

February 2003 Groundfish Science Review: Murdoch McAllister

Groundfish Science Review

Reviewer:

Murdoch McAllister

4 Hamilton House 26 Aldis Street

London, England

14 Feb. 2003

Preamble

My review consists of an executive summary and then responses to the questions posed in the terms of reference. Due to the limitations of time, I was unable to address a number of the specific questions in the depth and detail that I had wished to. I was also not able to find the time to specifically review the assessments for individual groundfish species, even the ones of greatest importance. Rather, I chose to focus on what I believed to be the areas where I could provide the most substantive comments. These were in particular with regard to key aspects of the methodologies and evaluations applied by the NEFSC to (a) address the trawl warp offset issue, (b) compute and select values for Fmsy and Bmsy, and (c) conduct projection evaluations. Due to the fundamental importance of providing a critique that was coherent and legible there was no time available to address all of the issues equally well. I was also not able to find time to address directly the written questions posed in to the reviewers at the end of the Groundfish Review and Assessment meeting in Durham Feb. 3-5. However, many of these questions will undoubtedly have been answered in various places within my review. The review is organized as follows. It begins with my Executive Summary that summarizes my main findings and recommendation. The main body of the work is my detailed review that addresses the points in the terms of reference one by one.

Executive Summary

1. Based on my review of the material that I was able to evaluate during this review process, it is my conclusion that most, but not all, of the science undertaken by the North East Fisheries Science Center (NEFSC) is by and large adequate to support the fishery management decisions that rely on this science. Most of the various conclusions reached by the NEFSC regarding the three key issues addressed by this review, according to my understanding, represent the best available science for the purpose of providing scientific management advice for the management of fisheries for New England groundfish.

2. There are a few places where some revisions to the methods applied are recommended before consideration by the NEFSC. The key revisions suggested immediately below, however, in my view are easily addressed and if a decision is made to carry out the revisions, these should be very easy and relatively quick to implement. It is conceivable that nearly the same conclusions could be reached after the suggested revisions have been carried out. However, on a stock-by-stock basis, the magnitude of differences between old and new results obtained cannot be predicted with any certainty before the recommended corrections are carried out. It is my view that if it is agreed to carry out some of my recommended revisions to the procedures, the high standard of quality control measures and scientific rigor practiced internally by the NEFSC will ensure that the new results obtained will be adequate for the purposes of fisheries management and of the highest scientific rigor achievable under present circumstances.

3. The recent trawl gear experiment and use of ten different types of analyses in the 2002 GARM Report, many of which conducted large numbers of tests using datasets for several individual stocks, taken together, provided a high likelihood of detecting any strong difference in survey catches that could arise from unequal warps being introduced in the 2000-2002 NEFSC bottom trawl surveys. The analyses were sufficient in type, number, scope, and scientific and statistical rigor to detect differences in survey catches arising from unequal warps and other survey problems. The statistical power calculations reported in the 2002 GARM report to evaluate detectable effect sizes clearly indicated this. That, in only a small fraction of the stocks evaluated small to moderate differences were statistically detected, and the direction of the difference, was not entirely consistent with a decrease in catching power in the offset warp gear, lends strong support to the conclusion that the trawl warp offsets from 2000-2002 did not systematically reduce the catching power of the trawl survey gear. In my view, the decision by the NEFSC to continue to use the unadjusted trawl survey data for New England groundfish stock assessment, is therefore scientifically justified. This conclusion does not necessarily rule out that there have been some serious changes in catching power of the gear for some of the species. However, the large number of evaluations, the rigor of the evaluations and the largely negative results obtained, give relatively little credibility to this possibility.

4. The sensitivity tests carried out to evaluate the implications of the trawl warp offsets the evaluation of stock status, and rebuilding plans adequately bounded the range of potentially introduced biases. The magnitudes of the potential biases considered were consistent with the range of potential biases, identified by the large number of evaluations carried out on the trawl data, experimental trawl data and the detectable effect size calculations.

5. In my view, most but not all of the methodologies currently used by the NEFSC to compute F_{msy} and B_{msy} provide an adequate scientific basis for fisheries management. The protocol that most seriously requires revision is the one used to evaluate the goodness of fit of the alternative stock-recruit functions to the data, and to select from the alternatives the model to determine F_{msy} and B_{msy} for the purposes of fisheries management. First, 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. Second, while there was an attempt to implement Bayes' Factor, the method of implementation was incorrect. 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 ($P(\text{data given model (i)})$), should be used to compute Bayes' factors. This is a goodness of fit measure integrated across the entire parameter space of the model, rather than at the posterior mode only, as given by the AIC. The marginal posterior probability of the data, $P(\text{data given model (i)})$, is easily computed using the AD MODEL builder software currently applied in this protocol and details are provide below to provide guidance for this calculation and for obtaining a proper interpretation the results.

6. In the protocol used to select a stock-recruit model for reference point determination, the NEFSC actually applied Bayes' factors to models that differed only in the type of prior pdf used. In contrast, the alternative models that can be compared with Bayes' factors cannot be considered to be alternative models if they differ only in the priors that are used. If the Bayesian statistical approach is to be adopted, it should be applied properly according to the body of theory already developed for its application. Thus, to be consistent with this body of statistical theory and methodology, 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.

7. In a number of different analyses conducted by the NEFSC, MCMC software was applied for statistical estimation (e.g., Brodziak 2003 and in the stock-recruit model analysis in Anon. 2002). Unless the results for different types of diagnostic methods are satisfactory, there is no adequate basis to know whether the MCMC results obtained have converged on the posterior distribution and are therefore reliable. However, no

convergence diagnostics were reported to have been applied, as they should have been for the MCMC results obtained (Gelman, et al. 1995). It is recommended that more than one diagnostic tool be applied to evaluate convergence, and that the ones provided in the WinBugs software be applied for this purpose.

8. Models with autocorrelation in recruitment should not be tested statistically against models without such correlation using the current time series of stock-recruit data. This is because the time series are too short to enable reliable statistical detection of autocorrelation. 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 F_{msy} , B_{msy} and in stock projection models.

9. The proxies for F_{msy} based on $F_{40\%}$ and $F_{50\%}$ for the various groundfish species appear to have sufficient scientific basis for the purposes of fisheries management based on considerable previous modeling and empirical work (Clark 1991, 1993; Mace and Sissenwine 1993, Mace 1994 cited in Anon. 2002).

10. 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 for use in fisheries management.

11. While the analyses and methodologies presented to this review as alternatives to the NEFSC have some conceptual and theoretical merits, the alternative methodologies were considerably less statistically rigorous, and in none of the instances provides a higher standard of science than that conducted by the NEFSC to address the scientific issues covered in this particular review. None of the analyses presented in their current form, as alternatives to the NEFSC analyses, provides an adequate scientific basis for the management of the New England groundfish fisheries. This is not to say that the alternatives are not valuable or have no future potential. Indeed some of them, (e.g., the analysis of the trawl experiment by Paul Starr), though less meticulous and rigorous, provide results largely consistent with those provided by the NEFSC. However, Starr's work and that carried out by other non-NEFSC scientists was admittedly carried out in haste, and the results provided cannot be considered to be credible for the purposes of fisheries management. Neither can they serve as a basis with which to challenge or question the credibility of NEFSC results. A great deal more effort and research would be required on these alternatives before they could possibly be considered to constitute an adequate scientific basis for providing fisheries management advice.

12. 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. 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. This modification is recommended because of the recent use of time-area closures for the management of New England groundfish fisheries. However, this extension is certainly not a necessary one since in my view the current ADAPT VPA and Age pro methods are currently adequate.

13. The ASPIC surplus production estimation and projection methodology has some serious methodological limitations. First, the software cannot currently incorporate a prior probability distribution for the intrinsic rate of increase (r) that could be derived from meta-analysis or demographic analysis, as can other more recent surplus production modeling methods (e.g., Myer and Miller 1999 and McAllister et al. 2002). This problem is a serious one because of the relatively small information in relative abundance time series typically available for jointly estimating r and carrying capacity (K). The imprecision in parameter estimates could lead to serious mis-assessments of stock status and seriously biased predictions of the stock's response to current and new management regulations. The ASPIC procedure uses only this single estimated value for r in its future projections. This procedure ignores the considerable uncertainty in estimates of r . The use of a probabilistic methodology for estimating r using demographic data or meta-analysis (Meyers et al. 1997, 1999), and a projection methodology that probabilistically accounted for uncertainty in r , should therefore be considered as a potential improvement to the current ASPIC methodology.

14. 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.

Groundfish Science Review By Murdoch McAllister

TERMS OF REFERENCE

1. TRAWL SURVEY ISSUES AND INFLUENCE ON MANAGEMENT ADVICE

Considering the results of the Groundfish Assessment Review Meeting (GARM), subsequent results from experimental trawl comparisons, and other appropriate information, provide an evaluation of the significance of potential differences in trawl survey catchability resulting from recently-discovered survey gear problems on management advice for groundfish stocks managed under the Northeast Multispecies Fishery Management Plan.

In responding, reviewers should consider the following:

- A. Are conclusions regarding use of 2000-2002 trawl survey data adequately supported by analyses reported by the GARM? Were analyses sufficient to detect differences in survey catches arising from unequal warps and other survey problems? Did the sensitivity analyses presented in the GARM report adequately bound the range of potential effects inferred from analyses of historical and comparative data? Did the GARM adequately characterize the uncertainties in estimated stock sizes and rebuilding mortality rates potentially arising from unequal warp offsets?

Each of these questions is addressed in turn below.

Are conclusions regarding use of 2000-2002 trawl survey data adequately supported by analyses reported by the GARM?

The main conclusion drawn in GARM (2002) regarding the use of trawl survey data in the stock assessment is that, "...the overall management advice is robust to variations in recent survey catch rates." Ten different studies to evaluate evidence for an intervention in NMFS trawl survey data, associated with the used mis-calculated trawl warps, were reviewed. These studies were described within the GARM and their implications for the introduction of bias in the stock assessments were evaluated. The studies, a brief summary of their findings and comments on these, are listed below.

1. Studies of rates of gear damage over time. Finding: "no significant change in frequency of trawl tows experiencing minor or major damage associated with the warp offset as compared to previous studies." Comment: Fig. 3.2.3 very clearly demonstrates that there was no meaningful difference before damage versus after,

with the proportions remaining very close to the long-term average of approximately 10% for both minor and major for the fall, spring and winter surveys.

2. Calculations of trawl geometry as a function of the warp offsets, by depth.

Finding: "Wing spread and head rope height did not vary appreciably with offsets that occurred in depths where groundfish typically occur (warp offset up to about 9 feet). "The net remained open with warp offsets up to 18 feet. Consistent trawl performance within this range of warp offsets is supported by the absence of detectable effects as indicated by the other information" considered regarding this issue. "Calculations based on geometry of the trawl in the offset condition ... and the postulated increase in the potential problem in relation to species catches-at-depth indicate that reductions on the order of 50% in trawl survey catches are implausible." Comment: Although I'm not qualified as a gear technologist, the modeling documented on p. 358-360 of the 2002 GARM report, appears to be a perfectly clear and logical way to make predictions about the potential effects of changes in trawl warp offset on changes in trawl survey catch at depth. The conclusions regarding a reduction of 50% being unlikely apply to a model that assumes that the warp offset effect is linear. The non-linear model considered predicted that reductions in catch should be considerably more evident at deeper stations. However, this prediction did not appear to be borne out by any of the empirical studies. Hence, these model results provide evidence against the hypothesis that the warp offsets caused reductions in catch in the order of 50% or more.

3. Patterns in mean/variance relationships in trawl survey catch data by stock.

Findings: "Empirical plots of catch data indicated no apparent differences in the variance compared to mean relationships for the species examined, and plots of the coefficient of variation ... of catches in numbers by survey stratum over time showed no obvious differences in pre-and post-warp offsets." There did not appear to be any meaningful changes in the survey CVs (CV Number / tow by strata) for any of the 20 species' pre-warp offset and post warp offset in figures 3.6.1-3.6.20).

4. Depth-at-capture information from pre-and post-warp misaligned cruises. Finding:

"There were no detectable differences in catch-weighted depth of capture of any species relative to the warp offset." Comment: The statistical tests carried out on pages 382-3 were appropriate, and appropriate Bonferroni correction factors were applied to adjust the alpha significance levels due to the large number of significance test carried out. The statistical power of the tests was not computed, to indicate the chance of correctly detecting a meaningful effect on e.g., catch-weighted average depth or conversely the detectable effect size in these tests. However, the reported absence of any consistent pattern in the results over the very large number of significance tests, suggests that the effects of warp offsets if any are minor.

5. Studies of the trends in abundance measures before and after the warp mis-markings. Finding: "There was no evidence for a trend in the direction of abundance

index changes associated with the warp offset, when comparing pairs of adjacent years.... While the evaluation of the changes in abundance indices is potentially confounded by underlying changes in resource abundance, the number of stock/index combinations showing positive increases in abundance was virtually identical between 1998-1999 and 1999-2000 (when the intervention was made). The abundance indices for the deepest dwelling stocks did not show differential reductions between years pre-and post-warp offsets." Comment: These results are credible and support the hypothesis that trawl warp offset had no substantial systematic cross-species effect on catching power. By themselves these results aren't entirely persuasive, but taken together with all of the other findings, they add to the weight of evidence.

6. Side-by side trawling experiments conducted by the Albatross and Delaware to estimate their relative fishing power, conducted before and after the warp mis-marking on the Albatross conducted in 2002, 1982, 1983, and 1988. Finding: "Estimates of fishing power coefficients ... were similar between vessels in experiments before and after the warp change on the Albatross IV There was only one statistically significant change in this ratio after the warp change in 10 species examined. In this one case, the ratio of Albatross to Delaware catch of yellowtail increased between the 1980s and 2002... Because these paired trawl studies were conducted simultaneously before and after the warp offset they are not confounded by underlying changes in the abundance of the groundfish stocks... For all species combined, the ratio of Albatross-Delaware catches was 0.88 before the warp offset and 0.91 after, suggesting negligible change." Comment: The analysis carried out was statistically rigorous and appropriately cognizant of the issues of statistical independence between hauls but lack of independence of species within hauls. Ten different fish stocks were used in this analysis, because there was only species where there were sufficient pairs of data were used for the analysis. The latter was an appropriate screening criterion for data to be included, although not in the analysis since it is important to ensure that the tests carried out will have acceptable statistical power. The results obtained appears to provide fairly strong evidence that across the stocks investigated, the catching power of the Albatross did not change as a result of the trawl warp offset in the 2001 and 2002 surveys. However, it does rule out the possibility that perhaps for stocks not investigated there could still be some important changes in catching power as a result of the trawl warp offset. Also, highly appropriately a statistical power analysis was carried out which indicated that the detectable effect size with the tests carried out was a difference of between 12% and 35%, if statistical power was to be 95%, alpha to be 5%, and a two-tailed test was to be carried out. It was thus appropriate to conclude that " large (greater than 40%-50%) reductions in catchability of the Albatross survey during the period of the warp offset are highly unlikely as they should have been detected."

