MANAGEMENT PROCEDURE; WHALING-ABORIGINAL

INTRODUCTION

The Scientific Committee of the International Whaling Commission (IWC) began the development of a Revised Management Procedure for commercial whaling (RMP) in 1987 (IWC, 1988). Following the completion of most of the scientific aspects of this in 1992 (IWC, 1993), the IWC Scientific Committee began to consider the development of an aboriginal whaling management procedure (AWMP – see Donovan, 1999). Aboriginal subsistence whaling is practised in several areas of the world, including: regions of Alaska where Eskimo communities hunt bowhead whales (*Balaena mysticetus*); regions of the Russian Federation where native peoples of Chukotka hunt gray (*Eschrichtius robustus*) and bowhead whales; and regions of Greenland where Greenlanders hunt minke (*Balaenoptera acutorostrata*) and fin (*B. physalus*) whales.

An AWMP is a set of rules whose main purpose is to determine an annual strike limit (effectively a hunting quota). The strike limit is based in part on the level of aboriginal subsistence need, which is established from time to time by the Commission. An AWMP includes the details of the data used (e.g. abundance estimates and catch data), how surveys should be conducted, and the rules used to determine a strike limit given the data. The core of an AWMP is its *Strike Limit Algorithm* (*SLA*), which calculates the annual hunting quotas given available data. Determination of a suitable AWMP is dependant on the establishment of objectives. The IWC's objectives for AWMP development (IWC, 1995) are to:

(1) ensure that the risks of extinction to individual whale stocks are not seriously increased by aboriginal subsistence whaling;
(2) enable aboriginal people to harvest whales in perpetuity at levels appropriate to their cultural and nutritional requirements (i.e. their 'need'), subject to the other objectives; and
(3) maintain the status of whale stocks at or above the level giving the highest net recruitment and ensure that stocks below that level are moved toward it, so far as the environment permits.

The IWC has given highest priority to objective 1 and lowest priority to objective 2.

Progress thus far in the development process is documented in the reports of a Standing Working Group established by the IWC Scientific Committee (IWC, 1997a; 1998; 1999). The development of an AWMP will culminate in the comparison of the performance of several candidate *SLA*s in terms of how well they are able to satisfy the three objectives. Simulation is used to evaluate the performance of a candidate *SLA* for a wide range of scenarios. The scenarios examine the implications of uncertainty about whale biology and population dynamics, the environment, the quality of research data, changes in need, and many other factors. For a specific scenario, the past dynamics and the 100-year future of a whale stock managed during that future period by a candidate *SLA* are simulated. The *SLA* provides annual strike limits based on historical and simulated data. The simulations assume that the strikes allowed by the *SLA* are removed as catch from the stock each year. The long-term impact of the catch on the stock and the degree of need satisfaction are assessed through a variety of summary statistics (e.g. see IWC, 1999). A set of 100 replicate simulations for the same scenario is called a trial.

Initial AWMP development has relied upon a small set of trials known as 'initial exploration trials', which are intended to:

> (1) aid in assessing the value of the performance statistics...for testing the adequacy of any [procedures...and]
> (2) provide an initial framework for scientists to begin the process of developing potential [procedures]....such trials should be seen as an exploratory tool for use at the beginning of the long-term AWMP development process (IWC, 1997b, p. 243).

AWMP development has now progressed to the point where informal comparisons among candidate *SLA*s may help developers identify strengths and weaknesses of particular approaches. It is, however, difficult to compare a set of candidate *SLA*s if they achieve different trade-offs among

[*] *Department of Statistics, Colorado State University, Fort Collins, CO 80523, USA*
[+] *CSIRO Division of Marine Research, GPO Box 1538, Hobart, Tas 7001, Australia.*
[‡] *Centrobe, Inc., 6700 Winchester Circle, Boulder, CO 80301.*

the management objectives. For comparisons to be most useful, *SLA*s should first be 'equivalence tuned'. Equivalence tuning is:

> a way to provide *SLA* developers with the opportunity to adjust their *SLA*s to strive towards a pre-specified balance of risk, satisfaction of need, and recovery (IWC, 1998).

This paper first overviews two alternative methods of equivalence tuning (*'depletion tuning'* and *'H-tuning'*). The specifications in IWC (1998) reflect an initial attempt by the IWC Scientific Committee at fully specifying the *H-tuning* concept proposed by Givens (1997). However, these specifications are incomplete and the additional specifications needed to apply *H-tuning* are provided in this paper. In the remainder of this paper, the conceptual advantages and disadvantages of *H-tuning* relative to *depletion tuning* are identified and compared.

