Review of species and size composition estimation for the western and central Pacific purse seine fishery

P.L. Cordue
Fisheries consultant
New Zealand

***For* Center for Independent Experts (CIE)**
November 2012

Executive summary

A review of the methods and procedures used by the Secretariat of the Pacific Community to estimate species and size composition of the three main tuna species (yellowfin, skipjack, and bigeye) in the purse seine fishery in the western and central Pacific Ocean was conducted by two independent CIE appointed reviewers. The review consisted of a pre-meeting review of documents, participation in a review meeting in Noumea, New Caledonia, from 22–25 October 2012, and the preparation of an independent report by each reviewer.

Estimating the species and length composition of tuna catches in the western and central Pacific Ocean is complicated by various aspects of the fishery. Ideally, sampling would be conducted on shore and/or estimates made from cannery receipts. However, for much of the fishery this is not possible because trans-shipping of fish is common and there are simply far too many ports to be covered. Skippers have been required to fill out logbooks on a set-by-set basis since the start of the fishery. However, logbook data are unreliable with regard to species mix in the sets as the skippers tend to ignore all but the main species in the catch.

Sampling at sea with observers seems to be the only viable option and the move to 100% observer coverage necessitates that a good sampling design is put in place. However, there are complexities associated with sampling at sea for this fishery. Catch is rapidly "brailed" out of the net and put into wells and, without the cooperation of the crew, there is little opportunity for observers to sample the fish. Samples have to be taken from brails in some fashion and there is undoubtedly some between-brail structure which must be dealt with by the sampling design.

The current design is for observers to take five fish from the top of each brail ("grab sampling"). The fish are identified by species and measured. Because of potential between-brail structure and preferential selection of fish by species and/or length, these data have been recognized to provide biased estimates of species mix and length frequency (apparently, small fish are underestimated, large fish overestimated; and hence skipjack numbers are underestimated and yellowfin numbers overestimated).

An alternative sampling method, "spill sampling", is being developed. In this approach every tenth brail from each set is sampled for species and length by opening the bottom of the brail and spilling the fish into a container (approximately 200–500 fish are sampled on each spill depending on the size of the fish). Paired spill-sample and grab-sample data have been collected on a number of trips since 2008. These data are now routinely used to "correct" the grab-sample estimates.

The observer program began in 1993 but the fisheries first developed in the early 1960s. Therefore, there is a long time period for which there is no observer data. The current method for estimating the catch histories in this period is by fitting GLM models to the observer data using explanatory variables from the logbooks. The models are then used to predict the species proportions in the sets which were not observed.

I agree with SPC scientists that the current grab-sampling design is inappropriate and does lead to biased estimates of catch by species and length. The development of a new sampling design based on spill-sampling is appropriate. However, there are many details of the design which need further research. Specifically the design needs to be developed by an experienced and mathematically and statistically skilled fisheries scientist. I have developed appropriate statistical models in this report for both the grab-sampling and the spill-sampling data and have shown how

the spill-sampling data can be scaled within set to produce a (nearly) unbiased estimator of number in each species-length class. However, there is still much to be done.

My main conclusions are:

- The grab-sample design leads to estimates of species mix and length frequencies which are biased.
- Spill samples, when properly scaled within set, can be used to produce (nearly) unbiased estimators of species mix and length frequencies.
- The current methods of correcting for the grab-sample bias using paired spill-sample and grab-sample data is based on an indefensible model and the "corrected" grab-sample estimates should be discarded (as should estimates of length frequencies and catch histories which are based on the faulty corrections of grab-sample data).
- Current methods of estimating catch histories for periods when there is little or no observer data are unnecessarily complicated and based on inappropriate models.
- Simpler, substitution methods can be used after defining appropriate strata by analysing the observer data at the set level.

My main recommendations are:

- A simulation model should be developed for testing alternative observer sampling designs with a view to implementing a design based on spill sampling.
- Given guidance on the required levels of precision for bigeye, details that need to be worked out include:
    - how to stratify within sets
    - how to stratify in time and space (extra sampling effort may be needed for bigeye; i.e., more than one observer on some vessels in some seasons/regions)
    - what average size of spill samples/frequency of brail sampling is best
    - how best to achieve a random/uniform distribution of selected brails
    - how best to scale set-level estimates to stratum-wide estimates
- The documentation of the design should include a full set of equations (and simulation results if needed) which show that estimators of length frequencies and catch by species are (almost) unbiased for individual sets and for strata.
- Historical data need to be reanalysed to produce defensible estimates of catch histories and length frequencies for use in stock assessment and for other purposes:
    - suitable post-stratification should be determined by model-based and descriptive analysis of the *uncorrected* grab-sample estimates at the set level;
    - paired grab and spill sample data should be used to correct *aggregated* species-length distributions and hence to corrected estimates of catch and length frequency in years with good observer coverage (uncertainty should be estimated by bootstrapping);
    - historical catches in years with little or no observer coverage should be estimated using appropriate stratification (on explanatory variables available in logbooks) and substitution of species-mix estimates from the same stratum type using the average from years that are likely to have had similar species mix (temporal trends in species mix within strata need to be carefully examined).

Background

A review of the methods and procedures used by the Secretariat of the Pacific Community (SPC) to estimate species and size composition of the three main tuna species (yellowfin, skipjack, and bigeye) in the purse seine fishery in the western and central Pacific Ocean (WCPO) was requested by SPC.

I am one of two Center for Independent Experts (CIE) reviewers who participated in the review, which consisted of a pre-meeting review of documents, participation in a review meeting at SPC, Noumea, New Caledonia, from 22–25 October 2012, and the preparation of an independent report by each reviewer. This report presents my findings and recommendations in accordance with the Terms of Reference (ToRs) for the review (Appendix 2, annex 2).

Review Activities

Pre-meeting

The documents for the review were made available electronically in a timely manner (Appendix 1). I carefully read the documents and then re-read the main review document (Lawson 2012). I noted that the model used in the correction of grab-sample data was completely wrong and I started work on developing an appropriate statistical model for grab and spill sample data.

During the meeting

As scheduled, the review meeting convened at SPC on Monday morning, 22 October 2012. Dr. Harley chaired the meeting (on the Monday) and suggested that after the presentations were completed that the agenda was pretty much up to the CIE reviewers. I suggested that the tentative agenda (Appendix 2, Annex 3) looked suitable except that we would not be preparing our reports at the meeting (as it suggested) but that we could certainly present some preliminary findings and recommendations at the final session of the meeting.

Monday was taken up with the presentations (Appendix 3). On Tuesday morning, Dr. Powers and I met and discussed what additional data and analyses we would like to see from the SPC scientists. We both believed it would be useful to see some summaries of raw data and some basic statistics to give us a better feel for the data (see Appendix 4). On Tuesday afternoon and Wednesday morning I continued to work on a statistical model of the grab and spill sample data and on ways to illustrate/explain why the "availability model" (Lawson 2009, 2012) was invalid.