7. Studies of standardized catch-rates from surveys conducted with mismatched warp compared to survey CPUEs from surveys with comparable spatial and temporal

coverage, and unaffected by the problem (e.g., Canadian trawl surveys and USA sea scallop surveys). In other words, the apparent trends in relative fishing power of bottom trawls used in NEFSC were computed using an index from NEFSC bottom trawl, DFO bottom trawl and NEFSC sea scallop survey data. Index trends were examined to determine if relative fishing power of NEFSC bottom trawls declined during 2000-2002 while mis-marked warps were used. Finding – "The frequency of species showing positive relative changes in abundance in Albatross surveys was nearly the same in the three years before (50%) and the three years after (54%) the warp change. For all species, the relative fishing power of the Albatross post-warp change was slightly, but not statistically significantly, greater than the comparison vessels." Comment: The analysis was rigorous in statistical methodology and this was specifically designed to maximize the information value of the analysis with the recognition of the low statistical power of statistical tests that used data for individual species only. For example, trawl survey data for twenty key groundfish species were included in the analysis and "as many species-survey comparisons as possible were included in the analysis and the statistical approaches used to analyze index trends accommodated all comparisons simultaneously because it would be difficult to detect a small or moderate size change in fishing power for any single species." Data were prepared to ensure that the comparisons were as meaningful as possible. For example, in cases where it was deemed that gear was inefficient for certain size range of a species (e.g., < 20 cm yellowtail), these size data were excluded from the analyses. Appropriate transformations of the data were carried out, e.g., log ratios of catch rates from two different surveys to produce indices of relative catching power. The standardization carried out on p. 438 was also appropriate to enhance the comparability of pre and post warp catch rates between the different surveys. In principle, the indices derived could be expected to cleverly cancel out year effects and effects. However, it was acknowledged that detected deviations from the mean of zero could be interpreted to reflect either changes in the reference survey (e.g., the DFO index or scallop index), or the NEFSC index after 2000. However, the prediction is that the catching power of the NEFSC index should lead to decreases in catching power relative to the reference surveys, since the surveys give no reason to suspect that there have been meaningful deviations in catching power before and after 2002. Due to time limitations, variance and statistical properties were not carried out. However, the overall trends in the relative fishing powers for each index were evaluated. In none of the comparisons were the predictions of results under the hypothesis of reduced relative fishing power born out. For example, the sign of the SLSCR values should become more negative with species mean depth. In contrast, "there was no obvious relationship between species mean depth and the sign of the SLSCR values during 2000-2001." Additionally, the number of species for which fishing power of NEFSC survey bottom trawls is below average should increase with the introduction of mis-marked warps. This hypothesis was not born out by the results (e.g., Fig. 3.9.1, 3.9.2); there appeared to be no meaningful change in this number before and after the NEFSC warp offset event in 2000.

8. Evaluation of evidence for difference in length distributions from survey catches pre-and post warp offset by evaluating the relative size compositions in Canadian and USA spring surveys (e.g., eastern Georges Bank). Finding: "Based on examinations of size distributions of cod and haddock, not only was there little difference in the proportions of large fish but there was little apparent difference in the entire size frequency distribution, by survey series in areas they overlap (northeast Georges Bank)." Comment: The length-frequency distributions plotted in Fig. 349 appear to be very similar between the various surveys and time periods evaluated. No discernable patterns regarding potential effects of warp offset changes in the NEFSC survey emerge based on visual inspection. However, the comparisons are made for only three, albeit important, species, cod, haddock, and yellowtail flounder. This small group of comparisons does not rule out the possibility of meaningful differences occurring for other important groundfish species.

9. Evaluations of monkfish size composition data collected on industry-based surveys and the winter 2001 Albatross survey. Finding: "Differences in the size composition of large monkfish between industry and Albatross winter surveys were minimal." Comment: This finding is very clearly demonstrated by the two very similar length frequency distributions in figure 3.3.4. However, it did appear that the central tendency of the length frequency distribution was just slightly (perhaps a few cm) less for the NEFSC winter survey. Such a difference is unlikely to be unimportant given that the central tendencies hover over about 50cm.

10. Evaluations of length compositions with data obtained by side-by-side trawling of the Albatross and Delaware in the spring 2002. Finding: "Size compositions from Albatross-Delaware paired towing experiments in spring 2002 also indicated no loss of large fish due to the Albatross warp mis-markings." Comment: Figure 3.3.5 provides strong support for this finding but again comparisons are made only for cod, haddock and yellowtail flounder.

Summary: In all ten tests, none of the results provided support for the hypothesis that the catching power for the majority of species decreased substantially as a result of the introduction of unequal warps. Rather, taken together, the results provided support for the hypothesis that only small to moderate changes in catching power, if any changes at all, could have resulted from the introduction of unequal warps. The conclusion "there is no indication of a systematic reduction in trawl survey fish catch efficiency due to trawl warp offsets" cannot be refuted by the results of any of the ten different analyses, and is well supported by the analyses. Based on the findings from the ten different sets of analyses, the GARM is well justified in its endorsement of "the nominal assessment calculations as the basis for management decision making." The conclusion that "the overall management advice is robust to variations in recent survey catch rates" is also supported by the ten analyses in the GARM for reasons stated below.

Were analyses sufficient to detect differences in survey catches arising from unequal warps and other survey problems?

Yes, as noted above, the analyses were sufficient to detect small to moderate differences in survey catches arising from unequal warps and other survey problems. For example, in the test 6 listed above, it is noted that a statistical power analysis was carried out which indicated that the detectable effect size with the tests carried out was a difference of between 12% and 35%, if statistical power was to be 95%, alpha 5%, and a two-tailed test was to be carried out. 95% is an adequate and appropriate level of statistical power, implying that if such effect sizes (12% to 35%, either positive or negative differences in catching power) had actually existed, they would have been detected correctly in 95% of the time. The use of ten different types of analyses, many of which conducted large numbers of tests using datasets for several individual stocks, taken together, provided a high likelihood of detecting any strong difference in survey catches that could arise from unequal warps being introduced in the 2000-2002 NEFSC bottom trawl surveys. The analyses were sufficient in type, number, scope, depth and statistical rigor to detect differences in survey catches arising from unequal warps and other survey problems.

Did the sensitivity analyses presented in the GARM report adequately bound the range of potential effects inferred from analyses of historical and comparative data?

Yes, the sensitivity analyses presented in the GARM report adequately bounded the range of potential effects inferred from analyses of historical and comparative data. As noted above, no meaningful differences in catching power were detected overall in the various analyses. Statistical power analyses indicated that the statistical tests carried out could have detected effects of between 12% and 35% with a statistical power of 95%. Thus, if effects had existed they should be of the order less than 10% to 35%. The choice of potential effect sizes of 10% and 25% decreases in catching power for sensitivity analyses in the VPAs and projection analyses were thus appropriate and justified. The value of 100% was outside of the range of plausible values for the effect sizes, but serves to demonstrate the effect on the analysis of a very large difference in catching power, should it have occurred. The use of only negative (and not) differences in catching power from the unequal trawl warps, and hence the corresponding positive changes in the survey abundance values and projection trajectories, is appropriate and justified since the main conjecture of concern is the possibility that the offset trawl warps caused a reduction in catching power. If it is found that offset trawl warps caused an increase in catching power for some stocks (e.g., as for yellowtail with an estimated 50% increase, Fig. 3.11.1), then it would be appropriate for sensitivity evaluations to be carried out that investigated the implications for the VPA stock assessment and projections of such an increase in catching power.

Did the GARM adequately characterize the uncertainties in estimated stock sizes and rebuilding mortality rates potentially arising from unequal warp offsets?

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.

B. Was the design and analysis of data from experimental trawl comparisons adequate to estimate the magnitude of differences resulting from the use of unequal trawl warps and other experimental treatments? Were estimates of the power of these experiments to detect statistical differences in fish catches between treatment and control survey configurations adequately described?

Was the design and analysis of data from experimental trawl comparisons adequate to estimate the magnitude of differences resulting from the use of unequal trawl warps and other experimental treatments?

The answer to this question is no, and this response is acknowledged in Fogarty (2003): "...the design of this experiment does not permit separation of the effects of the trawl warp offset from other gear characteristics.". In other words, if the intent was to estimate the magnitude of differences in catches resulting from the particular effects that could result solely from the use of unequal trawl warps, the experimental design was non-ideal. However, if the intent of the experimental design was to estimate the magnitude of differences resulting from the *simultaneous use of unequal trawl warps and other deviations from the "optimal" trawl survey gear configuration* (e.g., the "worst case scenario" or "deviant" gear), then the design and analysis of data from the experimental trawl comparisons appear to be adequate. This is explained further below. The use of three different depth strata was appropriate for the evaluation of depth-induced effects of the deviant gear configuration. The three depths used in the experiment appeared to be reasonable choices for this evaluation because they are representative of three typical depth zones at which the gear operates, they are expected to include the different suites of species found at the different depths typically surveyed, and the differences in depths are large enough to be able to detect the differences in "deviate" gear performance that could be expected to occur at different depths.

The use of the F/V Sea Breeze for side-by-side tows served as an appropriate "control" to help to estimate the differences between the two types of gears evaluated. This is in the respect that side-by-side tows by the Sea Breeze when available would act as a simultaneous baseline control, despite the basic differences in catching power between the two vessels. The use of trawl mensuration equipment and a camera mounted on the head-rope of the trawl was appropriate in order to qualitatively and quantitatively assess gear performance at depth. However, the use of the camera could potentially have modified gear performance and altered the potential effects of the deviant gear configuration with respect to the ideal survey gear configuration. The use of some closed areas for some of the spatial sampling sites was appropriate because it helped to ensure that fish densities were sufficiently high to test gear performance when species of interest were present in reasonable densities.

The protocol for randomized blocking of pairs of tows between the control and treatment nets appears to be sound and to control for the effects of time of day and tide, which can strongly affect gear performance and species catchability. The only downside of implementing the pair of tows 25 hours apart is that fish are typically mobile and move around in clumps of similarly aged individuals. Thus, the weather conditions, size composition and even the species composition could shift quite dramatically from one day to the next and give rise to very different catch compositions for the two tows within the randomized block. The occasional side-by-side tow between the Albatross and Sea breeze during the experiment should provide a means for comparing between day variability, and within day variability in catches at a give location, and to quantify the amount of between day variance in catches per site.

The use of increasing trawl warp offsets between port and starboard at increasing depths is appropriate because this captured the expected effects of depth on the difference in trawl warps between the starboard and port trawl warps. However, with the experimental design used, it does not appear to directly distinguish *the effects of increased trawl warp offsets* with *the effects of other features of the deviant gear*. The potential "primary" effects of differences in trawl warp between the starboard and port sides of the deviant net were confounded with potential "secondary" effects of depth on other aspects of gear performance. The potential modifications to effects of trawl warp offsets to match the mis-marked warps used during 2001-2002 surveys that could result from using in the deviant net cannot be known with this design. In other words, it will be not possible to evaluate how the additional modifications to the deviant gear; e.g.: (a) an already used versus new net, (b) trawl doors considered to be performing poorly, and (c) backstraps with no swivels and intentionally twisted two times versus backstraps with swivels and no twists, modified the effects of the offset trawl warps. A key feature of the design was that as depth increased, trawl warp differentials were increased in the deviant gear. However, the potential secondary effects of other non-ideal aspects of the gear rigging (a, b, c) on catch could also have been enhanced at increased depth. In a worst-case scenario, the potential "primary"

effects of the trawl warp offsets could have been cancelled by potential "secondary" effects from the other deviant aspects of the "deviant" gear configuration. An improvement would have been to implement two or three different trawl warp offsets at each depth keeping all other aspects of the gear the same for the deviant gear and then in the statistical analysis to treat the offset difference as a covariate. This would have made it possible then to directly estimate the effects on catches of different trawl warp offsets at depth and at different depths. In the view of trying to estimate the effects of trawl warp offsets on catches, it would then have been better to apply only one type of modification to the experimental treatment net, e.g.: the treatment net should have been modified to be different from the control net only in terms of the amount of the difference in trawl warp lengths on the starboard and port sides and all other aspects between the control and treatment nets kept the same as much as possible.

The statistical methodologies used by NMFS to evaluate the potential differences in catching power between the control and deviant nets appear to be appropriate for the purpose. The methodology and analysis reported in Brodziak (2002) to evaluate differences in catch rates between the control and deviant gears for each species – gear combination is suitable and adequately implemented. There are however a few issues to be consideration. A Bayesian statistical approach was implemented to analyze the data and test the hypothesis of there being no difference in catch rate between the control and treatment for each species area combination. It is mentioned that a thinning rate of two was applied to compute the various statistics required for the Welch test. It is well known that successive draws from MCMC chains can be highly correlated. Thus, it is highly unlikely that a thinning rate of two would create a set of 25,000 independent draws from the posterior. However, even if the chain correlations have not been eliminated, some correlation in chain results should be acceptable for the purposes of estimating quantities such as the posterior mean and variance. Some other thinning rates should be applied to test the sensitivity of test results to the rate of thinning. For example thinning rates of four and eight should also be applied to see whether the test results are sensitive to the thinning rate. Additionally, it is conventional to apply statistical diagnostics to evaluate whether convergence has been achieved in the set of draws taken from a Markov Chain. The WinBUGs software provides some of these diagnostics and at the very least these should have been carried out to evaluate whether convergence had been achieved. However, it was not reported whether convergence tests had been applied.

In all, 47 different Welch tests were carried out on the various species area combinations. Out of these 47 tests, 6 were found to be significant and Brodziak (2002) mentions that this is roughly 2.5 times the number expected by chance alone. Under this interpretation, each of the 47 tests was considered as independent from the rest. However, this assumption is not strictly correct, since many of the species are caught in the same hauls. In other analyses presented in the review, it has been argued that each haul may instead be considered to be statistically independent, and

that the haul should be considered the independent statistical-experimental unit. Given the lack of independence between the results for some of the species within each depth category, the number of independent tests should actually be less than 47, though it is not at face value possible to quantify the exact number of independent tests for the purpose of evaluating the proportion of tests with positive results relative to the alpha value chosen. In other studies, a Bonferoni correction factor was applied to readjust the critical values for hypothesis tests to take into account the large number of tests carried out in a single set of analyses. The Bonferoni correction was not applied in this paper, and it perhaps not due to the lack of independence in results between tests for each depth category.