THE EQUIVALENCE TUNING METHODS

Let $A_t^{i,j}$ (B) denote the strike limit for year t for replicate j of trial i for *SLA* 'A', given that removals up to year t–1 have been determined by *SLA* 'B'. Also let $\hat{C}$ denoted the equivalence tuned version of *SLA C*.

Tuning can be achieved by modifying the values for the internal parameters of the *SLA*. However, it is also possible to tune an *SLA* by defining the strike limit using a formula that involves the strike limits from the *SLA* and other quantities (Givens, 1997; 1999). For example, one might consider a tuned *SLA* providing strike limits:

$$\hat{C}_t^{i,j} = \beta_0 + \beta_1 C_t^{i,j} + \beta_2 R\hat{Y}_t^{i,j} \qquad (1)$$

where $C_t^{i,j}$ is the strike limit provided by the *SLA* for year t for replicate j of trial i and $R\hat{Y}_t^{i,j}$ is an estimate of the replacement yield for year t for replicate j of trial i (provided by the *SLA* or some other estimation procedure). If one adopts this approach, the β_i constitute the tuning parameters.

Depletion tuning

Tuning of a set of *SLA*s can be achieved by selecting the values for their tuning parameters so that each achieves the same median final depletion for a certain trial. This approach to tuning (henceforth referred to as *depletion tuning*) was used during the final stages of the development of the RMP (IWC, 1992).

H-tuning

Depletion tuning is straightforward to specify and apply. However, it may represent a narrow view of equivalence tuning for AWMP development because it does not explicitly refer to the need satisfaction and risk objectives (Givens, 1998a). Givens (1997; 1999) developed an alternative tuning method (*H-tuning*) that may be more relevant to management. Two candidate *SLA*s would be considered equivalence-tuned when the values for their tuning parameters are selected to resemble a 'gold standard' series of strike limits as closely as possible. The 'gold standard' (referred to as *H*) is a series that might be desired if stock status and productivity were known without error at all times. *H* is therefore 'an *SLA*, which represents a particular balance among risk, need satisfaction, and recovery and which operates in the idealised case when the parameters of the [simulation] programme are known exactly' (IWC, 1998). Clearly the use of *H* still permits depletion to play a role in equivalence tuning. However, *H-tuning* also ensures that equivalence tuning also considers need satisfaction and stock recovery.

H was defined by IWC (1998) as:

$$H_t = \min\{Q_t, \tilde{H}_t\} \quad \tilde{H}_t = \begin{cases} 0 & if\, N_t^{1+} < 2{,}000 \\ 0.8RY_t & if\, 2{,}000 \le N_t^{1+} < MSYL^{1+} \\ 0.9MSY & if\, N_t^{1+} > MSYL^{1+} \end{cases} \qquad (2)$$

where:

H_t is the strike limit for year t based on H;
Q_t is the need for year t;
N_t^{1+} is the number of whales aged 1 or older at time t;
MSY is the maximum sustainable yield;
RY_t is the replacement yield during year t; and
$MSYL^{1+}$ is the number of whales aged 1 and older at which MSY is achieved.

In the following, the ideal strike limit for year t for replicate j of trial i will be referred to as $H_t^{i,j}$.

A *SLA* is H-tuned by selecting values for its tuning parameters to minimise a distance measure (the sum of squared deviations is used in this paper) between $H_t^{i,j}$ and the corresponding strike limits by the *SLA*, denoted $C_t^{i,j}$, over every year of every replicate over a set of equivalence tuning trials. If the sum of squared differences for trial i is denoted Z_i, then *H-tuning* involves finding the values of the tuning parameters to minimise the quantity $Z = \sum_i Z_i$.

Now Z measures the similarity of a candidate *SLA* to H when the removals up to year t-1 are based on the data used for *H-tuning*. We define $\hat{Z}$ as the similarity of the *SLA* to H that would be achieved during simulated implementation. One might consider alternative ways to measure the amount of tuning achieved because the values of Z and $\hat{Z}$ may be different depending on how good the *SLA* was before tuning:

$$R_{fit}^2(\hat{C}) = 1 - \frac{Z}{\sum_i \sum_{j=1}^{100} \sum_{t=0}^{99} \left(H_t^{i,j}(H) - \overline{H}\right)^2} \quad \text{and}$$

$$R_{imp}^2(\hat{C}) = 1 - \frac{\hat{Z}}{\sum_i \sum_{j=1}^{100} \sum_{t=0}^{99} \left(H_t^{i,j}(H) - \overline{H}\right)^2}$$

where $\bar{H}$ is the mean of $H_t^{i,j}$ over all tuning scenarios, replicates, and years. $R_{fit}^2(\hat{C})$ measures the proportion of the variation in H explained during the tuning of the candidate *SLA* and $R_{imp}^2(\hat{C})$ actual simulated implementation. $R_{fit}^2(\hat{C})$ and $R_{imp}^2(\hat{C})$ are somewhat analogous to the use of the coefficient of determination in simple linear regression. Although the calculation of $R_{fit}^2(\hat{C})$ does not require any simulation beyond that needed for tuning, it is clear that $R_{imp}^2(\hat{C})$ is the more relevant measure.