On Wednesday afternoon the SPC scientists presented the analyses they had done in response to our requests. Also, Dr. Hampton joined the meeting for the first time. In order to clarify some of the results from the presentations and other issues I asked if anyone in the room had been at sea on a commercial purse seine vessel. As nobody had, Dr. Hampton suggested that we ask some appropriate SPC staff to come to the meeting and answer some questions. We were able to talk to two staff who had both worked as observers and who now worked with the observer data and/or on training observers.

Both SPC staff noted that bigeye were sometimes seen to float in the net and that they were therefore likely to be contained in the early brails. They both agreed that there was potential species/length structure across brails. The first person interviewed suggested that there was also structure within brails, with larger fish tending to be near the top. The second person was not quite so sure about this but didn't discount it entirely. He did note that the spill samples tended to come out of the middle of the brail (i.e., "funnelled" down the middle), rather than just coming from the bottom of the brail. Neither person thought that there was much possibility of observers accumulating grab samples from multiple brails and measuring them all at the end of the set rather than doing it between brails as they were grabbed (the crew are apparently keen to get fish into the wells as quickly as possible).

On Wednesday night, while working on a draft document of my preliminary findings and recommendations, I realized that I wanted to see some additional analysis of the spill samples to see if they were distributed uniformly with regard to brail weight and brail order. This is important if each spill sample is to be considered as a random sample from a *randomly* selected brail. I put in an additional request focussed on this issue (Appendix 4, Annex 2).

On Thursday morning I continued work on my preliminary-findings document for presentation in the afternoon. When the meeting reconvened we first considered the results from my additional requests. However, my main request had not been addressed because the spill-sample database did not contain the information on which brail (e.g.,1st ,2nd ,3rd, etc.) had been sampled. Dr. Lawson indicated that SPC would be updating the database to include this information which was on the spill-sample forms but which had not been entered.

The meeting concluded with the presentation of our preliminary findings and recommendations. I presented my draft document on screen and then Dr. Powers talked through some of his findings and recommendations. We both noted that our final reports may contain somewhat different findings and recommendations depending on the results of analyses that we may conduct after the meeting.

Post-meeting

I continued to work through the details of my recommendations: how to do the catch sampling in the future; how to correct the grab-sample data; and how to produce catch and length frequency estimates at different spatial and temporal scales given different levels of data availability. I spent a lot of time on deriving an *unbiased* estimator of catch numbers (by species and/or length class) at the set level for spill samples. I wanted to be sure that I knew how the data should be used before recommending that future sampling be based on spill sampling.

Summary of findings

Estimating the species and length composition of tuna catches in the WCPO is complicated by various aspects of the fishery. Ideally, sampling would be conducted on shore and/or estimates made from cannery receipts (which does happen for the Japanese fleet in recent years). However, for much of the fishery this is not possible because trans-shipping of fish is common and there are simply far too many ports to be covered.

Sampling at sea with observers seems to be the only viable option and the move to 100% observer coverage necessitates that a good sampling design is put in place. However, there are complexities associated with sampling at sea for this fishery. Catch is rapidly "brailed" out of the net and put into wells – there is little opportunity for observers to sample the fish without disrupting the fishing operation. Samples have to be taken from brails in some fashion and there is undoubtedly some between-brail structure which must be dealt with by the sampling design. The review has two terms of reference. The first calls for a review of the current methods and asks for recommendations on how to improve them. The second term of reference asks for recommendations for a future observer sampling design and any experimental work that may be needed in support of the design.

The two ToRs are considered below.

1. *Evaluate and provide recommendations on the statistical methods used to estimate species and size composition (skipjack, yellowfin and bigeye tuna) in the purse seine fishery, with particular attention to the following issues:*
    a. *The need for, and approaches to, simultaneous estimation of catches for all three species*
    b. *The need for, and approaches to, simultaneous estimation of the size and species composition.*
    c. *Approaches to take into account factors that influence the size and species composition, e.g., season, location, set type, and vessel flag.*
    d. *Approaches for interpolation (e.g., statistical versus substitution) where sample data are low or absent, especially for the estimation of historical catches.*
    e. *Approaches to characterize uncertainty in estimates of the catch by species;*
    f. *The ability to provide reliable estimates of catch by species at different levels, e.g. the set, the trip, vessel flag, assessment region.*
    g. *Approaches to analyze the paired spill and grab sample trials.*

This term of reference requires consideration of the current methods that are used to provide the estimates of species and size composition in the purse seine fishery for the three main tuna species. The various factors to take into account are listed. The factors do not provide coherent headings for presentation of my findings so I have used my own headings. Because the issues are very technical I have provided a summary of my findings before we get into the mathematics.

Summary for ToR 1

I first consider the current sampling design which uses "grab samples" from each brail of each set. I develop an appropriate statistical model for the grab samples and show that between-brail structure will introduce a bias in the estimates of proportion by number in a species-length class (e.g., 20-25 cm skipjack). Because of the small number of samples taken from each brail, and the human selection of the fish, it is not possible to construct an unbiased estimator.

The spill samples are then considered and a statistical model is presented for those data. The natural estimator of proportion-by-number in a species-length class is shown to have the same bias as the grab-sample estimator (when there is no selection-preference). However, because of the large sample sizes in each spill sample it is legitimate to weight the data from each sampled brail with an estimate of the number of fish in the brail. The weighted spill-sample estimator (under the assumptions of the statistical model) is shown to be (nearly) unbiased as an estimator of numbers within each species-length class.

The current methods of correcting grab-sample data using paired grab and spill samples are then reviewed. The model underlying the correction method is shown to be wrong in almost every respect (intuitively it makes no sense; the model ignores sample size, the probabilities of selecting fish from a given length class, and brail structure; and it contains a mathematical error). I present two defensible methods of correcting the grab sample data. One method is quick and easy and is recommended for short term use to correct *aggregated*-estimates of species-length frequencies. The second method is complex but it should provide better estimates of the parameters needed to correct grab-sample data at the set-level.

Having dealt with the methods of sampling data at the set level I then consider how grab-sample data can be used to provide aggregate estimates of species and length composition (i.e., at different spatial and temporal scales, and different classes of vessel or types of sets). I recommend that various models (GLMs included) and descriptive analyses be used to explore the data and to define appropriate strata (i.e., post-stratification). For each stratum, the grab-sample data can then be scaled within sets and added across sets to provide an uncorrected species-length frequency. This is then corrected using estimates of selection-preference from appropriate paired grab and spill sample data. Finally, it is scaled up to a stratum-wide species-length frequency using an estimate of the total number of fish caught in the stratum. Estimated catch by species is obtained by multiplying the number in each species-length class by the corresponding mean fish weights. Uncertainty in the estimates is obtained by bootstrapping from the raw data.