The results in Brodziak (2002) nonetheless indicate that for five out of the six significant results, the deviant gear resulted in a negative effect on catch rate, supporting the idea that the deviant gear might have been less efficient for some of the species. It is appropriately acknowledged that the low number of tows per area limits the statistical power of the tests. Detectable effect sizes, however, were not computed for the various species at each depth.

The MANOVA analysis carried out in Fogarty (2002) to analyze the density estimates obtained by the two gear configurations, takes into account the lack of independence between species at a given location because it combined into a single statistical model survey results for all species and hauls at each given depth category. The results by species for each haul were appropriately modeled using a gear type effect, a species effect, an area effect and a gear by species interaction term. The univariate analyses were also appropriate and correctly implemented. The multivariate analyses conducted identified an overall gear effect and in three of the univariate analyses, a significant negative effect was found for the deviant gear. Yet with all results considered in Fogarty (2003), it was correctly concluded that the analyses did not permit the rejection of the "null hypothesis of no gear effect for the complex of regulated groundfish species".

The statistical method to test for differences in length frequency distributions between the control and deviant gear configurations was the K-S test (Nies 2003). Nies (2003) recommends that some test other than the K-S test be applied, because this test does not take into account lack of independence between individual fish caught within a single haul as it should. Distributions with large sample size will appear to be statistically different even with relatively small differences in shape. However, due to the contagious nature of fish distributions, and the covariance in age and size structure in fish schools, the apparent statistical differences from the K-S test applied to the experimental data will almost surely be spurious. Therefore, the results of this statistical analysis are questionable and should not be taken seriously. It seems unlikely however that sufficiently powerful tests using the experimental data could be found to statistically detect meaningful differences in length frequency between the two gear configurations, due to the relatively small number of tows in the experiment

and the high degree of variability in length frequency distributions obtained from haul to haul.

Paul Starr's analysis of the experimental data: Because the analysis was hastily carried out, this was a rather crude but still credible analysis. For example, it did not appear that the data were $\log(x+c)$ transformed for the ANOVA analyses as they were in the analyses conducted by the NEFSC. Additionally, the analyses did not take into account the pairing of hauls by the Sea Breeze and Albatross, as did the analyses undertaken by the NEFSC. Nonetheless, the analysis provided the same general findings as the more detailed analyses carried out by the NEFSC and none of the results obtained were at variance with those obtained by analyses carried out by the NEFSC.

Were estimates of the power of these experiments to detect statistical differences in fish catches between treatment and control survey configurations adequately described?

In Brodziak (2002), detectable effect sizes in terms of differences in catch rates were not computed for the various species at each depth. Where the null hypothesis was not rejected, it is recommended that the minimum detectable effect sizes be computed. Statistical power and detectable effect sizes were not computed in Fogarty's analysis, but the analyses were sufficiently powerful to detect a gear effect in the MANOVA so a power calculation was not necessary here. However, for the instances in which there was a failure to reject the null hypothesis in the univariate analyses, it is recommended that the minimum detectable gear effect size be computed.

C. Advise on the significance of differences in species composition and relative catch rates resulting from side-by-side tows performed by commercial and government vessels in the recent trawl experiment with respect to model- and index-based estimates of stock size and fishing mortality rates.

For several of the species, and depending on the area, the mean catch rates were considerably higher for the Sea Breeze than the Albatross as revealed by the statistics in Fogarty (2003) and Starr (2003). According to Fogarty's (2003) Table 1, in instances where both vessels caught a species in all 16 of these instances, for Area 1, the Sea Breeze had a higher catch rate. For Area 2, in 10 of the 18 instances, the Sea Breeze had the higher catch rates. For Area 3, the Sea Breeze had higher catch rates in twelve of the 17 instances. Thus, in two of the three areas, the Sea Breeze had predominantly higher catch rates than the Albatross. The higher catch rates for the Sea Breeze was most pronounced for a number of the skate species, such as barndoor skate. However, for the commercially most important fishes, such as cod and haddock, the differences in catch rates, irrespective of the area, were in the order of half or double with the Albatross having higher catch rates in a number of instances.

With regards to differences in species composition between the two vessels, the Albatross tended to catch fewer species but the species missing tend to be those that are commercially unimportant. In Area 1, out of a total of 22 species, for the Albatross control net, six species were absent while for the Sea Breeze, one species was absent. In Area 2, six species were absent for the Albatross control net and four for the Sea Breeze. In area 3, four species were absent for the Albatross control net and two for the Sea Breeze. Thus, for some of the commercially less important groundfish species, such as the skates, the Albatross is markedly less efficient than the Sea Breeze and had a larger number of absent species than the Sea Breeze. This indicates that for some of the commercially less important species, the catchability of the Albatross survey vessel is considerably lower than the Sea Breeze. It is conceivable that when catchability becomes extremely low, then a trawl survey may fail to index trends in abundance for that species because capture events may be reflecting random events in, e.g.: species distribution and gear performance, rather than trends in abundance. If it is important to track trends in abundance of these less important species, then it would be appropriate that some other survey methodology be designed that could reliably track trends in abundance for these species.

However, if the catch rate were only half or a quarter, possibly even a tenth as much for the survey vessel as the commercial vessel, it could still be the case that the trawl survey still serves as a reliable index of relative abundance. Providing that under even low levels of abundance, the species appears regularly in the trawl tows, the survey is carried out in a consistent manner over time, and survey catchability does not change systematically over time, then even with very low catchability, the survey should still provide a reliable index of abundance over time and be useful in stock assessments – either model based or index-based.

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. For example, for a given coefficient of variation (CV) for positive observations, how high should the proportion of zeros in the trawl survey hauls be, before it could be concluded that the trawl survey protocol used does not provide a reliable index of abundance? To what extent does this vary with the statistical model used to model the trawl observations? For example, will a delta-lognormal model be more robust to high proportions of zeros than a simple log ($x+c$) model where x is the observed haul density and c is a constant? I also recommend that simulations using age-structured operating models be undertaken to evaluate the reliability of indexed-based estimates of stock size and fishing mortality rates for a variety of scenarios that cover plausible combinations for the survey CV for positive observations and proportion of zeros in trawl survey catches.

Additionally, I recommend that the standard gear used in the current NEFSC groundfish trawl survey design not be altered since it continues to provide credible relative indices of abundance for many of the groundfish species and a valuable time

series of other biological data. The survey has provided a device for measuring relative abundance of many different species and the specifications of the sampling protocol have been kept reasonably constant over a long period of time. When changes have occurred to the protocol, rigorous scientific tests have been applied, to ensure that the survey can continue to provide reliable time series of relative abundance. 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.

D. Comment on the precision of model-based calculations of stock size and fishing mortality rates in relation to variability in trawl survey catches and other sources of information included in assessments. Are the methods used for incorporating uncertainty into management advice sufficient? How should other sources of uncertainty (e.g., model selection, estimates of total removals) be incorporated?

Comment on the precision of model-based calculations of stock size and fishing mortality rates in relation to variability in trawl survey catches and other sources of information included in assessments.

It is presumed that term "variability" used in this term of reference is intended by the authors of it to mean error variability, or variation in survey indices other than the variability that represents actual variation in fish abundance. The precision of model-based calculations of stock size and fishing mortality rates reported in the 2002 GARM report are comparable to the precision in many other modern fisheries stock assessments, for example, those carried out by ICES and ICCAT. The precision in the stock assessment calculations takes into account variability in the trawl survey catches in a standard manner, common to other fisheries stock assessments. For example, a conditional non-parametric bootstrap was applied to quantify uncertainties in the VPA stock assessments. This is a rigorous and commonly applied statistical methodology to quantify uncertainty in stock abundance and fishing mortality estimates that results from variability in the input data. Error variability in the catch-age data is taken into account in the conditional non-parametric bootstrap because, in effect, the conditional non-parametric bootstrap procedure randomly re-samples these data together with the trawl survey-derived abundance indices.

The potential effect of any changes in error variability in trawl survey indices is to some extent dampened in model-based assessments that use catch age data because the models are constrained to incorporate the information in the catch-age data also. Thus, the year class strength patterns in the catch-age data will be reflected in the model output, even in the face of error variations in the trawl survey abundance indices. However, in the last few years modeled, the catch data have less influence

because the amount of the catch age data used to determine cohort strength decays as the current year is approached. If error variability in the trawl survey indices increased over the last few modeled years, the increase could render the assessment of current abundance and fishing mortality rates less reliable and could reduce the precision in stock biomass and fishing mortality estimates.

However, it should be noted that the 2002 GARM conducted analyses of the annual trawl survey abundance index CVs (standard deviation divided by the mean) before and after the trawl warp offsets were introduced in 2000. It was found that the CVs did not systematically increase or decrease with this event. Therefore it can be concluded that the error variability in the trawl survey indices did not systematically increase with the introduction of the trawl warp offset in 2000. These results, taken together with the other trawl warp offset analyses reported in the 2002 GARM report and the recent trawl experiment suggest that the reliability and precision of model-based estimates of abundance and fishing mortality rates, should be not be diminished as a result of the introduction of the trawl offset in 2000-2002.

Based on my review of the stock assessments and understanding of the methodology, I have no reason to doubt that the precision of model-based calculations of stock size and fishing mortality rates use appropriate methodology to take into account variability in the trawl survey results.

Are the methods used for incorporating uncertainty into management advice sufficient?

It is presumed that the term "management advice" in this term of reference is intended to mean stock assessment results in the form of estimates of stock biomass and fishing mortality rates, and stock projections under alternative fishery management control options, e.g., stock rebuilding plans. "Management advice" is also taken to mean fisheries scientists' interpretations of the stock assessment results for the purposes of advising fisheries managers about trends in abundance and fishing mortality rates and the potential consequences of alternative fisheries management measures that could be implemented. One of the key methods used by the NEFSC for incorporating uncertainty includes the conditional non-parametric bootstrap. The method is designed to account for uncertainty in stock size and fishing mortality estimates that result from sampling error variability in the catch-age and trawl survey abundance indices. This method could be argued to be sufficient for incorporating uncertainty from random error variability in the data.

The methods for incorporating uncertainty used by the NEFSC also include stock assessment modeling of alternative scenarios for key stock assessment model assumptions. In the 2002 GARM this method was applied to account for uncertainty in stock assessment results due to potential changes in trawl survey catching power with the introduction of the trawl warp offsets in 2000. As mentioned above, the

scenarios chosen appear to adequately bound plausible levels of change in catching power, as indicated by the multiple modeling and empirical studies that have been carried out. This scenario-based method has been applied to account for uncertainty in e.g., assumptions about future recruitment, and the rate of natural mortality. It could also be applied to assess uncertainty in management advice that could arise from uncertainty in estimates of discards (e.g., Chen 2003). It is a very clear and simple approach for accounting for and conveying uncertainty in fisheries management advice arising from uncertainty over basic stock assessment model assumptions, and could be argued to be sufficient for accounting for and conveying the effects of uncertainty in key stock assessment model assumptions on fisheries management advice.

Yet another method that the NEFSC has applied for incorporating uncertainty in management advice is the use of retrospective analysis. Retrospective analysis effectively models the e.g., recruitment estimates and fishing mortality estimates, for the last several years (e.g., 3, 5, 7 years from the present year) of the assessment when the stock assessment data (e.g., catch-age and biomass indices) for previous years (e.g., the last 3, 5, 7 years) are successively removed from the assessment and projection model. This analysis is particularly useful for pointing out potential stock assessment biases if consistent retrospective patterns occur. For example, a negative bias in projected fishing mortality rate (F) might be suggested if, under a given catch removal policy, the projected Fs from a seven year-from-present cut-off in data were much less than those from a 5-year cut-off, and the projected Fs from a 5-year cut-off were much less than those from a 3-year cut-off, and the same pattern occurred with a comparison of a 3-year cut-off and the present assessment with no cut-off in data. Because retrospective analysis provides a powerful diagnostic tool to detect retrospective patterns, it may also be employed to diagnose the degree of reliability, and hence uncertainty in fishing mortality and recruitment estimates provided in the current and recent assessments. Retrospective analysis is thus yet another tool that the NEFSC apply to help fisheries scientists interpret the degree of reliability, and hence uncertainty in present and recent stock assessment results and management advice.

The Agepro stock assessment modeling software that is used to evaluate the potential consequences of alternative fisheries management options provides a methodologically consistent means to incorporate the uncertainties modeled in the ADAPT VPA estimates of age structure, fishing mortality rates and potential changes in trawl survey catchability with the introduction of the 2000 trawl warp offset, and to project these into the future under the various management options considered. This software is sufficient for accounting for uncertainty in management advice derived from it because it utilizes the conditional non-parametric bootstrap, and has been applied for retrospective analysis and scenario-based modeling of key stock assessment model assumptions. Brodziak explained in his presentation at the review that the Agepro methodology does not explicitly account for implementation

uncertainty, e.g., uncertainty in how successfully and accurately the fisheries management policies adopted will eventually be implemented. In fact stock assessment models used in most stock assessments don't explicitly account for implementation uncertainty. However, it is conceivable that Agepro could be developed further to model implementation uncertainties. This use would be advisable if further research suggests that doing so could increase the accuracy of model projections from Agepro. The methodologies applied by the NEFSC to take into account uncertainty in the provision of management advice thus appears to be largely sufficient to account for the various important sources of stock assessment uncertainty.

The NEFSC has also recently adopted methods to incorporate model selection uncertainty in the provision of management advice. This strategy comes mainly in the form of protocols to take into account uncertainty over the structural form of the stock-recruit function used to do model projections. This protocol is evaluated in detail under the next section biological reference points. There are some potential improvements to the details of this protocol that could be implemented that would enhance its statistical consistency and rigor. With these minor improvements, it could be argued that the NEFSC's method to take into account uncertainty in the form of the stock-recruit function is also sufficient.

How should other sources of uncertainty (e.g., model selection, estimates of total removals) be incorporated?

As argued above, the NEFSC already employs methods capable of incorporating uncertainty in model selection, and data inputs such as uncertainty over estimates of catch removals. It is sufficient that these additional sources of uncertainty be incorporated in the scenario-based modeling approach that is already routinely applied to take into account uncertainty in the development of management advice from stock assessment modeling.