REVISED TECHNICAL SPECIFICATIONS FOR *H-TUNING*

H-tuning involves minimising the sum of squared differences between the strike limit series given by *H* and that given by *C*. However, there are three possible ways of defining this sum for a given trial i, and the end product of the tuning process can be shown to differ among these.

(1) Run a trial in which $C_t^{i,j}$ is removed each year, and the $H_t^{i,j}$ are calculated in each year from the past history of this simulation. In this case, both H and the *SLA* react to the removals specified by the *SLA*; only the *SLA* strike limits are removed and *H-tuning* then measures the squared difference between $H_t^{i,j}(C)$ and $C_t^{i,j}(C)$, i.e.:

$$Z_i = \sum_{j=1}^{100} \sum_{t=0}^{99} \left(H_t^{i,j}(C) - C_t^{i,j}(C)\right)^2 \tag{3a}$$

(2) Run a trial where $H_t^{i,j}$ is removed each year, and the $C_t^{i,j}$ are calculated in each year from the past history of this simulation. In this case, both H and the *SLA* react to the removals specified by H; the *SLA* strike limits are never removed and *H-tuning* measures the squared difference between $H_t^{i,j}(H)$ and $C_t^{i,j}(H)$, i.e.:

$$Z_i = \sum_{j-1}^{100} \sum_{t=0}^{99} \left(H_t^{i,j}(H) - C_t^{i,j}(H)\right)^2 \tag{3b}$$

(3) Run a trial where $H_t^{i,j}$ is removed each year and calculated on the basis of this simulation. A second trial is then run where $C_t^{i,j}$ is removed each year and calculated on the basis of this simulation. In this case, H reacts to itself and the *SLA* reacts to itself, so *H-tuning* measures the squared difference between $H_t^{i,j}(H)$ and $C_t^{i,j}(C)$, i.e.:

$$Z_i = \sum_{j=1}^{100} \sum_{t=0}^{99} \left(H_t^{i,j}(H) - C_t^{i,j}(C)\right)^2 \tag{3c}$$

For any of these three options, *H-tuning* consists of minimising $Z = \sum Z_i$, where the sum is taken over each of equivalence tuning trials. Z is minimised with respect to the tuning parameters in the *SLA* to be tuned. IWC (1998) appears to endorse Equation (3a). However, Equation (3c) is intuitively appealing because, *a priori*, the ideal *SLA* (H) must be independent of the *SLA* selected by the IWC, and because an *SLA* must react to the catches it allows. Furthermore, Equation (3c) would also be preferred from a control-theory point of view.

It should be noted, however, that none of the above three options avoid the difficulty that when the tuning parameter values that yield $\hat{C}$ are estimated using the modelling approach exemplified by Equation (1), the match between the *SLA* and H is optimised on the basis of the realisations of the *SLA* and H that were observed. However, when the tuned *SLA* is then applied in practice with the estimated tuning parameters, the strike limits are different. Since the tuned *SLA* provides different strike limits from the untuned version originally used as data for tuning, a form of predictive extrapolation is occurring.

NUMERICAL EXAMPLES

Comparing alternative *H-tuning* methods

The merits of the three alternative formulations of Z (Equations 3a–3c) were investigated using the prototype *SLA* of Givens (1998b), the trial scenarios from IWC (1998), and the August 1997 version of the IWC's simulation software. Table 1 shows some results for three tuned versions of this *SLA*. *H-tuning* was based on a 'very optimistic' and a 'very pessimistic' scenario with respect to whale stock conservation[1]. The prototype *SLA* was tuned using a variant of Equation (1):

$$\hat{C}_t^{i,j} = \beta_0 + \beta_1 C_t^{i,j} \tag{4}$$

where $C_t^{i,j}$ is the strike limit from the prototype *SLA*. Table 1 shows results from tuning (Z and R_{fit}^2) and from implementing the tuned *SLA* ($\hat{Z}$ and R_{imp}^2). It appears that Equation (3c) yielded the best results because it led to the largest R^2 values.

