CSIRO
MARINE RESEARCH

ISSUES IN THE SELECTION OF FINAL TRIALS FOR TESTING SBT MANAGEMENT PROCEDURES AND FOR THE PROCESS OF SYNTHESIZING RESULTS FROM THE SIMULATION TESTING

T. Polacheck
D. Kolody
M. Basson

CCSBT-ESC/0309/27

Table of Contents

List of Tables

List of Figures

Abstract

This paper discusses a number of issues related to the selection of final trials for testing SBT management procedures and to the process of synthesizing results from the simulation testing across a range of scenarios. The paper discusses implications from the robustness testing from the second stage of trials in terms of specifying final trial specifications. It also presents some additional results from the conditioning process with alternative parameter specifications for the operating model. The paper also discusses issues related to the stock/recruitment function and selectivity curves in the operating model. Finally, it provides further examination of issues related to synthesis of simulation testing across a wide range of scenario. The paper develops one possible approach for doing based on the approach used recently and successful by the IWC. The approach is provided as one possible framework and if found useful the details would require careful consideration and refinement. The paper emphasizes that critical importance at this stage of developing an agreed on approach for the synthesis process. Without an agreed on approach, it becomes extremely difficult to make further progress on refining the suite of candidate management procedures given that the tunning parameters within any decision rule allow for any decision rule to obtain a wide range of trade-offs in performance along the catch versus stock status axes.

Introduction

The CCSBT in 2001 agreed that the development of a management procedure was among the highest priority work for the scientific committee (Anon. 2001). A two year timetable which ends in March 2004 was adopted for completion of the technical work and testing process. The timetable foresees that the SAG meeting in August/September 2003 would develop the selection of the final trials for testing candidate management procedures. This would enable the next management procedure workshop scheduled for March 2004 to finalize its work and provide recommendations to the Commission on the selection of a management procedure. This paper addresses a number of issues relevant to the selection of the final trials and related issues that are important to resolve in order to complete the management procedure development work by March 2004. In particular, there is a close inter-relationship between the selection of final trials and the procedure that will be used to synthesize the results from these trials into a recommendation. Thus, the current paper addresses a number of technical issues related to the conditioning and projection process. It also proposes one possible approach with a detailed framework for the synthesis process.

Robustness Testing

Results of Robustness trial "tick tests"

On the basis of the 25 robustness "tick test" trials (and the MCMC scenario that we included in this group) defined with the intent of refining the final set of operating model scenarios, we would be inclined to suggest that the management procedure (MP) results are potentially sensitive to almost all of these alternatives. Figure 1 illustrates fairly typical results from two management procedures based on the ACRLRT decision rule (see Polacheck et al 2003b for a description of this rule). Any conclusions are obviously dependent on the definition of sensitive, which is somewhat arbitrary at this point (see below). Taking into account this caveat, we note the following:

- With the exception of the constant TAC decision rule, sensitivities tended to be qualitatively similar across a wide range of decision rules (see Polacheck et al 2003b).

- Over a range of tuning parameters (or control parameters) within a decision rule, different MPs varied in their sensitivities, presumably in relation to the yield/biomass trade-off.
- For all of the decision rules explored in Polacheck et al (2003b), MPs were fairly robust to the alternative specifications for selectivity change (H30M10Q0_SC and HM10Q0_SC) and the fecundity-based spawning biomass assumptions (H30M10Q0_Fec and H30M10Q0_Fec) – although we caution that the fecundity assumption was handled in a rather ad hoc fashion and probably did not represent the extent of plausible uncertainty in this dimension.
- It is probably necessary to carry the remaining uncertainty dimensions contained in these robustness trials into the final trials (or at least until there is further clarification on the management objectives and the process for synthesizing across scenarios of varying plausibility) in order to ensure a comprehensive and robust set of tests have been completed.
- Depending upon the synthesis process and the refinement of the management objectives, the importance of maintaining this range of uncertainty dimensions may be reduced. Similarly, which of these uncertainty dimensions are carried into the final trials (e.g. as robustness test or full crosses) and the range of values to consider within each dimension will also depend, at least in part, on the process adopted for synthesising results across the range of scenarios considered.

Implications from robustness testing

There are two rather different aspects of the robustness tests. The one aspect is from the point of view of the operating model (OM), and particularly the choice of scenarios to include in the next/final set of trials. The key question is whether the main parameters or assumptions to which the dynamics of the OM is sensitive have been identified, and hence whether the main uncertainties have been taken into account. This question should therefore be addressed by considering results over a wide range of management procedures (MP). Results from the robustness trials can identify which axes, in addition to those in the base set (i.e. the 18 scenarios with combinations of h, m and Q), should be considered in further trials, and which could potentially be 'dropped'.

The second aspect is from the point of view of a specific management procedure and its robustness against uncertainty. Here the question is more a matter of whether the specific MP is sensitive to a parameter or assumption.

Evaluation from the point of view of the OM

With regard to the first aspect the main issues are:

- how do we judge whether results are sensitive to a scenario or not?
- IF results are not sensitive, do we 'drop' that scenario/axis of uncertainty?
- IF results ARE sensitive, do we consider additional values along that axis, or only the original 'robustness' trial?

How do we judge sensitivity?

Irrespective of whether viewed from the OM or MP point of view, it is not straightforward to assess whether results are sensitive to a particular scenario (robustness trial) or not. This is because:

1. we have not defined what a 'pass' or 'fail' is in terms of each performance measure viewed separately
2. there are 19 performance measures (7 pertaining to the catch and 12 to the biomass) and we have not ranked or prioritised these, or specified a way of synthesising 'fail's and 'passes' over the 19 performance measures, also
3. we do not yet have the final set of criteria with respect to which the final set of MPs will be evaluated.

It is in fact rather difficult to see how we can make progress without the final set of performance measures, unless that final set is in fact identical to the current set of performance measures. This is because, without the final set, we may wrongly or prematurely ‘drop’ some of the robustness scenarios.

With regard to the first question, we have used the following criteria as defaults in the graphics ('percent of MPs whose evaluation criteria differ substantially') to indicate whether results from the robustness trial differs substantially from its comparable base case run(see Eveson 2003):

- For the 11 catch and biomass performance statistics, a substantial difference is defined as more than a 10% change in the median value relative to the reference case, or more than a 20% change in the range from the 10^{th} to 90^{th} percentile;
- For the TAC-related measures (already expressed as percent occurrences) a substantial difference is defined as an absolute change of more than 10% ;
- For the 6 robustness criteria, a substantial change is simply whether or not the result changes between the robustness model and the reference model (i.e. the robustness model passes the criteria where the reference model fails, or vice versa).

Clearly, other levels (percentages) can be used, and this needs to be discussed and agreed upon. It is, however, important not only to consider changes in the mean of a performance measure, but also changes in the variance or range of the measure.

With regard to synthesising over the 19 measures, a simplistic approach could be simply to count how many of the 19 performance measures show a substantial change (assuming we have defined that in some way as suggested above) over a wide range of MPs. If we decide to use this approach, however, we need to recognise that this would imply equal weighting of all the performance measures (e.g. a substantial change in the 5-year average catch has the same weight or importance as a substantial change in the Min(By:B2002) criterion). Such an equal weighting of performance measures is unlikely to reflect the actual importance of the different measures in terms of management’s actual objectives.

An alternative approach would be to weight each performance measure and form a weighted ‘pass’ or ‘fail’ score. Weighting of the 19 different performance measures could be a difficult task, but results so far suggest the possibility of simplifying the problem. On the basis of Figure 1 (and the set of similar figures in Polacheck et al 2003b), it appears that the most sensitive performance indicators are the biomass ratio measures (B2020:Bmsy, B2020:B1980, B2022:B2002) and the minimum biomass measure (Min(By:B2002)). What we mean by 'most sensitive' is that there are few, if any, cases where one of the other biomass-related performance measures shows a substantial change, but the above-mentioned ones do not. This suggests that we could consider focusing on a subset of the biomass performance measures (e.g. B2020:Bmsy, B2020:B1980, B2022:B2002 and Min(By:B2002)).

There also appears to be reasonable consistency and similarity between different MPs with regard to which indicators and for which robustness scenarios there are substantial differences in results (for the robustness trial and its comparable 'base case' run). The catch indicators also tend to be correlated with the biomass indicators (except, of course, for the "constant catch" MP). The indicator of catch variability (AAV) is not necessarily correlated with the others, but one could argue that this indicator reflects more the nature of the MP than of the underlying stock dynamics.

A much reduced set of performance could therefore be considered when evaluating whether results are sensitive to the robustness trial or not.

If results are NOT sensitive to a robustness trial

All MPs we investigated were fairly robust to the alternative specifications for selectivity change (H30M10Q0_SC and HM10Q0_SC) and the fecundity-based spawning biomass assumptions (H30M10Q0_Fec and H30M10Q0_Fec). This may suggest that these two assumptions, or axes of uncertainty, could be ignored in further trials. We do, however, caution that the fecundity assumption was handled in a rather ad hoc fashion, and there may be a need to revisit this assumption.

If the final set of performance measures are contained in the current set, then it would be reasonable to ignore 'insensitive' robustness trials in the next set of trials. If new performance measures are introduced, particularly if they are very different from the existing ones, then it would be prudent to continue to include even the 'insensitive' robustness trials in the next set.

If results are sensitive to a robustness trial

The MPs we considered appear to be sensitive to all but two of the robustness trials. This suggests that we may have to retain most of the robustness scenarios, and we need to consider how to go forward in the light of this. One option would be to include the same 'retained' robustness trials as "tick tests" in the next set of trials. Although this approach would address the issue from the MPs point of view, it does not necessarily address the issue of adequately covering the range of uncertainties from the OM point of view. In order to address that issue, it may be necessary to include additional scenarios with intermediate (or more extreme) values of the relevant parameter along the new axes of uncertainty identified by the robustness trials. The parameter values that were chosen for these robustness trials tended to be chosen on a rather ad hoc basis without any substantial consideration of their relative plausibility or what constituted the full plausible range. This was because it was hoped or anticipated that the results would not be sensitive.

In order to ensure that we have included the main axes of uncertainty in the final trials, it may be necessary to ask the following questions:

- would the existing robustness trial be sufficient in further testing?
- or should additional parameter values be considered?

In any case, it would seem necessary to carry the uncertainty dimensions represented by remaining sensitivity trials along until we get further clarification on the management objectives and the process for synthesizing across scenarios of varying plausibility. These factors will likely downgrade the importance of some operating models and robustness criteria.

Information Content in Fishery Data generated by the Operating Model

In conditioning the operating model, we have taken the approach of imposing a number of alternative parameter values (eg range of steepness values) that have a different quality of fit to different likelihood components. In doing this, we often introduce systematic deviations between predictions and observations. We need to be careful that the transition from the historical data to the simulated data does not provide too much information in the nature of the discontinuity. For example, if a very high steepness value is imposed in the operating model, the historical data would indicate that several recent recruitment years must have been coincidentally low. But if the operating model begins simulating future recruitment independently of this historical discrepancy, it is likely that one could distinguish between high and low steepness operating models with very few observations. In some cases, the operating models attempt to account for this fact by introducing a lag(1) auto-correlation (eg CPUE), but we really need to make sure this is done correctly for all data (including the age structure).

