Section 1: Trawl Surveys

Use of 2000-2002 Survey Indices for Stock Assessment

Group 1

- The NEFSC should continue to use the trawl survey data for New England groundfish stock assessment unadjusted, because such use is scientifically justifiable.
  Payne Summary Point 1, page 3
- There appears to be no systematic change in trawl survey performance in the period covered by the offset trawl warps.
  Bell, page 2
- Conversion coefficients, for use in assessments, should be determined for stocks for which the catch results from optimal and suboptimal gear settings differed most
  Payne Summary Point 5, page 4
- However, given that there is no consistent differences between the "optimal" and "worst case" trawls (i.e. some positive, some negative), there are no grounds for modifying the survey results except for those where significant changes were detected. In these instances, conversion coefficients should be determined.
  Bell, page 6

The NEFSC agrees with the conclusions of the reviewers that effects of trawl warp offsets are minimal and likely within the magnitude of survey variability for most species. The recommendation to use the unadjusted trawl survey indices for regulated groundfish stocks has been followed.

The species where significant differences were detected are not regulated groundfish species (i.e., sea scallop, herring, *Loligo*, smooth skate and winter skate). Calculating conversion coefficients for these species using only the data from the experiment would not be technically appropriate due to the multiple gear modifications in the experimental treatment and the associated concerns raised by all reviewers. Additional experimentation at different depths and warp offsets is necessary to estimate calibration coefficients, and might be possible with cooperative research projects.

Group 2

- The sensitivity tests carried out to evaluate the implications of the trawl warp offsets for the evaluation of stock status and rebuilding plans adequately bounded the range of potential introduced biases.
  Payne, Summary Point 1, page 3
- Yes, the 10, 25 and 100% were reasonable magnitudes. Although, the magnitudes were all in the same direction, I acted under the assumption that the mis-configured gear had lower efficiency. In light of the number of instances where

this assumption was not true, -10% and -25% perturbations would have been of interest.
Mohn, page 3

- Yes, the sensitivity tests carried out in the VPA analyses and projections using the 10%, 25% and 100% decreases in catching power adequately characterized the uncertainties in estimated stock sizes and rebuilding mortality rates arising from unequal warp offsets. The only minor improvement would have been to also consider scenarios in which increases in catchability resulted from the unequal warp offsets, because this outcome also appeared to be a possibility for perhaps a few of the stocks from some of the analyses (e.g., as for yellowtail with an estimated 50% increase, Fig. 3.11.1) and from the more recent trawl warp offset experiment.
  McAllister, page 13
- The sensitivity analyses performed did cover an adequate range of reduced catchability for the species examined. The subsequent trawl experiment has demonstrated that reductions in catchability in the region of 100% are highly unlikely, while changes in catchability of 10-25% are in the region of survey variability.
  Bell, page 5

The NEFSC agrees that the sensitivity analyses performed adequately bound the potential range of effects in the direction of decreased catchability in the eight surveys affected by trawl warp offsets. The finding that offset warps might have increased catchability for some species was unanticipated, and appears both in the analyses of historical data and in the trawl warp offset experiment. The NEFSC will investigate the need to produce sensitivity analyses for increased catchability for those species where there is evidence for increased catchability.

Group 3

- If the various survey gear protocol issues remain controversial, then an experiment specifically designed to detect these effects should be undertaken.
  Payne, Summary Point 2, page 3
- If the various survey gear protocol issues remain controversial, then an experiment specifically designed to detect these effects will need to be undertaken.
  Cook, page 5
- Future experiments of the same nature, which would be valuable, need clear objectives and sufficient time to evaluate objectives individually, not to be undermined by more and more perturbations being added to the initial experimental objective (in this case the effect of warp offset).
  Payne, page 11

The NEFSC agrees that the design of the November experiment did not allow us to address gear treatment effects (individual faults) raised in each of the reviewer's reports. However, the decision to carry out a 'worst case scenario' experiment was made at the insistence of some

members of the '*Ad hoc* Experimental Design Working Group' on 22 October, 2002. The experiment was still useful in bounding the problem, assuming that the individual gear changes were additive and not masking each other. If the GARM and ALB/DE gear comparison data analyses were not available, a more scientifically defensible experimental design would have been vigorously pursued. The NEFSC notes that the R/V Albatross-F/V Sea Breeze experiment was not a research vs. commercial experiment, but rather an assessment of research survey performance using two gear configurations. Some analyses of R/V Albatross vs. F/V Sea Breeze catches were performed (such as comparisons of CVs), but these were external to the original experimental design.

Issues related to Survey Variability/Consistency

Group 4

- Survey gear should perform as consistently as possible. Further, the commercial gear in the experiment yielded less sampling variability, and the reasons for this should be investigated.
  Payne, Summary Point 3, page 3
- However, it is desirable for any survey gear to perform as consistently as possible, and if there are lessons to be learned from the performance of the commercial gear in reducing sampling variability they should be investigated.
  Cook, page 6
- Consistency in catchability through time is the most important factor in a survey. Low catchability is not necessarily a problem.
  Bell, page 2
- The power of the gear will be most influential in those species or size of species that are rarely caught in the traditional gear but are well sampled by a different (perhaps commercial) net.
  Mohn, page 5
- While the current skipper will have learned techniques for gear deployment over time, it would be preferable if, when a replacement is required, someone with commercial fishing experience could be utilized.
  Bell, page 7

The variability in catches from the commercial gear in the Survey Trawl Experiment was lower than that of the #36 Yankee trawl. However, the average CV's of catches from each of the vessels over all species analyzed were very similar: 19.3 for the F/V Sea Breeze vs. 21.8 for the R/V Albatross. Thus, although it was mentioned in the Starr report and at the peer review that the commercial vessel produced more consistent catch rates, this is not necessarily supported by the data. In fact, compared to general trawl survey data, experimental results for both vessels were remarkably consistent. The NEFSC is conducting additional analyses of the data from the experiment in order to determine whether the variability in catches were significantly different on a species-by-species basis.

The formal set of bottom trawl protocols has been revised and subjected to peer review. NEFSC

and vessel operations will adhere to these protocols and ensure consistency in the gear used in future surveys. The NEFSC is committed to continued collaboration with fishing industry stakeholders in order to utilize experience of commercial operators to reduce sampling variability in survey gear. This may include the use of additional personnel with commercial fishing experience.

Group 5

- An evaluation of the ability of the survey to detect population signal above the inherent noise should be conducted for those assessments most dependent on survey indices of abundance.
  Payne, Summary Point 4, page 4
- It would be useful for an analysis to be conducted that evaluated the ability of the survey to detect population signal, above the inherent noise in the survey, for those assessments most dependent on survey indices of abundance.
  Cook, page 6
- I recommend that detailed simulation modeling be undertaken to address the point, at which the trawl survey no longer serves as a reliable index of abundance for low catchability species.
  McAllister, page 18

The NEFSC will undertake these analyses for the appropriate species. However, assessments that are primarily dependent on survey indices generally do not have independent sources of data to determine a true population signal. The NEFSC has previously identified several species and species groups which we acknowledge are not sampled adequately with the standard bottom trawl survey gear. These include many flatfish species and monkfish. For the flounders, we have a directed survey utilizing a 'flatfish net' that samples those species more efficiently. We have also conducted NEFSC/Industry cooperative surveys to monitor monkfish. The NEFSC is committed to initiating new directed sampling programs through cooperative research projects and other initiatives to produce scientific information required for stock assessments and management.

The NEFSC recognizes that fishery independent surveys represent a limited source of information for some species. However, for many of these species, survey data represents the only source of data available to assess stock status. In these cases, it is imperative that the NEFSC effectively communicate the quality of data to decision makers.