The tuned *SLA* was applied to a trial that was intermediate in terms of stock conservation between the two used for tuning. This trial was introduced to reduce the influence of the tuning trial set on two summary statistics that assess performance relative to the two key IWC objectives of avoiding excessive stock depletion and satisfying aboriginal need[2]. Final depletion is the stock size in the final year of the simulation expressed as a fraction of carrying capacity while total need satisfaction is the total catch allowed by the *SLA* expressed as a fraction of the total aboriginal need over the 100-year simulation period. High values are preferred for both statistics. The performance of the *SLA* tuned using Equation (3c) is slightly closer to that of H than those for the *SLA*s tuned using the other two definitions for Z_i (Table 1).

Iterated *H-tuning*

A question that arises is whether the strike limits $C_t^{i,j}$ should be calculated anew for every tested value of the tuning parameters or whether one could obtain a fast, approximate result by retaining the original strike limit series throughout the minimisation process. Clearly, if the tuning parameters used are internal to the *SLA*, the former approach is required. However, if the tuning parameters are external to the *SLA* (as in Equation 1) then it is possible to apply the other approach. Denote $\hat{Z}$ as the sum of squares achieved when the *H-tuning* objective function is updated at each iteration of the minimisation routine so that strike limits for the candidate *SLA* are based on the latest values of the tuning parameters. The problem with using $\hat{Z}$ rather than Z is that this puts the *SLA* simulation inside the minimisation loop - a process that may require prohibitive simulation. Basing *H-tuning* on Z

Table 1

Results for H-tuning using three alternative methods for generating the data used for tuning. The last two columns show the 5, 50, and 95% percentiles for the final depletion and total need satisfaction statistics for a trial specified in the text. The percentiles for H for this trial are (0.69, 0.69, 0.69) for final depletion and (0.86, 0.87, 0.89) for total need satisfaction.

Equation	Z/100000	R_{fit}^2	Z/100000	R_{imp}^2	Final depletion	Total need satisfaction
3a	163	0.511	151	0.527	(0.60, 0.71, 0.76)	(0.75, 0.84, 0.99)
3b	249	0.223	195	0.390	(0.70, 0.75, 0.78)	(0.73, 0.78, 0.86)
3c	148	0.537	149	0.535	(0.60, 0.69, 0.75)	(0.76, 0.86, 0.99)

[1] These are trials B3 and B7 of IWC (1998).

[2] These statistics are labelled $D_1(1+)$ and N_1, respectively, by IWC (1998).

rather than on $\hat{Z}$ can reduce computing time by three or four orders of magnitude while introducing only a 4-10% approximation in the objective function in some examples (Givens, 1999).

Improvement in mimicking H might be achieved efficiently by iterated *H-tuning* since the removals given by the first tuning result in different performance than was used to find the first tuning. This iterative approach might be considered as a strategy to achieve results resembling what would have been achieved by minimising $\hat{Z}$, without the simulation effort required to do so directly. Table 2 examines the improvement in performance that results for iterating the *H-tuning* process for the *SLA* in Table 1. Only the first iteration appears to be necessary in this example.

An example based on the Punt-Butterworth *SLA*

Table 3 lists summary statistics for the optimistic and pessimistic trials[3] for H and several different tunings of the Punt-Butterworth *SLA* (Punt and Butterworth, 1997). *Depletion tuning* was achieved by varying the values of an internal tuning parameter and was applied twice, once for each trial; it not being clear how to depletion tune for more than one trial simultaneously. The targets for *depletion tuning* were expressed in terms of the depletion of the mature female component of the population, and were set at the levels achieved by H. The *H-tunings* of this *SLA* were achieved by applying the model:

$$\hat{C}_t^{i,j} = \beta_0 + \beta_1 C_t^{i,j} + \beta_2 t C_t^{i,j} + \beta_3 t^2 C_t^{i,j} \qquad (5)$$

where $C_t^{i,j}$ was obtained using the original Punt-Butterworth *SLA*, and t is time. The selection of the values for β involved minimising the sum of squared deviations from H using Equation (3a) to define Z_i.

The first two rows of Table 3 illustrate the trade-off implicit with *depletion tuning*. When the *SLA* is tuned to the pessimistic trial, performance in terms of need satisfaction is relatively poor for the optimistic trial. Conversely, when the *SLA* is depletion tuned to the optimistic trial, performance for the pessimistic trial is extremely poor in terms of resource conservation. It is noteworthy that the lower 5% points for the performance statistics for the pessimistic trial are lower than those for H even when the Punt-Butterworth *SLA* is tuned to match the median final depletion for H. This is a reflection of the noise in the data available for use by the *SLA*.