2. BIOLOGICAL REFERENCE POINTS

Review the fishing mortality and biomass targets and thresholds established for the 20 groundfish stocks included in the Northeast Multispecies FMP. Consider the adequacy of technical analyses supporting estimates of F_{MSY} , B_{MSY} or their proxies, as provided in the *Report of the Working Group on Re-Estimation of Biological Reference Points for New England Groundfish Stocks* (the “Report”). Comment on issues related to the simultaneous achievement of B_{MSY} values for the groundfish complex.

In responding, reviewers should consider the following. Of particular note, the NEFMC's Science and Statistical Committee recommended that additional work was needed “...specifically to explore the implications of the uncertainty in the stock recruitment

relationship.” For this reason, more specific questions are included in order to add clarity to the issues to be addressed by the reviewers.

A. Comment on the technical basis for the estimation of F_{MSY} and B_{MSY} , and choices regarding the use of parametric (Beverton-Holt, Ricker, other candidate models, etc.) and non-parametric stock-recruitment relationships applied to yield per recruit estimates, surplus production models, or proxies for biomass and fishing mortality rate targets and thresholds.

According to the most recently formulated guidelines for developing reference points for fisheries management for New England groundfish (Anon. 2002), age-structured modeling methods and data are recommended to be the methods of choice to estimate F_{msy} and B_{msy} for each groundfish stock. These methods may be used providing that there are sufficient age-structured information and data to permit the estimation. Previously, a non-age-structured surplus production model called ASPIC had been applied to the trawl survey time series of biomass indices for many of the stocks. The reasons for moving to age-structured models as the method of choice are clearly outlined, and by my judgment are adequately justified in Anon. (2002). However, in some instances, reference points may still be estimated using a surplus production model. This model may be used if the recruitment variability is judged to be relatively low for the stock of interest, and if the time series of trawl survey abundance indices is judged to be long enough and with sufficient contrasts over time (e.g., winter flounder, Anon. 2002).

Where time series of stock and recruitment data are judged to be reasonably informative for the estimation of a stock-recruit model, ADAPT VPA estimates of recruitment and stock biomass are applied to estimate the parameters of the stock-recruit functions used for the estimation of F_{msy} and B_{msy} . This procedure was elaborate and used some sophisticated approaches for estimation and model selection. In fact there were a large variety of stock-recruitment model functional forms evaluated for each of the stocks with age-structured data. These model forms included various permutations of the Ricker and Beverton-Holt models, and models with and without autocorrelation in the stock-recruit function residuals. There were also permutations that included the use of, or exclusion of, prior pdfs for the average recruitment at unfished long-run equilibrium. The prior was established by taking the average and standard deviation of the largest recruitments in the time series or hind-casted recruitment estimates (Anon. 2002); that is estimates of recruitment that were obtained from the back-projected cohorts at the beginning of the VPA time series. In addition, where available, priors derived from meta-analyses of other similar stocks, were used for the Ricker alpha and Beverton-Holt steepness parameters (e.g., from papers by Myers et al. 1999). My comments on this protocol are as follows:

- While it appears to be perfectly reasonable to try to evaluate whether autocorrelation is present in the residuals from a fit of a stock –recruit function to

data, time series statistical theory suggests that estimation of autocorrelation patterns cannot be reliably obtained from a time series that is less than approximately 50 years long. Yet in all instances, the time series of VPA-derived stock-recruit data are much shorter than 50-years. Chris Legault confirmed that in all cases, the autocorrelation model was rejected. The problem is that there could easily have been false negatives. Due to the relative shortness of the time series, a meaningful autocorrelation could really have been present, but the short time series and observations errors in the data could not enable for this to be statistically detected. Thus, autocorrelation was excluded from all baseline estimates of B_{msy} and F_{msy} . In contrast, in many other assessments sensitivity tests are conducted to evaluate the sensitivity of important stock assessment model quantities such as key biological reference points to autocorrelation in recruitment residuals, even if it cannot be statistically detected in the historical time series. In projections, it is commonly found that when autocorrelation is included in recruitment residuals, where the biological risks and risks to the fishery are higher, and the F_{msy} is lower than in calculations when it is not included. This is due to the occurrence of series of negative residuals. In some stock assessments, e.g., at ICCAT, autocorrelation in recruitment residuals (e.g., with a value for the autocorrelation coefficient of 0.5) is assumed in the base case, even if it has not even been statistically detected in statistical analyses of historic data. The reason to presume autocorrelation in recruitment residuals has often been under the acknowledgment that patterns of environmental determinants of recruitment strength (e.g., oceanographic current patterns) where there are long time series of precise observations are very commonly found to be auto correlated. It is thus recommended that even if autocorrelation is not detected statistically, that calculations of B_{msy} and F_{msy} be evaluated for their sensitivity to plausible values for autocorrelation in recruitment residuals. It is also recommended that unless the stock-recruit time series is at least 40 years or some to be agreed acceptable length, that there be no attempt to statistically detect autocorrelation in stock-recruit model residuals. If such an exercise is to take place, then only the posterior distribution for the autocorrelation at lag 1 should be computed and evaluated. The autocorrelation model should not be formally compared with stock-recruit model forms without it, since with a short time series, the model with the autocorrelation will almost surely be rejected even if autocorrelation is actually present in deviates from the "true underlying long-term stock-recruit function". The issue of whether to assume some form of autocorrelation in baseline estimates of F_{msy} and B_{msy} needs to be formally addressed and that all available biological and environmental and oceanographic information be consolidated and reviewed to develop recommendations about the consideration of autocorrelation in baseline calculations of F_{msy} and B_{msy} and also in the AgePro projection model.

- 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.

- A Bayesian approach is applied to quantify parameter uncertainty and to quantify the relative plausibility of alternative stock-recruitment model forms. The Bayesian approach is coming to be the method of choice in fisheries stock assessment and is highly appropriate for quantifying uncertainty in stock-recruit parameters and for utilizing information from meta-analyses of similar stocks to provide probabilistic information on key parameters when the data at hand aren't particularly informative about those parameters. It is also becoming more commonly applied in stock assessment to evaluate the relative plausibility of alternative mathematical forms for stock assessment models given the available stock assessment data. I have however some concerns about the technical and conceptual aspects of the methods applied, particularly with the methods for evaluating the goodness of fit to the data of the parametric stock-recruitment model. These concerns are as follows.

(1) AIC was applied to evaluate the relative goodness of fit of each model alternative to the stock-recruit data. Yet the measures of goodness of fit included both a likelihood function and Bayesian prior pdfs for the stock-recruit model parameters. AIC is designed specifically for frequentist statistics where only the likelihood function of the data (and possibly some boundary constraints) is included in the statistical measure of model goodness of fit. I have not yet seen a published work that demonstrates that AIC is also appropriate for Bayesian statistical models. The interpretation of the AIC result for Bayesian statistical models is not straightforward and questionable. There are added complications with what's considered to be "alternative models" where goodness is to be cross-compared, and in some "model alternatives" non-informative priors are applied and, in the otherwise same model, informative priors are applied for the same parameter. As indicated above for each given stock, "24 different" stock-recruit models were evaluated. In this set there were for example Ricker models with and without an informative prior for slope and some models with and without informative priors for Unfished R. There were also some alternative priors for unfished R that were compared as "alternative models" with the alternative priors coming from either direct VPA estimates or from the mean and SD of Hindcast values. Take for example two "alternative" models with and without an informative prior for Ricker slope. If the informative prior for slope is consistent with the upper quartile of the posterior given by the data and a non-informative prior for the slope, the value obtained for the likelihood function will be lower than for the instance with a non-informative prior for slope. A Bayesian interpretation of this result would be that the data are updating the prior to some

extent, but it is unclear how correct either the prior or data are. Yet it is understood that the posterior result is an improvement on either the informative prior by itself, or the likelihood function of the data used together with a non-informative prior. The paradox of AIC would be that the result with the non-informative prior would be given a lower AIC rating, when the basic Bayesian interpretation is that the result with the informative prior should be the superior result.

Because AIC is designed specifically for Frequentist statistics, and not Bayesian statistics, I recommend that AIC not be used for model evaluation purposes in the Bayesian statistical modeling. Instead, I recommend that a conventional Bayesian statistical criterion for model comparison be used instead. This is not BIC, or the Bayesian information criterion. The BIC is analogous to the AIC but makes some different adjustments according to the number of parameters and data points. Most importantly, and rather paradoxically, the BIC is actually designed for use in frequentist statistical models only and like the AIC, is not designed for use in Bayesian statistical models where prior pdfs are included together with the likelihood function in the statistical model. I recommend instead that Bayes' factors be computed and applied instead (Raftery and Kass 1995; Parma 2002). If MCMC is used, this involves computing the harmonic mean of the likelihood function from the converged MCMC chain (Parma 2002 did this using the AD Model builder which was also used by the NEFSC). This gives an estimate of the Bayesian probability of the data integrated across all possible parameter values for each model alternative ($P(\text{data given model } i)$). The ratio of $P(\text{data given model } 1) / P(\text{data given model } 2)$ is called Bayes' factor (BF) and this is the conventional Bayesian measure of goodness of fit that is designed for evaluation of Bayesian statistical models. It is recommended that Bayes' factor be calculated for each model alternative for a given set of stock-recruit data with the simplest model treated as the reference model. If the value of Bayes' factor becomes very high for the alternative model, e.g., over a value of 19, then it might be argued that there is strong evidence for the alternative model over the reference model. E.G.: this would imply a Bayes' posterior probability of 95% for the alternative model and 5% for the reference model. In conventional statistics where an alpha of 0.05 is common as a reference point for hypothesis testing (giving a 5% rate for false positives), this BF of 19 would be analogous and consistent with NEFSC choice of $\alpha = 0.05$ for other analyses addressed by this review. Any Bayes' factor value between 1 and 19, would simply imply that though the evidence tips favorably for the alternative model, it could easily be due to random chance in the observation process. The use of Bayes' factors to decide upon a model for reference point calculation should be given closer attention.

(2) It appears to be questionable in a Bayesian sense to calculate Bayesian posteriors (or Bayes' factors) for alternative models where it is simply the prior pdf for a given model parameter in the same model that is being changed. This

was done for each set of stock-recruit data. With Bayes' posteriors for alternative models, it is not the Bayes' posterior for the same model but with different priors that is of interest to compare. It is instead, the Bayes' posterior for each structurally different model, e.g., Beverton-Holt vs. Ricker. Here, the chief comparison should be between the alternative models, not the same, e.g., Beverton-Holt, model that has different prior pdfs for its steepness parameter. It is recommended instead, that where an informative prior for the slope or steepness parameter is available for the stock of interest, e.g., from the Myers et al. (1999) meta-analysis work, that this be used as the baseline prior for the slope or steepness parameter. The most justifiable prior for the average unfished recruitment parameter should also be chosen beforehand. Alternative priors for each parameter e.g. for the average recruitment, should only be used as trial values to test the sensitivity of the Bayes' factor for models 1 and 2 to the priors chosen for key parameters.

- An MCMC algorithm is applied to estimate the posterior distribution for the stock-recruit parameters. While an apparently appropriate level of thinning is applied – 1 draw in every hundred, the length of the burn-in period was not reported, as it should have been. The degree of autocorrelation remaining in a typical run with 1 in every 100 draws taken should be indicated to verify that autocorrelation in the resulting samples was sufficiently small. 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). The application of such diagnostics are essential to diagnose whether the results obtained from MCMC have converged on the posterior, and at least a few different types of diagnostics tests should be applied for this. WinBUGS, though not used in this application provides some very useful and reliable diagnostics for this purpose. 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.
- The ASPIC (Prager 1994, 1994) surplus production model is used in some instances to compute B_{msy} and F_{msy} . Problems are appropriately pointed out with the use of surplus production models, e.g., for growth overfished populations or populations with high variability in recruitment. However, where this is not the case, there will still be estimation problems if there is only a one-way trip in the data (McAllister et al. 2002). A Bayesian approach to computing a prior for the intrinsic rate of increase, r , will help in such situations, providing that estimates of life history parameters are available. The approach taken by Myers et al (1997, 1999) to compute the intrinsic rate of increase from life history parameters could also be applied using a probabilistic framework. Thus, wherever there are declines in relative abundance indices, or an instance with an increase as result of reduced catches, the use of the informative prior for r in the

Bayesian statistical modeling of the surplus production model (Myer and Millar, 1999) should produce more reliable results and be used instead of the Frequentist approach that is currently applied that relies (perhaps unreasonably) in there being information in the data about both r and the carrying capacity (K). An informative prior for r could improve reference point estimation considerably in cases where the population is not growth overfished and the biomass trend has some predictable response to a change in catch removals. Thus, I recommend that where appropriate, a Bayesian surplus production model be applied to estimate B_{msy} and F_{msy} that incorporate a prior for r derived from demographic data. The Bayesian state space modeling approach in Myer and Miller (1999) also offers an elegant approach to dealing with both process error and observation error within the same statistical model. However, while an attractive approach, it is not essential that it be adopted since a deterministic model with observation error has often been found to perform adequately even when put through the litany of simulation tests undertaken by the IWC.

- The technical basis for non-parametric stock-recruitment relationships applied to yield per recruit estimates to estimate a B_{msy} proxy appears to be technically sound. To the best of my knowledge, the use of $F_{40\%}$ as a proxy for F_{msy} for most stocks and $F_{50\%}$ for redfish appear to be well supported by the existing literature (Clark 1991, 1993, Mace and Sissenwine 1993; Mace 1994; Dorn 2002 – cited in Anon. 2002). My recommendation that could help to objectively test the validity of the non-parametric methodology would be to simulation test the non-parametric methods using an operating model approach similar to that used by the IWC. This would be to test the potential bias and imprecision in this methodology to come up with a B_{msy} proxy. Some various underlying stock-recruit model function forms could be chosen for the operating model, such as the Beverton-Holt, Ricker, or two-step or two-line "non-parametric" models. Recruitment variability could be simulated to replicate observed error variation; autocorrelation in recruitment residuals could also be modeled. Particular values could be chosen for growth, selectivity and natural mortality. Thus, the true underlying "operating" model of the system could be known and known input values chosen for the operating model parameters. The operating model and parameter values chosen would then be applied to give the "true F_{msy} " and " B_{msy} ". Following this, "observed" stock-recruit datasets could be generated using this "operating model" and "observation error" model. The accuracy of the non-parameteric B_{msy} proxy estimation procedure together with the various F_{msy} proxies could then be simulation tested for their accuracy and precision. This could thus help to inform scientists about the reliability and potential risks of applying the non-parametric methods for the purposes of fisheries management the New England groundfish stocks.
- The technical basis for the index-based assessments, to identify relative F threshold and relative F target, from my brief review of the method as described

in Anon. (2002) appears to be satisfactorily derived and to produce potentially useful results for the purpose of guiding fisheries management decisions with respect to these reference points. However, my recommendation again would be to simulation test the protocol using an operating model approach as suggested for the bullet point above. This would be to test the accuracy and precision of the estimates of the management reference points, and current status with respect to them. The motivation for this suggestion is the notion that no estimator of fisheries management procedure should be implemented unless it has been thoroughly simulation tested using an operating model approach (similar to that used in the IWC) and found to perform adequately well under the conditions foreseeable for the fishery resource of interest.

