Peer Reviews Report for

"A Pilot Study of a New Sampling Design for the NOAA Fisheries Access Point Angler Intercept Survey"

Reviewed by
Dr. Lynne Stokes, Southern Methodist University
Dr. Michael D. Larsen, Larsen Statistics and The George Washington University

Introduction

This document combines the comments provided by two different peer reviewers of the 2012 Report entitled "A Pilot Study of a New Sampling Design for the NOAA Fisheries Access Point Intercept Survey" by Breidt, et al. The first section provides the general comments of each of the two reviewers, and the remainder of the document lists specific reviewer comments without identifying the source of each comment.

General Comments of the Reviewers

Reviewer #1:

Introduction: A pilot study was conducted in 2010 in North Carolina to test the feasibility of implementing a new sampling design for the Access Point Angler Intercept Survey, a component of the Marine Recreational Fisheries Statistics Survey (MRFSS) of the NOAA Fisheries Service. The new design was compared to the traditional MRFSS sampling design. This review discusses the quality and completeness of the report on the pilot study.

In order for results from sample surveys to be representative of a population, the survey must be designed and implemented following certain principles. First, it is important to identify the target population and how the population can be enumerated or listed, either explicitly or implicitly. Second, it is important to know the probabilities that units in the population are included in the sample. These probabilities can be influenced by the structure of the population and the sampling scheme. Stratification can be used to increase precision of estimators as well as to ensure adequate coverage of subpopulations. Cluster sampling often reflects the structure in the population and can be utilized to reduce costs and for practical considerations, but often at the cost of decreasing precision of estimators. Third, adequate plans need to be in place to deal with likely problems in survey data collection, including refusal to participate, difficulty in answering questions, resource limitations, and other challenges. When the target population is well defined and probabilities of sample inclusion for members of the population are known, the sampling strategy is an example of probability sampling and can be the basis for scientific statements about the population.

The changes proposed and implemented for the Access Point Angler Intercept Survey are consistent with professional scientific survey practice. These changes, if implemented in an

optional way, are expected to yield estimates that are improved in terms of less bias and reduced variance in this survey.

Summary: The report (Breidt et al. 2012) provides a thorough and professional evaluation of the pilot study. Numerical and graphical presentations are sufficient for comparing results. Unfortunately, results using the new design are not clearly superior to the results using the previous design with weight adjustment. In particular, sample yield is lower using the new design. In part this reflects the requirement that interviewers not substitute sites or fishing modes as they see fit. This is an important change for reducing selection bias. As in most surveys, it is not really possible to measure the amount of bias, because it could only be compared to a much more rigorous and large effort.

Despite the lower sample yield, the recommendation that seems most reasonable is to continue with comparisons of the new design with the original design. The authors have given several suggestions for improvement. These should be considered carefully before proceeding to other areas of application. Importantly, one should consider changes in stratification, cluster definition, and allocation to improve performance of estimates. Additional analysis of the pilot data could provide some guidance about adjusting the specific implementation of the new design. The new design has the potential to produce much better survey estimates. For that promise to be realized, additional work will be necessary in refining the approach.

The changes proposed and implemented for the Access Point Angler Intercept Survey are consistent with professional scientific survey practice. These changes, if implemented in an optional way, are expected to yield estimates that are improved in terms of less bias and reduced variance in this survey.

The report (Breidt et al. 2012) provides a thorough and professional evaluation of the pilot study. Numerical and graphical presentations are sufficient for comparing results. Results using the new design are not clearly superior to the results using the previous design with weight adjustment. The recommendation, however, is to continue with comparisons of the new design with the original design. The authors have given several suggestions for improvement. These should be considered carefully before proceeding to other areas of application. Importantly, one should consider changes in stratification, cluster definition, and allocation to improve performance of estimates. The new design has the potential to produce much better survey estimates. For that promise to be realized, additional work will be necessary in refining the approach.

It is recommended that one study correlations among survey outcome variables and information available at each site. Some of this information could be quantitative and other information categorical or dichotomous. Thinking broadly about potential relationships could guide useful data collection. Even if not used in a statistical model to improve estimation, it still might be interesting to examine associations of auxiliary variables with outcomes.