The results for the H-tuned variant lie between those for the two depletion tuned variants. It is noteworthy that the trade-offs achieved by the H-tuned variant are better than can be achieved by simply changing the value of the internal tuning parameter in the Punt-Butterworth *SLA*. It is important to separate tuning and performance issues in Table 3. *Depletion tuning*, as currently envisaged, can only consider a single trial. This may seem attractive for equivalence tuning because of its simplicity, but equivalence is only useful if it is achieved at a performance level that is relevant for *SLA* comparison. Table 3 shows, as expected, that an injudicious choice for the trial used for *depletion tuning* and/or the target level can be problematic. In contrast, *H-tuning* is designed to achieve a reasonable balance among several trials which should preclude the type of problem evident in Table 3 for the 'Depletion tuned to optimistic trial' results.

One might also consider whether these tuning methods are useful for optimising *SLA* performance. The final row in Table 3 ('Manual') lists results for a variant that was tuned manually through the introduction of several additional tuning parameters. This variant is more risk-averse than the H-tuned variant, and might therefore be preferred. Although not carried out for this example, the level of risk-aversion of the H-tuned *SLA* can be set at any desired level using the full optimisation approach of Givens (1997; 1999) by weighting the tuning trials. Table 3 shows that the simple class of models used for the *H-tuning* example (Equation 5) did not contain a member whose performance was superior to that

[3] The trial scenarios and simulation framework used here are based on Appendix 3 of IWC (1998), except that density-dependence was assumed to relate to the total (1+) rather than mature female component of the population.

Table 2

Results of an iterative H-tuning experiment, for the original *SLA* and three subsequent iterations ('It.'). The ideal performance of H is also shown. Simulation performance is shown by the 5, 50, and 95% quantiles for the final depletion and total need satisfaction statistics. The Z and $\hat{Z}$ entries are scaled by 100,000. In the zero iteration, C was manually depletion tuned to achieve a reasonable risk/catch/recovery balance for the two scenarios (optimistic and pessimistic).

It.	Z	R^2_{fit}	$\hat{Z}$	R^2_{imp}	Optimistic: total need satisfaction	Optimistic: final depletion	Pessimistic: total need satisfaction	Pessimistic: final depletion
0	n/a	n/a	235	0.268	(0.70, 0.74, 0.95)	(0.87, 0.92, 0.93)	(0.61, 0.66, 0.69)	(0.39, 0.44, 0.47)
1	148	0.537	149	0.535	(0.77, 0.95, 1.00)	(0.83, 0.86, 0.92)	(0.60, 0.67, 0.73)	(0.34, 0.42, 0.48)
2	145	0.548	145	0.548	(0.81, 0.97, 1.00)	(0.83, 0.85, 0.91)	(0.60, 0.67, 0.73)	(0.34, 0.42, 0.47)
3	144	0.551	152	0.524	(0.81, 0.97, 1.00)	(0.83, 0.84, 0.91)	(0.61, 0.67, 0.75)	(0.33, 0.41, 0.47)
H					(1.00, 1.00, 1.00)	(0.83, 0.83, 0.84)	(0.57, 0.58, 0.58)	(0.51, 0.51, 0.52)

Table 3

Performance statistics (final depletion and total need satisfaction) for H and several variants of the Punt-Butterworth *SLA*. Final depletion in this table relates to the size of the mature female component of the population.

	Optimistic trial				Pessimistic trial			
	Final depletion		Total need satisfaction		Final depletion		Total need satisfaction	
Variant	5%	Median	5%	Median	5%	Median	5%	Median
H	0.472	0.531	1.000	1.000	0.383	0.406	0.518	0.523
Depletion tuned to pessimistic trial	0.722	0.767	0.547	0.572	0.365	0.406	0.490	0.520
Depletion tuned to optimistic trial	0.472	0.531	1.000	1.000	0.000	0.000	0.849	0.899
H-tuned	0.481	0.540	0.941	0.977	0.163	0.239	0.535	0.691
Manual	0.536	0.608	0.812	0.904	0.241	0.392	0.367	0.506

which could be achieved manually. The difficulty of selecting an appropriate model is a drawback of the *H-tuning* approach if one uses the modelling approach of Givens (1997; 1999).

COMPARISON OF TUNING METHODS

The best *SLA* can differ depending on the equivalence tuning method

Suppose that there was an objective way in which the preferred *SLA* among candidates could be selected once all the candidates have been equivalence tuned. We show below that the best *SLA* could depend on how the candidate *SLA*s were equivalence tuned.