Finally, the methods of reconstructing catch histories for each species are considered. Current methods fill in gaps in the data (e.g., a stratum with no observer sampling) using GLMs which predict species composition using explanatory variables available from the logbooks. These models are dubious because of various technical flaws and inappropriate levels of scale (the trip level or higher levels of aggregation are used, but many of the processes occur at finer spatial and temporal scales). In years when there is little or no observer data, species mix estimates (for strata defined by explanatory variables from the logbooks) should be borrowed from appropriate years where there are data.

Current observer sampling design

In general an observer-sampling design has to deal with two levels of sampling: how to sample the fish at the tow/set level; and how to allocate observers to vessels in order to obtain adequate samples within (pre-defined) strata. In the WCPO purse seine fishery, the move to 100% observer coverage means there is now little need to consider the allocation of observers to vessels.

The "design" at the set level specifies that an observer is to "grab" 5 fish from each brail (they are asked to take the 5 fish which are nearest to them). In practice, observers may take anywhere from 1 to about 20 fish per brail (see Appendix 4, Annex 3). The reasons for taking fewer fish are probably to do with how much time the crew gives the observer to sample the fish (i.e., not much) and the size of the fish (i.e., if they are quite large it may be difficult to take all 5 quickly). "Enthusiasm" is the only reason that has been suggested for why an observer would take more than 5 fish per brail.

The grab-sampling data for a set are now routinely "corrected" to account for what has been termed the "grab sampling bias" (GSB) and potential preferential selection of different sized fish by the observers (the "human" factor). The main justification for naming the GSB seems to be that the taking of 5 fish per brail is a number-based selection method which is thought inappropriate. Lawson (2009, 2012) gives an example which he calls a "thought experiment" where there are equal weights of large yellowfin and small skipjack in a set and they are brailed such that all the yellowfin are in the early brails and all the skipjack are in the later brails (with each brail having the same weight). Taking 5 fish per brail gives a combined sample (across all brails) of 50% yellowfin and 50% skipjack by *number* (which is wrong) and this also leads to the wrong estimate of species proportion by weight.

Dr. Lawson has concluded that the main problem with the grab sampling is something to do with a *number* of fish being taken. In his presentation, he said: "Thus, catches are estimated from species compositions determined from the weight of the fish, but the sampling protocol is based on the number of fish. This is the basic problem." His preoccupation with this belief seems to derive from Lawson (2008) where he simulated grab sampling and spill sampling for three scenarios with a mix of small skipjack and large yellowfin. He did not simulate brail structure, but sampled at random from the population with either a prescribed number of fish (grab sampling) or a prescribed weight of fish (spill sampling). He compared the spill samples and grab samples on the basis of coverage (which ranged from 0.2% to 2% – grab sample coverage being by number and spill sample by weight). He showed that there was a bias when estimating the proportion of skipjack in the catch (by weight) for grab sampling when the coverage level was low. His spill-sampling showed little or any bias. Hence, his conclusion: "weight is good" and "number is bad".

Unfortunately he didn't notice that the only reason the spill samples showed little bias was because they had larger sample sizes by number. The comparison should not have been by coverage, but by average sample size. It would then have been apparent that taking a random number of fish (by sampling a fixed weight of fish) was not advantageous. Intuitively, it seems obvious that once the fish are taken there is no extra information in how many were taken – it is known what lengths and weights they are, whether there be 30 or 100 is not relevant. The simulations merely showed that there is some bias (caused by non-linearity) in the estimator of species proportion by weight when a small number of fish is sampled at random.

The basic problem with the grab-sample "design" is that the design was never thought through properly. Taking 5 fish per brail was never a good idea but not because it was a fixed *number* of fish. The problem is that the design takes no account of potential structure between brails. The design appears to be based on the naïve assumption that all brails are somehow homogeneous and that taking 5 fish randomly from each of b brails is equivalent to taking a random sample of $5b$ fish from the set.

Returning to the example in the "thought experiment", it is easy to see how the correct estimate can be arrived at from the grab samples. One simply has to recognize that each *brail* is being randomly sampled – it is as if each brail is a stratum – and that before samples from different strata are combined they must be scaled-up within the respective strata. So, as the samples from each brail are either 100% yellowfin or 100% skipjack, it should be estimated that there are the same number of brails containing yellowfin and skipjack (and they all have the same weight) and hence the estimate of species proportion is 50% yellowfin and 50% skipjack by *weight* (which is true). Taking a sample of fixed *weight* from each brail and combining all the samples without scaling within brail also gives the correct answer. However, sampling by weight is not the solution. Change the example so that the yellowfin brails weigh half that of the skipjack brails; sampling by fixed weight then gives 50% yellowfin and 50% skipjack by weight whereas in reality yellowfin is 1/3 by weight and skipjack 2/3 by weight.

On separate days, Dr. Harley and Dr. Hampton both offered the same tentative defence of the 5-fish-per-brail design saying that it was aimed at providing "good estimates of catch over a wide area" (i.e., not at the set level but for an estimate of catch by species over a much broader scale e.g., by quarter and 5 degree square). However the grab-sample design does introduce a bias at the set level (as we will see below) and the problem with a bias at the set level is that it tends to be systematic and of the same sign (positive or negative) for *most* sets. In terms of design, it is better to aim for very little bias at the set level rather than to hope that the errors from biased estimators will somehow cancel out for aggregated estimates.

Rather than specific examples, let us now move on to statistically modelling the grab sample data.

For a specific set, assume that there are b brails and n fish are sampled by an observer from each brail. Assume that species and length have been divided into s species-length classes and consider the kth class (e.g., 20-30 cm skipjack).

Let,

W_i = the weight of the ith brail
N_i = the number of fish in the ith brail
N_{ik} = the number of fish in the kth class in the ith brail
p_{ik} = the proportion of fish (by number) in the kth class that are in the ith brail.

Assume that the observer has a potential species-length preference (rather than selecting at random) given by non-negative parameters $r_1, \ldots, r_s$.

Let,

X_{ik} = the number of fish sampled from the ith brail that are in the kth class
$\mathbf{X_i}$ = $(X_{i1}, \ldots, X_{is})$

We shall assume that the observer's preferences modify the number-based probabilities of random sampling with replacement (i.e., assuming that the probability of selecting a single fish from within a class is related to the proportion of fish in the class by *number* and that n is small compared to each N_i).

It follows that,

$\mathbf{X_i} \sim$ **Multinomial**$(n, r_1 p_{i1}, \ldots, r_s p_{is})$.

Given the typically small sample sizes per brail, the only calculation that we can use (in practice) to estimate the species-length mix in a set, is to combine samples across brails. We will look at the expected value of the proportion of fish (by number) in the kth class.

Let, Y_k be the total number of fish in the sample in the kth class:

$$Y_k = \sum_{i=1}^{b} X_{ik}$$

It follows that,

$$\mathrm{E}(Y_k) = \sum_{i=1}^{b} \mathrm{E}(X_{ik}) = \sum_{i=1}^{b} n r_k p_{ik} = n r_k \sum_{i=1}^{b} \frac{N_{ik}}{N_i}$$

Hence, for the estimator of proportion by number in the kth class,

$$\mathrm{E}\left(\frac{Y_k}{nb}\right) = r_k \frac{1}{b} \sum_{i=1}^{b} \frac{N_{ik}}{N_i} = r_k \overline{p}_k$$

where $\overline{p}_k$ is the average proportion of fish in the *k*th class per brail.