In the development of the ASCURE decision rule, we attempted to identify if unrealistically strong data signals were coming out of the data that could be exploited effectively by an MP. We think that this effect is evident in the operating model results, as illustrated in Figure 2. In this figure, we have plotted the results of a regression in which the SBT depletion level in 2002 was predicted using a linear regression on the operating model generated value for CPUE for ages 10+ in 2003. Shown in Figure 2 is the relationship between the observed and predicted depletion levels from this regression. Clearly, the one year of CPUE values was highly informative (R-squared = 0.6). The high and low depletion levels correspond to the different mortality assumptions (M05 vs M10 and M15). It seems highly implausible that we would actually be able to determine the current state of depletion to the degree indicated above, given only a single observation of age 10+ CPUE in 2003. We also found that there was a weaker relationship for predicting MSY from early CPUE but it was still substantial. Ultimately, a useful feedback control decision rule will depend on the new data for discriminating stock dynamics, but we will get a false sense of security if unrealistic information is available (even if it is not explicitly recognized, MPs may take advantage of it). It is important in specifying the final set of trials that we avoid providing unrealistically informative data to the candidate MP's.

Additional Conditioning Results

Polacheck and Kolody (2003a) presented a large number of results of the conditioning process with alternative parameter specifications for the operating model developed for the first stage testing. Given that no substantive changes were incorporated into the structure of the operating model as a result of the Second Management Procedure Workshop, only limited additional exploration of the behaviour of the operating model was undertaken. Results are presented here for differential weighting of the early size data and for runs in which the number of years in which selectivity changes were allowed was varied and/or the penalty given to temporal changes in selectivity was allowed to vary. Note results are only presented for the central natural mortality vector (vector 2).

Changing Weights on Early Size Data.

Figures 3 and 4 present estimates for the SSB and recruitment when the early size data (i.e. pre 1965) were further down weighted by either a factor of 0.1 or 0.5 relative to the default values. The primary effect was to smooth out some of the estimates of recruitment prior to the start of the fishery, higher estimates of the initial spawning biomass (up to ~20%) and lower estimates of current spawning biomass (up to 17%). This changed the absolute estimates of current depletion levels (B_{2002}/B_0) only slightly (by up to ~3.5%). The largest changes in the estimates of current biomass were seen with high steepness values. The values of the objective function (Table 1) in this case are not really comparable. However, as might be expected, the down weighting of the early size data always results in a better fit to the stock and recruitment function (i.e. suggesting that these early data are not compatible with an underlying stationary Beverton and Holt relationship). The magnitude of the estimated changes in the current spawning biomass suggest that the performance of management procedures (particularly in terms of catch levels relative to current ones) may be sensitive to the weight given to these early data.

Changing the penalty on temporal change in selectivity

Seven alternative options for the penalty function for changes in selectivity over time were explored. These are described in Table 2. Trends in the estimates of SSB and recruitment were somewhat sensitive to these different options (Figures 5 and 6). The conditioning results appear the most sensitive to allowing increased number of selectivity changes particularly in the early longline fishery on the spawning ground and the sensitivity is greatest for higher values of steepness (Tables 2 to 4). Thus, allowing for three selectivity changes (one every 5 years – options 6 and 7) results in a decrease in the estimate of current spawning biomass of 20% for a steepness value of 0.8. As would be expected, allowing for these changes results in a substantial decrease in the value of the objective function (by up to ~100 – Table 5). Also, as might be expected the main improvements are in the fits to the corresponding size data and in the stock-recruitment function. If the objective function was treated in a likelihood context, this decrease would not be considered significant. Thus, for the options 6 and 7 there are 93 additional parameters. However, these parameters are not independent and similar changes could probably be achieved with a more concise parameterization. Overall, the results suggest that the conditioning of the operating model is sensitive to the degree of temporal consistency in separability that is assumed, particularly in the early years of the longling fishery.

Selectivity Relationships

Figures 7 and 8 display the typical selectivity functions estimated by the operating model for the main longline fishery targeting SBT (i.e. fishery 1 (longline fisheries on the feeding grounds) and the two longline fisheries on the SBT spawning grounds - i.e. fishery 4 (Japanese) and fishery 5 (Indonesian). In all cases, the selectivity function is estimated to be highly domed both on the feeding grounds and spawning grounds, particularly for the first 30-40 years of the fishery. Also, the estimated selectivity functions have shifted towards older ages in the 1990's. Particularly for the older ages, these patterns raise several questions:

1. What biologically could account for the large differences in vulnerability among older age classes when given the relatively small differences in size (e.g. more than a factor of 10 in selectivity between ages 14 and 17 with less then average of 9 cm difference in length)?
2. What factors either in the stock or the fishery could have lead to the large changes in selectivity on the feeding grounds in the Japanese longline fishery (e.g. why in the 1990's have fish around age 15 become the most vulnerable?)
3. What factors either in the stock or the fisheries account for the large differences in the selectivity functions for Japanese and Indonesian longline fishery on the spawning grounds (For example, in the Japanese fishery, age 13 fish are estimated to be the most vulnerable while in the Indonesian fishery they are relatively invulnerable (less

then 10% of age 23). In contrast, fish older then age 18 are estimated to essentially have completed escaped the Japanese fishery while ages 18-23 are estimated to be the most vulnerable in the Indonesian case).

One interpretation of the estimated selectivity patterns is that historically there was a large “cryptic” component or refuge for older age fish that has subsequently been collapsing. Alternatively, another interpretation is that the model is being allowed to “over-interpret” the information on age of older animals contained in the size data, particularly given the uncertainties about the growth curve (especially for cohorts born prior to 1960), the representativeness of the sampling and measurement errors (e.g. conversion of weights to lengths and length measurements by non-scientific staff- crew). Even if the length data for larger animals are not given much weight in the objective function, there is very little else (if anything) within the objective function to constrain the selectivity functions from having steep domed shapes for older ages. It is perhaps worth noting that these questions about the plausibility of selectivity patterns for older ages are analogous to the problems encountered in the estimation of the plus group in past VPA assessments based on cohort slicing (e.g. Polacheck et al 1996, 1998).

It appears that the conditioning results may be highly sensitive to alternative, less domed shaped selectivity patterns. Thus, reducing the maximum age in the Indonesian fishery for which a selectivity term is estimated to 15 (i.e. for all ages 15 and older selectivities are assumed to be equal) resulted in substantial differences in the resulting estimates of SSB and recruitment (Figure 9 and 10). This resulted in a flat selectivity for older ages (Figure 11), but also in a substantially worse fit in terms of the objective function (Table 6). In all cases, some of the other components of the objective function besides the fit to the Indonesian longline fishery were substantially worse fit. In particular, for low steepness, the fit to the stock recruitment curve was very poor, while at high steepness the fit to longline fishery 1 size data as well as to the stock recruitment account for about half of the difference in the value of the objective function. It is interesting to note that the relative quality of fit to the SR function is reversed among scenarios, such that high steepness is favoured when Indonesian selectivity is constant for ages 15+.

In terms of the operating model and its conditioning for use in testing of management procedures, we question whether a wider range of hypotheses/model structures need to be considered that would result in what might be arguably considered to be more plausible selectivity patterns? Such hypotheses are not easily explored within the current operating model structure[1] and it seems reasonably clear that to impose such hypotheses/structures would results in degradation to the fit to estimated catch at length distributions for the longline fisheries.

Stock Recruitment Relationship

Currently the CCSBT MP operating model uses a Beverton and Holt stock recruitment relationship and is conditioned assuming this model is true and that the SBT stock was at

---

[1] Thus, in Polacheck (2003), changing the smoothness penalty for selectivities from 3rd to 2nd differencing had minimal effects on the results and the estimated selectivities were still estimated to be highly domed. Similarly attempts to impose a more flat selectivity pattern on older ages in the Japanese longline fishery within the current operating model structure by lowering maximum age for which a selectivity was being estimated only increased the sharpness of the dome – e.g. given the current weight being given to the size frequency data for larger fish, the combination of the data and model structure appear to only allow for highly domed selectivity patterns.

equilibrium at the beginning of the conditioning process (1931). In the initial and second stage trials, a range of values were adopted for the steepness parameter in order to capture the uncertainty related to the productivity or relative amount of compensation in the SBT stock. Although the objective function used in conditioning the operating model tends to favour low steepness, there was a general agreement that the objective function should not be taken as a literal likelihood function and as such did not provide a reliable measure of the relative plausibility for the different steepness values. Further, it is well recognized the SBT “one-way trip” historical SBT historical trajectory (i.e. continuously declining parental biomass) does not provide a highly informative time series for estimating productivity levels. Nevertheless, comparisons from the conditioning process of the actual estimates of the spawning stock biomass and recruitment to the estimated Beverton and Holt stock recruitment curves indicate a very poor fit for the higher steepness values. This raises the question of whether high levels of productivity for SBT are necessarily incompatible with the historic data for SBT or whether a problem may exist with the functional form of the stock-recruitment curve. This question is addressed further in this section.

Figure 12 compares the estimates of SSB and recruitment for the period from 1965-1997 from the conditioning process for the three different fixed values of steepness used in the second stage trials for the natural mortality vector 2. The years 1965-1997 are displayed because these are the years for which there is actually some reasonable level of support in the data for the estimates (e.g. the CPUE tuning indices begin in 1969; prior to 1965 non-feeding ground catches dominated and the catch at length data prior to 1965 is uncertain). What is clear in Figure 12 is that as the value of steepness is changed the overall shape of the SSB/recruitment relationship is similar and the relative pattern of the individual estimates remains largely unchanged. Thus, if only the points in Figure 12 are considered, they would suggest that the stock and recruitment relationship has been essentially linear over most of the history of the fishery. As the imposed steepness is increased, the slope of the apparent underlying linear relationship increases (Figure 13). However, there is no indication of increasing curvature or concavity in the relationship. Essentially, the pattern of recruitment and SBB estimates is fixed by the data and there is only uncertainty about the scale or their absolute magnitude (and relatively little about the scale for recruitment for a given mortality vector). When a Beverton and Holt stock-recruitment curve is imposed on these estimates, the best fit will be one that has low steepness (i.e. essentially linear). Thus, as can be seen in Figure 14, the higher the steepness, the poorer a predictor the Beverton and Holt curve provides for the stock and recruitment estimates over years with the most support in the data.