Fig. 1 illustrates this result in the context of the trade-off between need satisfaction and depletion avoidance that is the dominant issue in *SLA* development. Panels (a) and (b) show time-parameterised trajectories of need satisfaction and stock depletion; the left endpoint of each curve corresponds to $t=0$, the right endpoint corresponds to $t=100$. The performance of an *SLA* is therefore summarised by the path between the endpoints. Panels (a) and (b) correspond to two different trials, either or both of which might be used to H-tune candidate *SLA*s. The dotted vertical line corresponds to a potential target for *depletion tuning*. Let us assume that *SLA* 'A' has already been depletion tuned, resulting in the curve labelled 'A'. *SLA* 'B' has not yet been tuned, and the range of possible tunings is shown by the gray shaded region. Only the edges of the region, labelled as the B1 and B2 tunings, need consideration. For Trial 1 (panel (a)), *depletion tuning* would lead to the selection of tuning B1 of *SLA* 'B', because this is the only tuning that exactly hits the target depletion level indicated by the vertical dotted line. Clearly B1 is inferior to 'A', so comparison of 'A' and 'B' after *depletion tuning* would result in the decision that 'A' is the better *SLA*. However, if *H-tuning* were used (based on Trial 1 only), B2 would be a tuning of *SLA* 'B' that is preferred over B1. In this case, comparison of 'A' and 'B' after *H-tuning* would lead to the conclusion that *SLA* 'B' is better. Thus, the decision about which *SLA* is better ('A' or 'B') depends on which tuning method is employed.

The same conclusions hold for hypothetical Trial 2 in panel (b) of Fig. 1. Since the same conclusion holds for both trials, minimising squared error across both trials (as would be done for *H-tuning*) does not eliminate this paradox. Panels (c) and (d) of this figure show possible strike limit trajectories corresponding to panels (a) and (b), respectively; these panels are provided only to display the problem in a more familiar context.

Clearly, if the *depletion tuning* process were to be stopped as soon as the target depletion was hit, without considering whether other sets of parameter values might also achieve the same depletion but at the same time yield a much better time-trajectory of strike limits, then this focus on a single objective (depletion) to the exclusion of others could lead to poor decisions.

The examples in Fig. 1 are conceptual, are not based on any real *SLA* and are exaggerated for visual clarity. It is probably unlikely that such extreme differences would occur in practice because productivity for whales is relatively