Group 6

- Consideration should be given to using estimates of survey CV to supplement current attempts to minimize the influence of survey variability, in survey-data-only stock assessments, through the use of running averages, thus producing a more risk-averse strategy.
  Payne, Summary Point 7, page 4
- Attempts to minimize the influence of survey variability in survey-only assessments have been made through the use of three-year running averages... A

precautionary approach to management might look to use estimates of survey CV to produce a more risk-averse strategy.
Bell, page 6

- Because estimation error can trigger a stock rebuilding response even when there is no need to do so, consideration should be given as to how assessment error should be handled within the management process.
Payne, Summary Point 8, page 4
- It does appear that estimation error can trigger a stock rebuilding response even when there is no need to do so. Given the potential economic impact of this, some thought should be given to how assessment error should be handled within the management process.
Cook, page 7

The NEFSC agrees that the use of running averages can result in delayed management response in both periods of stock decline (increasing the risk of stock collapse) and when stocks are increasing (increasing the risk of forgone yield). However, if point estimates, which are influenced by normal survey variability are used, the risk of inappropriate management actions is quite high. In fact, running averages of survey indices are used by both Councils as a way to ensure that trends in surveys are providing the adequate ratio of signal to noise. This is, as recognized, a trade-off. To implement a survey program that is able to simultaneously reduce survey variability for all 50 commercially important species where survey results are used for assessment purposes is unrealistic. Issues related to the consideration of sampling error as it pertains to index-level assessments and management practices based thereon should be addressed primarily in the management arena.

The issue of how estimation error is dealt with in the management process relates not only to trawl survey data but all elements of index and analytic stock assessments. Typically, estimates of confidence intervals are provided on control parameters of interest from stock assessment results (e.g., current fishing mortality rates and spawning biomasses). These confidence intervals provide a range of feasible values primarily dictated by variability in survey catch numbers at age used to tune the assessments. As noted in the GARM report, the actual confidence intervals are likely wider due to unknown potential biases in catches, sampling variation in catch at age, underestimated discards, and other factors. If managers intend to attain, on a continuing basis, F_{MSY}, then there is no margin for estimation error, and on average, the target will be exceeded ½ of the time and fishing will be below the target rate the other ½ of the time. Since F_{MSY} is considered a limit reference point that should not be exceeded, this management strategy will produce results that do not meet the limit reference point concept for a substantial period, and therefore lower target fishing mortality rates have been recommended.

The other major source of estimation error that can be considered by managers is retrospective bias. In many stock assessment arenas, there are persistent biases that occur wherein current year fishing mortality rates are underestimated and biomasses overestimated. These biases are revealed in subsequent stock assessments when more data on the age composition and survey catch rates of various year classes accumulate. In some cases the biases on F and biomass are

reversed. These biases may result from a number of causes, but "missing" catches are the most prevalent suspect. In these cases, managers may want to guard against exceeding the F reference points by setting F lower (by the degree of retrospective bias), if the bias is a relatively stable as a proportion of the fishing mortality rate, and has been persistent (and is thus predictable). These issues are appropriately addressed by the PDT and New England Fishery Management Council.

Group 7

- Moreover, the need to communicate what constitutes an accurate and precise survey to stakeholders, and the difficulty in doing so, requires more attention. *Mohn, page 1*

This is an issue that has been raised in the past and has been addressed in many ways. NEFSC staff have participated in a number of forums and classes in order to explain the survey and how it fits into the assessment process. We routinely host industry representatives onboard our research vessels and now have a program to provide compensation for fishing industry participants. As a result of the Survey Trawl Experiment Workshop in January, 2003 the NEFSC will increase industry's involvement in the preparation of post cruise reports and will meet with industry representatives during the survey season each year. In addition, the NEFSC in conjunction with the New England and Mid-Atlantic Fishery Management Councils has established a Fisheries Independent Surveys Advisory Committee with representation including council members, industry stakeholders, academic scientists, and NEFSC scientists.

<u>Changes to the Survey Time Series</u>

Group 8

- If the NEFSC surveys are subjected to redesign with the involvement of stakeholders (all reviewers consider this measure to be unnecessary), independent scientists with a knowledge of survey design must be included to ensure that scientific standards and data-series continuity are not compromised. *Payne, Summary Point 6, page 4*
- The long time series of fishery-independent relative abundance indices, that the survey provides, are fundamental to the stock assessments that are carried out for the New England groundfish fisheries and the same sampling protocol should be maintained in order to enable reliable stock assessments to be continued to be carried out. *McAllister, page 19*
- Therefore, if survey design needs to be reviewed in light of recent experience, and the reviewers concluded that there was not conclusive evidence to do so, it must be done with the involvement of stakeholders. Further, independent scientists with a knowledge of survey design should be included at an early stage if redesign is being countenanced, to ensure that scientific standards are not compromised and that the results of any research survey conducted with amended design are comparable with those of the current design. *Payne, page 12*

- If there is a move to redesign the NMFS surveys with the involvement of stakeholders, I strongly recommend that independent scientists with knowledge of survey design are included to ensure that scientific standards are not compromised.
  Cook, page 6
- No gear will sample all species, so compromises must be made in their selection (general groundfish trawl or flat fish trawl or shrimp or scallops…). Once the choice is made, it should be used as long as possible, with routine mensuration to assure consistency.
  Mohn, page 1
- For a trawl survey to be a useful index of abundance, it must be comparable over a period of time. Its power is of less importance as long as the species of interest are adequately sampled.
  Mohn, page 3

The NEFSC agrees that the current survey design and time series provides useful information for science and management interests, and will not immediately alter the currently employed methodology without careful consideration. However, the NEFSC also recognizes the evolving needs of management and science as outlined in the Findings and Recommendations of the October 2002 Trawl Survey Workshop, and is committed to implementing a strategic process to refine and develop new survey systems. These survey systems will utilize enhanced capabilities of a new research vessel, experience gained through 40 years of surveying, and gear and net mensuration technology to upgrade surveying systems in the near future. A strategic design process will be implemented that includes involvement of fisheries scientists, managers, and a diverse group of stakeholders. Products from this process including the sampling design, gear selection and rigging, new survey protocols, and plans to transition from the current time series to the new survey systems will be subjected to a scientific peer review prior to implementation.

Section 2: Biological Reference Points

Group 9

- Most methodologies used by the NEFSC to compute FMSY and BMSY are adequate, but the protocol used to evaluate the goodness of fit of alternative stock/recruit functions to the data, and to select alternative models to determine FMSY and BMSY for the purposes of fisheries management, needs to be revised.
  Payne, Summary Point 11, page 4

The reviewers commented on both general and specific aspects of the approaches used to estimate biological reference points (e.g., fishing mortality rates and stock biomasses generating maximum sustainable yields). Chairman Payne's overall comment noted that most methods employed by the NEFSC to estimate F_{MSY} and B_{MSY} are adequate but revisions to goodness of fit criteria and procedures for model selection are necessary in the few cases where parametric stock-recruitment models are used. The qualification in this statement is based on detailed

comments by one reviewer (McAllister, comments numbered 5-9 and 15-16 in his report). Because of the potential for these comments to influence reference point determination necessary in the short term for management decision making, we next provide a specific detailed response to these concerns in the section *Issues Related to Goodness of Fit and Stock-Recruit Model Selection*. Additionally, we amplify on a number of issues where detailed information are already available with which to comment upon the implications of recommended research on management outcomes for groundfish stocks. Bell noted that there may be reasons to suspect that Ricker-type stock recruitment functions would be appropriate for cod stocks, based on observations of cannibalism elsewhere. These issues are therefore explored in detail, using data and published studies pertaining to USA cod resources. Last, we comment on how recommended longer-term research issues could be addressed given known resources.

Issues Related to Goodness of Fit Criteria and Stock-Recruitment Model Selection

Comments regarding the use of various model selection criteria for stock-recruitment functions apply in only three cases: Gulf of Maine cod, Georges Bank cod and Southern New England winter flounder. These are the only stocks for which parametric stock-recruitment functions were used to derive estimates of B_{MSY} and F_{MSY}. This section responds to some of the issues regarding the appropriateness of various model selection procedures (which remains a vigorous ongoing debate). The section concludes with a sensitivity analysis of the impacts on management advice of choosing alternative models, as proposed by the reviewers.

Group 10

- AIC (or BIC) should not be used as model selection criterion for the Bayesian statistical models used. Instead, Bayes' factor or Bayes' posterior is method of choice for evaluating the goodness of fit of Bayesian statistical models to the available data.
  McAllister, page 4
- The AIC value was incorrectly applied to compute the marginal posterior and Bayes' factor for each alternative model. Instead, the marginal probability of the data, given each stock-recruit function (P (data given model (i))), should be used to compute Bayes' factors.
  McAllister, page 4

These comments concerned the appropriateness of the application of the AIC (Akaike's Information Criterion) for stock-recruit model selection. In particular, it was suggested that other "Bayesian" (in contrast to "frequentist") procedures were preferable for this application. Here it is helpful to understand that "Bayesian" and "frequentist" refer to two differing viewpoints about how to handle randomness or chance in statistical estimation procedures. For Bayesians, both data and parameters estimated from data are viewed as being randomly distributed. For frequentists, data are randomly distributed but parameters are not. Instead frequentists view parameters as fixed constants with a true but unknown value. The Bayesian approach allows researchers to include prior knowledge, subjective experience, or information from other studies in estimation. Bayesian estimation updates prior beliefs about parameter

values with the likelihood of observed data. The frequentist approach is based on how frequently one would expect to obtain an experimental result given the analysis was repeated many times. In contrast to the Bayesian approach, frequentist estimation does not include prior beliefs and is based solely on the likelihood of the observed data. While both statistical approaches have inherent strengths and weaknesses, neither has been proven superior in a scientific inference. Nonetheless, when prior beliefs are vague and uninformative, Bayesian and frequentist estimation procedures often produce similar results.