Table 7 presents the values for the objective function and its various sub-components for the three different fixed steepness values for natural mortality vector 2[2].As has been noted previously, there are not very large differences in the fit to data components of the objective function over the range of steepness values. In fact the fits except for CPUE and for longline fishery 2 catch-at-size tend to improve with increasing steepness. However, the penalty for the stock recruitment function (*Sg R*) increases at a faster rate resulting in an overall best fit for the lower steepness values. This suggests that reasonably high levels of productivity (i.e. the extent to which recruitments have been above the replacement line) are actually compatible with the historical data but not when expressed in terms of a Beverton and Holt stock-recruitment. Because of the strong linear relationship between the estimates of SSB and

---

[2] Note that values in this table differ slightly from those provide in the notes distributed with SBTProj4 software for H55 scenario. The reason for this is not clear.

recruitment, a Beverton and Holt model with high steepness results in a poor fit, with implications of high auto-correlation or regime shifts.
The other feature of potential concern in the fitted stock and recruitment estimates relates to the historical variability observed prior to 1969 (the year that CPUE is considered reliable). The conditioned model suggests that there was a period of very poor recruitment 5-7 years before the start of the fishery, followed by very favourable recruitments that lasted for ~10-20 years (Fig. 15). Recruitments for 10 years preceding the fishery were about 50% below R_0, and recruitments for the first 10 years of the fishery (1951-1960) are estimated to have been 37 to 73% above R_0. The first 20 years of the fishery are about 13 to 50% above R_0 across the range of steepness values considered. In essence, the conditioning process is saying that what sustained the stock through most of the history of exploitation were these large recruitments, and not the underlying surplus production in the stock. However, the early data are highly uncertain: we do not know very much about the catch-at-size sampling procedures in the Japanese longline fishery prior to 1965 and the actual catch size/age composition might have changed substantially during this developmental period. We are highly uncertain about this period, but any attempt to downweight this early data (short of removing it) has a rather limited effect, because there is no other data to influence the early dynamics (except the stock-recruitment curve), and the model might be substantially wrong (if selectivity was highly variable or the actual length-at-age relationship was different (for density dependent reasons). We would suggest that these arguments are sufficient for defending a stock-recruitment relationship scenario in which the recruits near the beginning of the fishery are constrained to near equilibrium levels. When combined with the earlier considerations about linearity, this would presumably allow us to explore internally consistent stationary stock-recruitment scenarios that are equally plausible at high and low steepness levels. However, we would also argue from observations of the existing stock and recruitment relationship fitting, in both the earliest and recent years, that some form of time series structure should also be recognized (and the current lag(1) auto-correlation may not be sufficient if it provides too much information to the MPs).

All of the above would suggest that there is substantially more uncertainty about the functional form of the past and thus future stock recruitment relationship for SBT than is currently represented in the operating model. The current Beverton and Holt stock recruitment relationship with high values of steepness is a parsimonious model with respect to the historical data. In order for these relationships to be plausible requires a high degree of autocorrelation in the error structure (and a lag(1) auto-correlation is probably not a sufficient description of this process) or alternatively there is an element of non-stationarity in the process. In these cases, the past observations do not provide a good predictor for the future dynamics. However, while such hypotheses may be worth considering, it would seem a higher priority to ensure that the simpler hypotheses of high productivity (without non-stationarity or high auto-correlation) are invoked in a manner consistent with the historical data .

It should be noted that by considering a range of values for steepness within the Beverton and Holt formulation that one is likely to cover a similar range of future recruitment scenarios that might be seen if alternative functional forms for the stock-recruitment relationship were used (although their relative frequency in the trials may differ substantially). As such, this may reduce the imperative for actually considering alternative forms. Nevertheless, there would be at least three reasons why some alternative forms may be worth including in the final set of trials:

1. The set of projected future recruitment estimates are more likely to be consistent with past trends (i.e. unless the error structure is captured appropriately, there is likely to be a discontinuity (e.g. "jump-up") in the recruitment estimates early in the projections with high steepness scenarios). This will not only be of dubious plausibility, but also potentially provides unrealistically good information with which MPs can distinguish productivity in the very near future (see above - section Information Content in Fishery Data Generated by the Operating Model).
2. Performance of management procedures with respect to performance indicators related to MSY will be different.
3. The assignment of weights to scenarios with higher productivity would be more straightforward (e.g. there would be less of a need to decide how to weight the contradictions between the lack of fit to the stock/recruitment function at higher steepness against the better fit to the actual data in these cases).

Synthesis of Results across Scenarios

There are a number of approaches that could be utilized for synthesizing results from different management procedures across a range of operating model scenarios and this question was discussed extensively at the Second Management Procedure Workshop (Anon. 2003, Polacheck and Kolody 2003b). One potential approach is to adopt an MCMC approach to integrate over parameter uncertainty. However, it was decided not to adopt a full MCMC approach for a number of reasons including the fact that "the likelihood is not considered to be reliable" (Anon. 2003). It was recognized that future meetings would need to decide on process for assigning weights and specific proposals were encouraged to be developed and submitted to the 2003 SAG meeting.

At the Second Management Procedure Workshop a two phase approach was suggested based on a specific elaboration of the "hybrid" approach presented in Polacheck and Kolody (2003b). The general idea behind the hybrid approach is to first eliminate management procedures that do not provide robust/acceptable performance in terms of stock status/conservation objectives and in the second step to choose the MP among those that remain that optimize performance. While there was general agreement on the merits of such a two phase approach, the approach leaves unresolved both the question of defining robust/acceptable performance and the question of what constitutes "optimal" performance among those MP that passed the first phase. In addition, resolution of these questions contains between the two issues. We recognize that managers will ultimately select the MP that best meets their shared objectives, but the range of MPs developed and presented by the scientists will provide a better range of options if the range of acceptable goals can be reasonably constrained at this stage. Without such constraints, it is not clear how meaningful options could be presented.

Within any management procedure, there will be a trade-off in performance relative to future stock status and catch optimization objectives. All decision rules being considered contain tuning (or control) parameters that when varied allow a continuum of management procedures in terms of performance along this trade-off axis. Thus, if highly restrictive criteria (in terms of management conservation objectives) are used to define robust/acceptable performance in the first stage, the second phase of optimization can focus primarily on the comparison of performance in terms of harvesting objectives. However, to the extent that the criteria used in the first stage constitute lower bounds of acceptable performance, optimization in the second phase will need to define "optimal" in terms of the trade-off between performance relative to conservation-related and catch-related performance measures (e.g. how much better performance on catch should be traded off against worse performance on conservation-related objectives). Defining and applying such a multi-dimensional optimality criteria to come up with the "optimal" procedure is difficult particularly when the probability/weights associated with alternative scenarios that are to be integrated across are not well defined and have a large subjective component. Experience in other fora (e.g. IWC) suggest that a more practical and viable approach may be to first "optimize" with respect to stock-status-related performance measures and secondarily optimize with respect to catch-related performance measures.

The following is a suggestion for a framework for consideration that might be utilized by the CCSBT Scientific Committee for synthesizing performance across a range of scenarios. This framework was developed based on recent experience within the IWC Scientific Committee. This Committee was faced with a similar problem of selecting a variant of its basic management procedure based on the results from a large number of scenarios (>200) covering a wide range of different and disparate hypotheses for the underlying dynamics and stock structure for the resource being considered. It should be emphasized that this framework is put forward as a draft for discussion. If considered useful as a general way forward, it will likely require considerable refinement and elaboration based on discussion within the Scientific Committee/SAG and experience gained in actually trying to implement it.

1. List the major axes of uncertainty that are represented in the different scenarios (i.e. steepness, mortality, CPUE/abundance relationships, etc).
2. Within each major axes (or dimension) of uncertainty define the explicit set of hypotheses to be tested (e.g. steepness values of 0.3, 0.55, 0.8, etc) Note that all hypotheses within a given dimension need not be crossed with all hypotheses from other uncertainty dimensions.
3. Based on the results of robustness tests eliminate hypotheses that "pass tick test"[3] (i.e. show no substantial difference in performance across the range of decision rules being considered[4]).
4. For each hypothesis within an uncertainty dimension (e.g. each mortality vector), rank it as to high, medium or low. Ranking of high, medium or low are meant to represent the relative plausibility that should be given to the different hypotheses. In this sense, a specific hypothesis needs to be considered within the context of the other hypotheses being considered and the extent of the overall uncertainty space that the hypotheses is meant to represent[5]. If there is substantive and unresolvable disagreement within the Scientific Committee regarding the ranking for a hypothesis, that hypothesis would be given an "unresolved" ranking. Within any uncertainty dimension, at least one hypothesis needs to be given a high ranking (or in the case where all hypotheses within an uncertainty dimension are given an unresolvable ranking, all hypotheses should be treated as having a high ranking).

---

[3] Note this step is not essential, but is intended to reduce the number of scenarios that need to be considered in the remaining steps. It is intended that whether this step is undertaken or not would not change the overall outcome.

[4] No substantial difference will need to be defined – e.g. less than a 10% difference in the mean value for relevant performance indicators or 20% difference in the CV.

[5] However, rankings are not meant as absolute probabilities and all hypotheses can be given an equal weighting. Ranking would involve consideration of the relative plausibility of an hypothesis based both on fits to the observed data in the conditioning process and on general considerations relative to SBT and other fish stocks.

5. For any scenario (e.g. H55M10Q0), give it an overall ranking based on its lowest ranking within any uncertainty dimension (i.e. consider the ranking for H=0.55, M=0.10 and Q=0). For a scenario in which one hypothesis within one or more of the uncertainty dimension was ranked as “unresolvable”, these ranking would be treated as medium (unless within a dimension all of the hypotheses were ranked as unresolvable – in which case it would be treated as a high ranking as indicated in step 4).
6. Consider any management procedure as acceptable if it passes, for example, the following **(Note the highlighted percentages are illustrative and figures would need to be agreed on in consultation with management):**
   - For scenarios ranked as high, that there is a 90% probability that the spawning stock in 2022 is 40% greater than it currently is or the spawning stock in 2020 is greater then the 1980 level[6].
   - For scenarios ranked as medium, that there is a 90% probability that the spawning stock in 2022 is 10% greater than it currently is.
   - For scenarios ranked as low, that there is a 90% probability that the spawning stock in 2022 is greater than 75% of its current level.
7. Consider any management procedure as acceptable if it AV statistic is less than xx or another statistic relating to the stability of the TAC.
8. From among those management procedures which are considered acceptable after step 7, rank them based on their mean catch among those scenarios ranked high.
9. Select the scenario with the highest mean catch ranking in step 8, unless there is agreement based on examination of the more detailed performance statistics (including the worm plots) that one of the lower rank procedures may be preferable.

The above framework implies that the conservation objectives (i.e. as defined in step 6) are viewed first, and catch objectives are viewed thereafter for MPs which pass the conservation objectives. This approach would not present managers with options in terms of the stock status/catch trade off. It might be worth considering whether step 6 above could be turned into a one dimensional criterion (e.g. by considering only the percentage values for the high rank scenarios). In this case, one could then express and present the overall results of a range of MPs in terms of a stock status/catch trade off (e.g. mean catch versus the percentage by which the spawning stock in 2022 is greater than it currently is, with a 90% probability).

The above 9 step approach attempts to capture a number of features of the management procedure development and evaluation process including the following:

1. A basic recognition that all scenarios are not equally likely, combined with a recognition that all assignment of absolute quantitative probabilities to different scenarios would be difficult and unproductive. In contrast, rankings into three general categories is more likely to achieve consensus.
2. That it is difficult to represent the entire uncertainty space within any uncertainty dimension and that is not feasible to undertake a full cross of all hypotheses across all uncertainty dimensions.
3. That mean or average performance is not an appropriate measure of “optimality” in terms of conservation-related objectives. That “optimality” in this context needs to take into account the range of potential outcomes and may be considered in terms of

---

[6] An alternative criteria that could be considered might be to consider the rebuilding level relatve to the maximum possible (i.e. under a zero catch). In such case, the criteria might be to have a 90% probability that the spawning stock in 2022 is above 40% of the level it would be under a zero catch scenario.

the risk/likelihood of achieving or not achieving management objectives as represented by specific performance measures.

4. Conservation related statistics can exhibit bi-modal distributions, in such situations an average or weighted average provides little guidance on actual expected outcomes.
5. That industry has placed a high weight on stability of catches and minimization of inter-annual fluctuations.
6. The approach is transparent and straightforward to apply.