We should note that the actual proportion of fish in the *k*th class for the set is ρ_k:

$$\rho_k = \frac{\sum_{i=1}^{b} N_{ik}}{\sum_{i=1}^{b} N_i} = \frac{\overline{N}_k}{\overline{N}}$$

where $\overline{N}$ is the average number of fish per brail and $\overline{N}_k$ is the average number in the *k*th class per brail.

The bias of the estimator of the proportion of fish in the *k*th class (by number) is therefore:

$$\text{Bias}\left(\frac{Y_k}{nb}\right) = r_k \overline{p}_k - \frac{\overline{N}_k}{\overline{N}}$$

Obviously, if r_k is "big" there can be a positive bias (which makes sense) or if it is "small" there can be a negative bias (when $r_k = 0$, the magnitude of the bias is just the true value of the proportion). However, even when there is no observer selection-preference, there is still a bias which is:

$$\overline{p}_k - \frac{\overline{N}_k}{\overline{N}} = \text{mean}\left(\frac{N_{ik}}{N_i}\right) - \frac{\text{mean}\left(N_{ik}\right)}{\text{mean}\left(N_i\right)}$$

That is, the bias (with no observer preference) is the difference between the mean of the ratios and the ratio of the means (across brails for the number of fish in the *k*th class and the number of fish in the brail). There are two obvious sufficient conditions for the bias to be zero: when there are the same number of fish in every brail; and when the proportion of fish in the *k*th class is the same in every brail. Neither of these conditions will be met in practice for any sets other than those with only a single brail (in which case we just have random sampling from the set and of course the bias is zero).

It is useful to look at an example to see how big the bias can be in the absence of observer preference.

Consider, Lawson's "thought experiment" where all brails are of the same weight and there are equal numbers of brails with large yellowfin and small skipjack. If w_Y is the weight of each yellowfin and w_J is the weight of each skipjack, then the biases for the species proportions are:

$$\text{Bias}\left(\text{yellowfin}\right) = \frac{1}{2} - \frac{1}{1 + \frac{w_Y}{w_J}}$$

$$\text{Bias}(\text{skipjack}) = \frac{1}{2} - \frac{1}{1+\dfrac{w_J}{w_Y}}$$

As the yellowfin are bigger than the skipjack, the bias is positive for yellowfin and negative for skipjack. Depending on what relative weights are used the biases can cover the range (-0.5, 0.5).

Lots of other examples could be done, but I will leave that to the reader. The scenario where the proportions are constant across brails provides a “dividing line”. On one side, where the N_{ik} trend up more quickly than the N_i there is a negative bias and on the other side there is a positive bias.

Since the brailing process is similar across sets and vessels and it probably introduces between brail structure, it seems likely that the use of grab samples in the standard calculation will lead to a systematic bias for many sets. The bias will feed directly into aggregated catch estimates which will therefore likely be biased at all spatial and temporal scales.

Clearly, the continued use of the grab-sampling design is far from desirable and the move to an alternative method is fully justified. We will now have a look at the “spill sample” method that is being trialled.

Spill samples

Starting in 2008 there have been commercial purse seine trips where two observers have been sent to conduct simultaneous grab sampling and “spill” sampling from each set. The paired data have been used to “correct” grab sample data prior to it being used to produce catch estimates and length frequencies (Lawson 2012).

The spill sampling method involves taking a relatively large sample from a brail by temporarily opening the bottom of the brail and hence “spilling” fish into a large container. A single spill is of the order of 200-500 fish. Currently every tenth brail is sampled, with the first brail sampled on a set being “randomly” chosen: a “first brail number” is kept note of by the observer and it is incremented by one from set to set; it starts at 1 for the trip and goes as high as 10 before being set back to 1; if the observer notices that a set will have less brails than the current “first brail number” then it is left to their discretion to choose which brail to sample (Lawson pers. comm.).

The larger number of fish sampled in each spill compared to the hand-selected grab samples of about 5 per brail is certainly an improvement. It opens up other possibilities for the calculation of estimates rather than just combining brail samples across the whole set. Also, and importantly, it removes the possibility of human-based preferential selection. There is still the question of whether it provides a random sample from a brail or not, as there is the possibility of structure within a brail. Also, sampling every 10th brail is perhaps not often enough and 200-500 fish per spill does seem a lot. There is also the possibility that the current use of “first brail number” is not providing a uniform spread of brails across position (1st, 2nd, etc.) or brail fullness (we saw one spill-sample form; the sample had been taken from brail 1 which was 1/3 full whereas all other brails were full or near to full).

We need to check the mathematics for spill samples – do we get unbiased estimators?

We will consider a specific set again and continue with the notation and methods used to model the grab samples. Initially, we will assume that a *single* brail is chosen at random for the spill sample which takes m fish (we will assume that m is constant across brails for now). For simplicity, we will assume that the spill sample is multinomial for the given brail and the probability of selection in the kth class is the proportion by number in the class in the brail (this assumption may not be true and should be checked; it will probably need experimental work).

Let,

Z_k = the number of fish in the kth class from the spill sample
$\boldsymbol{Z}$ = $(Z_1, \ldots, Z_s)$

We are assuming,

$\boldsymbol{Z}$ **|** ith brail chosen ~ **Multinomial**$(m, p_{i1}, \ldots, p_{is})$

The natural estimator for the proportion of fish in the kth class in the set is Z_k / m. We can calculate its expected value using a standard conditional probability result ($\mathrm{E}[X] = \mathrm{E}[\,\mathrm{E}(X \mid Y)\,]$).

We have,

$$\mathrm{E}\left(Z_k\right) = \mathrm{E}\left[\mathrm{E}\left(Z_k \mid \ i\text{th brail chosen}\right)\right] = \sum_{i=1}^{b} \frac{1}{b} m p_{ik}$$

Therefore,

$$\mathrm{E}\left(\frac{Z_k}{m}\right) = \frac{1}{b}\sum_{i=1}^{b} p_{ik} = \overline{p}_k$$

This is the same result as for the grab samples (with no selection preference) and the natural spill-sample estimator has the same bias as the grab-sample estimator. However, this formulation did assume that the number of fish in the spill was independent of which brail was chosen and this is unlikely to be true. Another way to model the spill-sample is to assume that the sample size varies across brails according to the mean weight of fish in the brail.

For example, let S be the fixed weight of the spill-sample container when it is filled with a sample and let w_i be the average fish weight in the ith brail, then we have (approximately):

$$m_i w_i = S \rightarrow m_i = \frac{S}{w_i} = \frac{SN_i}{W_i}$$

It is only an approximation because there will be some random variation in the spill-sample size depending on which fish fall in. In any case, we can modify our equations to let m be a random variable whose value depends only on which brail is chosen.