While other procedures could have been employed, AIC is clearly appropriate for frequentist inference based on the likelihood function of the data (Burnham and Anderson 1998). The observation that AIC is inappropriate for Bayesian inference neglects the fact that this approach was used to select from the set of highest posterior density model estimates (e.g., maximum likelihood estimates). This was done because it was recognized that only the best estimates would be directly used for projections and management advice. In fact, the general approach used for stock-recruitment model selection could be described in either Bayesian or frequentist terminology. Thus, the emphasis on Bayesian inference is not necessary for interpreting the results.

Furthermore, the suggestion that the measures of stock-recruitment goodness-of-fit (e.g., calculated AIC values) included prior probability values in these applications is not technically correct. The AIC was (correctly) evaluated solely on the basis of the likelihood function, a technical point that could easily have been missed given the large amount of documentation. Last, we note that use of BIC (e.g., Bayesian Information Criterion) for stock-recruitment model selection for the Gulf of Maine cod, Georges Bank cod and Southern New England winter flounder produces identical results to AIC-based model selection

Group 11

- If NEFSC adopts a Bayesian statistical approach to select a stock-recruit model for reference point determination, it should first decide (Beverton-Holt or Ricker) the baseline set of priors deemed the most appropriate reflection of existing knowledge of model parameters for each stock-recruit model form, and then evaluate the alternative functional forms using Bayes' factor.
  Payne, Summary Point 15, page 4
- If the Bayesian statistical approach is to be adopted ...the NEFSC should first decide for each stock-recruit model form (Beverton-Holt or Ricker), on the baseline set of priors, that it deems to be the most appropriate reflection of existing knowledge about the model parameters. Then only the alternative functional forms should be evaluated using Bayes' factor, not the same functional forms but with different priors.
  McAllister, page 4

This reviewer (in his report, and as captured in Payne's summary point) suggested that neither AIC nor BIC are appropriate for evaluating model selection in a Bayesian context. This contradicts the advice of Kass and Raftery (1995), who point out that

"BIC gives a rough approximation to the logarithm of the Bayes factor, which is easy to use and does not require evaluation of prior distributions. It is well suited for summarizing results in scientific communication."

Kass and Raftery also point out that

"Model comparisons based on AIC are asymptotically equivalent to those based on Bayes factors. But this is true only if the precision of the prior is comparable to that of the likelihood."

In the context of typical stock-recruitment data, estimates of key quantities like steepness are imprecisely determined when estimated solely on the likelihood function – a point repeated by several of the peer reviewers. This lack of precision is why prior distributions based on meta-analyses of many stock-recruitment data sets were used in some models to estimate steepness parameters. In fact, the assumption that the precision of the prior distributions used for steepness is comparable to the precision based on the likelihood of the data is very plausible for groundfish stock-recruitment data. As a consequence, it may be inferred that either AIC or BIC provide a reasonable approximation of the Bayes factor in this estimation problem. Since AIC and BIC produce identical results, it can be concluded that use of Bayes factors for model selection would not be likely to produce different results.

Nonetheless, the use of Bayes factors was proposed by McAllister (above, from his report, page 5) as an alternative model selection method. Application of Bayes factors may be useful when it can be assumed that the marginal density of the data under each model is proper (has a non-zero integral over the parameter space). However, this assumption was questionable for several stock-recruitment data sets. In particular, the inadequacy of this assumption was a reason why extrinsic model selection criteria were applied. Notwithstanding this comment, we examined the sensitivity of biological reference point estimates to the selection of different stock-recruitment models for the three stocks that used parametric stock-recruitment relationships to set B_{MSY} and F_{MSY}: Gulf of Maine Cod, Georges Bank Cod, and Southern New England Winter Flounder (as noted in the introduction to this section).

This reviewer agreed with the approach of first removing biologically implausible models and observed that the criteria that were applied *"...were applied in a consistent and appropriate manner"* :

Group 12

- The hierarchical criteria for comparing parametric stock-recruitment model fits listed on the lower half of p. 23 in Anon. (2002) and top of p. 24 Anon. (2002) up to point #6 appear to be perfectly sensible criteria to apply to evaluate whether the estimation results obtained are plausible for a given fit of a stock-recruit model alternative to the stock-recruit data. From my review of the various results

> presented, it appears that these criteria were applied in a consistent and appropriate manner.
> *McAllister, page 24*

These six explicit criteria for model evaluation were: (1) parameter estimates must not lie on the boundary of their feasible range of values, (2) the estimates of MSY lies within the range of observed landings, (3) the estimate of B_{MSY} is not substantially greater than the non-parametric proxy estimate, (4) the estimate of F_{MSY} is not substantially greater than the value of F_{MAX}, (5) the dominant frequencies for the autoregressive parameter, if applicable, lie within the range of one-half of the length of the stock-recruitment time series, and (6) the estimate of recruitment at B_{MSY}, the maximum spawning stock size proxy input to the stock-recruitment model, is consistent with the value of recruitment used to compute the non-parametric proxy estimate of B_{MSY}. Given the application of these stated hierarchical criteria for identifying plausible fits, the three stocks with parametric stock-recruitment relationships (two cods and SNE winter flounder) had either one (SNE WF) or two (both cod stocks) acceptable models from which the best model must be selected. The full range of models examined including the naming conventions, model functional forms (Ricker, Beverton Holt), and the use of various prior information based on stock-recruitment models fit to the same or similar species elsewhere is given in attached Table 1 (page 30.) We describe the sensitivity analyses of alternate model choices for each stock below.

For the Southern New England winter flounder stock, there was only one acceptable model based on the agreed-upon criteria. Thus, the use of AIC, BIC, or the Bayes factor is irrelevant.

For the Gulf of Maine cod stock, there were two acceptable models: BH and PRBH. These models are both Beverton-Holt curves and differ only by the latter using a prior on unfished recruitment generated from the VPA time series of recruitments. If the AIC-based model selection criterion was ignored and the PRBH model was selected instead of the BH model as the best model, then point estimates for F_{MSY} would change only slightly from 0.23 to 0.24, B_{MSY} would change from 82,800 mt to 65,500 mt, and MSY would change from 16,600 mt to 13,900 mt. By way of comparison, the empirical-based estimates for GOM cod are F_{MSY} =F40% = 0.17, B_{MSY} = 87,600 mt and MSY = 13,700 mt.

For the Georges Bank cod stock, there were only two acceptable models: PRBH and PRABH. These models are both Beverton-Holt curves with priors on unfished recruitment generated from the VPA time series of recruitments and differ only by the latter using an autoregressive AR[1] error term. The autoregressive parameter (phi) in the PRABH model was estimated as 0.02 (essentially phi=0). If the PRABH model was selected instead of the PRBH model, there would be no difference in the biological reference points. In particular, F_{MSY} estimates are identical F_{MSY} = 0.18, B_{MSY} estimates differ by less than 1% (217,000 mt vs 216,000 mt), and MSY estimates differ by less than 1% (35,200 mt vs 35,100 mt).

In the reviewer's opinion, the most justifiable prior for the average unfished recruitment should be chosen beforehand (McAllister, page 27). This neglects the fact that empirical Bayesian approaches (alternatively, penalized likelihood approaches in a frequentist view), as were used in

the WG report, have been advocated because it is appropriate to allow observed data to have some role in choosing informative prior distributions (Carlin and Louis 2000). Furthermore, this prior choice would eliminate the necessity of choosing between two models for Gulf of Maine cod because only one of the two acceptable models would be allowed.