It should be emphasized that the above approach is presented as one possible framework and, if found useful, the details in particular would need careful consideration. However, it is critical at this stage to develop an agreed on approach for the synthesis process. Without an agreed on approach, it becomes impossible to make further progress on refining the suite of potential management procedures given that the tuning parameters within any decision rule allow any decision rule to obtain a wide range of trade-offs in performance along the catch versus stock status axes.

Literature Cited

Anon. 2001. Report of the Eight Annual Meeting. CCSBT. 15-19 October. Miyako, Japan.

Anon. 2002. Report of the third Stock Assessment Group Meeting. 3-7 September, 2003. Canberra, Australia. CCSBT.

Anon. 2003 Report of the Second management procedure workshop. CCSBT. 7-9, 12, 14-15 March. Queenstown, New Zealand.

Eveson. P. 2003. An update of the graphics used for evaluating the performance of candidate management procedures for southern bluefin tuna. CCSBT-ESC/0309/23.

Polacheck, T. and D.Kolody. 2003a. The Behaviour and Fit of Alternative Operating Model Specification for Testing the Performance of Southern Bluefin Tuna Candidate Management Procedures. CCSBT-MP/0304/7

Polacheck, T. and D. Kolody. 2003b. Synthesising Performance of Candidate Management Procedures Across Different Operating Model Scenarios. CCSBT-MP/0304/7.

Polacheck, T., D. Ricard, P. Eveson, M. Basson and D. Kolody. 2003a. Results from Initial Testing of Some Candidate Management Procedures for Southern Bluefin. . CCSBT-MP/0304/6.

Polacheck, T. D. Ricard, P. Eveson, M. Basson, D. Kolody and J. Hartog. 2003b. Results from further testing of candidate management procedures for southern bluefin tuna. CCSBT-ESC/0309/29.

Polacheck, T, A. Preece and N. Klaer. 1998a Assessment of the Status of the Southern Bluefin Tuna Stock Using Virtual Population Analyses – 1998. CCSBT/9807/17.

Polacheck, T., A. Preece, A. Betlehem and K. Sainsbury. 1996. Assessment of the status of the southern bluefin Tuna Stock using virtual population analysis - 1996. CCSBT/SC/96/26.

Table 1: Minimum value for the best fit to the objective function and its various components when the effective sample size for the early length frequency data (i.e. pre 1965) is decreased relative to the default.

Steepness	Relative SS	LL1	LL2	LL3	LL4	IND	SURF	CPUE	Tags	Sel.Ch	Sel.sm	Sg.R	Total
0.30	0.1	214.59	49.72	54.32	136.27	40.67	100.21	-46.65	12.58	28.12	53.92	-48.39	595.36
	0.5	236.05	49.75	79.00	165.80	40.52	100.42	-47.17	12.68	32.62	55.30	-40.96	684.02
	1.0	256.09	49.76	104.14	192.85	40.42	100.58	-47.35	12.73	37.85	57.19	-36.14	768.13
0.55	0.1	215.12	49.94	54.00	136.58	40.25	99.50	-43.12	11.58	28.42	53.83	-46.04	600.05
	0.5	235.88	50.00	77.76	165.62	39.89	99.68	-43.84	11.66	32.45	55.18	-36.70	687.57
	1.0	255.53	50.03	102.58	191.54	39.74	99.72	-44.24	11.76	37.60	55.92	-29.14	771.05
0.80	0.1	215.95	50.25	53.74	136.63	40.70	98.84	-41.98	11.23	28.69	53.83	-39.63	608.26
	0.5	236.23	50.29	76.94	164.66	39.84	99.05	-42.50	11.39	32.56	55.18	-28.95	694.70
	1.0	254.76	50.13	100.86	187.52	39.57	98.95	-43.08	11.63	37.88	57.02	-19.58	775.65

Table 2: Description of the 8 options considered for the penalty function for the temporal changes in selectivity.

Option	*Description*
1	Default
2	Decrease penalty assigned to all selectivity changes by a factor of 2
3	Decrease penalty assigned to fishery 1 selectivity changes by a factor of 4
4	Allow for selectivity changes every 2 years in fishery 1
5	Allow for selectivity changes every 2 years in fishery 1 and decrease penalty by a factor of 4
6	Allow for selectivity changes in fisheries 3 and 4 in 1960, 1965 and 1970
7	Allow for selectivity changes in fisheries 3 and 4 in 1958, 1963 and 1968
8	Allow for selectivity changes every 2 years in fishery 6

Table 3: Relative change in the estimates of current and initial spawning biomass relative to the default specification for the different options for the penalty function for temporal changes in selectivity in Table 2.

	Current SSB			Initial SSB		
Option	H=0.30	H=0.55	H=0.80	H=0.30	H=0.55	H=0.80
1	1.00	1.00	1.00	1.00	1.00	1.00
2	1.03	1.03	1.04	1.01	1.00	1.00
3	1.01	1.02	1.02	1.01	1.00	0.99
4	0.97	0.93	0.88	1.02	1.02	1.02
5	0.99	0.94	0.86	1.02	1.02	1.01
6	0.91	0.83	0.80	1.07	1.07	1.04
7	0.92	0.84	0.81	1.11	1.09	1.05
8	1.00	1.01	1.01	1.00	1.00	1.00

Table 4: Relative change in the estimated depletion level (B_{2002}/B_0) for the different options for the penalty function for temporal changes in selectivity in Table 2.

	Current SSB		
Option	H=0.30	H=0.55	H=0.80
1	0.17	0.16	0.18
2	0.17	0.17	0.19
3	0.17	0.17	0.19
4	0.16	0.15	0.16
5	0.16	0.15	0.15
6	0.14	0.13	0.14
7	0.14	0.13	0.14
8	0.17	0.16	0.18

Table 5: Minimum value for the best fit to the objective function and its various components for the different options for the penalty function for temporal changes in selectivity in Table 2.

Steepness	Option	LL1	LL2	LL3	LL4	IND	SURF	CPUE	Tags	Sel.Ch	Sel.sm	Sg.R	Total
0.30	1	256.09	49.76	104.14	192.85	40.42	100.58	-47.35	12.73	37.85	57.19	-36.14	768.13
	2	240.32	49.81	102.88	192.28	40.27	96.22	-46.99	13.71	21.23	55.03	-36.62	728.12
	3	233.46	49.70	106.11	192.19	40.27	99.99	-46.60	13.42	23.72	57.83	-37.43	732.65
	4	207.32	48.93	104.02	189.88	42.26	100.14	-58.19	8.94	45.24	62.50	-36.72	714.33
	5	188.91	48.27	104.64	187.38	42.73	99.76	-60.22	8.80	24.45	64.76	-37.70	671.80
	6	253.76	49.66	85.89	140.91	40.72	100.18	-45.70	12.49	40.36	62.56	-40.59	700.24
	7	254.00	49.68	84.06	143.89	40.77	100.21	-45.63	12.48	42.61	64.41	-46.20	700.29
	8	255.95	49.35	104.23	191.85	37.67	94.53	-47.31	12.27	40.58	58.38	-36.60	760.90
0.55	1	255.53	50.03	102.58	191.54	39.74	99.72	-44.24	11.76	37.60	55.92	-29.14	771.05
	2	240.19	50.06	100.91	190.57	39.37	95.14	-44.17	12.65	21.00	54.04	-28.38	731.39
	3	233.57	49.95	104.02	190.55	39.26	98.64	-44.09	12.36	23.72	57.00	-28.60	736.39
	4	207.76	49.42	102.12	188.71	41.16	99.11	-57.46	9.15	44.72	62.79	-32.06	715.42
	5	189.91	48.77	102.28	186.27	41.29	98.41	-59.83	9.03	24.51	65.32	-32.26	673.70
	6	253.33	49.86	85.19	139.96	40.57	99.47	-42.07	11.58	40.57	62.98	-39.67	701.77
	7	254.14	49.89	84.16	143.78	40.68	99.44	-42.08	11.55	42.54	64.46	-44.20	704.36
0.80	1	255.17	49.64	102.76	190.88	37.00	93.90	-44.15	11.37	40.18	57.01	-29.68	764.08
	2	254.76	50.13	100.86	187.52	39.57	98.95	-43.08	11.63	37.88	57.02	-19.58	775.65
	3	239.55	50.11	99.06	185.90	39.13	94.28	-43.33	12.54	21.18	54.98	-17.76	735.66
	4	232.90	49.97	102.06	185.24	39.04	97.63	-43.45	12.28	23.90	57.75	-16.69	740.63
	5	207.91	49.89	100.82	185.87	40.70	98.10	-57.39	9.46	45.04	62.91	-23.87	719.44
	6	190.50	49.18	100.68	183.05	40.66	97.15	-60.10	9.41	24.73	65.57	-23.51	677.31
	7	253.96	50.21	84.48	139.89	41.41	98.78	-41.23	11.18	41.05	63.24	-34.22	708.74
	8	254.25	49.74	101.01	186.84	36.81	93.35	-42.96	11.28	40.35	58.06	-19.94	768.79

Table 6: Minimum value for the best fit to the objective function and its various components when the maximum age in the Indonesian longline fishery for which selectivity is allowed to vary is reduced from 30 to 15 for various values of steepnes.

Steepness	Max ages	LL1	LL2	LL3	LL4	IND	SURF	CPUE	Tags	Sel.Ch	Sel.sm	Sg.R	Total
0.30	30	256.09	49.76	104.14	192.85	40.42	100.58	-47.35	12.73	37.85	57.19	-36.14	768.13
	15	260.31	48.86	99.24	184.99	68.42	98.89	-36.45	13.09	43.03	57.15	43.69	881.22
0.55	30	255.53	50.03	102.58	191.54	39.74	99.72	-44.24	11.76	37.60	55.92	-29.14	771.05
	15	265.96	49.18	99.12	190.09	65.54	98.60	-41.35	12.55	42.17	57.18	2.68	841.74
0.80	30	254.76	50.13	100.86	187.52	39.57	98.95	-43.08	11.63	37.88	57.02	-19.58	775.65
	15	265.59	49.58	99.23	189.67	63.74	98.23	-42.62	11.65	41.96	57.15	-7.54	826.64

Table 7: Minimum value for the best fit to the objective function and its various components for different values of steepness (natural mortality was assumed to equal vector 2 in these results).