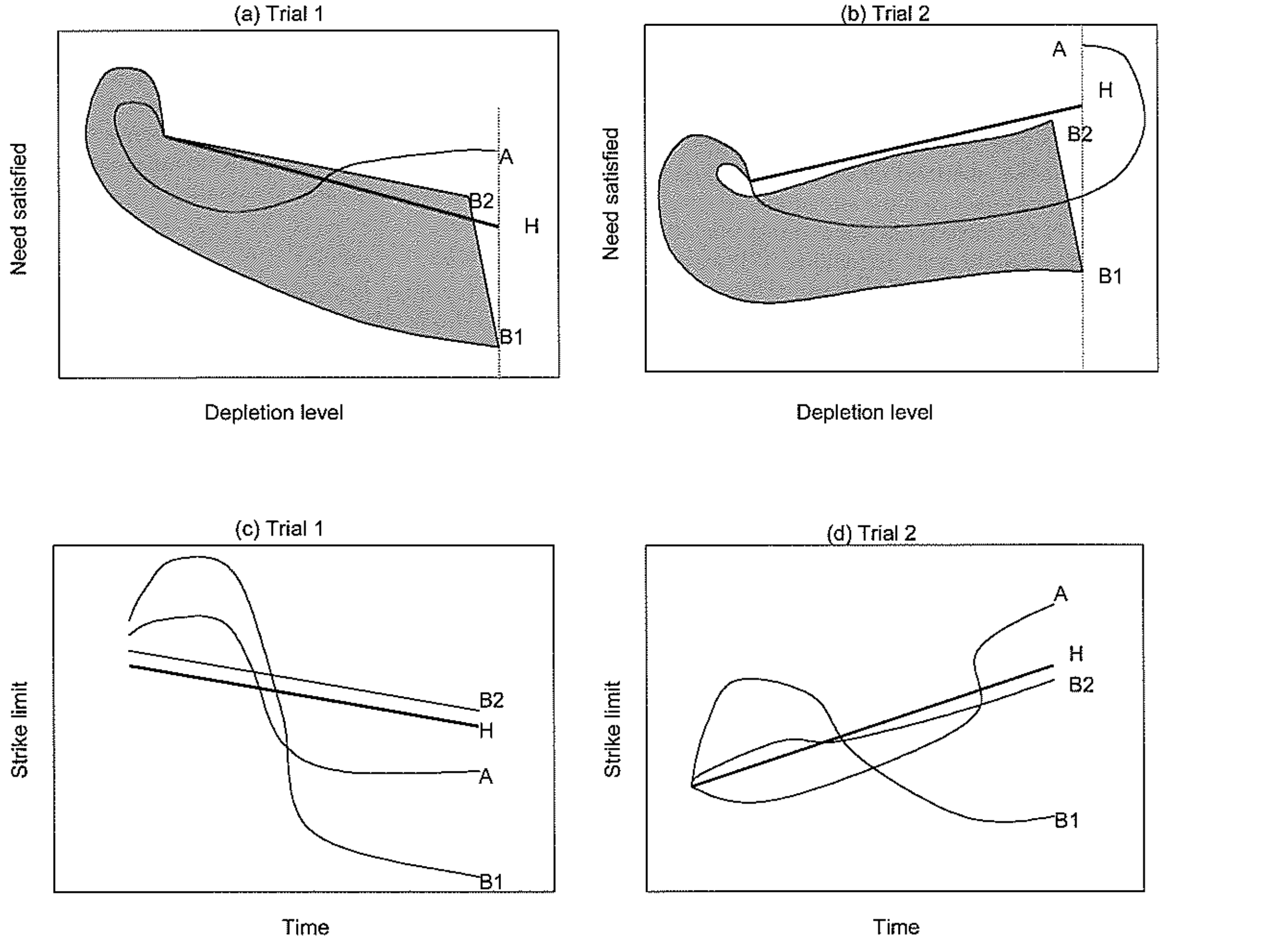

Fig. 1 Panels (a) and (b) show time-parameterised curves of need satisfaction and stock depletion for two hypothetical *SLA*s, 'A' and 'B', and for the ideal *H*. The shaded regions correspond to a range of possible tunings of 'B', with the edge curves B1 and B2 representing two extreme options. Panels (c) and (d) show strike limit trajectories which might correspond to the curves in (a) and (b), respectively.

small so there is a fairly 'linear' trade-off between depletion and catch. Furthermore, for any realistic *SLA*, there may be many combinations of parameter values that can 'hit' a target depletion level. For example, the *SLA* proposed by Punt and Butterworth (1997) has seven parameters that can be manipulated to achieve different balances among the management objectives. For this reason, it may be unrealistic to consider the curves labelled 'A' as the 'best' versions of *SLA* 'A'.

When is an *SLA* H-tuned?

It is almost always[4] possible to depletion tune an *SLA* because the target is simply a median final depletion. One potentially troubling aspect of *H-tuning* is that no *SLA* can be 'perfectly' tuned. The performance of any candidate *SLA* must be inferior to that of *H* because *H* is based on the true values for the population parameters whereas an actual *SLA* must estimate these from the data. This raises the important question of how to determine whether an *SLA* has been adequately equivalence tuned if *H-tuning* is used. A way to partially resolve this problem is to define a 'baseline' *SLA* and to define an *SLA* as 'tuned' if it achieves at least a (pre-specified) increase in R^2_{imp} compared to the 'baseline' *SLA*. However, this is an arbitrary way of solving the problem.

Use of time in *H-tuning* models

In principle, tuning approaches based on Equation (1) could have as one of the variables time, t (e.g. see Equation 5). However, this is problematic because, for some trial scenarios, need may be an increasing function of time. An *SLA* tuning which involves time as a factor would presumably adjust the raw strike limit in a time-dependent manner. Such a tuned *SLA* may perform poorly on trials where need is constant (or decreasing) through time.

This problem is fundamentally caused by using too few trials when *H-tuning*. For instance the choice of only two tuning trials in the above examples could fail to adequately span the range of testing space in some situations. A solution would therefore be to use more trials when tuning, including some with time-independent need trajectories. However, a simpler solution is simply to prohibit the use of the time for *H-tuning*.

Setting the target level for *depletion tuning*

An arbitrary choice for the *depletion tuning* target (e.g. 0.6) might result in comparison of *SLA*s that have been tuned to a performance balance that does not reflect a reasonable balance of the competing goals of aboriginal whaling management, or that does not reflect the operating conditions the *SLA* will face in reality. One might consider that a good target level for *depletion tuning* would be *MSYL*. Punt (1999) shows ranges and relationships for *MSYL* that are not entirely compatible with the Scientific Committee conventional wisdom, suggesting that basing a depletion target on *MSYL* may be prone to difficulties.

It seems sensible to set the *depletion tuning* target equal to that achieved by *H*, in order to ensure that the target is justified and reasonable. Since aboriginal whaling is conducted on stocks of diverse status and recovery potential, it is important to know that *SLA*s for these stocks are compared after being tuned to a balance of management objectives that might be reasonably expected for each stock. *H* provides an idealisation of this.