These sensitivity analyses thus show that choosing a less likely alternative model would not cause changes in estimated F_{MSY} or B_{MSY} for two stocks (SNE winter flounder and GB cod) or cause changes in F_{MSY} for the remaining stock (GOM cod). Only the B_{MSY} value for GOM cod can potentially be changed if an alternative model selection method that chooses a less likely model was employed (e.g., 62,500 vs. 82,800 mt).

Group 13

- More than one diagnostic tool should be applied to evaluate convergence on posterior distributions, rather than simply relying on the MCMC software for statistical estimation.
  Payne, Summary Point 16, page 5
- Furthermore, although it is claimed that the algorithm was run for 500,000 iterations, and this seems like many, the methods, if any, that were used to test or diagnose for convergence were not reported as they should have (Gelman et al. 1995)...This is a serious omission, and results cannot be taken to be reliable unless such diagnostics have been applied and found to consistently indicate convergence.
  McAllister, page 27

As shown, one reviewer asserted that the lack of reported convergence diagnostics for the MCMC calculations is a *'serious omission*". This comment must be interpreted in light of its practical, not just theoretical importance.

First, the MCMC calculations were used to characterize precision of the posterior distribution of key management quantities, such as MSY, not to provide point estimates of these quantities. Indeed, the only practical significance of the MCMC calculations was to provide Bayesian confidence intervals for key parameters. Since management of New England groundfish is based on point estimates of key parameters, not their variance, it is moot whether the lack of convergence diagnostics is an important point.

Second, if convergence diagnostics had been calculated, it is very likely that the robust thinning rate of accepting only 1 out of every 100 MCMC samples would exhibit no substantial autocorrelation. This expectation is supported by Congdon (2001). Thus, while it was an omission to not document the convergence of the sequences, this was neither serious nor particularly relevant given the practical uses to which the MCMC calculations are applied.

Use of Ricker (overcompensation)-Type Stock-Recruitment Model Forms

Group 14

- If the issue of model selection (Beverton-Holt vs. Ricker) remains on the table, then model validation needs to be addressed more fully, and more divergent models ought to be tested (e.g. non-parametric deterministic and probabilistic).
  Payne, Summary Point 14, page 4

There were no cases where a Ricker curve was used to calculate parametric MSY-based reference points since none of the Ricker functions examined passed the series of hierarchical model tests. In practice, it is often impossible to discern between Beverton-Holt and Ricker curves based solely on statistical goodness-of-fit criteria (Brodziak 2002). Nonetheless, least squares estimation procedures combined with AIC criteria, similar to those used in this report, have been found to have an inherent bias towards selection of Ricker curves when the actual curve was Beverton-Holt in recent simulation studies (DeValpine and Hastings 2002). Thus, strict adherence to goodness-of-fit criterion to choose a parametric model could be very misleading and it is very important to apply common sense when judging the adequacy of fisheries models (Schnute and Richards 2000).

Group 15

- There may be *a priori* biological reasons for assuming an over-compensatory stock-recruit function (cannibalism, spatial interference between adults and progeny, etc.), and this is acknowledged in the report although this is not taken through to implementation. Cod are known to be cannibalistic for a number of stocks, and it is a reasonable assumption that the same occurs in the Gulf of Maine and Georges Bank stocks. The acceptance of the Beverton-Holt type relationships for these stocks therefore appears to be choosing the wrong model albeit for the right statistical reasons.
  Bell, page 10

While the Ricker model fits were rejected in all cases on other technical grounds it is worth wile to consider this potential mechanism in more detail. Accordingly, cannibalism in the primary New England groundfish stocks was examined further. Food habits data collected during spring and autumn NEFSC surveys during 1973-1997 show that the observed incidence of cannibalism in cod and haddock is very low. Out of 12,305 Atlantic cod stomachs examined, only 16 contained cannibalized cod (<0.2%) and the average percent composition by weight of the cannibalized cod was less than 0.1%. Similarly, out of 3,537 haddock stomachs examined off the Northeast USA, only 1 contained cannibalized haddock (<0.1%) and the average percent composition by weight of cannibalized haddock was less than 0.1%. For benthic feeding flatfishes, such as yellowtail and winter flounder, the incidence of cannibalism was virtually nil. Thus, the observed data on groundfish food habits do not support the hypothesis that cannibalism is an important mechanism for overcompensatory stock-recruitment dynamics in primary New England groundfish stocks (at least over the range of stock sizes observed since the early 1960s).

Low rates of cod cannibalism are not necessarily the case elsewhere in the world. For example cod cannibalism in the North Sea appears much more frequent, as the percentage of cod in cod stomachs was <1-13% (by weight) depending on predator size group and season (Daan 1983). The average number of cod prey per 100 cod predator stomachs there was 31-104 individuals, again depending on predator age (Daan 1983). Cod in the North Sea grow at rates equivalent to those on Georges Bank, and, up until the last several years, were exploited at similar fishing mortality rates (albeit with smaller sizes at first entry into the fishery). Thus, the apparent much greater frequency of cannibalism there is based on comparisons of feeding data derived from predator (cod) populations with similar population dynamics (see also pertinent results concerning cod cannibalism from a preliminary multispecies VPA that was constructed for Georges Bank and is summarized in a later section for this report (*Consider More Complex Models of Technological and Biological Interactions to Evaluate the Feasibility of Multiple Reference Points*, beginning on page 17).

Results from Ricker models fit by the Working Group (NEFSC 2002a) calculated values of F_{MSY} that substantially exceeded F_{MAX}. For example, of the four Ricker model fits for Gulf of Maine cod, two had estimates of $F_{MSY} = 2.0$ (the boundary condition), and two had $F_{MSY} = 0.6$. The calculated F_{MAX} value for the stock is, by comparison, 0.26. For this to be true, it must be the case that growth overfishing is relatively unimportant in contrast to recruitment underfishing, which is simply the notion that high numbers of spawners reduce intraspecific juvenile survival through some overcompensatory density-dependent mechanism. One possible mechanism for strong density-dependent intraspecific interactions is cannibalism. However, as noted above, data suggest this mechanism is not especially important for cod, haddock, and benthic-feeding flatfish. Furthermore, the fits for the Ricker models computed for the two cod stocks imply estimates of F_{MSY} that are inconsistent with the general life histories for these stocks (about three to four times or greater than the calculated $F_{40\% MSP}$ values). Thus, there does not appear to be evidence to support, *a priori*, a Ricker-type stock recruit model for cod stocks, and the results of the Ricker fits are inconsistent with management reference points recommended for all of the other groundfish stocks.

Simulation Testing for Survey-Index Related Reference Points

Group 16

- A simulation study, using age-structured operating models, should be undertaken to validate the survey index method in a management context, to investigate its sensitivity to *ad hoc* assumptions and to evaluate the potential biases and imprecision in the results.
  Payne, Summary Point 12, page 4
- I recommend that a simulation study be undertaken to validate the survey index method in a management context and to investigate its sensitivity to *ad hoc* assumptions.
  Cook, page 10
- The index-based approach to stock assessment and projections has some appealing conceptual merit. However, to ensure that it provides an adequate

scientific basis for fisheries management advice, it should be simulation tested using age structured operating models to evaluate the potential biases and imprecision in the results obtained.
McAllister, page 6

Several reviewers recommended that the methods for the development of index-based reference points should be tested with simulations. We concur with the recommendation and additional, more extensive, simulation testing will be conducted.

However, a considerable amount of analytical and simulation work occurred during development that validated the method for a number of simple examples. These results were provided at SARC 35 (NEFSC 2002b), where extensive simulation tests and comparisons of the index methods to VPA results were made for stocks wherein both data sets were available (these results were available, but not included in materials reviewed by the Committee). These simulation results showed that the model for selecting index reference points that result in stock replacement can find and recover the true equilibrium relative fishing mortality rates associated with population stasis. We recognize that the model's primary utility follows from the behavior of populations without strong density dependent processes (see results from SARC 35; NEFSC 2002b). I n these populations, population growth is controlled by recruitment and its linear dependence on spawning stock biomass. Since a large number of stock assessments and population indicators verify the overfished status of New England groundfish stocks, the applicability of the approach to other stocks was judged to be appropriate. As the index methodology relies on pooled estimates of abundance and catch, it is likely to be affected by problems similar to surplus production models. In particular, very strong year classes and abrupt changes in fishing mortality rates will induce persistent transient conditions, especially when fishing mortality is low.