Steepness	LL1	LL2	LL3	LL4	IND	SURF	CPUE	Tags	Sel.Ch	Sel.sm	Sg.R	Total
0.30	256.09	49.76	104.14	192.85	40.42	100.58	-47.35	12.73	37.85	57.19	-36.14	768.13
0.55	255.53	50.03	102.58	191.54	39.74	99.72	-44.24	11.76	37.60	55.92	-29.14	771.05
0.80	254.76	50.13	100.86	187.52	39.57	98.95	-43.08	11.63	37.88	57.02	-19.58	775.65

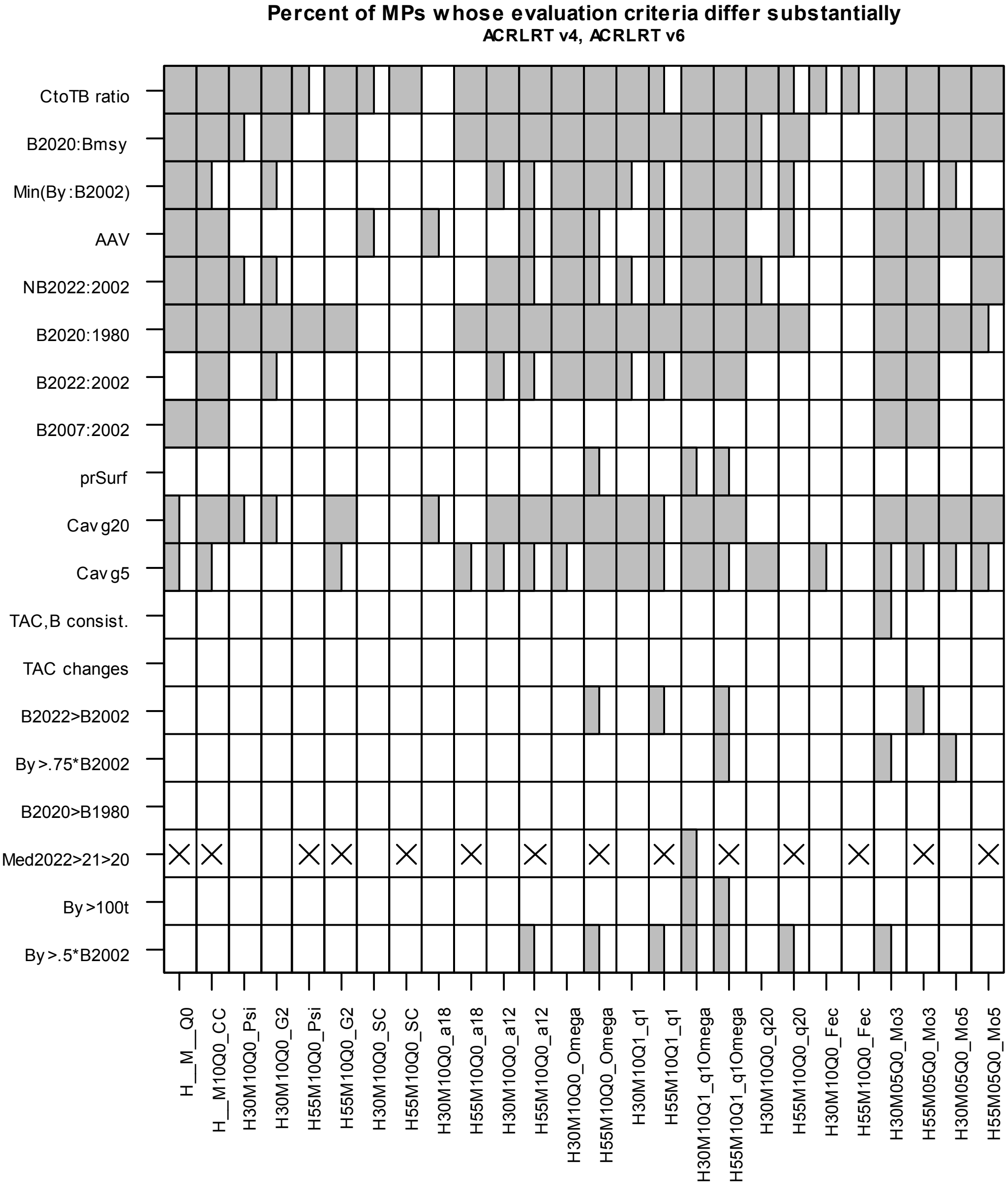


Figure 1: Illustration of typical results to robustness "tick test" trials. "Substantial" differences (defined in Polacheck et al 2003b) between the robustness operating model and the most similar reference case model are shaded. "X" indicates criteria is not applicable.

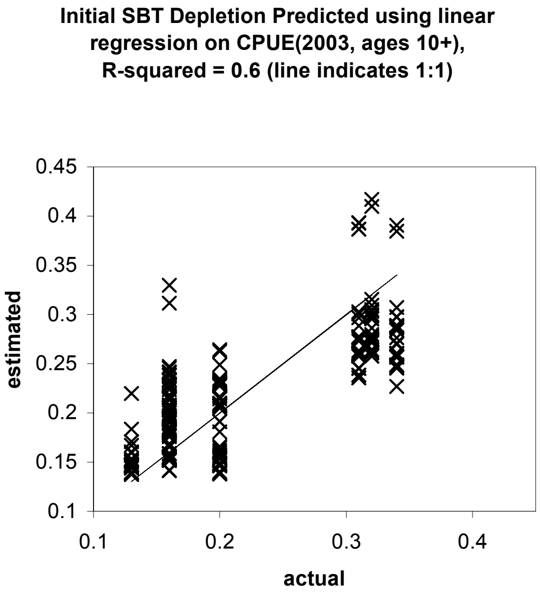

Initial SBT Depletion Predicted using linear regression on CPUE(2003, ages 10+), R-squared = 0.6 (line indicates 1:1)
estimated
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.1
0.2
0.3
0.4
actual

Figure 2:

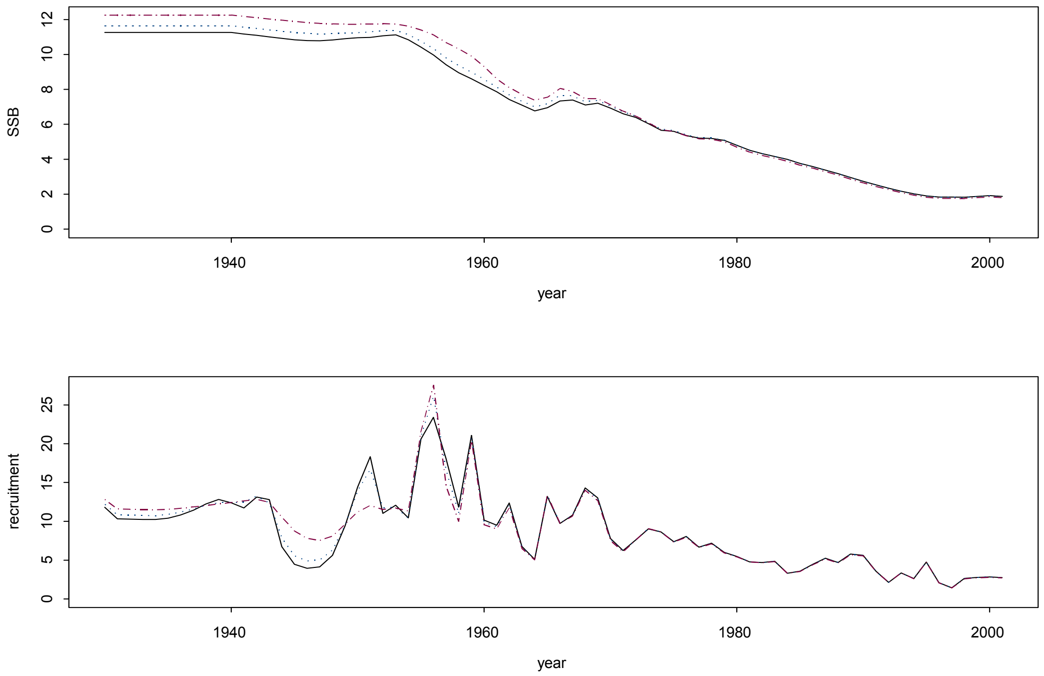


Figure 3: Comparison of the best fit estimates of temporal trends in SSB and recruitment for steepness equal to 030 for differen assumed relative sample size for the early (pre 1965) size data (solid line is the default, dotted line represents a decrease of 0,5 and the dashed line a decrease of 0.10).

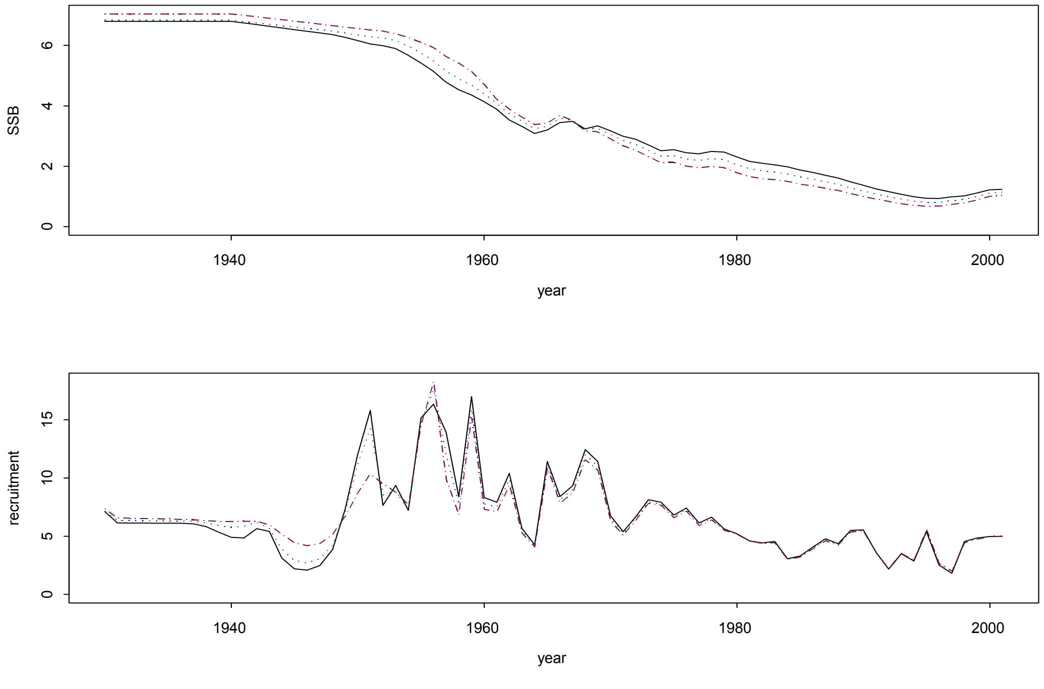


Figure 4: Comparison of the best fit estimates of temporal trends in SSB and recruitment for steepness equal to 080 for different assumed relative sample size for the early (pre 1965) size data (solid line is the default, dotted line represents a decrease of 0,5 and the dashed line a decrease of 0.10).