One can envision cases where a good *SLA* has performance inferior to a competitor at the exact level chosen for equivalence *depletion tuning*, while having superior performance at most other relevant tuning levels. In this admittedly pathological case, insistence on a single, exact *depletion tuning* target would lead to a mistaken preference for the generally inferior *SLA*. Multiple targets should be considered when using *depletion tuning* as was the case during the final stages of RMP development (IWC, 1992). *H-tuning* can avoid this problem altogether.

Which equivalence tuning approach is most appropriate?

If the behaviour of a *SLA* changes smoothly with the values for its tuning parameters, then it is reasonable to prefer the simple and familiar method of *depletion tuning* in most cases (although care needs to be taken when selecting the trial and/or target level). However, if need satisfaction is to considered explicitly in tuning then *H-tuning* for equivalence may be preferred. *H-tuning* can also improve *SLA* performance as a by-product of equivalence tuning. *Depletion tuning* puts the lengthy simulation program within an optimisation loop, whereas *H-tuning* can be conducted much faster. Therefore, if speed rather than exactitude is a consideration (as might be the case during the initial development process when each developer may have several candidate *SLA*s), *H-tuning* might be preferred.

ACKNOWLEDGEMENTS

Partial support for Givens and Bernstein was provided by the North Slope Borough (Alaska), the State of Alaska (through the Alaska Department of Community and Regional Affairs) and the National Atmospheric and Oceanic Administration (through the National Marine Mammal Laboratory to the Alaska Eskimo Whaling Commission). We thank David Poole, Department of Statistics, University of Washington, Seattle WA 98185 USA, Greg Donovan and two anonymous referees for their valuable comments.

REFERENCES

Donovan, G.P. 1999. Editorial. *J. Cetacean Res. Manage.* 1(1):ii-v.

Givens, G.H. 1997. Separable multicriterion design and performance optimization of AWMPs. Paper SC/49/AS4 presented to the IWC Scientific Committee, September 1997 (unpublished). 21pp.

Givens, G.H. 1998a. AWMP development and diverse prototypes. *Rep. int. Whal. Commn* 48:483-95.

Givens, G.H. 1998b. Further investigations of AWMP *Strike Limit Algorithms* and their optimisation. Paper SC/50/AWMP5 presented to the IWC Scientific Committee, April 1998 (unpublished). [Available from the Office of this Journal].

Givens, G.H. 1999. Multicriterion decision merging: competitive development of an aboriginal whaling management procedure. *J. Am. Stat. Assoc.* 94:1003-4.

International Whaling Commission. 1988. Comprehensive Assessment Workshop on Management. *Rep. int. Whal. Commn* 38:163-70.

International Whaling Commission. 1992. Report of the Scientific Committee, Annex D. Report of the sub-committee on management procedures. *Rep. int. Whal. Commn* 42:87-136.

International Whaling Commission. 1993. Report of the Scientific Committee. *Rep. int. Whal. Commn* 43:55-228.

International Whaling Commission. 1995. Chairman's Report of the Forty-Sixth Annual Meeting, Appendix 4. IWC Resolution 1994-4. Resolution on a Review of Aboriginal Subsistence Management Procedures. *Rep. int. Whal. Commn* 45:42-3.

International Whaling Commission. 1997a. Report of the Scientific Committee. Annex I. Report of the Workshop on the Development of an Aboriginal Subsistence Whaling Management Procedure (AWMP). *Rep. int. Whal. Commn* 47:192-202.

International Whaling Commission. 1997b. Report of the working group on aboriginal whaling managment procedures, Annex P. *Rep. int. Whal. Commn* 47:243-9.

[4] 'Almost' because the median final depletion may not be achievable even if the strike limit was set equal to zero for all years.

International Whaling Commission. 1998. Report of the Scientific Committee. Annex I. Report of the Standing Working Group on the Development of an Aboriginal Subsistence Whaling Management Procedure (AWMP). *Rep. int. Whal. Commn* 48:203-36.

International Whaling Commission. 1999. Report of the Scientific Committee. Annex F. Report of the Standing Working Group (SWG) on the Development of an Aboriginal Subsistence Whaling Management Procedure (AWMP). *J. Cetacean Res. Manage. (Suppl.)* 1:157-78.

Punt, A.E. 1999. A full description of the standard Baleen II model and some variants thereof. *J. Cetacean Res. Manage. (Suppl.)* 1:267-76.

Punt, A.E. and Butterworth, D.S. 1997. Preliminary evaluation of a (generic) strike limit algorithm for aboriginal subsistence whaling with comments on performance statistics and simulation trials. Paper SC/49/AS7 presented to the IWC Scientific Committee, September 1997, Bournemouth (unpublished) 22pp. [Available from the Office of this Journal].