- Comment on Butterworth et al.'s (2003a,b) contributions: The age-structured surplus production modeling approach outlined by Butterworth et al. (2003a,b) appears to be sound but the none of the results obtained are credible due to the haste in which the analysis was carried out and the relatively small amount of attention given to the formulation of plausible model structures, inputs and statistical assumptions. The estimates of stock-recruit model steepness, F_{msy} , B_{msy} and Current Stock Biomass/ B_{msy} provided for both of the cod stocks (Gulf of Maine and George's Bank) for example have no credibility whatsoever for reasons given below. Due to their lack of credibility, these results also do not provide a meaningful basis with which to question the credibility of the stock assessment-modeling results obtained by the NEFSC for these two cod stocks. Issues in the Butterworth et al. papers that require scrutiny include finding appropriate approaches to deal with (1) the various inaccuracies in the time series of catch biomass data due to discarding and under-reporting, (2) modeling the vulnerability-at-age patterns in the trawl survey data, (3) the potential inaccuracies of using Pope's approximation, (4) the appropriateness of the likelihood functions used (e.g., the likelihood of the catch-age data (A.25) is atypical and ad hoc), (5) identifying an appropriate set of estimable parameters, (6) identifying appropriate values for the magnitude of the variance in stock-recruitment model deviates, (7) the abundance of each age class in the initial year of the model, (8) a suitable model for the survey constant of proportionality and fishery catchability, (9) the potential inaccuracies of temporal changes in growth rate and fecundity at age over the long time series modeled, (10) the choice of an appropriate starting year for the stock assessment model, and so on. Given the very limited contrast in the X-variate (spawner biomass) of the stock-recruit data for the Gulf of Maine and George's Bank, there is no empirical basis to reliably estimate two or three stock recruit parameters simultaneously as has been attempted. Given the lack of data anywhere near the origin for Gulf of Maine cod there is really no reliable basis to estimate the slope or steepness of the stock recruit model for Gulf of Maine cod, as has been attempted. Given the generally increasing trend in the George's Bank cod stock-recruit data, it also appears that there is little empirical basis to estimate the long-run average unfished

recruitment and to be able to discriminate statistically whether over-compensation is plausible. It appears advisable then to either (a) use an informative Bayesian prior for the slope or steepness of the stock-recruit function based on meta-analyses such as Myers et al. (1997) or (b) identify a single baseline value for steepness consistent with knowledge of cod population biology. Before any such approach could be considered as a candidate for stock assessment modeling of New England groundfish, it would need to be very thoroughly simulation tested using an operating model approach that the first author is so infinitely well familiar with. Only if the approach was found to have satisfactory estimation performance over a range of plausible detailed operating models, could it then be considered as a possibly acceptable alternative to the existing stock assessment approaches undertaken by the NEFSC for New England groundfish stock assessment.

- **Are the Working Group assumptions (growth, maturity ogive, natural mortality, partial recruitment) appropriate for estimating a B_{MSY} proxy, which establishes a minimum biomass threshold and a rebuilding target?**

Yes, to the best of my knowledge, they are. The current values are obtained under relatively low fish stock sizes. If any of these were to show marked density dependence at abundances close to B_{MSY} then the values derived from the analyses could be biased. However, there is absolutely no evidence to indicate any density dependent effects on these parameters, and it is reasonable to assume that they should not change substantially even if there are some density dependent effects on them. One thing is certain – it is extremely complicated to try to include density dependence in growth rates in a population dynamics model and there are a variety of ways of doing so. Even if it could be done, the practical utility of the results obtained would be questionable due to the lack of data to ground-truth any such exercise.

- **Comment with reference to specific species on whether the use of Beverton-Holt type stock-recruitment curves, as opposed to the use of dome-shaped (Ricker type) curves, represent reasonable scientific judgment employing sound methodology and appropriate data sources. Is there a theoretical or practical basis to detect overcompensation (Ricker curve) from the stock-recruitment curve for each groundfish species based on the magnitude of the intrinsic rate of population increase (r) and the carrying capacity (K) parameter estimates from ASPIC production models?**

The Bayesian statistical methodology outlined in Anon. (2002) provides a scientific-statistical basis for judging the appropriateness of the Beverton-Holt versus Ricker stock-recruit functions for Gulf of Maine Cod, Georges Bank Cod, and Southern New England Winter Flounder. As pointed out above, the methodology applied is an ad hoc combination of Bayesian and Frequentist statistical approaches. Due to the lack of a sound statistical-theoretical basis for the model selection procedure applied, the

biological reference point results obtained for the groundfish stocks to which the protocol has been applied are in my view questionable. Simple modifications to the procedure for model selection are suggested above to provide a statistically consistent Bayesian methodology appropriate to the Bayesian MCMC estimation procedure already adopted by the NEFCS for reference point estimation (Anon. 2002). When applied properly, this should make it possible to compute the posterior probability for the presence of overcompensation in the stock-recruit function for these three stocks, given the best available data.

It is also recommended that the simulation testing approach applied to test the reliability of the detection of the appropriate stock-recruit function be redone to use stock-recruit data that more accurately represent the range of stock-recruit observations seen for these three stocks than the ones presented in the simulation evaluation reported by Chris Legault at the CIE review meeting February 4, 2003. The simulated stock-recruit data presented by Chris Legault appear to have included stock-recruit data points nearer to the origin than are seen for example by the Gulf of Maine Cod (e.g., Fig. 3.1.7, Anon. 2002). The simulated dataset appeared to be more similar to that for Georges Bank Cod (Fig. 3.2.6), which has more observations at lower stock sizes. For the Gulf of Main cod, this potential misrepresentation in the patterns of the data could misrepresent the ability of the estimation procedures applied to detect the correct stock-recruit function. It is thus advised that the simulation testing procedure be redone to adequately represent the key features in the actual observed stock-recruit data such as the nearness of points to the origin on the spawner axis.

The simple population biology equations presented in Crecco (2003) for estimating stock recruit functions based on the Gompertz and Logistic age-aggregated surplus production models for Gulf of Main cod appear to be conceptually sound. If the parameter inputs to the equations, e.g., r , and K could be estimated accurately, and the uncertainty in the parameter estimates and other input values to the calculations taken into account in the calculations, then this approach could potentially provide an alternative basis to estimate form of the stock recruit function for Gulf of Main Cod and other species and to evaluate the degree of over-compensation in the stock-recruit function. However, in all of the calculations applied in Crecco (2003), the inputs are treated as known without error and the outputs are also treated as known without error. For example, the approach uses point values for fishing mortality rates and stock biomass values derived from VPA stock assessments as inputs to the estimation of surplus production model parameters. Point estimates of r and K are then derived from the resulting plots of surplus production versus stock abundance. The results are stock-recruit functions with varying amounts of overcompensation. However, the extent to which uncertainty in the parameter and data inputs to this procedure affect the range of plausible extents of over-compensation needs to be evaluated. I would recommend that a Bayesian estimation approach be applied to quantify a probability distribution for the predicted stock recruit functions suggested by this modeling approach. This should take into account uncertainty in all inputs to the calculations in a rigorous probabilistic framework. Uncertainty in the values for F , stock

size, and the estimates of r and K and other key parameters should all be taken into account in the calculations. This would then provide a probabilistic approach to evaluating the probability that there is meaningful overcompensation in the stock-recruit function for the species of interest. More comments on Crecco's approach are provided further below.

- **Could alternative non-equilibrium production models for groundfish species be examined for estimating F_{MSY} and B_{MSY} thresholds?**

Yes, as mentioned above, a Bayesian surplus production model (e.g., based on Meyer and Millar 1999; McAllister et al. 2002, and Myers et al. 1999) could be considered for this purpose. The reasons for this are explained above.

- B. Comment on the justification for changing the overfishing threshold to $F_{40\%}$ (the proposed proxy for most groundfish stocks) from $F_{20\%}$ that generally defined overfishing before Amendment 9, or from the F_{MSY} estimates in Amendment 9? Are the proposed proxies for F_{MSY} (e.g., $F_{40\% \text{ MSP}}$ for Georges Bank haddock, $F_{50\% \text{ MSP}}$ for Acadian redfish, etc.) more appropriate to achieve MSY, given the groundfish stock dynamics? Are the proposed proxy reference points overly conservative or too liberal for a fishing mortality threshold that complies with the Magnuson-Stevens Act?**

Based on my understanding of the scientific evaluations that support these choices of proxies, the choices of $F_{40\%}$ and $F_{50\%}$ as proxies for F_{msy} are adequately justified (Clark 1991, 1993, Mace and Sissenwine 1993 cited in Anon. 2002). These various studies to my understanding provide a sufficient theoretical and plausible empirical basis for the F_{msy} proxy values derived to be used for the purposes of fisheries management under the Magnuson-Stevens Act. These proxies for F_{msy} provide the benefit of being fixed values and not estimated and subject to estimation error variation due to the updating of stock assessment and reference point estimation procedures over time. The instability in fisheries management reference point values and management regulations appears to be a very serious problem given the continual updates the estimation methodologies and modeling procedures and the updates to the data series, and considering the relatively small amount of information in the stock-recruit data plots about the key stock-recruit function parameters that determine the reference points. It is thus recommended that the potential merits of using such F_{msy} proxies instead of estimates of F_{msy} from stock-recruit data and priors derived from meta-analysis (Myers et al. 1999) be formally reconsidered. Simulations that applied an operating model approach could be used to evaluate which of the two approaches could be expected to lead to the most desirable management outcomes given present uncertainties in the data and the form of the stock-recruit function for the three stocks for which the parametric approach is currently applied. There does not appear to be sufficient information available to assess whether the proposed proxy reference points either overly conservative or too liberal for a fishing mortality threshold that complies

with the Magnuson-Stevens Act. Simulation modeling could be applied to help to evaluate this question.

- **Reconstruction of the theoretical S-R curve can be done indirectly for each groundfish species by merging results (YPR, SSB/R) from the Thompson-Bell yield-per-recruit model and expected equilibrium yield (mt) from various stock production models. Are the resulting F_{MSY} values similar to the $F_{40\%}$ values (e.g. for haddock) from the Y/R curve? Is $F_{40\%}$ a suitable proxy for F_{MSY} under these conditions?**

Crecco (2003) illustrated this surplus production model- YPR approach using data for Gulf of Maine cod and using the Gompertz and logistic surplus production models. In Crecco (2003), F_{msy} for Gulf of Maine Cod under the logistic production model fitted in the paper is 0.50 (Table 3). Under the Gompertz model fitted within the paper F_{msy} is 0.70. Using the 2001 NEFSC estimates of r and K , Crecco's value for F_{msy} is 0.25. These values compare with and F_{msy} of 0.225 based on the Beverton Holt model and $F_{40\%}$ of 0.166 for Gulf of Maine cod in Anon. (2002). Thus it appears that the F_{msy} values provided by Crecco's (2003) estimations are at odds with the current NEFSC estimates of F_{MSY} .

The reasons for this discrepancy are unclear but possibly due to the different algorithm applied by Crecco (p. 5, 2003) to estimate the surplus production model parameters, r and K . Crecco (2003) derived from the NEFSC 2001 ADAPT VPA outputs, fishing mortality estimates, surplus production and stock biomass (Figs. 1 and 2, Crecco 2003) and estimated r and K for the two different surplus production models by fitting the surplus production functions to the derived surplus production and stock biomass data. This is a considerably different approach to the estimation of surplus production model parameters than the ASPIC method used by NEFSC.

The values that Crecco obtained for r were 1.02 and 0.73 for the logistic and Gompertz models, respectively. These appear to be extraordinarily high values for r for cod, especially when compared to the meta-analysis results obtained for the parameter for cod stocks in Myers et al. (1997). Furthermore, Crecco's values for r are markedly higher than the values obtained for Gulf of Main Cod by the NEFSC using ASPIC (0.23). These markedly higher values for r and lower values for K obtained by Crecco (2003) are no doubt responsible for Crecco's much higher estimates of F_{msy} . It is somewhat relieving to see that when Crecco used the NEFSC estimates of r and K , the value he obtained for F_{msy} (0.25) was quite similar to the values obtained by the NEFSC for F_{msy} (0.225).

Thus, it appears that providing that the values for r and K are accurately estimated, then Crecco's surplus production model – YPR approach could be a useful approach to taking into account age-structured processes in the estimation of F_{msy} . It would also be necessary to ensure that recruitment variability was not high and that there

was no chronic growth overfishing for the stock of interest to ensure that the surplus production model estimation approach provided valid estimates of the parameters r and K .

One other troublesome aspect regarding the use of surplus production models is that there is a large variety of alternative functional forms for surplus production and this will give rise to uncertainty over the shape of the stock recruit model obtained from a surplus production – YPR calculation. The logistic and Gompertz are only a few. Others include the Pella and Fletcher surplus production models that allow for a range of different inflection points for MSY. As Crecco clearly demonstrated, the stock-recruit function derived from the surplus production model – YPR calculations can vary considerably depending on the functional form of the surplus production model that is considered (Figs. 3 - 5, Crecco 2003). And from the appearance of the fits of the different surplus productions to the surplus production data (Figs 1 and 2, Crecco 2002), there appears to be relatively little ability in stock biomass data to be able to distinguish between alternative functional forms for surplus production. Therefore, just as for the approach that fits a stock-recruit function to noisy stock-recruit data, the surplus production model YPR method to obtaining a stock-recruit function offers no additional facility to allow statistically rigorous discrimination between alternative functional forms for the stock-recruit function.

It is therefore my conclusion that the surplus production YPR approach, while conceptually interesting, does not offer any substantial benefits over the current general approach applied by the NEFSC to the determination of F_{msy} reference points for New England groundfish. If there is interest to further develop the methodology, it should be developed to more adequately take into account and model uncertainties in the model inputs and parameter values and uncertainties over surplus production model structure. Additionally, as mentioned above, the methodology should be simulation tested using a rigorous operating model approach mentioned above to more objectively evaluate the potential statistical performance of the methodology if it was to actually implemented to estimate F_{msy} . The approach should be found to provide reasonably unbiased and precise estimates of F_{msy} under a variety of plausible scenarios for error structures in the input data and alternative hypotheses regarding surplus production model structure – particularly when the user does not know the true form.