However, fishing mortality rates have generally been very high except during the last five years or so. We note that observation error will likely be a dominant concern for full scale testing. Our future work will focus on the influence of such error on the ability to recover relative Fs at replacement. However, simulation testing as previously reported at SARC 35 (NEFSC 2002b) is sufficient to clarify most issues raised in reviewer's comments.

Calculation of Intermediate Biomass targets, e.g., the Highest Biomasses Observed

Group 17

- An adaptive approach to using harvest management to control biomass is eminently sensible in the light of uncertainty, and should be pursued.
  Payne Summary Point 19, page 5
- In the above circumstance, the proposal by the working group in the Report (page ix) to adopt an adaptive approach to biomass management seems eminently sensible, and I would recommend that it is followed.
  Cook, page 11

- An adaptive strategy which intermittently reassesses F targets without reference to B_{MSY} would be more estimable (for example a variation on the F strategies proposed in Shepherd 1981), and a definition of rebuilt that is not dependent on theoretical levels of biomass.
  Mohn, page 5

Owing to uncertainties in recruitment associated with higher spawning biomasses (a consequence of driving the stocks towards the origin of the stock-recruitment curve through persistent overfishing), a number of *ad hoc* approaches were suggested by the reviewers (e.g., Cook, Mohn) for setting more modest biomass targets within the documented histories of stock sizes. These proposals include the highest observed biomass level, arbitrary percentage increments above the current biomass, and variations thereof. While these *ad hoc* approaches may be useful as elements of a management strategy to eventually achieve B_{MSY}, they cannot be viewed as alternative *estimates* of B_{MSY}. This is because the circumstances that may have contributed to the previous biomass maxima are likely irrelevant to the current fishery management program. For example, the highest observed spawning biomass for Georges Bank haddock occurred in 1962, at 199.5 thousand tons. For the decade prior to 1962, the average fishing mortality was 0.35, with the age at 50% selection of about 2.1 years. This contrasts with the recommended F_{MSY} value of 0.26 combined with the current age at entry of about 3.1 years. If fished under the current management program, the resulting 1962 spawning biomass could have been 290-340 thousand tons, substantially higher than the calculated B_{MSY} (250 thousand tons).

Additionally, if results of the suggested *ad hoc* procedures are applied <u>as if they were estimates of B_{MSY}</u>, significant inconsistencies between the biomasses and the fishing mortality rates arise, which become problematic for stock projections and associated scientific advice. For example, the F and biomass reference points for three stocks were based on a Beverton-Holt stock-recruitment function. If a lower *ad hoc* B_{MSY} value is substituted, then the calculated fishing mortality rate and the stock-recruitment relationship are no longer relevant. This becomes particularly problematic in projecting stock biomasses, since the projection models (AGEPRO) require the specification of an underlying recruitment function to evaluate the probability of stock rebuilding. Given the above considerations, the suggested *ad hoc* approaches cannot be scientifically justified as alternative estimates of B_{MSY}.

<u>Consider More Complex Models of Technological and Biological Interactions to Evaluate the Feasibility of Multiple Reference Points</u>

Group 18

- Can all 19 stocks be moved simultaneously towards B_{MSY} through single-species management? Given current knowledge of the complexity of biological interactions and the ecosystem in which the stocks exist, the reviewers doubted

whether the question could be answered satisfactorily. Single-species MSYs are not good indicators of multispecies MSY, so caution will be needed in the choice of B_{MSY} target. There could well be value in assessing interactions in multispecies fisheries, using spatially heterogenous models such as Iceland's BORMICON.
Payne, Summary Point 20, page 5

- As well as biological interactions, models are needed to assess technical interactions in multispecies fisheries.
Mohn, page 10
- An evaluation of technical interactions in the mixed fishery should be undertaken to investigate the consistency of multiple MSY targets.
Payne, Summary Point 24, page 5
- I would recommend that an evaluation of technical interactions in the mixed fishery be undertaken to investigate the consistency of multiple MSY targets.
Cook, page 13

Current assessment approaches rely primarily on single-species stock assessments to provide evaluations of stock status, projections of stock rebuilding, and (in the context of management scenario analysis) evaluations of the effects of management measures. In the case of technological interactions models, NEFSC scientists have developed a wide array of generalized tools to conduct such analyses (Brown et al. 1979; Murawski 1984; Overholtz 1985; Overholtz and Murawski 1985; Murawski and Finn 1986; Murawski et al. 1991; Overholtz et al. 1995). However, the management systems employed by the Council do not incorporate controls that are amenable to the application of these methods. For example, models and systems for optimal area- and fleet-based fishing effort (days at sea) control were described in the 1980s, but the days at sea control systems implemented by the Council do not explicitly specify area- or fleet-based limits. Similarly, TAC-based models for technological interactions, and fleet-based effort controls could be evaluated with these developed methods, were the Council to use them. The one area wherein technological interactions models are used effectively is in management scenario analysis relative to closed areas and DAS allocations. The "GAMS" model used by the PDT (Walden 2001) is a tool to evaluate the effects of alternative fishery closures relative to a suite of desired single-species fishing mortality rate targets.

Some work has been done on biological interactions relative to the evaluation of interspecies predation rates (Overholtz et al. 1991; Tsou and Collie 2001). The results of a preliminary MSVPA constructed for Georges Bank (Tsou and Collie 2001) showed that for ages one and older the predation mortality rates on groundfish species examined (cod, haddock, yellowtail flounder) were 0.2 or smaller (declining significantly with age). They conclude there were no meaningful differences in calculated fishing mortality rates between single- and multispecies VPA results, since predation was limited in groundfishes primarily to ages two and younger, and catches of these groundfish generally start at age three. Furthermore, they found no clear pattern of increasing rates of cannibalism with increasing cod biomass on Georges Bank, even over a wide range of cod biomasses. The calculated fractions of cannibalism contributing to cod diets over the historical period were extremely low. In the future, models such as these will become more important in evaluating biomass trade-offs and the feasibility of community-wide stock

recovery plans. The historical patterns of groundfish growth rates in relation to population and groundfish community biomasses can also be examined for evidence of density dependence (which could be a mechanism that might limit productivity at high stock sizes). Examination of these data for Georges Bank cod, haddock and yellowtail flounder indicated no obvious changes associated with stock biomasses. Growth rates since the 1960s have been remarkably consistent, varying without trend (these data were presented to the peer reviewers). As part of NMFS' Stock Assessment Improvement Plan (NMFS 2001) upgrades to these assessments to routinely include inter-species predation and technological interactions are planned. Advancements in this area would require data and resources associated with "tier three" (most advanced) assessment capabilities.

Stock-Recruitment Model Validation

Group 19

- The stock/recruit models used in projections should be validated against historical observations of stock dynamics.
  Payne, Summary Point 21, page 5
- I would recommend that the s-r models used in projections are validated against historical observations of stock dynamics.
  Cook, page 12
- The methods used to derive non-parametric stock/recruit functions to approximate B_{MSY} should be simulation tested with a variety of underlying operating models for stock/recruit processes, to test the robustness and accuracy of the methodology.
  Payne, Summary Point 13, page 4

Cook's recommendation is a useful one, and work on these analyses has commenced. Preliminary analysis indicates that modeled trajectories for some stocks are above the actual stock path, and for others below. This exercise is, however, complicated by the fact that, by using the actual Fs and fishing patterns, and modeled recruitment, actual recruitment events that resulted in increased fishing mortality rates (e.g., for yellowtail flounder and haddock year classes) were uncoupled in model simulations. Thus, these models require a complex set of additional assumptions regarding the link between stock and fishing effects for adequate interpretation of these trajectories. Perhaps a more appropriate simulation testing environment would be to generate a set of known population trajectories and to project them under model and sampling (assessment) uncertainty, as was done in the National Research Council's testing of stock assessment models (NRC 1998a; Restrepo 1998). To this end, NMFS is developing a generalized simulation modeling environment ("POPSIM") for testing the adequacy of both stock-recruitment models and alternative stock assessment models. POPSIM is nearing completion and beta testing, and will be incorporated into the stock assessment toolbox. Simulation testing of the adequacy of stock-recruitment formulations will then be more rigorous.

Some S/R model validation based on known (simulated) population data, fit with alternative stock-recruit functions, has been previously reported. These results were presented to the

reviewers, and were developed primarily to understand the consequences of the wrong choice of S/R model under "known" conditions (e.g., Dr. Legault's presentation to the reviewers).