Reviewer #2:

I am very positive about the work that was done on the pilot project and the changes that put the Intercept survey on a more scientific footing. I am in full agreement with the recommendation that the new design be adopted. The report is clear and thorough.

I have just one suggestion that would require any significant changes to the report, and then a few other suggestions that would require more minor changes.

Specific comments of the reviewers

1. *Time of day stratification:*

Time of day stratification versus sampling at peak intervals: This suggestion is quite important to eliminate interviewer judgment about when to interview. Without control in time of day stratification, there could be substantial bias in estimating mean catch and number of fishing trips.

Is interviewing assigned at night? It would seem prudent to not sample in the middle of the night. Or to sample with a low frequency in the middle of the night. Perhaps the four time frames are not enough. What if you excluded 12-4am and divided the remaining 20 hours into 5 slots each of 4 hours? You could still have two low intensity sites in a cluster.

2. *Geographic stratification:*

Geographic stratification versus sampling across entire state: This suggestion is important for ensuring coverage of the whole state and allocating interviewers to areas.

Effective and creative selection of strata could lead to efficiency and adequate information for estimation in some sub-state areas. Stratification should be implemented separately in each state. It could reflect NOAA Fisheries Service divisional units in addition to major state-level geographic features.

Heavy stratification into small strata could lead to gains in efficiency. One would want to track and estimate interviewer effects if few interviewers are collecting information in a single strata or across strata.

3. *Clustering sites:*

In the old design sites were not clustered together, and interviewers were allowed to visit other sites. Allowing interviewers to choose sites invalidates the assumptions of probability sampling and would be likely to lead to bias in estimation. Not clustering sites together could lead to high variability in probabilities of selection if selection is based on activity level. Although not related to bias, high variability in probabilities can be disadvantageous due to associated high uncertainty in estimators. Grouping low activity sites together and visiting all grouped sites should stabilize probabilities of selection, but still allow high rates of sampling among high

activity sites. This seems like a good compromise. Control of routes and time spent by interviewers is a critical addition to the design.

Some high activity sites might be included with certainty every survey year due to their importance to state-level fishing. Other sites might appear every few years in surveys. It would be a good idea to review information used to create probabilities of selection every year.

Given the availability of GIS and other technologies for recording time and location, it probably would be worthwhile to design measures of performance of interviewing assignments. That is, if you can collect some process information about travel and ease of collecting information by site, then it might be possible to improve resource planning in the future.

4. ***Sampling Frame and Probabilistic Sampling; Elimination of Opportunistic Sampling; Angler Trip Counts:***

Specifying a formal sampling frame and implementing a probability sampling scheme is a major advance over the former method. Statistical models can be useful for improving efficiency and removing some amount of bias, but it is best to start with a probability design. Use of models to improve efficiency can still be considered, but one is relying on models to a much smaller degree if the starting point is a probability sampling design.

Eliminating opportunistic switching of fishing modes removes a potential source of bias.

Getting accurate numbers of angler trips ending in a given interval is important as stated in the report for assessing secondary probabilities of selection. It will be important to assess whether adequate resources are available for both counting anglers finishing trips as well as interviewing in sites. It might be quite difficult for a single individual to interview effectively and count over a broad area at the same time. According to Appendix A, page 28 (page 33 of PDF)

> At sites with low activity you should be able to both count and conduct interviews at the same time. At sites with moderate-to-high activity you will alternate between counting and conducting interviews by the hour. Be sure to record the start and stop time for the time spent counting and the time spent interviewing as two separate sampling periods, even if you do not switch sites."

It is recommended that an experiment be conducted to compare the current new plan (simultaneous counting and interviewing) with dedicated counting plus interviewing. The latter likely will require a second person.

5. ***Issuing and Completing Assignments:***

The new design requires issued assignments to be completed without rescheduling. The former design allowed interviewers to reschedule. As stated in the report, "Eliminating assignment rescheduling greatly reduces the possibility of a nonresponse bias that could result from a failure to obtain observations from some of the selected assignments. It also eliminates possible temporal undercoverage biases that could result from the rescheduling of assignments." It is the opinion of the reviewer that these statements most certainly are true.