With regards to the NEFSC estimates of $F_{40\%}$ and F_{msy} , these are in most instances not identical. But neither are the estimates glaringly different. My recommendation for there to be a formal evaluation to more widely use the $F_{40\%}$ and $F_{50\%}$ proxies for MSY instead of the estimated F_{msy} due to the problems of estimation of F_{msy} from very noisy data.

C. Evaluate evidence for density-dependent regulation of population size (e.g., simultaneous occurrence of various stocks at higher population sizes,

predator-prey, and growth rate information) for the groundfish complex. Are potential non-stationary stock dynamic processes (i.e., environmental variations in recruitment survival) and/or trophic limitations adequately accounted for in estimates of B_{MSY} ? Is there evidence that B_{MSY} values estimated for the 20 groundfish stocks cannot be simultaneously achieved?

In the information available and presented, no evidence was presented that suggested for any species that density dependent regulation from interactions from other species could occur over the range of observed fish densities. Paul Rago's presentation that showed species density ellipses provided no results to suggest that low density of one species could result from high density of others. Chris Legault's presentation also failed to demonstrate the presence of species interactions that could affect the population growth of any one species over the fish densities historically observed. Non-stationary stock dynamic processes do not appear to be accounted for in estimates of B_{MSY} . Whether they should or not given the current paucity of information about such processes for New England groundfish is another question. Attempts to do so could easily lead to more problems than by choosing not to model them, particularly due to the enormous uncertainty over just how such processes might operate and how to model them reasonably when trying to estimate B_{MSY} . If there is no empirical basis to suppose that such density dependent regulation of population size is important, then Ockham's razor should apply with regards to the models used. The methods used are thus well justified in not accounting for such complex processes. There appears to be no evidence available to suggest that B_{MSY} values estimated for the 20 groundfish stocks cannot be simultaneously achieved.

3. STOCK REBUILDING AND RELATED PROJECTIONS

The Sustainable Fisheries Act requires that various resources be rebuilt to B_{MSY} in no more than 10 years, unless life history attributes of individual stocks dictate a longer rebuilding period (e.g., Georges Bank cod, Acadian redfish, etc.). Considering the uncertainty in stock dynamics and the ability to achieve target rebuilding fishing mortality rates for all stocks in the complex simultaneously, comment on stock projection methodology used to advise on management strategies intended to achieve stock rebuilding goals.

In responding, reviewers should consider the following:

- A. Evaluate the adequacy of projection methods used to guide the attainment of B_{MSY} , specifically focusing on estimates of uncertainty in starting stock sizes,

recruitment, and implementation uncertainty in the attainment of target fishing mortality rates. Comment on potential biases and precision of stock projection methodologies.

Most of the projection methods used appear to be adequate overall and represent the best available science. Attempts have been made to account for processes that could lead to bias such as discarding. The Agepro method, which is the method of choice by the NEFSC for New England groundfish is well documented and rigorously constructed. It is well designed to take into account parameter uncertainty with the application of the conditional non-parametric bootstrap method. It also interfaces well with the ADAPT VPA stock assessment model and the bootstrapping output produced by ADAPT VPA to take into account parameter uncertainty. It appears, however, that in choosing a single base-case scenario to run the projections, that uncertainty over model structure is not evaluated. It would appear to be important to evaluate the sensitivity of projections to alternative plausible stock-recruit function forms, the possibility of autocorrelation in stock-recruit model residuals, even if they are not statistically detected in analyses of historic data, and values for the rate of natural mortality. Such sensitivity evaluations should of course implement the changes consistently in both the ADAPT VPA assessment and the AgePro model to ensure consistency of the modeled assumptions in the inputs to the AgePro model. Agepro and ADAPT VPA are set up to evaluate all such things; the only thing that needs to be done is to allow the scientists to find the time to be able to perform such important sensitivity tests for the various groundfish stocks.

In cases where ASPIC is used, only the maximum likelihood estimate of the intrinsic rate of increase, r , is used. This ignores uncertainty in the value for r , and this is a very serious omission. The ASPIC estimation and projection software should thus be revised to model the uncertainty in the estimates of r and the implications of this uncertainty for model projections.

The ability of the index-based projection software to reasonably model future trajectories of the age-structured populations about which its trying to make predictions should be evaluated using the operating model approach that's been already mentioned extensively above. Due to the gross simplifications of this modeling approach, its reliability should be tested before it is to be used to provide management advice about, e.g., the chance that various alternative stock rebuilding plans will achieve Bmsy in a timely manner. Yet, in Anon. (2003) p. 37, it is reported that "Results of this approach, summarized in Table 4.1.2, suggest a reasonable degree of coherence with rebuilding schedules and catch projections derived from more complicated age-structured models." This is some assurance that the method can produce credible results, but the issue over the precise conditions under which the method can be expected to work versus not work remain to be adequately delineated. A key issue to resolved, for example, is the number of years into the future that the

approach could be deemed to be reliably projected for each of the groundfish species evaluated.

B. Are stock projection methodologies sufficient to distinguish the relative merits of various management scenarios?

The AgePro software appears to be sufficiently detailed to be able to do this. The only additional sophistication that could be added would be to develop a spatially disaggregated version of Age pro to take into the effect of heterogeneity in the spatial distribution of the stock at different stages in its life history, spatial heterogeneity in patterns in fishing effort, and the potential implications of the time-area closures already implemented and that could potentially be implemented for the future population dynamics of the various groundfish populations and the resulting spatial dynamics of the New England groundfish fleet.

If the alternative management scenarios include ones such as the alternative fishing mortality rate control policies that Butterworth presented in his presentation on alternative "phase-down" policies, it does not appear that fisheries management controls are yet available to so precisely and accurately implement such fine tuned controls on fishing mortality rates on a stock by stock basis. It is most likely that the magnitude of implementation errors is large enough to prohibit such fine tuned controls. However, the principle of a phase-down approach to implementing large reductions in fishing effort certainly has its merits with regards to socio-economic considerations. It is thus up to NEFSC whether it wishes to formally consider such alternatives if it is required that large reductions in fishing mortality rate take place for given fish stocks.

C. The Magnuson-Stevens Act requires that overfished stocks be rebuilt to a biomass level consistent with producing the maximum sustainable yield from the fishery. Is there a scientific basis for arguing that an intermediate biomass target meets that requirement?

It is not possible to answer this question because it appears to be a policy matter.

Added References

Kass, R. E. and A. E. Raftery. 1995. Bayes factors. *J. Am. Stat. Assoc.* 90: 773–795.

McAllister, M. K., E. K. Pikitch and E. A. Babcock. 2001. Using demographic methods to construct Bayesian priors for the intrinsic rate of increase in the Schaefer model and implications for stock rebuilding. *Canadian Journal of Fisheries and Aquatic Sciences* 58:1871-1890.

February 2003 Groundfish Science Review: Murdoch McAllister

Meyer, R. and Millar, R.B. 1999. BUGS in Bayesian stock assessments. *Can. J. Fish. Aquat. Sci.* **56**: 1078-1086.

Myers, R. A., K. G. Bowen, N. J. Barrowman. 1999. Maximum reproductive rate of fish at low population sizes. *Canadian Journal of Fisheries and Aquatic Sciences* 56: 2404-2419.

Myers, R. A, G. Mertz, and P. S. Fowlow. 1997. Maximum population growth rates and recovery times for Atlantic cod *Gadus morhua* *Fishery Bulletin* 95: 762-772

Parma, A. M. 2002. In search of robust harvest rules for Pacific halibut in the face of uncertain assessments and decadal changes in productivity. *Bulletin of Marine Science* **70**: 423-453.

APPENDIX 1

STATEMENT OF WORK

Consulting Agreement between the University of Miami and
Dr Murdoch McAllister

January 7, 2003

Introduction

This document presents terms of reference for peer review of the stock assessment and population dynamics science supporting the New England Fishery Management Council's (NEFMC's) Northeast Multispecies Fishery Management Plan (FMP). Specifically, the review will focus on three major terms of reference:

- Effects on the accuracy and present usefulness of trawl survey data due to uneven trawl warps and other recently-discovered gear-related trawl survey problems. These evaluations will be based on gear testing cruises and related workshops conducted during autumn 2002 as well as any other information available to the reviewers.
- Estimates of stock biomass and fishing mortality targets and thresholds for the complex of stocks comprising the groundfish resource, and,
- The adequacy of projections of stock rebuilding to achieve the biomass targets, consistent with time frames as mandated under the Sustainable Fisheries Act.

These three focus areas were originally proposed by the staff of the NEFMC. Specific comments appropriate to three terms of reference provide guidance to the review committee recognizing that reviewers are likely to be unfamiliar with the specifics of the Northeast Multispecies FMP and, the provisions of the Sustainable Fisheries Act. Overall, the terms of reference generally concentrate on the adequacy of the science currently available to support fishery management plan development.

For each subject area, a brief objective statement is provided to give an overall context for the terms of reference to the reviewers. Within these subject areas, specific questions are provided with the intent of providing a minimum set of questions to consider in formulating the group's responses. It is envisioned that the reviewers' responses will take the form of detailed reviews of the information and conclusions reached in the various

February 2003 Groundfish Science Review: Murdoch McAllister

supporting documents and verbal presentations made to the group, along with their own summaries and opinions regarding the adequacy of existing science in supporting fishery management decisions. The reviewers are encouraged to pay particular attention to alternative methods presented by independent experts, if any, in concluding whether the conclusions of the Report, or other approaches, represent the best science available.

2. Trawl survey issues and influence on management advice

Considering the results of the Groundfish Assessment Review Meeting (GARM), subsequent results from experimental trawl comparisons, and other appropriate information, provide an evaluation of the significance of potential differences in trawl survey catchability resulting from recently-discovered survey gear problems on management advice for groundfish stocks managed under the Northeast Multispecies Fishery Management Plan.

In responding, reviewers should consider the following:

- A. Are conclusions regarding use of 2000-2002 trawl survey data adequately supported by analyses reported by the GARM? Were analyses sufficient to detect differences in survey catches arising from unequal warps and other survey problems? Did the sensitivity analyses presented in the GARM report adequately bound the range of potential effects inferred from analyses of historical and comparative data? Did the GARM adequately characterize the uncertainties in estimated stock sizes and rebuilding mortality rates potentially arising from unequal warp offsets?
- B. Was the design and analysis of data from experimental trawl comparisons adequate to estimate the magnitude of differences resulting from the use of unequal trawl warps and other experimental treatments? Were estimates of the power of these experiments to detect statistical differences in fish catches between treatment and control survey configurations adequately described?
- C. Advise on the significance of differences in species composition and relative catch rates resulting from side-by-side tows performed by commercial and government vessels in the recent trawl experiment with respect to model- and index-based estimates of stock size and fishing mortality rates.
- D. Comment on the precision of model-based calculations of stock size and fishing mortality rates in relation to variability in trawl survey catches and other sources of information included in assessments. Are the methods used for incorporating uncertainty into management advice sufficient? How should other sources of uncertainty (e.g., model selection, estimates of total removals) be incorporated?

3. Biological reference points

Review the fishing mortality and biomass targets and thresholds established for the 20 groundfish stocks included in the Northeast Multispecies FMP. Consider the adequacy of technical analyses supporting estimates of F_{MSY} , B_{MSY} or their proxies, as provided in the *Report of the Working Group on Re-Estimation of Biological Reference Points for New England Groundfish Stocks* (the “Report”). Comment on issues related to the simultaneous achievement of B_{MSY} values for the groundfish complex.

In responding, reviewers should consider the following. Of particular note, the NEFMC’s Science and Statistical Committee recommended that additional work was needed “...specifically to explore the implications of the uncertainty in the stock recruitment relationship.” For this reason, more specific questions are included in order to add clarity to the issues to be addressed by the reviewers.

- A. Comment on the technical basis for the estimation of F_{MSY} and B_{MSY} , and choices regarding the use of parametric (Beverton-Holt, Ricker, other candidate models, etc.) and non-parametric stock-recruitment relationships applied to yield per recruit estimates, surplus production models, or proxies for biomass and fishing mortality rate targets and thresholds.
- Are the Working Group assumptions (growth, maturity ogive, natural mortality, partial recruitment) appropriate for estimating a B_{MSY} proxy, which establishes a minimum biomass threshold and a rebuilding target?
 - Comment with reference to specific species on whether the use of Beverton-Holt type stock-recruitment curves, as opposed to the use of dome-shaped (Ricker type) curves, represent reasonable scientific judgment employing sound methodology and appropriate data sources. Is there a theoretical or practical basis to detect overcompensation (Ricker curve) from the stock-recruitment curve for each groundfish species based on the magnitude of the intrinsic rate of population increase (r) and the carrying capacity (K) parameter estimates from ASPIC production models?
 - Could alternative non-equilibrium production models for groundfish species be examined for estimating F_{MSY} and B_{MSY} thresholds?
- B. Comment on the justification for changing the overfishing threshold to $F_{40\%}$ (the proposed proxy for most groundfish stocks) from $F_{20\%}$ that generally defined overfishing before Amendment 9, or from the F_{MSY} estimates in Amendment 9? Are the proposed proxies for F_{MSY} (e.g., $F_{40\% \text{ MSP}}$ for Georges Bank haddock, $F_{50\% \text{ MSP}}$ for Acadian redfish, etc.) more appropriate to achieve MSY, given the groundfish stock dynamics? Are the proposed proxy reference

points overly conservative or too liberal for a fishing mortality threshold that complies with the Magnuson-Stevens Act?

- Reconstruction of the theoretical S-R curve can be done indirectly for each groundfish species by merging results (YPR, SSB/R) from the Thompson-Bell yield-per-recruit model and expected equilibrium yield (mt) from various stock production models. Are the resulting F_{MSY} values similar to the $F_{40\%}$ values (e.g. for haddock) from the Y/R curve? Is $F_{40\%}$ a suitable proxy for F_{MSY} under these conditions?
- C. Evaluate evidence for density-dependent regulation of population size (e.g., simultaneous occurrence of various stocks at higher population sizes, predator-prey, and growth rate information) for the groundfish complex. Are potential non-stationary stock dynamic processes (i.e. environmental variations in recruitment survival) and/or trophic limitations adequately accounted for in estimates of B_{MSY} ? Is there evidence that B_{MSY} values estimated for the 20 groundfish stocks cannot be simultaneously achieved?

4. Stock rebuilding and related projections

The Sustainable Fisheries Act requires that various resources be rebuilt to B_{MSY} in no more than 10 years, unless life history attributes of individual stocks dictate a longer rebuilding period (e.g. Georges Bank cod, Acadian redfish). Considering the uncertainty in stock dynamics and the ability to achieve target rebuilding fishing mortality rates for all stocks in the complex simultaneously, comment on stock projection methodology used to advise on management strategies intended to achieve stock rebuilding goals.