Using conditional expectation again,

$$\mathrm{E}\left(\frac{Z_k}{m}\right) = \mathrm{E}\left[\mathrm{E}\left(\frac{Z_k}{m} \mid \ i\text{th brail chosen}\right)\right]$$

But, if the ith brail is chosen, then $m = m_i$ and Z_k is **Binomial**(m_i, p_{ik}) so,

$$\mathrm{E}\left(\frac{Z_k}{m}\right) = \frac{1}{b}\sum_{i=1}^{b}\frac{1}{m_i} m_i p_{ik} = \overline{p}_k$$

and we have the same bias as before.

The way to get an unbiased spill-sample estimator is to weight the spill sample by an estimate of the number of fish in the selected brail. Because of the large sample sizes we can use the actual sample to get a mean fish weight and then, from the weight of the brail, an estimate of the number of fish in the brail (note, brail fullness and brail capacity are both recorded, so brail weight can be estimated). The sample size is not that crucial. I wouldn't do the weighting with 5 fish, but would be happy to do it with 50 fish.

Let,

W = the weight of the selected brail
$\overline{w}$ = the mean fish weight in the selected brail

and estimate the number of fish in the selected brail using

$$\hat{N} = \frac{W}{\overline{w}}$$

We will estimate the number of fish in the kth class in the *set* (and we do this by scaling the estimated proportion by the estimated numbers of fish in the selected brail and the number of brails).

We have,

$$\mathrm{E}\left(\frac{b\hat{N}Z_k}{m}\right) = \mathrm{E}\left[\mathrm{E}\left(\frac{b\hat{N}Z_k}{m} \mid \ i\text{th brail chosen}\right)\right] = \mathrm{E}\left[\mathrm{E}\left(\frac{bW_i\left(Z_k \mid i\text{th brail}\right)}{m_i\overline{w}_i}\right)\right]$$

where m_i is still the number of fish in the spill sample from the ith brail and $\overline{w}_i$ is the mean fish weight in the sample from the ith brail.

Because of the large sample sizes there is little correlation between the mean fish weight in the spill sample and the number of fish in the kth class in the sample. Also, because of the large sample size, the expected value of the reciprocal of mean fish weight is very close to the reciprocal of the expected value.

That is,

$$\mathrm{E}\left(\frac{bW_i\left(Z_k \mid i\text{th brail}\right)}{m_i\overline{w}_i}\right) \approx \frac{bW_i\,\mathrm{E}\left(Z_k \mid i\text{th brail}\right)}{m_i\,\mathrm{E}\left(\overline{w}_i\right)} = \frac{bW_i m_i p_{ik}}{m_i w_i} = b\frac{W_i}{w_i}p_{ik} = bN_i p_{ik}$$

and

$$\mathrm{E}\left(\frac{b\hat{N}Z_k}{m}\right) \approx \sum_{i=1}^{b} \frac{1}{b} bN_i p_{ik} = \sum_{i=1}^{b} N_i \frac{N_{ik}}{N_i} = \sum_{i=1}^{b} N_{ik}$$

which is exactly what we want.

Important points to note. It would be nice to get the mean fish weight correct (but the errors associated with brail-fullness and brail capacity will probably dominate). Ideally, the sample would be weighed, but the weight could be estimated using set-appropriate length-weight relationships (relationships which are likely to be correct for the particular set given the "type" of set it was – which could be defined by factors including season, year and region as well as associated/unassociated – see the recommendation for collecting length-weight data). Also, the scaling-up to estimate the set numbers has to be done in the prescribed way using an estimate of numbers in the selected brail. It must ***not*** be done using an estimate of numbers in the set (i.e., set weight divided by mean fish weight). The scaling may seem counter-intuitive, but it is what is needed to balance out the errors on *average* – some are under, some are over, but we get the average right – the whole point of getting an unbiased estimator.

When more than one brail is randomly selected for spill sampling then the samples should be combined by scaling each sample in the prescribed manner and then taking the arithmetic average within each species-length class. This will be an (almost) unbiased estimator just like the estimator from each single brail.

To turn a species-length distribution (by number) into an estimate of catch by species, the numbers in each species-length class are multiplied by the corresponding mean fish-weights and summed across length within species. Again, the mean fish-weights for each species-length class should be set-appropriate. Note, the sum of the catch estimates across species will not exactly equal the set weight estimated by the observer or the captain. But, they are just estimates based on eye-balling the catch.

If there are systematic differences between the estimates of total catch from the sampling data and observers eye-ball estimates it could be due to the length-weight relationships used or, more likely, bias in the observers eye-ball estimation procedure. The latter would be a worry as it affects the brail-weighting procedure as well.

Correcting grab-sample estimates using paired spill-sample data

Current approach

The paired grab-sample and spill-sample data are potentially valuable for correcting catch and length frequencies estimates that were based only on grab samples. Lawson (2009, 2012) developed a method for making corrections.

From Lawson (2102):

"The selectivity bias was estimated from the paired grab and spill samples using the model developed in Lawson (2009):

$$n_{jk} = N_{jk}A_j + \varepsilon \qquad (1)$$

$$= \frac{W_k T_{jk}}{\bar{w}_j} A_j + \varepsilon \qquad (2)$$

where n_{jk} is the number of fish in length interval j selected by a grab sampler from set k; N_{jk} is the "true" number of fish in length interval j in set k ; A_j is the probability that a grab sampler will select a fish from among the N_{jk} fish, which can be considered as the *availability* of a fish to be selected; W_k is the total weight of set k; T_{jk} is the "true" proportion of fish of length interval j in set k, in terms of weight, determined from the spill sample taken from set k; $\bar{w}_j$ is the average weight of fish of length interval j; and ε is a random variable of mean zero. Note that when estimating *availability*, each length interval is considered independent and treated separately; the same approach is taken when correcting the historical grab samples with the estimates of *availability* using equations (5)–(10) below."

The spill-sample data, for each set, are substituted into equation (2) as if they were the true values. In the original approach the A_j were estimated directly using a linear model and assuming the errors were normal (Lawson 2009). In the latest approach "availability" is estimated as a continuous function of length by using a cubic spline (see Lawson, 2012). Note, availability is being modelled as a function of length only and is assumed species independent.

I have presented this model in detail so that I can explain the many serious flaws that it contains.

Consider equation (1) first. This says that the expected number of fish from the sample in length class j depends only on the number of fish in the length class within the set and a fixed scalar which is constant from set to set. Two vital parameters are not mentioned: the sample size and the total number of fish in the set. To illustrate why these parameters are important, consider some simple examples.

For simplicity, consider a set where there is only 1 brail and assume that the observer has no length-selection preference. If the observer takes 6 fish as opposed to 3 fish then the expected number of fish in any length class is doubled – sample size matters. Now suppose that the brail contains 10 fish in the jth length class and some number of other fish. If there are 90 other fish then the probability of selecting a fish in the jth length class is 0.1 and if 5 fish are selected the expected number is 0.5 However, if there are 990 other fish then the probability of selecting a fish from the jth length class is 0.01 and for 5 fish the expected value is 0.05. The probability of selecting a fish in a length class depends on the *proportion* of fish in the length class, not just the number of fish in the class.