6. ***Interviewing limits:***

The previous design had a cap on interviews. Instead, the pilot design has interviewers utilize the full time frame for sampling. There are a couple of possibilities motivating this change. First, for a given number of clusters, a larger sample size within each cluster is better. Second, interviews of a higher quality provide better information. The change in policy obviously should increase the number of interviews in some clusters. It might also positively impact interview quality as long as the interview period is not too long and tiring. If one knows that one can stop interviewing after a given number of interviews, then one might do the required number of interviews quickly in order to be done with the assignment. Requiring interviewing to continue for a given period of time eliminates the incentive to finish quickly.

It is recommended that available technology (GIS, computer time/date stamps, etc.) be used to monitor interviewers if there is any question about the legitimacy of data collection reporting.

7. ***Eligibility for Interviews:***

As I understand the report, the new design allows interviewing children under age 5 and individuals who are returning from a contest. These considerations are beyond my expertise to evaluate. Of course, interviewing children usually requires parent/guardian permission and attention to what the children might or might not be able to reasonably answer. I could see one defining the ultimate sampling unit either as an individual angler or as a group of anglers together (in a boat, in a group such as a family with small children). There could be advantages and disadvantages to both arrangements in terms of ease of interviewing, clarify of definition of unit, and value of information in estimation.

The inclusion of tournaments is specified on page 30 (PDF page 35) of Appendix A, but I see no mention of ages of interviewees. Given that children under age 5 were not included before, it is likely important to add a section stating that they should be interviewed. Procedures for interviewing children (e.g., parental/guardian permission, presence of parent/guardian) should be included.

8. ***Complete vs. Incomplete Beach/Bank Interviews***

The inclusion of incomplete angling trip under the old design probably was problematic for estimation. Even if anglers were asked what percent of their trip was completed it would have had some degree of speculation. It is likely as stated in the report that those fishing longer would have a higher probability of being selected for an incomplete trip and also have a higher number of fish, thereby producing bias. Although it makes finding someone to interview harder (they have to be done fishing), this change likely removes a source of bias.

9. ***Questionnaires and Data Forms***

These seem to have been adapted appropriately for the new design. Are any thoughts being given to electronic data capture on, for example, a smart phone, GIS position recording, or time/date stamp to automate the recording of some information?

10. ***Estimation methodology***

A prime advantage of a probability sampling design is that design-based estimation should produce unbiased (or nearly unbiased in the care of ratio estimation) estimates of desired quantities. It is appropriate to switch to a design-based estimation method.

One could consider the use of statistical models in order to utilize auxiliary information available at all sites. It is recommended that one study correlations among survey outcome variables and information available at each site. Some of this information could be quantitative and other information categorical or dichotomous. Environmental: area of lakes, length of fishing bank, etc. Population: population size within 10, 25, and 50 miles of site, percent of population nearby below poverty and below 200% of poverty, etc. Administrative: is it a state park? Is a state permit required? Is hunting allowed at the site? Is there a swimming beach? It is uncertain what will be predictive, but thinking broadly about potential relationships could guide useful data collection. Even if not used in a statistical model to improve estimation, it still might be interesting to examine associations of auxiliary variables with outcomes.

11. ***Comments on some survey questions***

The introduction to the potential respondent could be rephrased. Instead of, Hi, I'm representing NCDMF, it would be better to not use an acronym in the opening contact. See page 33 (PDF page 38) of the Appendix A (NOAA 2011).

The screening question (recreation) is as follows (page 33, PDF page 38 of the Appendix A, NOAA 2011):

> Was the primary purpose of your trip today for recreation, that is, for fun and relaxation, or was it to provide income either from the sale of fish or from the sale of the fishing opportunity?

The screening criterion might be better stated, or additional instructions could be provided to interviewers. How would someone be classified who is fishing to have something to eat? Presumably some people fish regularly in order to supplement their diet. I would assume as recreation. What if they trade some fish to someone for vegetables or wild game? Does this qualify as recreation or income?