In responding, reviewers should consider the following:

- A. Evaluate the adequacy of projection methods used to guide the attainment of B_{MSY} , specifically focusing on estimates of uncertainty in starting stock sizes, recruitment, and implementation uncertainty in the attainment of target fishing mortality rates. Comment on potential biases and precision of stock projection methodologies.
- B. Are stock projection methodologies sufficient to distinguish the relative merits of various management scenarios?
- C. The Magnuson-Stevens Act requires that overfished stocks be rebuilt to a biomass level consistent with producing the maximum sustainable yield from the

February 2003 Groundfish Science Review: Murdoch McAllister

fishery. Is there a scientific basis for arguing that an intermediate biomass target meets that requirement?

Schedule

The independent peer review is to be completed by March 1, 2003. In order to meet that deadline, the following review format and timeline is proposed.

3-5 February: Public workshop (including participation of independent reviewers) on the GARM Report and report of biological reference points during this week.

6-8 February: Independent reviewers meet in executive session to discuss results from the two workshops and supporting documentation.

10-14 February: Independent reviewers prepare their individual reports and submit them to the summarizer.

17-21 February: Summarizer prepares his/her report summarizing findings of individual reports prepared by panel members, which will be made available to the public.

The February 3–5, 2003 public workshop will begin with an introduction followed by a series of presentations summarizing the various documents presented to the panel. Open comment periods will allow for additional scientific input from various members of the public regarding additional analyses and comments. Peer reviewers will interact with agency and independent scientists and members of the public to ask appropriate questions and discuss results.

Specific

The consultant shall be provided with all background material required to prepare for the review, and the consultant shall attend the February 3 – 5, 2003 workshop, the February 6 – 8, 2003 executive session, and to develop an individual, non-consensus report that shall be submitted for final summarization. The report shall also be submitted to the Center for Independent Experts as a review report.

The consultant's duties shall not exceed a maximum total of 14 days: Several days prior to the workshop for document review; the three-day workshop; the three-day closed door session; and several days following the meeting to complete the workshop and executive session report. The reports are to be based on the consultant's findings, and no consensus reports shall be accepted.

February 2003 Groundfish Science Review: Murdoch McAllister

The consultant's duties include:

1. Reading all background material provided;
2. Participating in the February 3 – 5, 2003 workshop on the Groundfish Assessment and Review Meeting (GARM) Report and report of biological reference points;
3. Participating in the February 6 – 8, 2003 executive session to discuss results from the two workshops and supporting documentation;
4. No later than February 14, 2003, submitting a written, nonconsensus report that is based on the results of the workshops and supporting documentation, the executive session discussions, and on the terms of reference described in the statement of work. The report should be submitted to the workshop summarizer and to the CIE¹; the CIE report should be addressed to the “University of Miami Independent System for Peer Review,” and sent to Dr. David Sampson, via email to david.sampson@oregonstate.edu, and to Mr. Manoj Shivlani, via email to mshivlani@rsmas.miami.edu.

¹ The written report will undergo an internal CIE review before it is considered final. After completion, the CIE will create a PDF version of the written report that will be submitted to NMFS and the consultant.

APPENDIX 2

TERMS OF REFERENCE FOR THE GROUND FISH SCIENCE REVIEW, FEBRUARY 3-8, 2003

Introduction

This document presents terms of reference for peer review of the stock assessment and population dynamics science supporting the New England Fishery Management Council's (NEFMC's) Northeast Multispecies Fishery Management Plan (FMP). Specifically, the review will focus on three major terms of reference:

- Effects on the accuracy and present usefulness of trawl survey data due to uneven trawl warps and other recently discovered gear-related trawl survey problems. These evaluations will be based on gear-testing cruises and related workshops conducted during autumn 2002 as well as any other information available to the reviewers.
- Estimates of stock biomass and fishing mortality targets and thresholds for the complex of stocks constituting the groundfish resource, and,
- The adequacy of projections of stock rebuilding to achieve the biomass targets, consistent with time frames mandated under the Sustainable Fisheries Act.

These three focus areas were originally proposed by the staff of the NEFMC. Specific comments appropriate to the three terms of reference provide guidance to the review committee, recognizing that the independent reviewers are likely to be unfamiliar with the specifics of the Northeast Multispecies FMP and the provisions of the Sustainable Fisheries Act. Overall, the terms of reference generally concentrate on the adequacy of the science currently available to support FMP development.

For each subject area, a brief objective statement is provided to give an overall context for the terms of reference to the reviewers. Within these subject areas, specific questions are provided with the intent of providing a minimum set of questions to consider in formulating the group's responses. It is envisioned that the reviewers' responses will take the form of detailed reviews of the information and conclusions reached in the various supporting documents and verbal presentations made to the group, along with their own summaries and opinions regarding the adequacy of existing science in supporting fishery management decisions. The reviewers are encouraged to pay particular attention to alternative methods presented by the independent experts, if any, in concluding whether the conclusions of the Report, or other approaches, represent the best science available.

TERMS OF REFERENCE

1. TRAWL SURVEY ISSUES AND INFLUENCE ON MANAGEMENT ADVICE

Considering the results of the Groundfish Assessment Review Meeting (GARM), subsequent results from experimental trawl comparisons, and other appropriate information, provide an evaluation of the significance of potential differences in trawl survey catchability resulting from recently discovered survey gear problems on management advice for groundfish stocks managed under the Northeast Multispecies Fishery Management Plan.

In responding, reviewers should consider the following:

- A. Are conclusions regarding the use of 2000-2002 trawl survey data adequately supported by analyses reported by the GARM? Were those analyses sufficient to detect differences in survey catches arising from unequal warps and other survey problems? Did the sensitivity analyses presented in the GARM report adequately bound the range of potential effects inferred from analyses of historical and comparative data? Did the GARM adequately characterize the uncertainties in estimated stock sizes and rebuilding mortality rates potentially arising from unequal warp offsets?
- B. Was the design and analysis of data from experimental trawl comparisons adequate to estimate the magnitude of differences resulting from the use of unequal trawl warps and other experimental treatments? Were estimates of the power of these experiments to detect statistical differences in fish catches between treatment and control survey configurations adequately described?
- C. Advise on the significance of differences in species composition and relative catch rates resulting from side-by-side tows performed by commercial and government vessels in the recent trawl experiment with respect to model- and index-based estimates of stock size and fishing mortality rates.
- D. Comment on the precision of model-based calculations of stock size and fishing mortality rates in relation to variability in trawl survey catches and other sources of information included in assessments. Are the methods used for incorporating uncertainty into management advice sufficient? How should other sources of uncertainty (e.g. model selection, estimates of total removals) be incorporated?

2. BIOLOGICAL REFERENCE POINTS

Review the fishing mortality and biomass targets and thresholds established for the 20 groundfish stocks included in the Northeast Multispecies FMP. Consider the adequacy of technical analyses supporting estimates of F_{MSY} , B_{MSY} or their proxies, as provided in the *Report of the Working Group on Re-Estimation of Biological Reference Points for New England Groundfish Stocks* (the "Report"). Comment on issues related to the simultaneous achievement of B_{MSY} values for the groundfish complex.

In responding, reviewers should consider the following. Of particular note, the NEFMC's Science and Statistical Committee recommended that additional work was needed "...specifically to explore the implications of the uncertainty in the stock recruitment relationship." For this reason, more specific questions are included in order to add clarity to the issues to be addressed by the reviewers.

- A. Comment on the technical basis for the estimation of F_{MSY} and B_{MSY} , and choices regarding the use of parametric (Beverton-Holt, Ricker, other candidate models, etc.) and non-parametric stock-recruitment relationships applied to yield per recruit estimates, surplus production models, or proxies for biomass and fishing mortality rate targets and thresholds.
- Are the Working Group assumptions (growth, maturity ogive, natural mortality, partial recruitment) appropriate for estimating a B_{MSY} proxy that establishes a minimum biomass threshold and a rebuilding target?
 - Comment with reference to specific species on whether the use of Beverton-Holt type stock-recruitment curves, as opposed to the use of dome-shaped (Ricker type) curves, represent reasonable scientific judgment employing sound methodology and appropriate data sources. Is there a theoretical or practical basis to detect overcompensation (Ricker curve) from the stock-recruitment curve for each groundfish species based on the magnitude of the intrinsic rate of population increase (r) and the carrying capacity (K) parameter estimates from ASPIC production models?
 - Could alternative non-equilibrium production models for groundfish species be examined for estimating F_{MSY} and B_{MSY} thresholds?
- B. Comment on the justification for changing the overfishing threshold to $F_{40\%}$ (the proposed proxy for most groundfish stocks) from the $F_{20\%}$ that generally defined overfishing before Amendment 9, or from the F_{MSY} estimates in Amendment 9? Are the proposed proxies for F_{MSY} (e.g. $F_{40\% \text{ MSP}}$ for Georges Bank haddock, $F_{50\% \text{ MSP}}$ for Acadian redfish) more appropriate to achieve MSY, given the groundfish stock dynamics? Are the proposed proxy reference points overly conservative or too liberal for a fishing mortality threshold that complies with the Magnuson-Stevens Act?
- Reconstruction of the theoretical S-R curve can be done indirectly for each groundfish species by merging results (YPR, SSB/R) from the Thompson-Bell yield-per-recruit model and expected equilibrium yield (mt) from various stock production models. Are the resulting F_{MSY} values similar to the $F_{40\%}$ values (e.g. for haddock) from the Y/R curve? Is $F_{40\%}$ a suitable proxy for F_{MSY} under these conditions?
- C. Evaluate evidence for density-dependent regulation of population size (e.g. simultaneous occurrence of various stocks at higher population sizes, predator-prey, and growth rate information) for the groundfish complex. Are potential non-stationary stock dynamic processes (i.e. environmental variations in recruitment survival) and/or trophic limitations

adequately accounted for in estimates of B_{MSY} ? Is there evidence that B_{MSY} values estimated for the 20 groundfish stocks cannot be simultaneously achieved?

3. STOCK REBUILDING AND RELATED PROJECTIONS

The Sustainable Fisheries Act requires that various resources be rebuilt to B_{MSY} in no more than 10 years, unless life history attributes of individual stocks dictate a longer rebuilding period (e.g. Georges Bank cod, Acadian redfish). Considering the uncertainty in stock dynamics and the ability to achieve target rebuilding fishing mortality rates for all stocks in the complex simultaneously, comment on the stock projection methodology used to advise on management strategies intended to achieve stock rebuilding goals.

In responding, reviewers should consider the following:

- A. Evaluate the adequacy of projection methods used to guide the attainment of B_{MSY} , specifically focusing on estimates of uncertainty in starting stock sizes, recruitment, and implementation uncertainty in the attainment of target fishing mortality rates. Comment on potential biases and precision of stock projection methodologies.
- B. Are stock projection methodologies sufficient to distinguish the relative merits of various management scenarios?
- C. The Magnuson-Stevens Act requires that overfished stocks be rebuilt to a biomass level consistent with producing the maximum sustainable yield from the fishery. Is there a scientific basis for arguing that an intermediate biomass target meets that requirement?

APPENDIX 3

REVIEWERS AND AGENDA GROUND FISH PEER REVIEW (GPR)

Public Meeting – 3-5 February 2003, New England Center, University of New Hampshire, Durham, New Hampshire

<http://www.necc.unh.edu/>

Independent Peer Reviewers (contracted through the Center for Independent Experts (CIE: University of Miami))

Dr Ewen Bell, Centre for Environment, Fisheries and Aquaculture Science, Lowestoft, England

Dr Robin Cook, FRS Marine Laboratory, Aberdeen, Scotland

Dr Murdoch McAllister, Imperial College, London, England

Dr Robert Mohn, Department of Fisheries and Oceans, Halifax, NS, Canada

Dr Andrew Payne (Chair/summarizer), Centre for Environment, Fisheries and Aquaculture Science, Lowestoft, England

Public Session Moderator

Mr Don Perkins, Gulf of Maine Aquarium

AGENDA - modified during meeting to accommodate participants' availability

Monday, 3 February

0900-1700 Public Session – Topic: **Trawl Survey Issues**

Background Documents:

-Report of the Workshop on Trawl Warp Effects on Gear Performance

<http://www.nefsc.noaa.gov/nefsc/publications/crd/crd0215/>

- Report of the Groundfish Assessment Review Meeting

<http://www.nefsc.noaa.gov/nefsc/publications/crd/crd0216/>

- Report of the Trawl Survey Experiment Workshop
available online

- Other contributed documents

Order of the Day:

Introduction of peer reviewers, presentation of terms of reference, and discussion of ground rules
(Moderator)

Formal Presentations:

- *An overview of trawl survey issues* – **Russell Brown** (30 minutes)
- *Intervention analyses to detect trawl warp offset problems, sensitivity analyses, scale of potential offset factors*- **Paul Rago** (1 hour)
- *Trawl warp and related experiments*- **Michael Fogarty** (1 hour)
- *An evaluation of Paul Starr's analysis of the fishing gear experiment*- **Doug Butterworth** (20 minutes)
- *Comparison of length composition data from trawl experiments*- **Tom Nies** (30 minutes)

Facilitated discussion regarding presented materials in relation to terms of reference (all)

Tuesday, 4 February

0800-1700 Public Session – Topic: **Biological Reference Points**

Background Documents:

- Report of the Working Group on Re-Evaluation of Biological Reference Points for New England Groundfish

<http://www.nefsc.noaa.gov/nefsc/publications/crd/crd0204/>

- Report of the Overfishing Definition Review Panel:

<http://www.nefmc.org/documents/overfishing/>

- Report of SAW 36

- Report of the Groundfish Assessment Review Meeting:

<http://www.nefsc.noaa.gov/nefsc/publications/crd/crd0216/>

- NEFMC Council Meeting Report for July 2002, summarizing Scientific and Statistical Committee review of re-estimated reference points

<http://www.nefmc.org>

(Go to "News and Motions," then click on "Council Reports")

- Other contributed documents

Order of the Day:

Formal Presentations:

- *Re-Evaluation of biological reference points: goals and objectives*- **Steven Murawski** (1 hour)
- *A Strategy to evaluate alternative stock-recruitment models*- **Christopher Legault** (30 minutes)
- *Evidence for density-dependence in species and ecosystem responses*- **Ralph Mayo** (30 minutes)
- *An age-structured production model based assessment and reference point evaluation for the Gulf of Maine cod stock*- **Doug Butterworth** (1 hour)
- *Decision analyses using biological reference points in evaluating groundfish stock status*- **Yong Chen** (30 minutes)
- *Overfishing thresholds (F_{MSY} , B_{MSY}) for New England groundfish from empirically based stock recruitment models*- **Victor Crecco** (30 minutes)
- *A general biological reference point working group model*- **Andy Applegate** (20 minutes)

Facilitated discussion regarding presented materials in relation to terms of reference (all)

Wednesday, 5 February

0800-1700 Public Session – Topic: **Projections of Stock Rebuilding**

Background Documents:

-National Standard Guidelines for Overfishing Definitions: Final Rule
http://frwebgate.access.gpo.gov/cgi-bin/getdoc.cgi?dbname=1998_register&docid=fr01my98-23.pdf

- AgePro Users manual:

- GARM Report Revised Projections
<http://www.nefsc.noaa.gov/nefsc/publications/crd/crd0216/>

- Other Contributed Documents

Order of the Day:

Formal Presentations:

- *NMFS National Standard Guidelines and Stock Rebuilding*- **Pamela Mace** (40 minutes)

February 2003 Groundfish Science Review: Murdoch McAllister

- *Projection Methodologies used to evaluate medium-term impacts-* **Jon Brodziak** (40 minutes)
- *A phased rebuilding strategy, using the cod stocks from Gulf of Maine and Georges Bank as examples-* **Doug Butterworth** (30 minutes)
- *Rebuilding strategies for three key stocks-* **Dave Lincoln** (30 minutes)

Facilitated discussion regarding presented materials in relation to terms of reference (all)

Thursday, 5 February – Saturday, 8 February

Executive Session – Invited Peer Reviewers and support staff person (Karena Jolles, New Hampshire Fish and Game Department)

Discuss issues raised at public workshop and in supporting documents. Develop strategy for completing individual reports and how summarizer will convert them to a final document.