Validation power is, of course primarily limited by the dynamic range contained in the observed S/R data set. In the case of New England groundfish, most of these observations are from stocks significantly degraded by chronic overfishing. Since most stock-recruit models would give very similar results for stocks existing on the left-hand limb of the stock-recruitment curve, power to discern functional forms will only increase as stocks rebound, which is the fundamental issue to be evaluated in an adaptive approach to stock rebuilding.

Criteria for Including Autocorrelation in Stock-Recruitment Models

Group 20

- There may be other objective criteria that are more appropriate for deciding when to include autocorrelation in models to determine F_{MSY} and B_{MSY}, and in stock projection models. Such criteria should be investigated.
  Payne, Summary Point 17, page 5
- It is recommended that consideration be given to the issue of determining other objective criteria that might be appropriate for determining when to include autocorrelation in models to determine Fmsy, Bmsy and in stock projection models.
  McAllister, page 5

The recommendation to consider alternative criteria for deciding when to include autocorrelation requires development of these other criteria and simulation testing of their impacts. Alternative criteria must have adequate theoretical as well as correlative basis (e.g., expected autocorrelation from life history or environmental attributes) because observed time series are generally so short. We agree that time series data are limited in most cases, but, compared to other fishery data sets, they are remarkably long. For example, the data set for Georges Bank haddock includes seven decades of stock and recruitment information. One way to increase the length of such data sets is to "hind-cast" stock and recruitment from VPA-survey series, or to use models that blend high quality recent information with available historical data. Some of these methods have been explored (e.g., Brodziak et al. 2001 hindcast recruitment and spawning biomasses back to 1963). Fishery oceanography programs in the NEFSC (e.g., GLOBEC) have as one focus the use of environmental data series with which to evaluate correlation with observed recruitment trends. These analyses will continue.

Evaluate Costs Associated with Potentially Erroneous B_{MSY} Values

Group 21

- If the current B_{MSY} reference points are adopted, the potential costs of adopting an erroneous value need to be evaluated.
  Payne, Summary Point 9, page 4

- If the current B_{MSY} reference points are to be adopted, I would recommend that an analysis is done to evaluate the potential costs of adopting an erroneous value.
  Cook, page 8

The recommendation to evaluate costs associated with potentially erroneous B_{MSY} values (in relation to Frebuild) falls in the policy arena due to the requirement to balance competing goals of catch and sustainability. Similarly, the repetition of the Working Group recommendation for adaptive approaches requires making policy decisions regarding implementation of rebuilding strategies. The term "costs" is presumed to mean both costs associated with attempting to achieve B_{MSY} values that may be too high, as well as benefits forgone due to B_{MSY} values that may be set too low. In the former case, the only time these costs would be incurred is when managers implement Frebuild values significantly lower than the F_{MSY} values, in order to meet B_{MSY} values in a specified time frame where the initial B_{MSY} values are set too high. In this regard, adaptive approaches that have been proposed for fishing at or below F_{MSY} for a period of years in the rebuilding program would seem to ameliorate the short-term costs of immediate reductions to fishing rates significantly below F_{MSY}. As for adopting erroneously low B_{MSY} values, this would, again, only be a significant consideration in the circumstances when F was allowed to remain above F_{MSY} for a period of time (e.g., a tapered F reduction strategy where F remained above F_{MSY} for part of the rebuilding period) and the long-term target was significantly below the true B_{MSY}. These types of analyses are primarily the province of the New England Fishery Management Council's Plan Development Team (PDT).

Adaptive Approaches to Achieving B_{MSY}

Group 22

- From a purely scientific perspective, the setting of an intermediate biomass target could be defended. In such a case, once a rebuilding F and trajectories were in place, this intermediate target could be used as a signpost to see if stock rebuilding was on track. If the rebuilding target were not on track, an assessment would be needed to determine what the causes are (e.g., were the original projections in error, was recruitment atypical, or was recruitment as expected but F not achieved).
  Payne, Summary Point 23, page 5
- From a scientific point of view, an intermediate biomass target could be defended. For example, once rebuilding F and trajectories were found, the biomass target could be used as a signpost to see if the stock were on track.
  Mohn, page 11
- Among other management scenarios that may be considered include the limiting of maximum interannual change in total catch. This would enable a more structured transition within the industry.
  Bell, page 12

Several reviewers reiterated the initial recommendation of the Reference Point Re-evaluation Working Group (NEFSC 2002a) to approach stock rebuilding towards BMSY as an adaptive

management/science partnership. Elements of a true adaptive management program involving science and management actions have been provided to the Council. Constraining fishing mortality to levels at or below F_{MSY} will provide the most important element of such an adaptive approach that cannot be mined through ever more sophisticated analyses of existing data – observations of the responses of stocks and ecosystems at relatively high stock sizes. Only then will the dynamic responses associated with the right-hand side of the production curves for species, complexes and ecosystems be more fully understood.

Use of Spawning Stock Biomass as a metric of Spawning Potential may Overestimate Stock Resiliency

Group 23

- Most stock-recruit functions use SSB as a proxy for stock reproductive potential. There is increasing evidence that use of this proxy can overestimate a stock's resilience to fishing pressure.
  Bell, page 8

The basis of this comment is that since SSB does not account explicitly for the effects of maternal experience or size on egg hatching success or larval viability, there may be excessive reliance on first time spawners (Trippel 1999; Wigley 1999). This mechanism is likely to have contributed to resource declines in many northwest Atlantic groundfishes, including some off New England. As demonstrated for Georges Bank cod (Murawski et al. 2001), failure to take these mechanisms into account can result in setting inappropriately high F reference points. These issues are of particular concern in setting reference points that are intended to avoid recruitment failure, as opposed to those intended to achieve F_{MSY}. Management reference points of the magnitude proposed for the New England groundfish resources (e.g., $F_{40\% MSP}$ or there abouts) should result in stable age distributions with a high fraction of multiple spawners and a higher proportion of large (old) spawners – more so than now (Murawski et al. 2001). These considerations were particularly important in developing the hierarchical criteria used by the Re-Estimation Working Group for evaluating candidate stock-recruitment model fits – and especially the criterion that computed F_{MSY} values were unlikely to be appropriate if they substantially exceeded F_{MAX}.

Section 3: Stock Rebuilding and Related Projections

Group 24

- Alternative means of modelling groundfish stock dynamics should be evaluated, and their results compared with present procedures, although the currently used ADAPT model is deemed by the reviewers to be scientifically sound. Reasons for differences in the outputs of the different models need to be sought.
  Payne, Summary Point 10, page 4
- Most of the stock assessment and projection methodologies currently applied by the NEFSC provide an adequate scientific basis for fisheries management. The ADAPT VPA and AGEPRO methodologies provide a rigorous and adequate basis for assessing stock biomass and fishing mortality rate, making projections,

evaluating the differences in potential consequences of alternative possible fisheries management policies, and for taking into account parameter and important model structure uncertainties.
Payne, page 16

The third Term of Reference for the peer reviewers was to comment on stock rebuilding and related projection methods.

The reviewers appeared to be satisfied that the scientific basis for stock projections was sound. As a direct implication, groundfish rebuilding trajectories derived from AGEPRO projections were deemed to be adequate.

Three long-term research issues were raised that could help to improve the projection method. These were: (i) including alternative sources of uncertainty in projections; (ii) ability to include trends in some parameters in projections; (iii) validation of stock-recruitment models against historical observations of stock dynamics.

The reviewers commented that alternative sources of uncertainty not accounted for in the current projection method

Group 25

- The AGEPRO software used to forecast rebuilding strategies... is well documented and rigorously constructed, and it interfaces well with the ADAPT VPA stock assessment model and the bootstrapping output produced by ADAPT VPA to take into account parameter uncertainty. It certainly allows for uncertainty in the estimates of initial population size, natural mortality and future recruitment, and has additional potential for examining autocorrelation in recruitment. While this covers some of the uncertainty likely in projections, there could well be uncertainty associated with variations in weight, maturity and fishing mortality that widen the confidence limits further.
  Payne, page 19

It is clear that the inclusion of variation in weight, maturity, and fishing mortality (F) could help to improve the characterization of uncertainty in projections. On the other hand, the AGEPRO model incorporates uncertainty in both initial stock size and recruitment, the two primary sources of future uncertainty given F. Thus, while the reviewers are correct to point out that including stochastic variation in life history parameters could be useful, the effects would likely be minor in comparison to those due to stock size estimation uncertainty and inherent recruitment variation. For fishing mortality, the question of an appropriate model of variability must be addressed prior to its inclusion in the projection methods. Variation in F is a difficult issue to address empirically given overall changes in groundfish management measures through time. Overall, the inclusion of alternative sources of variability in projections is a useful long-term research recommendation

One reviewer also commented on the inability to include trends in some parameters in AGEPRO projections:

Group 26

- Of more concern is the inability for the software to cope with trends in parameters. Systematic changes in parameters such as weight- and maturity- at age can cause significant bias in stock projections.
  Bell, page 12

It is certainly true that some bias could result if directional changes in weight or maturity at age occurred and were not included in the projection model. However, many complex factors can affect fish growth and maturity rates including, but not limited to, oceanographic condition, population density, and multi-species interactions. Thus, a key difficulty in addressing this issue is developing a reliable predictive model of future trends in fish weights and maturity. Ongoing monitoring of fish growth and maturity rates is a standard component of NEFSC stock assessments and also part of an adaptive approach to rebuilding groundfish stocks. Overall, evaluating possible trends in life history parameters is useful long-term research recommendation.