12. ***Comments on results and recommendations***

In the pilot study, the original design resulted in more interviews per assignment and per hour. Estimates of mean catch rates were not very different under the two methods. If measured simply by estimated sampling variance, the results using the new method are not superior. As Breidt et al. (2012; page 9) notes,

> The estimates generated from the MRFSS sampling design were more precise than the estimates generated from the Pilot design largely because a greater number of sampling assignments were completed under the MRFSS design.

One can speculate, however, that the new design actually could be better in terms of selection bias and other problems. As in most surveys, it is not really possible to measure the amount of bias, because it could only be compared to a much more rigorous and large effort.

> It should be noted, however, that the potential for non-sampling errors was much larger under the MRFSS than under the new design, which is not reflected in these precision comparisons. (Breidt et al 2012; page 9)

One also can speculate that the performance of the new design can be improved.

> Although sampling under the new design in this study yielded a much larger percentage of completed assignments with no angler trip interviews and a much smaller number of interviews per positive assignment, changes in the allocation of sampling across sampling strata could greatly reduce these differences. (Breidt et al 2012; page 9)

Importantly, one could try to analyze the existing pilot survey in order to better inform choices about the new sample design.

13. ***Allocation of sample***

The report gives suggestions on how to improve. The 2nd and 3rd recommendations are critical. One should study when, where, and for which mode the MRFSS survey gained such large amounts of sampled units. Time stratification and whether to exclude the middle of the night are topics that need consideration.

One should also identify resources before allocating sample. It sounds like resources were woefully inadequate in some areas. One can implement legitimate probability sampling schemes with controls connected to geography and resources, and more effort apparently is needed in this direction.

14. ***Splitting high intensity sites***

Among issues to study in the future, splitting high intensity sites so that size is not so large among the largest probably is a good idea. Handling a mix of very small and very large sites is difficult when a limited amount of time is available for each and there are substantial travel times to the next site. It is recommended that stratification become finer and clusters be formed to be more uniform in size. Recommendations for future research #9, 12, and 13 likely should be moved up in priority. One aspect of this is time stratification. Surely one can omit 12-4am and do a better job. One could consider splitting the remaining twenty hours into five 4-hour blocks. One then could still visit two low intensity sites (up to 2 hours each) in a single cluster if they are geographically close enough.

15. ***Adjusting allocation to mode, or allowing mode switching***

In Breidt et al. (2012), it is noted that

> For the Pilot, assignments were allocated evenly across the four modes in each state subregion: Man-made (MM), Beach Bank (BB), Private/Rental (PR), and Charter (CH). Allocation of mode-specific assignments within each state subregion and day type (i.e. kind of day) was determined monthly.

It seems to this reviewer that the actual modes used would not be close to even in every site. Being able to switch modes would have given the original survey a big advantage. Perhaps one can learn from the actual data collected (pilot versus original design) how mode restrictions impacted data collection.

16. ***Performance of the proposed new design***

The proposed new design and implementation rules undoubtedly produce estimates whose properties can be evaluated with much more precision than those of the MRFFS. I also believe that with adjustments and improvement, the new design will produce estimates that are, on average, better than those of the MRFFS. However, that message does not come through strongly enough in this report, in my opinion. The productivity measures (interviews per hour, etc.) and the standard errors reported (Figures 4,5,9 and 10) always show the pilot at a disadvantage. There are mentions of possible biases in estimates from the MRFFS, but then in other places the new estimation method developed by Breidt et. al. are described (p. 13) as approximately design unbiased.

The only statements about standard error are vague but suggest that even if sampling hours were the same, the Pilot would have been better but would still compare unfavorably to MRFFS. (p. 60 "If the number of PSU's observed in the Pilot design had been increased to match the number of assignments in the MRFFS design, the variances of the mean catch rates would have been lower." but "If both designs had completed the same number of assignments, the MRFFS design would still likely have provided estimates with greater precision") But how much better? The report doesn't make it easy for the reader to make a fair comparison. It would be useful for evaluating to have something like design effects for the two methods, since there are competing factors make the Pilot both less efficient (lower interviews per assigned hour) and more efficient (less variable weights, as mentioned on p. 54) It isn't clear (to me) how to do this precisely from the data, but even some rough measure of "variance per sampler hour" would be helpful.