The thinking on ignoring sample size is perhaps due to considering examples where 5 fish are always taken from each brail (as per the protocol) and thinking that as the number of brails increases so does the sample size and the number of fish in each species-length class. However, for a fixed number of fish in a given species-length class and 5 fish taken from each brail, sample size can still change between sets when there are a different number of brails. Consider two sets which both have 100 fish in the jth length class and 900 other fish. In the first set suppose that the j-class fish are evenly distributed amongst 10 brails each containing a total of 100 fish. Each brail contains 10 j-class fish. The expected number of j-class fish in the sample is $10 \times 5 \times 0.1 = 5$. Now consider a second set where the j-class fish are evenly distributed amongst 20 brails, each containing 50 fish. There are 5 j-class fish in each brail, so the expected number of j-class fish in the sample is 20 × 5 × 0.1 = 10. Sample size really does matter – even if it is always 5 fish sampled per brail and the total number of fish in the set is fixed.

Between-brail structure is also ignored by the model. This is perhaps less serious than ignoring sample size and the number of other fish, because in practice it may be sufficiently random to be accommodated in the error terms. However, if it is systematic (such as large fish tending to be brailed first) then it should be modelled. Consider an extreme example. As in the previous two examples, suppose there are 100 j-class fish and 900 other fish. Also suppose there are 10 brails and 5 fish are taken per brail – but allow different numbers of fish per brail. In one set suppose all the j-class fish are in the first brail and the other fish are in the remaining 9 brails. The expected number of j-class fish in the sample is 5. In another set suppose that the j-class fish are split equally between the first two brails and the other fish are in the remaining 8 brails. The expected number of j-class fish in the sample is now 10. We didn't change the sample size or the number of j-class fish or the total number of fish or the number of brails – just the between-brail structure.

If the A_j in equation (1) are assumed to be constant from set to set then all of the systematic structure from sample size, the number of other fish, and between-brail structure goes into the errors. This defeats the whole purpose of statistical modelling – which is to find a model which adequately describes the structural features of the data so that errors are "random" (without structure).

Equation (2) introduces a mathematical error. The total set weight divided by the mean fish weight (in the set) is the total number of fish in the set. This needs to be multiplied by the proportion of fish in the jth length class, by *number*, to obtain the number of fish in the jth length class. Instead, equation (2) wrongly uses the proportion of fish in the jth length class by *weight*.

The substitution of the spill-sample data for the set's length composition and mean weight is yet another flaw in the approach as the substitution of random variables for unknown parameters introduces additional error structure which is not captured by the single additive error (ε). Given all of the mistakes made in the model it doesn't matter whether a cubic spline is fitted or a linear model is used – the resulting estimates are of no value.

How to do it properly

Earlier in this report I presented proper statistical models for the grab-sample data and the spill-sample data. These are the type of models that should be used, with the paired data, to estimate the unknown parameters including the observer selection-preferences (the r_k which modify the usual multinomial probabilities).

There are two main ways this could be done.

The complicated method is to model the data at the set level and for each set to estimate the length and species composition given a brailing process which preferentially removes some species and/or size of fish. In the simplest case the selectivity of the brail would be assumed constant over all sets. However, more than one brail selectivity could be modelled. Likewise, the observer selection-preference parameters could be assumed constant over all sets or they could be assumed to vary over a small number of classes. Most of the unknown parameters come from defining the species and length composition of each set. These parameters could be incorporated into the model using a number of species-length classes or a mixture of normal distributions. Maximum likelihood estimation would be used and it would need to be implemented in automatic-differentiation software such as ADMB (Fournier 2012). It may be possible to use the unpaired grab-sample data in this analysis as well, since the change in species and length structure across brails may help estimate some of the parameters. Analysing the data in this way should give an understanding of how much of the bias in the grab-sampling data is attributable to between-brail structure and/or selection-preference. It may then be sensible to correct the grab-sample data at the set level.

The simple approach below is the one I recommend in the short-term to produce defensible estimates of species mix and length frequencies (to replace the existing unreliable estimates). However, I would not recommend that the parameter estimates from this approach be used to correct the grab-sample data at the set level. Rather, the method is there as a short-term fix to apply *average* corrections to *aggregated* species-length frequencies.

The paired data can be used to produce aggregated estimates of species-length composition (total *numbers* in each species-length class) for the grab-samples alone and for the spill-samples alone (scaled and aggregated using the methods described in this report). The species-length compositions can then be modelled as multinomial for the spill samples and "modified-multinomial" for the grab samples. The "modified multinomial" is the model I have already used for the grab samples – the normal probabilities are modified by the selection-preferences (the r_k). For this setup, the maximum likelihood estimates can easily be found analytically (using Lagrange multipliers to deal with the constraints).

To be precise:

Suppose that the aggregated species-length distribution (by *number*) for the spill-sample data over a number of sets is given by $\boldsymbol{X}$:

$$\boldsymbol{X} \quad \sim \quad \textbf{Multinomial}(n_x, p_1, \ldots, p_s)$$

where n_x is an assumed effective sample size and p_k is the proportion by number in the kth class for the aggregated catch (i.e., the total catch over the sets which were aggregated)

Also, suppose that the aggregated species-length distribution (by *number*) for the grab-sample data over a number of sets is given by $\boldsymbol{Y}$:

$$\boldsymbol{Y} \quad \sim \quad \textbf{Multinomial}(n_y, r_1 p_1, \ldots, r_s p_s)$$

where n_y is an assumed effective sample size and r_k is the observer selection-preference parameter for the kth class (although here, it is also hoped to be capturing the average between-brail effects as well).

The maximum likelihood estimators of the p_k and r_k are very natural:

$$\hat{p}_k = \frac{X_k}{n_x} \quad \text{and} \quad \hat{r}_k = \frac{Y_k / n_y}{X_k / n_x}$$

The p_k are just nuisance parameters and can be ignored – what matters are the estimates of the r_k and these are just the ratios of the proportions in the aggregated grab and spill samples (Lawson 2008 looked at estimates very similar to these and referred to them as "empirical estimates").

Applying the grab-sample correction
There are two steps to correct an aggregated species-length composition estimated only from grab samples.

First, a selection of paired spill and grab sample data should be chosen to reflect the types of sets which went into the grab-sample-only aggregation. The selection preferences can then be estimated as above. The second step is simply to divide the proportion in each species-length class by the estimated selection-preference.