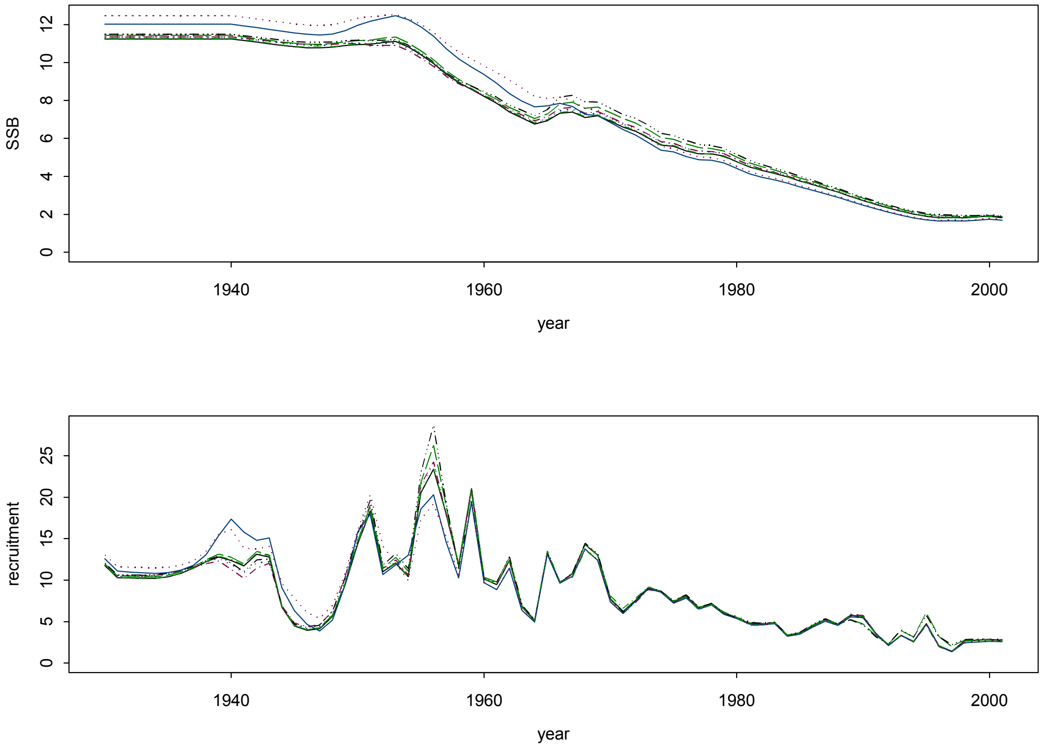


Figure 5: Comparison of the range of best fit estimates of temporal trends in SSB and recruitment for steepness equal to 0.30 for the different options in Table 2.

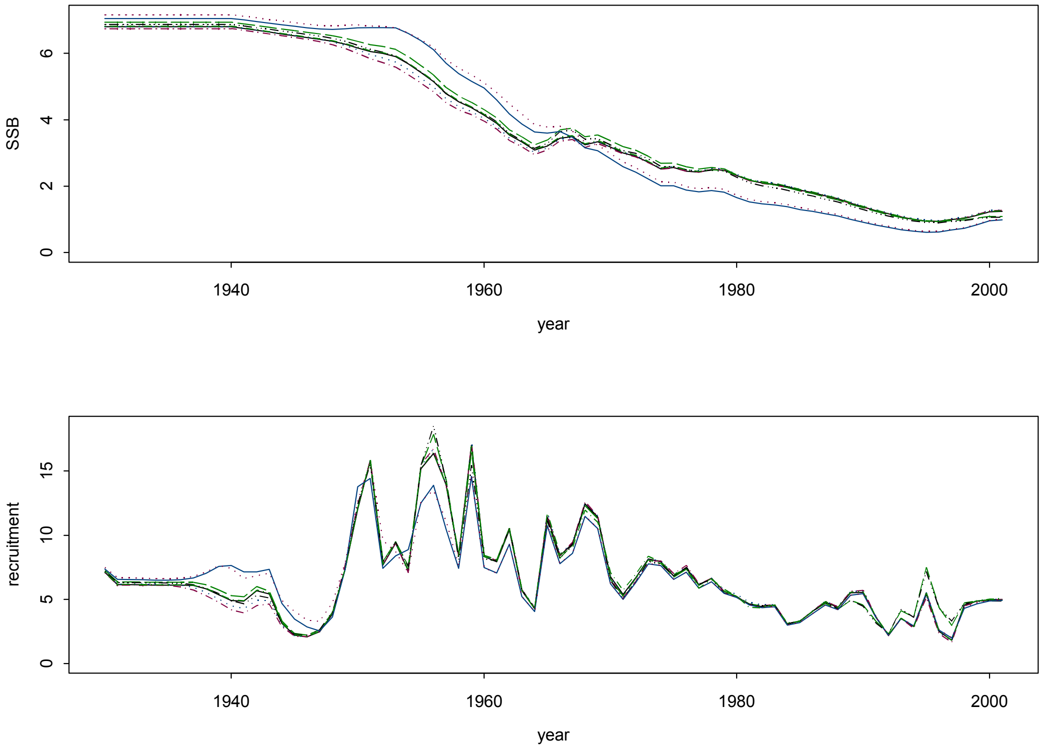


Figure 6: Comparison of the best fit estimates of temporal trends in SSB and recruitment for steepness equal to 0.80 for the different options in Table 2.

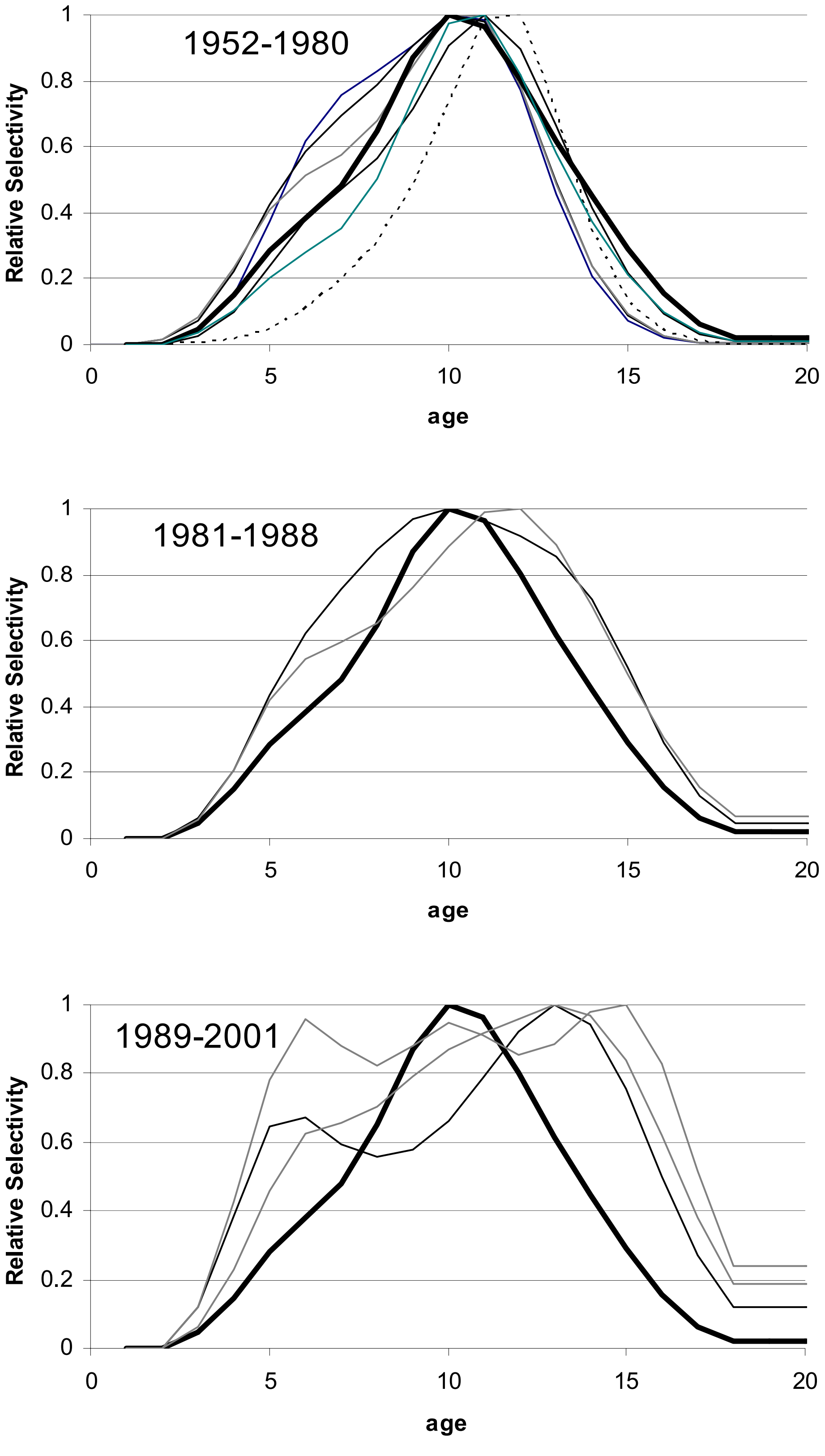


Figure 7: Comparison of the estimated relative selectivity function for the longline fishery on the feeding grounds. Results are shown for steepness 0.55 and natural mortality vector 2. All selectivities have been standardized to the maximum in a year. The solid line in each figure is the selectivity curve for the 1977-80 period.

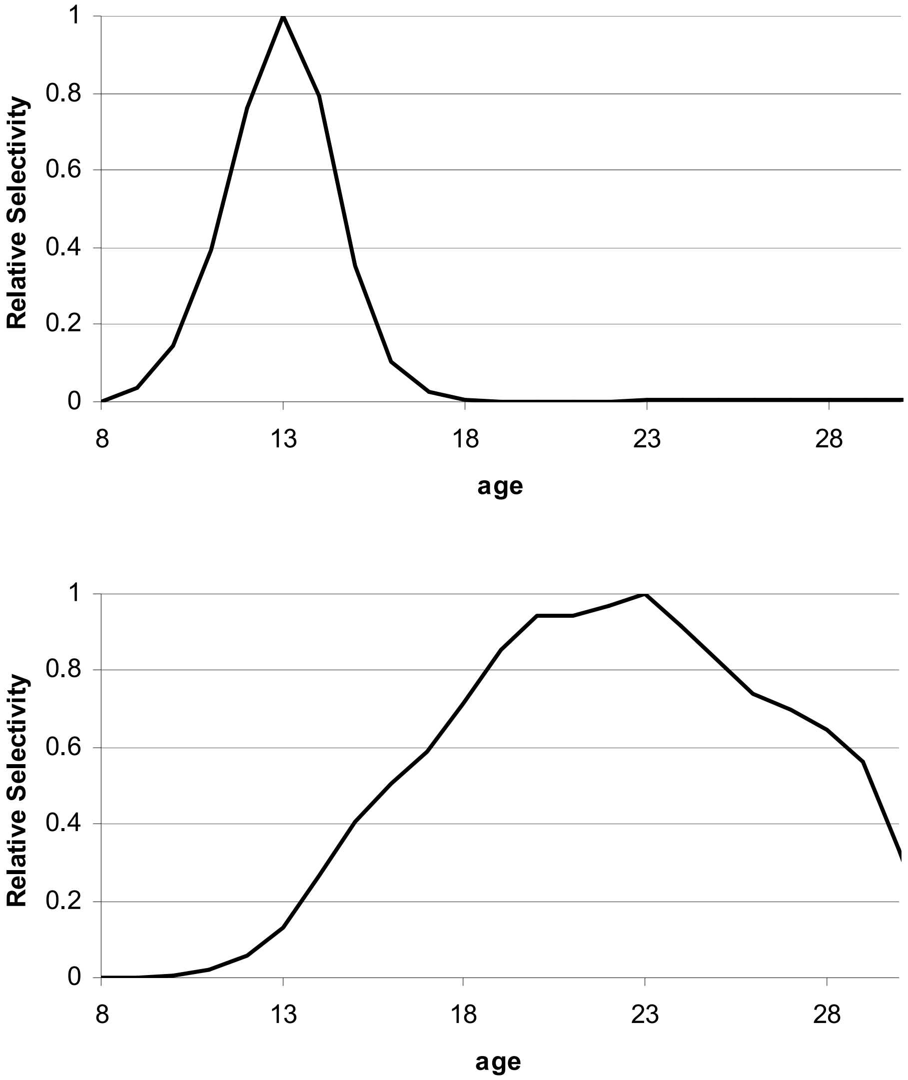


Figure 8: Comparison of the estimated relative selectivity function for the longline fishery on the spawning grounds. Upper panel is for the Japanese fishery (primarily pre 1970) and the lower panel is for the Indonesian fishery (primarily in the 1990's). Results are shown for steepness 0.55 and natural mortality vector 2. All selectivities have been standardized to the maximum in a year.