Another analysis that would be useful to see would be a prediction of how much the design effect could be improved after implementation of some of the suggested improvements in allocation to the various strata. It seems that this could be calculable for at least a few major species or total catch, to give an idea of the potential value of the new method.

And of course this doesn't even consider the possible residual biases in the MRFFS data. The similarity of both sets of estimates would seem to suggest that there is little bias left in the newly weighted MRFFS, but then we get statements like this one on p. 8 :"… the potential for non-sampling errors was much larger under the MRFFS…" So does the statement on p. 8 refer to the "old MRFFS" or does it mean that even the "new MRFFS" is believed to still have substantial non-sampling errors. If the latter, it would be helpful to briefly describe what kinds of non-sampling errors the authors think still remain. Presumably, the authors feel there are smaller non-sampling errors than with just the original MRFFS estimators. Is that the interpretation of the authors? I find that surprising, although it doesn't mean that even if it was true in NC in this year

it will always be true. Does this mean then that the average catch per trip really does not vary much by time of day (night vs. day), or at alternative sites in the same vicinity (chosen by MRFFS samplers)? Or is it that the precision was so low we couldn't tell there was a difference? More discussion of what evidence there is about this would be helpful to explain the advantage of the new design to those who might not be so convinced of the superiority of a strictly probability sampling design on theoretical grounds alone.

17. ***Interpretation of results:***

I would like to see more interpretation of results, or explanation of most likely causes, when there are differences between MRFFS and Pilot results. Here are some places where that would be informative:

a. p. 8 The summary states that the proportion of anglers reachable by the CHTS is comparable for all but beach/bank mode. Is there any intuition or speculation about why that would be true? Also, I believe it is true that an angler is not in the CHTS frame if he or she does not have a landline phone. If that is correct, it should be noted as part of the description of "the effort ratio estimator."

b. p. 37 Do you have any speculation on why the jagged curve for the pilot? is it just because of the small sample size, or could it be because of the rules for clusters that have people driving from one place to another at certain hours?

c. p. 38 It is hard to know what to make of the differences seen in Table 6. What is the difference between reported and observed? Were these fish chosen because they are ones that are particularly common? How is the average # of fish computed? Is it average per angler overall, or average per angler that targeted or has any of that fish? Do these differences seem reasonable or explainable (e.g., are some fish caught more at night, for example, and so may be more likely to be seen in the Pilot?)

d. p. 47 The fact that the MRFFS estimates are higher when there is a difference is described as "interesting." Is there a reason to believe this is a real difference and indicates a bias on the part of MRFFS data, due to either the noncoverage of night fishing or some other reason?

e. Table 7. Does it make sense that these estimates would differ due to noncovearge of night fishing in MRFFS, or is there some other reason? Maybe only locals fish from the beach at night?

f. On p. 7 and again on p. 38 are mentions that the telephone survey (CHTS) is used along with the intercept data to come up with total catch estimates. I believe these descriptions are too cryptic for a reader who is not already familiar with the estimation method to understand. Even a brief explanation that total catch is estimated as (total trips by coast county anglers) *(catch per trip for all anglers) *(# of all anglers/# of coastal county anglers) would be helpful, and would help elucidate some of the comments about why bias can result if certain assumptions (e.g., avg. catch the same for coastal and noncoastal) do not hold.