That is, from grab-sample only data (no paired spill-sample data for these sets), we have an aggregated length-species composition which we model as:

$$\boldsymbol{Y} \sim \textbf{Multinomial}(n_y, r_1q_1, \ldots, r_sq_s)$$

and we want to estimate the q_k (note, I am running out of symbols so I am reusing X and Y). We have the maximum likelihood estimates of the r_k and we know that the maximum likelihood estimates of the r_kq_k are simply the proportions in the species-length distribution (Y_k/n_y), so we have the maximum likelihood estimator for each q_k:

$$\hat{q}_k = \frac{1}{\hat{r}_k}\frac{Y_k}{n_y}$$

Obtaining aggregate estimates

Estimates are wanted over different levels of aggregation. No doubt there are routine reports providing estimates by 5 degree squares and quarter and flag. Estimates on larger spatial scales are used for stock assessment and are split by fleets and associated or unassociated sets.

Plenty of observer data
In years when the observer coverage is sufficient to put adequate sample sizes in each stratum then I assume that individual set estimates are combined (in some fashion) and scaled-up to each stratum and then added across strata. There are different ways to do this. How it was most recently done for length frequencies is described in Lawson (2012). I didn't notice any documentation on *exactly* how the *aggregated* catch by species is estimated from the observer data.

The best approach, for scaling up catches and length frequencies, will differ depending on whether corrected set level data are available or not. In the short term I have suggested that only *aggregated* grab-sample data be corrected. (There is some hope of getting reasonable *average* corrections, but at the individual set level these may not be appropriate.).

To estimate a stratum-wide species-length frequency from the grab-sample data I suggest the following approach. Uncorrected grab-sample data are scaled-up within set (using the estimated *number* of fish in the set) and then added across sets to provide an aggregated species-length frequency. This is then corrected using stratum-appropriate estimates of observer selection-preferences (see the earlier section). Scaling up to stratum numbers is then done by multiplying the proportions within each species-length class by the total number of fish caught in the stratum (estimated using the species-length frequency to get the mean fish weight for the stratum and dividing total stratum catch by the mean fish weight). Effective sample sizes for the species-length frequencies can be estimated by bootstrapping from the raw data.

The stratum-wide catch by species is then estimated from the stratum-wide species-length frequency. The numbers in each species-length class are multiplied by stratum-appropriate mean fish-weight for each species-length class. The uncertainty in the estimates can be estimated by bootstrapping from the raw data.

Little or no observer data

The current method for dealing with strata with little or no observer data is to use predicted catches from logbooks based on GLMs which predict species mix (Lawson 2012). The models have typically been run at the trip level (or higher levels of aggregation) to avoid correlations between errors at the set level. However, Hoyle (2012) did do some exploratory work using a binomial model for skipjack proportion at the set level. He used the "corrected" proportions from the grab sample data and assumed an effective sample size of 10 for each set.

Models at the trip level are inappropriate even for exploratory analysis. We would expect that the underlying processes are occurring on a much finer spatial and temporal scale than that allowed by an analysis of trips. A trip can last 3 months and cover large areas. Aggregation of effects is therefore a certainty and to attribute an effect to just the mode of an explanatory variable is not sensible. Lawson (2012) used modes for some variables and also a weighted average, for latitude and longitude: "The latitude and longitude assigned to each stratum was the average latitude and longitude of the location of sets in the stratum, weighted by the catch". This is inappropriate; for example, a stratum might contain two (or more) clusters of widely separated locations and the location attributed to the catches would then end up being somewhere between the clusters.

Exploratory analysis should be at the set level. Data should not be aggregated in some fashion and then submitted to a GLM as "replicates" – the models used by Dr. Lawson all do some level of aggregation and should be discarded. The model used by Hoyle (2012) is at the set level but is really just for illustrative purposes – showing that there are other ways to look at this data rather than the way that Dr. Lawson did. There is a problem with the type of model that Dr. Hoyle was using. It treated the response variable for a set as a hit or miss on skipjack (a binomial model has a response variable which is either 0 or 1). He was able to use the binomial model by assuming that each set had a sample size of 10 (and each sample was either skipjack or another species). In reality, the "response variable" for each set is the vector of species proportions in the observer sample – it is not a one dimensional object and the effective sample size is not known.

Philosophically, I am not comfortable using (dodgy) models to predict catch by species for a set which was not observed. I think a much simpler approach is appropriate when the required observations do not exist.

A good long term approach to dealing with uncertain historical catches for the three species may be to use a multi-species stock assessment model which has the option, in some years, to estimate species proportions in a given total catch (i.e., parts of the catch history would be non-species specific and just include the total catch across all three species). However, until this model is developed and tested and shown to work well, I would recommend that some simple substitution rules be used to estimate catch histories.

The observer data can be used with whatever explanatory variables are generally available, in the logbooks, to understand which variables are useful for explaining variation in species composition across sets. GLMs and tree regressions can be used together with just plain old descriptive analysis to get an understanding of how best to post-stratify the data (don't let the tree-regressions run wild, as they are prone to do). For these sorts of analysis I would be happy to even use random variables as explanatory variables even though doing so does violate the error assumptions (e.g., total set weight may be useful in explaining some variation in species proportion – that can legitimately be used as a post-stratifying variable even though it shouldn't technically be used as an explanatory variable in a GLM). I am thinking that for some years, where there is good observer coverage that finer spatial and temporal scales than what are currently used could be defensible.

For years in which there is little or no observer data some of the defining factors of the strata are predetermined for particular requirements: e.g., for stock assessment, catches by quarter, associated/unassociated sets, and Multifan-CL areas. However, additional factors which are available from logbooks may come out of the analysis of the observer data (e.g., regional differences in species proportion within the Multifan-CL areas; or differences due to total catch). To estimate species proportion in any given stratum I suggest that an average of estimates from "like" years is used. For some years, when there is adequate observer data in a similar stratum then substitution from that stratum could be used.

By "like" years I mean years in which the species proportions are likely to be similar. It will be important to look at trends in species-proportion estimates across years within strata. If there is a strong trend then clearly early-year estimates should not be made by averaging estimates from later years. Rather, an average of the closest years should be used.

In the long term it would be good to use Bayesian methods to estimate species proportions for all years. In the years with good observer coverage, a uniform Dirichlet distribution would be used as a prior and the updates would be done using the multinomial data from observer sampling. For early years, instead of borrowing the estimates from "like" years, the posteriors from those years would be used as a prior for the earlier years. In the early years the priors could be updated using the logbook estimates of species mix if some sensible statistical assumption could be arrived at for these data (i.e., in terms of how much information they actually contain about species mix).

2. *Based on the findings of (1) above provide recommendations for:*
    a. *Protocols for the sampling of species and size composition by scientific observers aboard purse seine vessels; and/or*
    b. *Recommendations for future experimental work that would lead to the determination of new sampling protocols.*

Grabbing of fish by observers from the top of each brail was never a good idea and the practice needs to be phased out as quickly as is practical. The use of spill samples in a thoroughly researched statistical design seems appropriate.