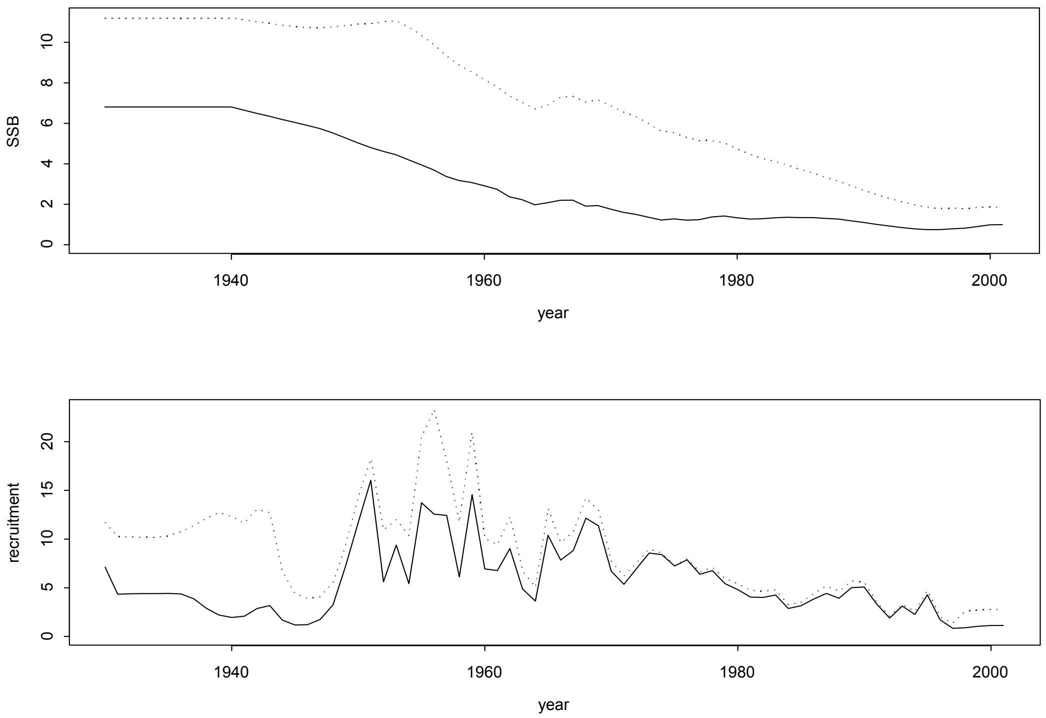


Figure 9: Comparison of the best fit estimates of temporal trends in SSB and recruitment for steepness equal to 0.30 two different values for the maximum age for age in the Indonesian longline fishery for which selectivity is allowed to vary (solid line is for age 30 and dotted line is for age 15).

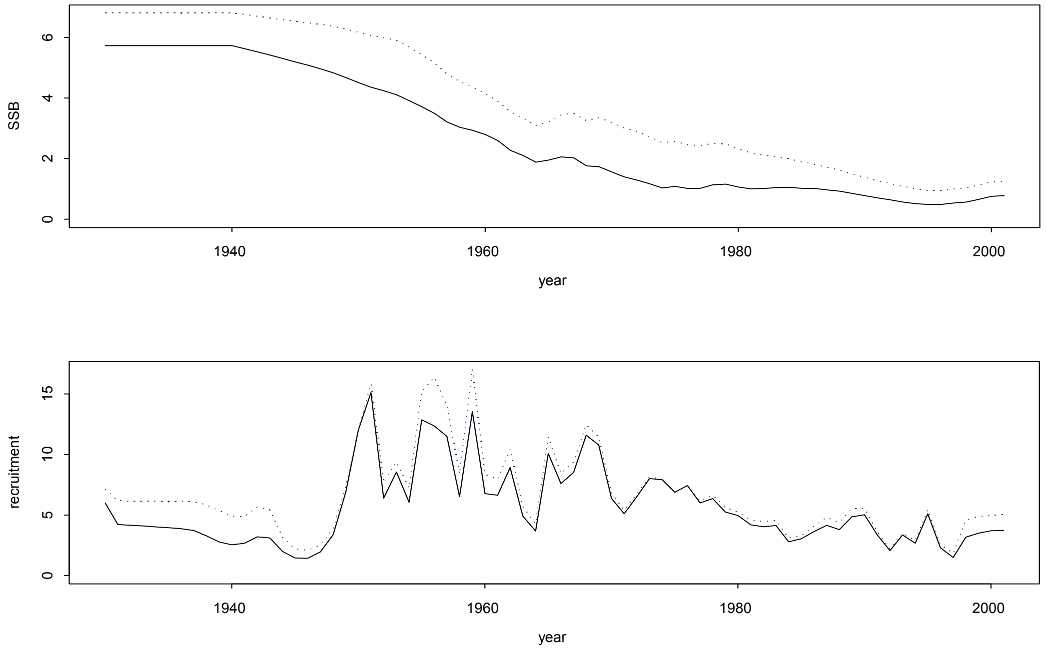


Figure 10: Comparison of the best fit estimates of temporal trends in SSB and recruitment for steepness equal to 0.80 two different values for the maximum age for age in the Indonesian longline fishery for which selectivity is allowed to vary (solid line is for age 30 and dotted line is for age 15).

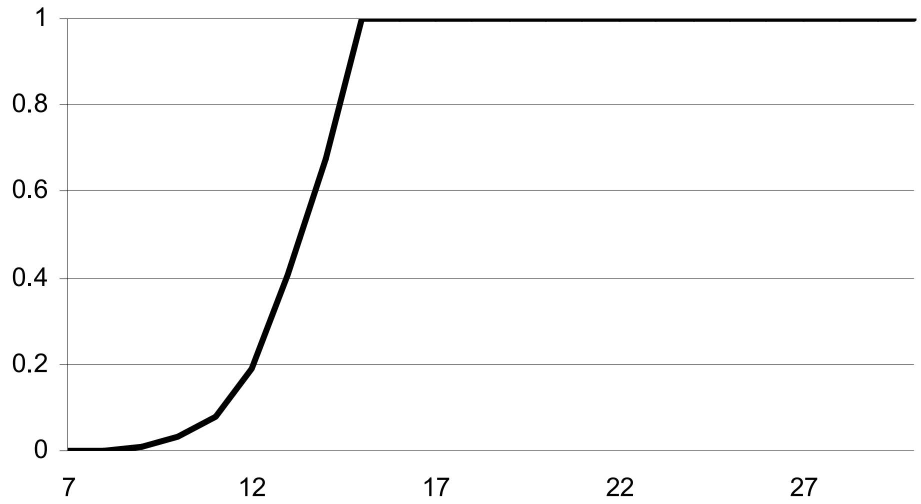


Figure 11: Example of the estimated selectivity curve for fishery 5 (the Indonesian longline fleet) when the maximum age for which selectivity is allowed to vary is set to 15. This figure should be contrasted with the lower panel in Figure 8. Results are shown for steepness 0.55 and natural mortality vector 2. All selectivities have been standardized to the maximum in a year.

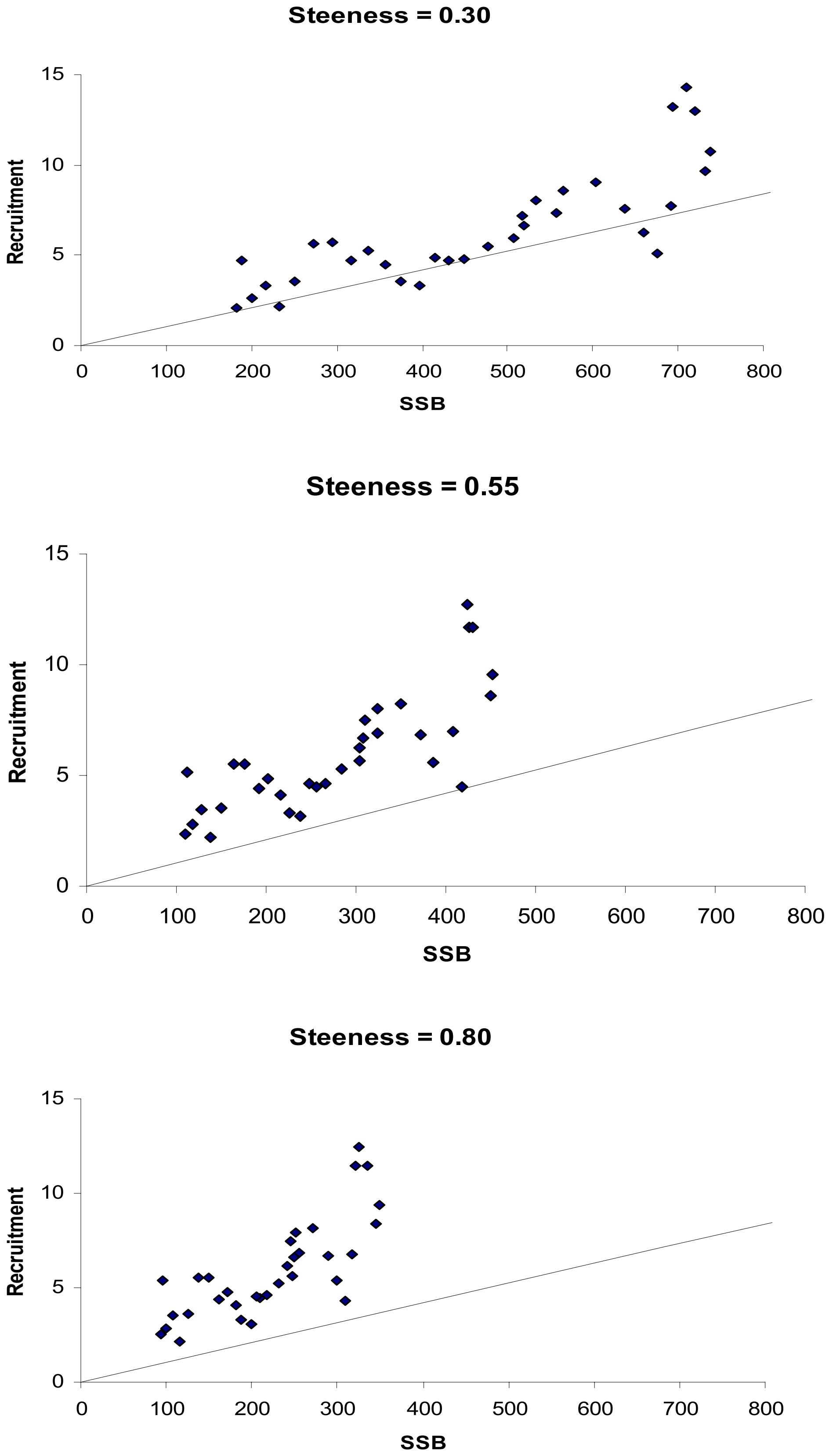
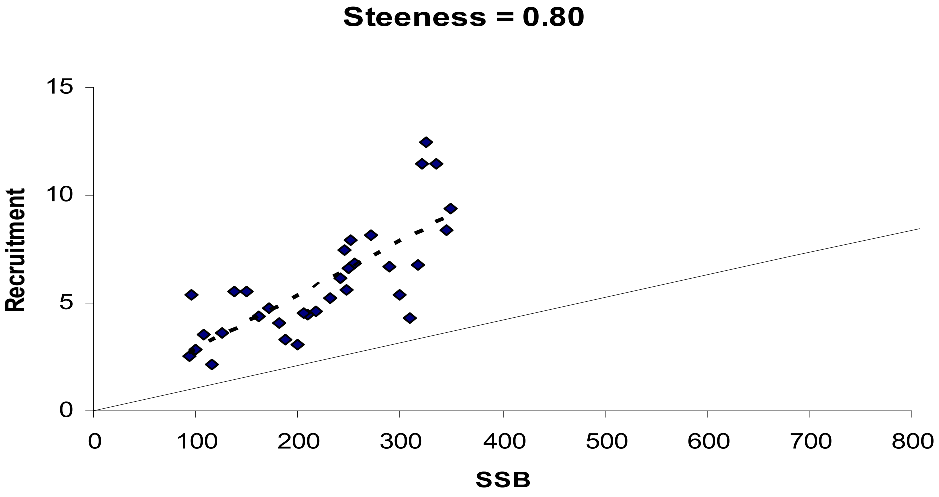
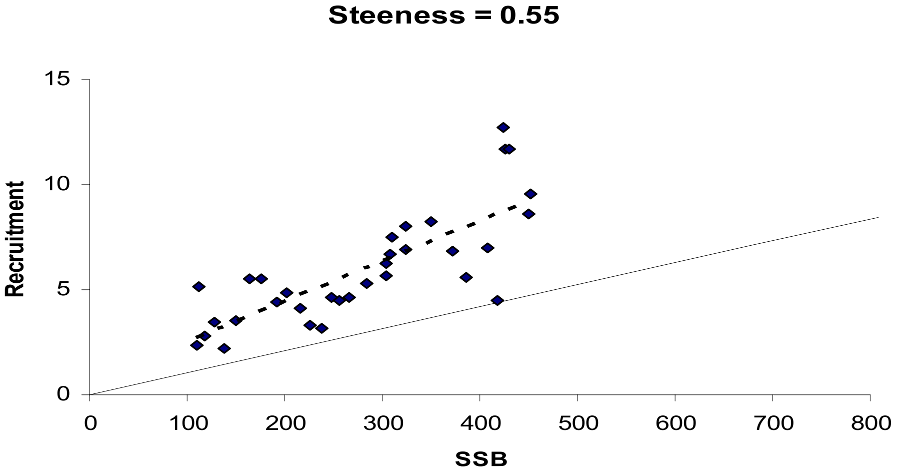