One reviewer commented that it would be useful to validate the stock-recruitment models used for projections:

Group 27

- I would recommend that an evaluation of technical interactions in the mixed fishery be undertaken to investigate the consistency of multiple MSY targets.
  Cook, page 12

Since the stock-recruitment models used for projections are estimated using stock assessment data there is a close correspondence between observed recruitment patterns and simulated recruitment trajectories. Nonetheless, the possibility of non-stationary stock-recruitment dynamics needs to be considered. In this context, validation of stock-recruitment models used for projections is a useful long-term research recommendation. Preliminary work on such forward projections indicates that for some stocks the selected stock-recruitment functions would produce optimistic rebuilding trajectories, while for others the paths are pessimistic, as compared with the actual stock histories. However, such analyses are quite complex, and their interpretation relative to historical fishing patterns is not straightforward. Work on this issue will continue.

Section 4: Stock Assessments

Consider Alternative Stock Assessment Model Approaches

Group 28

- Alternative means of modelling groundfish stock dynamics should be evaluated, and their results compared with present procedures, although the currently used ADAPT model is deemed by the reviewers to be scientifically sound. Reasons for differences in the outputs of the different models need to be sought.
  Payne, summary point 10, page 4

- It appears that if a new model is proposed an evaluation takes place (at a SARC or GARM) and the selected model becomes the one upon which the assessment, BRPs and projections are made...Another approach would be to maintain the non-selected model, using it as an estimate of model uncertainty to more fully capture uncertainty in risk in projections...In the present context, VPA models are used, and the criteria for replacement or joint resource desciption need to be investigated an codified.
  Mohn, page 2
- Alternative non-equilibrium production models should be investigated, and other types of models as well.
  Mohn, page 9
- There was not the available time within this review process to perform an in-depth analysis of why the model should appear so flexible and I would therefore recommend that investigations are made into the ADAPT and ASPM model differences.
  Bell, page 11

The peer reviewers proposed several recommendations to enhance the process through which stock assessments are conducted for New England groundfish stocks. The main focus of the recommendations was to apply several assessment models to the data as a means of determining the extent to which model uncertainty may contribute to the overall assessment results. The other focus was to perform simulation testing of any model which may be used as a basis for management advice.

The principal method employed by the NEFSC to derive estimates of stock size and fishing mortality on an age-structured basis is Virtual Population Analysis (VPA). Stock assessments for New England groundfish, based on this methodology have been previously validated in peer review (NRC 1998b). This method is widely used by stock assessment scientists in both North America and Europe to determine the state of numerous groundfish stocks across the North Atlantic. VPA has been thoroughly tested and evaluated in many scientific fora, and the results from this model have been used by scientists in national institutes and regional fishery bodies (Anon 2002a; 2002b; NRC 1998 a; b) to provide assessment advice to fishery managers for several decades.

While the same basic VPA model is widely employed, methods of calibrating the results to independent observations of trends in population size (often based on research vessel survey data) may differ. Various VPA calibration methods have also been tested (Anon. 1988), and the results compared in a rigorous peer-review setting in both national and international fora. The often preferred method of evaluating the performance of VPA or any assessment model is by comparison of model results to a known state of nature. This process is accomplished by first constructing a simulated population whose age structure, population size, catch history and fishing mortality rates are known exactly and then comparing estimates of these parameters, whether they are based on age-specific models, age-aggregated models, or index based approaches, to the known states (e.g., Restrepo 1998). The performance of any assessment model has been shown to vary depending on the quality of the input data, and the internal dynamics of various fish stocks.

The assessment review process at NEFSC has evolved over time to the point where assessments used to provide management advice are subjected to independent peer review at several levels (e.g., Anon 1998). While each assessment is initially completed by an individual scientist, the final product is often the result of several modifications to the initial analysis obtained by consensus of a working group or committee of peers such as the SAW or TRAC. Assessments performed on an annual basis to meet management needs are generally updated by addition of the most recent information on catch and stock size trends. While these assessments are always reviewed by a body of peers, the review is generally less comprehensive than the benchmark reviews to which all assessments are subjected periodically (generally every 3 to 4 years). The most recent stock assessment update occurred when 20 groundfish stock assessments were reviewed at the Groundfish Assessment Review Meeting (GARM) in October 2002 (NEFSC 2002c).

When an assessment benchmark review is conducted, the performance of the baseline assessment model is subjected to a more rigorous set of analyses, including a comparison with alternative assessment models as recommended in Anon (1998) . These include both age-aggregated and age-structured methods such as the ASPM and other similar models which employ a very different approach and may include different fundamental assumptions to estimate stock size and fishing mortality than VPA. Several of the NEFSC VPA-based assessments have been evaluated in this manner under a benchmark review process. Within this setting, it is common to have several scientists who are familiar with various models as well as the basic fishery and survey data to work as a group to ensure that the assumptions and the basic data employed in each of the assessment models reflect conditions in the fishery (e.g., changes in selectivity over time) and in the population (changes in growth, maturation, etc.) This ensures that any differences in the results are due to the model and not the assumptions or the input data.

Results obtained by employing several models may provide insight into the extent of 'model uncertainty', that is, the extent to which the assessment results may differ due to the choice of model. For example, some models may provide more specific information on the quality of the catch data or the fishery and survey selectivity factors, while others may account for changes in the spatial distribution of fishing effort. In these cases, the diagnostics from each model can be rigorously reviewed by the working group and a consensus achieved that reflects the most reliable and biologically meaningful results.

As part of the Marine Fisheries Stock Assessment Improvement Plan (NMFS 2001), NMFS has developed a National Fisheries Toolbox which provides the analyst the ability to perform simulation modeling and to execute various assessment models when assessing a stock, including an Age Structured Production Model (ASPM) and a 2-Box VPA (as suggested by one reviewer: "The only modifications to the methodology would be to consider extending the stock assessment-modeling framework to become a two-area or multi-area model, such as with the VPA Two-Box method developed recently by Clay Porch in the SEFSC Miami Lab." *McAllister, page 6*.)

Alternative assessment models have been, and continue to be employed as part of the benchmark assessment peer review process at NEFSC. The alternative model results presented to the groundfish peer reviewers in February 2003, however, were not produced in an open scientific

forum as described above, thus their relevance could not be immediately ascertained to the level of rigor and transparency to which the VPA results have been previously subjected. The reviewers were therefore, unable to properly evaluate why the results differed from the NEFSC assessment results, causing one reviewer to state: "If configured similarly, one might reasonably expect the two approaches to yield similar results" (*Cook, page 9*).

The reviewers also considered it important that consistency be maintained between models used to determine stock status and those used to derive biological reference points. This was an important reason for revising reference points (NEFSC 2002a), since inconsistencies in reference point determination and stock assessment methods were numerous. The assessment method employed at NEFSC provides direct output of spawning stock biomass and recruitment which are incorporated in the estimation of biomass and fishing mortality reference points and for projections. As one reviewer noted: "Most of the stock assessment and projection methodologies currently applied by the NEFSC provide an adequate scientific basis for fisheries management. The ADAPT VPA and Age-pro methodologies provides a rigorous and adequate basis for assessing stock biomass, and fishing mortality rates, doing projections, evaluating the differences in potential consequences of alternative possible fisheries management policies, and for taking into account parameter and important model structure uncertainties". *(McAllister, page 6).*

References

Anon. 1988. Report of the Workshop on Methods of Fish Stock Assessment, Reykjavik, 6-12 July, 1988. ICES CM 1988/Assess:26

Anon. 2002a. Report of the ICES Advisory Committee on Fishery Management, 2002. ICES Cooperative Research Report No. 255.