18. ***Clarifications Requested:***

a. p. 4 When the report says that the Pilot compared the performance of the new sampling design and estimates of catch per trip, etc. with the "traditional MRFFS sampling design" (p. 4), does this mean that the comparisons are with the old design but the new estimators developed by Breidt et al.? The remainder of the report seems to suggest the comparisons are always to the new analysis method for MRFFS, but the use of the word traditional made me wonder as I was reading the executive summary.

b. How does the new design obtain nearly 3 sites per assignment? (2.46, from p. 8). Does that mean that most assignments are 3-site clusters? It might be good to have a summary somewhere of how many units in the frame are of each size.

c. What does the word positive mean in this statement on p. 10 "provide higher average number of interviews per *positive* assignment." Also on p. 61.

d. p. 11 #6. It is not clear in what way cluster formation would be modified. More two-site clusters? Why would that happen?

e. p. 11 Recommendation # 10. Apparently there was some indication that the site list had some flaws. Could these be clarified?

f. p. 12 #12. Clarify what is meant by "explicit" and "implicit"

g. p. 23 "at least one psu was selected from each interval…" Clarify that this means "per month" (I think).

h. p. 36 The x-axis scales are different on these plots. I don't think they should be. In fact, why should n't they all be from 0 to 24?

i. p. 39 I don't' understand the difference in the last two sentences of the paragraph before Table 7.They seem contradictory, but then I don't know what "fishing area" means.

j. Figures 4 and 5 should have a label of Pilot and MRFFS to be consistent with the other descriptions, rather than NC and MRFFS.

k. p. 10 I do not understand the meaning of #1 under future recommendations.

l. p. 53 The word *inappropriate* seems wrong. It is not that samplers are making bad choices, but that they are making choices at all that is wrong with MRFFS.

m. p. 69 The need for frame maintenance assignments is mentioned a couple of times, but no information on the reason for this suggestion is mentioned. Was there some problem with the list of sites that was discovered during the pilot?

n. Probability sampling is referred to as probabilistic sampling on p. 5. Is this meant to imply something different than what we usually mean by probability sampling?

o. p. 15. Why are there quotes around the statement under item #3? I presume because it is quoting from the previous report, but it should say so.

p. pp. 52, 53,55 and several other places There are comments about the fact that one advantage of the Pilot covering all times is that it eliminates potential bias due to different catch rates at night. But it also eliminates potential bias caused by different coastal county anglers' percentage at night. I believe this should be emphasized.

q. p. 52 "It was not necessary to require samplers to regularly work overtime." Was this a problem with MRFFS? If so, then state this advantage.

19. ***Other comments:***

a. Executive summary and p. 13 mentions that the MRFFS survey was "…not providing coverage of fishing trips ending on private property or at night." Much discussion was devoted to the issue of improving night fishing, but no other discussion of private property was included in the report. It seemed a bit like a "bait-and-switch" as I kept looking for discussion of innovations for the private property issue. Maybe this should be removed from the overview.

b. p. 12 # 11 The idea of basing pressure measures on # of fish landed seems a good one, especially if this allocation is very different than the one being used. It would be even better if the variability in fish landings could be used, though maybe mean and variance of fish landings are positively correlated.

c. p. 68 Suggestion #8. I like this idea a great deal. In fact, the potential for cost savings if the entire effort estimation were to be moved to the intercept survey is large. With the new design, my guess would be that number of trips could be more accurately estimated from the intercept survey than from telephone or mail, due to the large measurement error and nonresponse of the effort survey. The only problem I see is the private access sites. How much of the total effort is on private access sites? Maybe a telephone survey would still be necessary but only to estimate this ratio rather than the current coastal county ratio estimated from the intercept. Why not just compare total public access estimates from telephone and the intercept survey rather than needing to estimate by site (thought this is perhaps not a discussion for this report)? If the intercept survey were to be used for both effort and catch, one might consider counting both arriving and departing anglers, since the spatial-temporal time frame would be appropriate for either. The two estimators would be correlated but not identical, so should provide a little extra information basically for free.

References

Breidt, F.J., et al. (2012). A Pilot Study of a New Sampling Design for the Access Point Angler Intercept Survey. Report to the Marine Recreational Information Program's (MRIP) Design and Analysis Work Group (DAWG).

NOAA. (2011). Appendix A. Field Procedures Manual. *Procedures Manual: Development and Testing of Alternative Sampling Design for the MRFSS Intercept Survey.* North Carolina Division of Marine Fisheries. National Oceanic and Atmospheric Administration National Marine Fisheries Service (NOAA Fisheries). February of 2011.