Commence the report drafting process individually and through debate.

Consult other participants for clarity purposes.

APPENDIX 4

BIBLIOGRAPHY CONSULTED/MADE AVAILABLE

1. Formal Documentation (received before or at the meeting)

Almeida, F. and L. Jacobson. Working Paper: Species Compositions from the NMFS/Industry Survey Trawl Study Conducted by the R/V *Albatross IV* and F/V *Sea Breeze* 28 October-6 November, 2002. 24 pp.

Almeida, F., and L. Jacobson. Species Size Compositions from the NMFS/Industry Survey Trawl Study Conducted by the R/V *Albatross IV* and F/V *Sea Breeze*, 28 October - 6 November 2002.

Almeida, Frank. Working Paper: Comparison of R/V *Albatross IV* and F/V *Sea Breeze* Catch during the NMFS/Industry Survey Trawl Study. Presence vs. Absence by Species. 9 pp.

Almeida, Frank. Working Paper: Composition of the R/V *Albatross IV* 'Other Catch' Component during the NMFS/Industry Survey Trawl Study Conducted 28 October-6 November, 2002. 5 pp.

Almeida, Frank. Working Paper: Cruise Report of the NMFS/Industry Survey Trawl Study Conducted by the R/V *Albatross IV* and F/V *Sea Breeze*, 28 October-6 November, 2002. 6 pp.

Brodziak, J. K. T. and P. J. Rago. AGEPRO Version 2.02 User's Guide. July 23, 2002. 107 pp.

Brodziak, Jon. Comparison of Average Catch Rates of 20 Species for Optimal and Worst-Case Scenario Net Configurations by Area. January 14-15, 2003.

Butterworth, D S, R A Rademeyer and E´ E Planganyi. An Age-Structured Production Model Based Assessment and Reference Point Evaluation for the Gulf of Maine Cod Stock. 41 pp. (3 pp. Addendum added)

Butterworth, D S, R A Rademeyer, E´ E Plaganyi. Results for Georges Bank Cod of Age-Structured Production Model Based Assessments Similar to those Conducted for the Gulf of Maine Cod Stock. 22 pp.

Crecco, Victor. Overfishing Thresholds (F_{MSY} , B_{MSY}) for New England Groundfish from Empirically-Based Stock-Recruitment Models. January 26, 2003. 21 pp.

Fogarty, Michael J. Analysis of R/V *Albatross IV* - F/V *Sea Breeze* Trawl Configuration Experiment. 9 pp.

Lovgren, Jim. Observations from the *Albatross IV* correctional cruise. February 5 2003. 4 pp.

February 2003 Groundfish Science Review: Murdoch McAllister

National Oceanic and Atmospheric Administration. 50 CFR Part 600 Magnuson Stevens Act Provisions; National Standard Guidelines; Final Rule. May 1, 1998. Federal Register 63(84): 24212-24237.

New England Fishery Management Council. Council Report. July 2002. 6 pp.

New England Fishery Management Council. Correspondence received by Council regarding the trawl gear survey information.

New England Fishery Management Council. Report of the Groundfish Overfishing Definition Committee. November 27, 2000. 12 pp.

Nies, Tom. Working Paper: Analysis of Catch-at-Length Data from the NMFS Industry Survey trawl Study Conducted by the R/V *Albatross IV* and F/V *Sea Breeze*. October 28 - November 6, 2002. 18 pp.

Northeast Fisheries Science Center Reference (NEFSC) Document 02-15. Report of the Workshop on Trawl Warp Effects on Fishing Gear Performance. October 2-3, 2002. 80 pp.

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

Northeast Fisheries Science Center (NEFSC)/Industry Cooperative Survey Gear Study 28 October-6 November, 2002. Source Document: Specifications for Construction of NEFSC Standard #36 Bottom Trawl.

Northeast Fisheries Science Center (NEFSC), National Marine Fisheries Service. Final report of the Working Group on Re-Evaluation of Biological Reference Points for New England Groundfish. March 19, 2002. 232 pp. + 163 pp. of Appendix 7.0.

Northeast Regional Stock Assessment Review Committee (36th SARC). Draft Advisory Report on Stock Status. January 2003. 50 pp.

Overfishing Definition Review Panel. Final report: Evaluation of Existing Overfishing Definitions and Recommendations for New Overfishing Definitions to Comply with the Sustainable Fisheries Act. June 17, 1998. 179 pp.

Restrepo, V.R. et al. Technical guidance on the Use of Precautionary Approaches to implementing National Standard 1 of the Magnuson-Stevens Fishery Conservation and Management Act. 1998. NOAA Technical Memorandum NMFS-F/SPO-31.

Starr, Paul. Memorandum: Analysis of NMFS Trawl Survey Data: R/V *Albatross IV* and F/V *Sea Breeze*. January 10, 2003. 16 pp.

Stauffer, Gary. NOAA Protocols for Groundfish Bottom Trawl Surveys of the Nation's Fishery Resources. December 16, 2002. 81 pp.

2. Presentation or illustrative material (received at the meeting)

Applegate, Andy. Handout: General Biological Reference Point Working Group Model.

Brodziak, Jon. Presentation: (Age-Structured) Projection Methodologies Used to Evaluate Medium-Term Impacts. February 5, 2003.

Brown, Russell W. Presentation: Issues with NOAA Fisheries Bottom Trawl Surveys Conducted.

Butterworth, Doug. Summary of Paul Starr's Analysis Presented to the trawl Experiment Workshop, January 14, 2003.

Butterworth, Rademeyer and Plaganyi. Updated Projections covering phased rebuilding.

Chen, Yong. Presentation: Decision analyses using biological reference points in evaluating groundfish stock status. February 2, 2003.

Correspondence Received by Council Regarding the Trawl Gear Survey Information

Fogarty, Mike. Presentation: Effects of Trawl Warp Offsets and Gear Configuration on Survey Catches.

Goudey, Clifford A. Letter to Paul Howard (NEFMC). Comments on the significance of the warp offset issue and on the utility of the recent R/V *Albatross IV* and F/V *Sea Breeze* comparison cruise in determining the possible sampling errors in recent trawl surveys. January 28, 2003.

Industry Stakeholder Concerns raised by those who participated in the September 25-27 experimental cruise, including a list of questions from fishermen. Handout.

Legault, Christopher M. Presentation: A Strategy to Evaluate Alternative Stock-Recruitment Models.

Lincoln, Dave. Presentation: Rebuilding Strategies vs. Catch.

Mace, Pamela M. Presentation: The implementation of National Standard 1 since the SFA. February 2003.

Mayo, Ralph. Presentation: Ecosystem Implications of Revised Biomass Targets.

Murawski, Steve. Presentation: Age-Specific Catchabilities Estimated for Four Stocks w/ ADAPT.

Murawski, Steve. Presentation: Reference Point Re-Estimation.

February 2003 Groundfish Science Review: Murdoch McAllister

O'Malley, James. From Science to Illusion: Mathematics in Fishery Management. In *Pacem in Maribus XXVI*, Halifax, November 29-December 3, 1998.

O'Malley, James D. Letter to Mr Ricks Savage. East Coast Fisheries Federation, Inc. May 16, 2002.

Rago, Paul. Presentation: Intervention Analyses to Detect trawl Warp Offset Problems for NMFS R/V Survey Indices from 2000-2002. February 3, 2003.

Stevenson, Barbara. Handout: Trawl Data for R/V *Albatross IV* and F/V *Sea Breeze*.

APPENDIX 5

CLOSING QUESTIONS POSED AND COMMENTS MADE TO THE PANEL BY PARTICIPANTS

Doug Butterworth

1. Have ADAPT assessments explored a sufficient set of sensitivities, for example in respect of alternative values of M , and what are the implications for reference point estimates?
2. Comment on the appropriateness of MSY -based management targets given the imprecision of the estimates and difficulties associated in particular with changes over time resulting from new data and changed methodologies.
3. Given the ASPM-based reference points for two cod stocks, albeit based on initial analysis, are appreciably different from those based on ADAPT methodology, how important is it that further ASPM-based results be developed to be taken into account in the next set of management decisions for these stocks?
4. Given that assessment method, with current data, appear unable to estimate parameters such as stock-recruitment steepness (and hence B_{MSY}) with great precision, what is the potential role of adaptive management towards improving such precision? (Note the relevance of this question in respect of scientific aspects of the TOR 3C.)
5. Is it important for the Albatross to survey efficiently (as well as comparably over time) to be able to use associated swept-area estimates of absolute abundance to "ground-truth" estimates provided by population model assessment methods?

Geoffrey Smith

1. Given the fact that B_{MSY} values are generally set at one half of the carrying capacity of the stock, is it unreasonable to assume that all 19 stocks of groundfish can be rebuilt to B_{MSY} simultaneously?
2. Do rebuilding strategies that allow continued over fishing in the near term pose a greater biological risk than those that reduce fishing mortality rates to levels at or below F_{MSY} ?
3. Is the question of the National Standard Guidelines requirement to rebuild overfished stocks to B_{MSY} in 10 years or less a scientific question or a legal and/or public policy question?

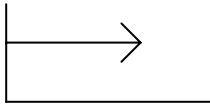
Priscilla Brooks

1. Fisheries management in the United States is governed by the Sustainable Fisheries Act and National Standards that dictate narrowly the parameters within which management plans are developed. Pamela Mace's presentation gave you a fairly thorough overview of the law and guidelines. I believe that you must keep in mind the legal reality in which we work and in which your report will be incorporated B_{MSY} and F_{MSY} must be estimated and stocks must

be rebuilt within 10 years, except in circumstances in which the natural history of the stock dictates more time. Given these realities, is the NMFS science related to the biological reference points, that is the GARM report, sound?

Jon Brodziak

1. Are the steepness parameters (h) values implied/estimated in the Butterworth production models for GOM cod credible, in the context of Myers et al. (1999. Maximum reproductive rate...CJFAS)?

$$h = 1 \Leftrightarrow R$$


SSB

Ron Smolowitz

1. What is the sensitivity of the trawl survey to towing speed changes over time?

Eric Smith

1. Perhaps a useful follow-on question is to ask "Is there justification, given scientific uncertainty in biological reference points and projection methodology, for setting a lower intermediate 10-year rebuilding target that can be adjusted upwards as the stock builds and our estimate of that value becomes more certain?" This better captures the essence of the Council's question/concern from a management standpoint. TOR 3C
2. Is a Ricker-type S-R curve more (or equally) justified relative to a B-H type curve for cod and haddock? TOR 2A, bullet #2

Phil Ruhle

Please look over NMFS protocol for groundfish surveys, recently developed.

1. The speed issue is of great concern but the gear used is also a problem. In all other surveys gear is well addressed but NEFSC survey net design and age is 40 years. The design has not been used by industry in 20 years.
2. Bottom contact on this gear is very lax as is all aspects of handling of this gear; this is shown in NEFSC protocol as compared to other science centers.

Pamela Mace

Note about the Precautionary Approach:

1. See page 11 of Technical Guidance for a statement about how the precautionary approach is appropriate to management decisions, but not to scientific estimation of assessment-related parameters and variables.

Andy Applegate

1. Which other analytical methods can be used to validate the reference point estimates and rebuilding projections given the heavy reliance on less robust and variable recruitment estimates? How do managers use the scientific advice while this effort is made?
2. Are there better methods within the context of the current National Standards to evaluate the performance of the plan and monitor rebuilding of a set of multispecies fisheries?

Tom Nies

Question on Trawl Experiment

1. Was the design of the experiment adequate to determine if errors in the trawl warp cable affected recent survey results?

Points to consider:

-The control net differed from the design of the survey net used for the past two years (ignoring the issue of warp length). Some differences: different doors, use of swivels on doors, different backstraps, different ground cable rigging.

-Experimental tows were all conducted either into or with the current. Survey tows are towed in the direction of the next station, without regard to current. The experiment never towed cross-current, and a poorly rigged net may tow differently in a cross-current.

-If, as suggested by Paul Rago based on Pennington's work, the effective sample size for frequency distributions is closely related to number of tows (as opposed to fish caught), were there enough tows to draw conclusions on catch at length/age?

-Is the assumption of a covariate relationship between the commercial vessel and Albatross catches justified by the analysis? (I have not seen the final paper by Dr. Fogarty).

-For the covariate analysis, how does the insertion of values for missing Sea Breeze catches affect the results?

Jim O'Malley

1. Is there evidence of any application of the precautionary principle in the assessments or rebuilding targets?
2. Is such an application legitimate in science?

David Frulla

1. If one manages towards B_{MSY} for every species in a mixed stock fishery at the same time, can this result in under utilisation of certain species? TOR 3

2. Can differing F reduction strategies accommodate considerations relating to a mixed stock fishery, economic consideration, and uncertainties related to significantly higher new reference points, while achieving the appropriate biomass target over the relevant rebuilding period? TOR 3