Anon 2002b. NAFO Scientific Council Reports, 2002.

Brodziak, J. 2002. In search of optimal harvest rates for west coast groundfish. North American Journal of Fisheries Management. 22:258-271.

Brodziak, J.T.K., W.J. Overholtz, and P.J. Rago. 2001. Does spawning stock affect recruitment of New England groundfish? Canadian Journal of Fisheries and Aquatic Sciences 58(2): 306-318.

Brown, B.E., J.A. Brennan and J.E. Palmer. 1979. Linear programming simulations of the effects of by-catch on the management of mixed species fisheries off the northeastern coast of the United States. Fishery Bulletin (U.S.) 76(4): 851-860

Burnham, K.P., and D.A. Anderson. 1998. Model selection and inference: a practical information theoretic approach. Springer-Verlag, New York.

Carlin, B.P., and T.A. Louis. 2000. Bayes and empirical Bayes methods for data analysis, 2nd edition. Chapman and Hall, New York.

Congdon, P. 2001. Bayesian Statistical Modelling. John Wiley, New York

Daan, N. 1983. Analysis of the cod stomach samples during the 1981 stomach sampling program. ICES C.M. 1983/G:61.

De Valpine, P., and A. Hastings. 2002. Fitting population models incorporating process noise and observation error. Ecological Monographs. 72:57-76.

Kass, R.E, and A.E. Raftery. 1995. Bayes factors. Journal of American Statistical Association. 90:773-795

Murawski, S.A. 1984. Mixed-species yield-per-recruitment analyses accounting for technological interactions. Canadian Journal of Fisheries and Aquatic Sciences 41(6): 897-916.

Murawski, S.A., and J.T. Finn. 1986. Optimal effort allocation among competing mixed-species fisheries, subject to fishing mortality constraints. Canadian Journal of Fisheries and Aquatic Sciences 43(1): 90-100.

Murawski, S.A., A.M. Lange and J.S. Iodine. 1991. An analysis of technological interactions

among Gulf of Maine mixed-species fisheries. ICES Marine Science Symposia 193: 237-252.

Murawski, S.A., P. J. Rago and EA. Trippel, 2001. Impacts of demographic variation in spawning characteristics on reference points for fishery management. ICES Marine Science Symposia 58: 1002-1014.

NEFSC. 2002a. Final Report of the Working Group on Re-evaluation of Biological Reference Points for New England Groundfish. Northeast Fisheries Science Center Research Document 02-04. 249 pp.

NEFSC 2002b. Report of the 35th Northeast Regional Stock Assessment Workshop (35th SAW), Consensus Summary of Assessments. Northeast Fisheries Science Center Research Document 02-14. 259 pp.

NEFSC 2002c. Assessment of 20 Northeast Groundfish Stocks through 2001. A report of the Groundfish Assessment Review Meeting (GARM), Northeast Fisheries Science Center, Woods Hole, Massachusetts, October 8-11, 2002. October 2002. Northeast Fisheries Science Center Research Document 02-16. 511 pp.

NEFSC 2003. Report of the 36th Northeast Regional Stock Assessment Workshop (36th SAW), Consensus Summary of Assessments. Northeast Fisheries Science Center Research Document 03-XX. xx pp.

NMFS 2001. Marine fisheries stock assessment improvement plan. Report of the National Marine Fisheries Service National Task Force for Improving Stock Assessments. Second Edition (revised). U.S. Department of Commerce, NOAA Tech. Memo. NMFS-S-F/SPO-56. 69 pp.

NRC (National Research Council) 1998a. Improving fish stock assessments. National Academy Press, Washington, D.C., 177 pp.

NRC (National Research Council) 1998b. Review of Northeast fishery stock assessments. National Academy Press, Washington, D.C., 128 pp

Overholtz, W.J. 1985. Managing the multispecies otter trawl fisheries of Georges Bank with catch optimization methods. North American Journal of Fisheries Management. 5: 252-260.

Overholtz, W.J., and S.A. Murawski. 1985. A preliminary assessment of management options for the Georges Bank multispecies trawl-fisheries with special reference to haddock and yellowtail flounder. NEFSC Woods Hole Lab. Ref. 85-08.

Overholtz, W.J., S.A. Murawski and K.L. Foster. 1991. Impact of predatory fish, marine mammals, and seabirds on the pelagic ecosystem of the northeastern USA. ICES Marine Science Symposia 193: 198-208.

Overholtz, W.J., S.F. Edwards, and J.K.T. Brodziak. 1995. Effort control in the New England

groundfish fishery: a bioeconomic perspective. Canadian Journal of Fisheries and Aquatic Sciences 52(9): 1944-1957.

Restrepo, V.R. (ed.). 1998. Analyses of simulated data sets in support of the NRC Study on Stock Assessment Methods. NOAA Technical. Memorandum NMFS-F/SPO-30

Schnute, J. T., and L. J. Richards. 2001. Use and abuse of fishery models. Canadian Journal of Fisheries and Aquatic Sciences. 58:10-17.

Trippel, EA. 1999. Estimation of stock reproductive potential: History and challenges for Canadian Atlantic gadoid stock assessments. Journal of Northwest Atlantic Fishery Science 25: 61-81.

Tsou, T.-S. and J.S. Collie. 2001. Estimating predation mortality in the Georges Bank fish community. Canadian Journal of Fisheries and Aquatic Sciences 58: 908-922.

Walden, J.B. 2001. Modeling the impact of proposed Amendment13 area closures. Presentation made to the Social Science Advisory Committee, New England Fishery Management Council, May 22, 2001, Gloucester Mass, (unpublished NEFSC mimeo)

Wigley, S.E. 1999. Effects of first-time spawners on stock-recruitment relationships for two groundfish species. Journal of Northwest Atlantic Fishery Science 25: 215-218.

Table 1. Definition of models used for fitting stock-recruitment data for New England groundfishes.

Name	Stock Recruitment Relationship	Auto-regressive	Priors			
					Unfished R	
	Steepness	Slope	VPA	Hindcast		
BH	Beverton & Holt					
ABH	Beverton & Holt	Yes				
PBH	Beverton & Holt		Yes			
PABH	Beverton & Holt	Yes	Yes			
PRBH	Beverton & Holt				Yes	
PRABH	Beverton & Holt	Yes			Yes	
P2BH	Beverton & Holt		Yes		Yes	
P2ABH	Beverton & Holt	Yes	Yes		Yes	
PRHCBH	Beverton & Holt					Yes
PRHCABH	Beverton & Holt	Yes				Yes
P2HCBH	Beverton & Holt		Yes			Yes
P2AHCBH	Beverton & Holt	Yes	Yes			Yes
RK	Ricker					
ARK	Ricker	Yes				
PRK	Ricker			Yes		
PARK	Ricker	Yes		Yes		
PRRK	Ricker				Yes	
PRARK	Ricker	Yes			Yes	
P2RK	Ricker			Yes	Yes	
P2ARK	Ricker	Yes		Yes	Yes	
PRHCRK	Ricker					Yes
PRHCARK	Ricker	Yes				Yes
P2HCRK	Ricker			Yes		Yes
P2AHCRK	Ricker	Yes		Yes		Yes

Model Name Decoder

Model names were built iteratively as more analyses were conducted (For example, see table 3.1.1 for Gulf of Maine cod). To decode the model name:

1. Start at the right, the last two letters are either BH (Beverton and Holt) or RK (Ricker), which distinguish the two possible underlying stock recruitment relationships.
2. If there is an A just before either BH or RK this means that an autoregressive error term was assumed in the model.
3. All the remaining models start with a P.

4. If the P is alone except for the letters already examined this means that the model assumed a prior for the steepness parameter in the Beverton and Holt model or the slope parameter in the Ricker model.
5. If the P is followed by R (not part of RK for the Ricker model), then the model assumed a prior for the unfished recruitment from the VPA data.
6. If the P and R are followed by HC, then the model assumed a prior for the unfished recruitment that was derived from hindcast data.
7. If the P is followed by 2, then the model assumed both a prior for unfished recruitment (either from the VPA data, no additional letters, or the hindcast data, HC) and a prior for either the steepness parameter in the Beverton and Holt model or the slope parameter in the Ricker model.