Figure 12: Estimates of stock and recruitment values for 1965-1997 from the conditioning process for three different values of fixed steepness for mortality vector 2. The solid line is the replacement line.





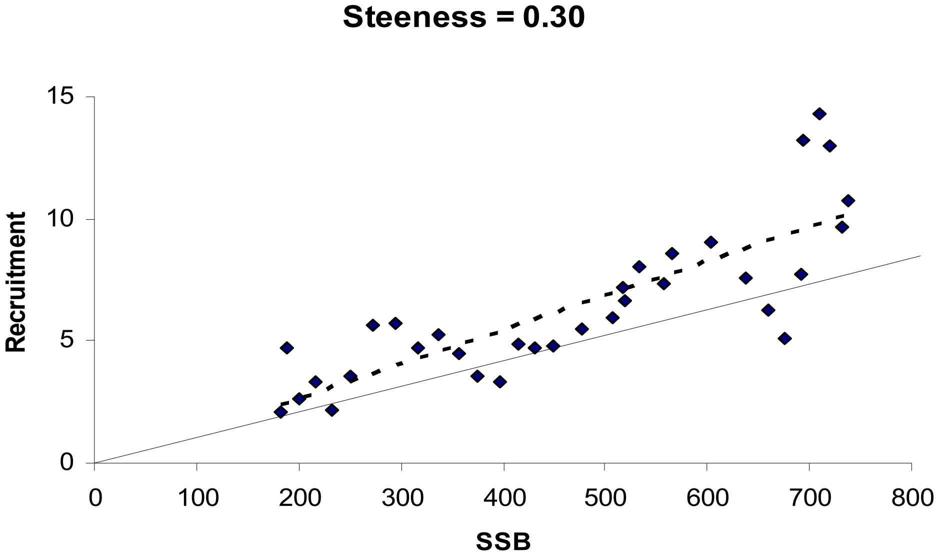


Figure 13: As per Figure 12 with the best fit linear trend add to the figure (dashed line).

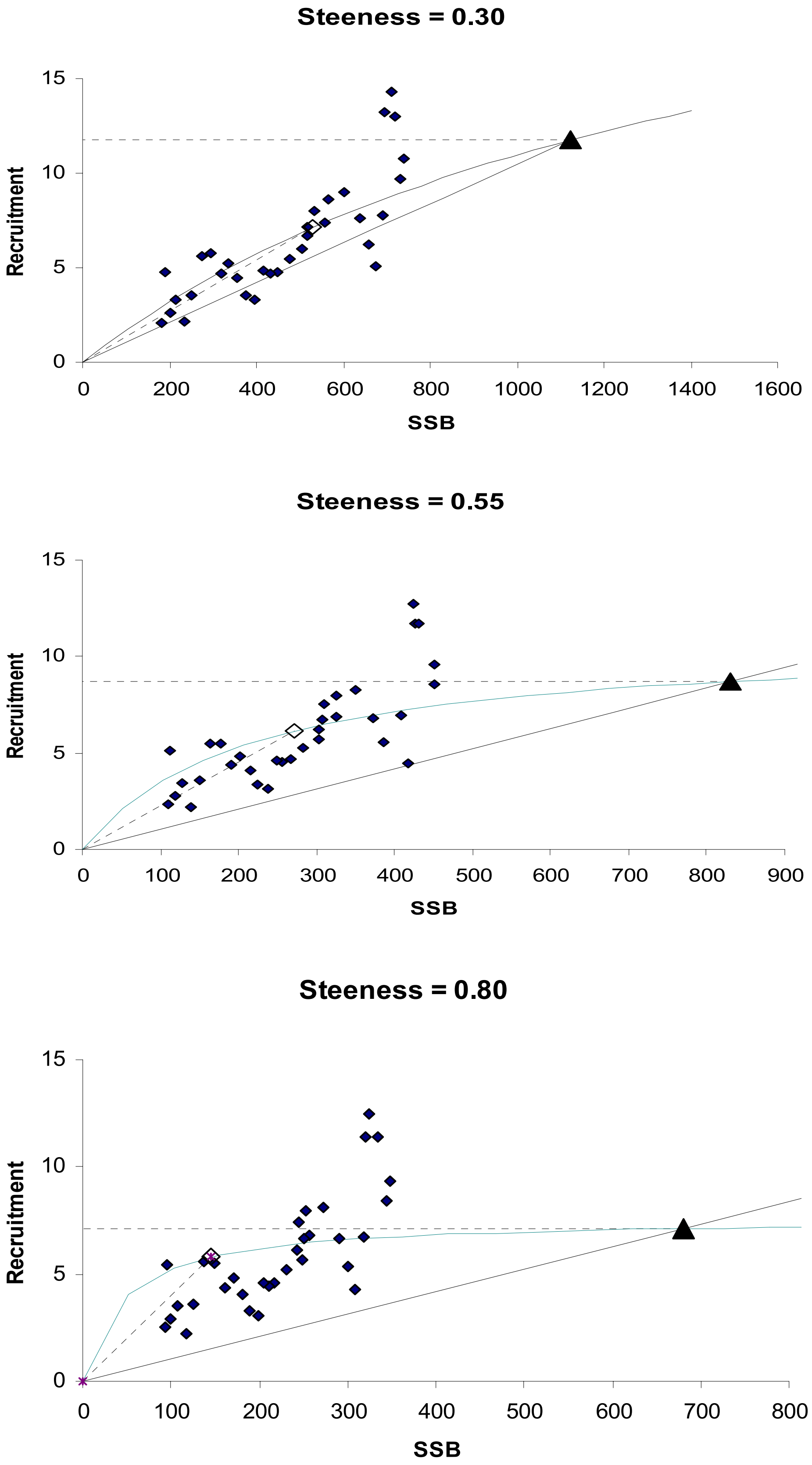


Figure 14: As for Figure 12 with the estimated stock and recruitment curve added (solid curved line), the estimated equilibrium point (solid triangle) and the point corresponding to MSY (open diamond).

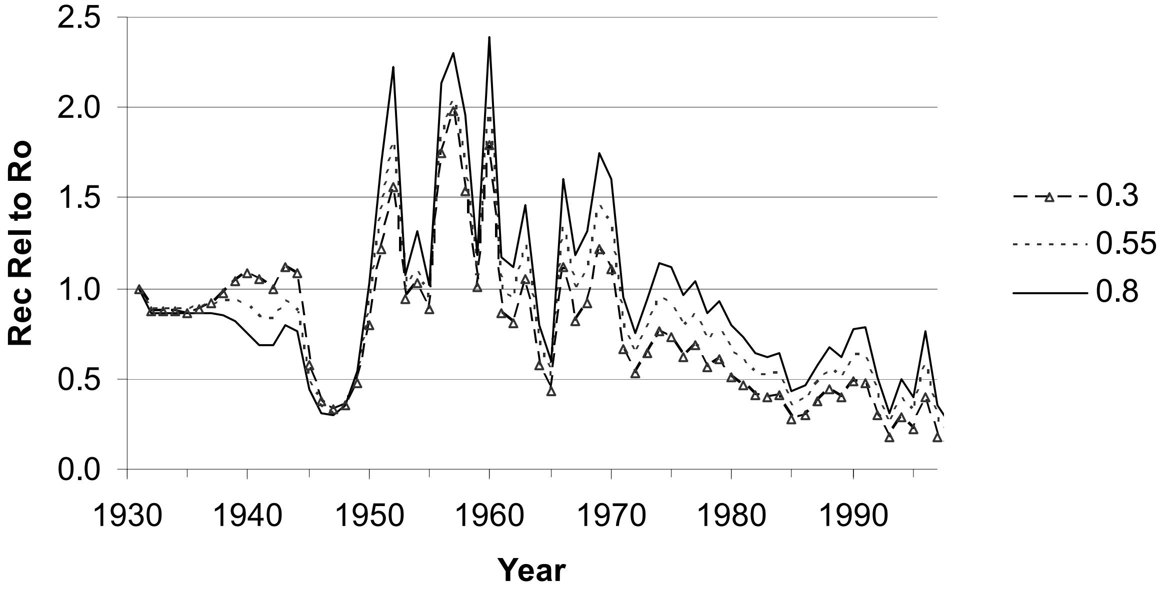


Figure 15. Best fit estimates of recruitment relative to equilibrium value (R_0). Natural mortality has been set to vector 2.