A proper design does not just involve "specifying protocols". The existing data must be thoroughly analysed and a design developed using theory and simulation (given advice on what levels of precision are required – especially for bigeye, which as a small part of the catch will be difficult to estimate precisely). This is not a job for a reviewer. I have given the basic equations that relate to the existing data and shown how spill-sample data should be scaled to provide unbiased estimates. That is the sort of thing that should have been done back in the 1990s before the observer program began (and then there would have been a very different design). But, it is still not enough. A good design will only be achieved if an experienced fisheries scientist with excellent mathematical and statistical skills is allowed to become fully familiar with the data, the behaviour of observers, and the practical constraints of sampling at sea on purse seine vessels. They can then build an appropriate simulation model, which they can ground-truth with existing data, and use it to test alternative designs before making a recommendation on a particular design.

The basic elements of a good design are probably going to involve spill sampling from randomly selected brails. In practice, a uniform spread of brails across brail position (i.e., 1st, 2nd, etc.) and brail fullness is what should be aimed for. The current approach of using a "first brail number" which is incremented by 1 up to 10 and then set back to 1 is probably not adequate. There are likely to be too many brails where the observer has to choose which brail to sample. The higher the first brail number gets the more likely it is that the next set will have fewer brails than the first one that is to be sampled (lots of sets have fewer than 10 brails – see Appendix 4). Also, taking a very big sample every 10 brails is unlikely to be best approach. Smaller samples (at least 50 fish) from more brails are likely to be better.

Stratification within a set is also likely to be desirable. If there is a trend of decreasing fish size across brails then blocking consecutive brails into strata and choosing a random brail within each stratum is likely to improve precision. The anecdotal evidence that bigeye sometimes float to the surface and are therefore more likely to be in the first brail should be investigated. If it is true then it might be that putting the first brail in its own stratum (for areas and types of sets that are likely to catch bigeye) could improve the precision of bigeye estimates. Indeed, the brail operator could be encouraged to brail the bigeye first (if they are floating).

The use of a single length-weight relationship for each species, in the analysis of the data is of concern. There can be large variations in the "fatness" of fish across time and space. It is not appropriate to be using the length-weight relationships which are used in the stock assessments. Instead, length-weight data should be routinely collected by some of the observers. It needs to be established how much variation there is in the relationships across time and space. Once this is known then an appropriate number of relationships can be estimated and applied to the corresponding data (e.g., perhaps by region within year).

With regard to experimental work there are a couple of questions which I think should be looked at. There is the possibility of within brail structure, perhaps with larger fish tending to be nearer the top. If this is the case, then spill sampling may not provide unbiased estimates of species-length composition from a given brail. So, some experiments looking at within brail structure would be a good idea.

A more subtle question is whether it is appropriate to model spill-samples as a multinomial distribution (even if there is no within-brail structure). Fish are certainly not selected one at a time and it may be that the nature of spilling (with clumps of fish falling into the container) lead to something other than the usual multinomial distribution where the probabilities are directly related to proportion by number.

Conclusions and Recommendations

The current grab-sampling design used for the WCPO tuna purse seine catch is inadequate and current methods used for analysing the data and producing catch and length frequency estimates are badly flawed and therefore inappropriate. The future sampling design should probably use spill-sampling, but the exact nature of the design needs to be determined by an experienced fisheries scientist with excellent mathematical and statistical skills after they have become familiar with the available data, the current sampling procedures, and the practicalities and peculiarities of sampling purse seine catches at sea.

In the short-term, species-mix and length frequencies estimates based on grab-sample data should be corrected at the aggregated level. In the medium to long-term, set-specific corrections could be applied after data are analysed at the set-level using models which estimate brailing selectivities as well as observer selection-preference parameters. I am recommending a short-term fix because there appears to be some urgency to correct the existing catch histories and length frequencies. Currently, they are based on corrections to grab-sample data which use an indefensible model and must therefore be considered very unreliable.

My main conclusions are:

- The grab-sample design leads to estimates of species mix and length frequencies which are biased.
- Spill samples, when properly scaled within set, can be used to produce (nearly) unbiased estimators of species mix and length frequencies.
- The current methods of correcting for the grab-sample bias using paired spill-sample and grab-sample data is based on an indefensible model and the "corrected" grab-sample estimates should be discarded (as should estimates of length frequencies and catch histories which are based on the faulty corrections of grab-sample data).
- Using multinomial models of the paired spill-sample and grab-sample data it is possible to correct aggregated grab-sample estimates of species mix and length frequency (using the maximum likelihood estimates of the parameters).
- Current methods of estimating catch histories for periods when there is little or no observer data are unnecessarily complicated and use inappropriate models.
- Simpler, substitution methods can be used after defining appropriate strata by analysing the observer data at the set level.
- With 100% observer coverage, levels of precision for the two main tuna species (yellowfin and skipjack) should be more than adequate. However, for bigeye, which is generally a small proportion of the catch, there may need to be some careful stratification in order to achieve a good level of precision.

My main recommendations are:

- A simulation model should be developed for testing alternative observer sampling designs with a view to implementing a design based on spill sampling.
- Given guidance on the required levels of precision for bigeye, details that need to be worked out include:
    - how to stratify within sets
    - how to stratify in time and space (extra sampling effort may be needed for bigeye; i.e., more than one observer on some vessels in some seasons/regions)

    - what average size of spill samples/frequency of brail sampling is best
    - how best to achieve a random/uniform distribution of selected brails
    - how best to scale set-level estimates to stratum wide estimates
- The documentation of the design should include a full set of equations (and simulation results if needed) which show that estimators of length frequencies and catch by species are (almost) unbiased for individual sets and for strata.
- Historical data need to be reanalysed to produce defensible estimates of catch histories and length frequencies for use in stock assessment and for other purposes:
    - suitable post-stratification should be determined by model-based and descriptive analysis of the *uncorrected* grab-sample estimates at the set level;
    - paired grab and spill sample data should be used to correct *aggregated* species-length distributions and hence to corrected estimates of catch and length frequency in years with good observer coverage (uncertainty should be estimated by bootstrapping);
    - historical catches in years with little or no observer coverage should be estimated using appropriate stratification (on explanatory variables available in logbooks) and substitution of species-mix estimates from the same stratum type using the average from years that are likely to have had similar species mix (temporal trends in species mix within strata need to be carefully examined).

References

Fournier, D.A. et al. 2012: AD Model Builder: using automatic differentiation for statistical inference of highly parameterized complex nonlinear models. *Optimization Methods and Software 27*: 233–249.

Appendix 1: Bibliography of supplied material

The following documents were supplied before the meeting. Also, during the meeting, some additional background papers were provided. The PowerPoint presentations were also provided (see Appendix 3).

Hampton, J. & P. Williams. 2011. Misreporting of purse seine catches of skipjack and yellowfin-bigeye on logsheets. Working Paper ST–WP–02.

Hoyle, S. 2012. Analyses of factors affecting species composition in purse seine catch. Draft document. 44 p.

Lawson, T.A. 2007. Analysis of the proportion of bigeye in 'yellowfin plus bigeye' caught by purse seiners in the WCPFC Statistical Area. Information Paper SC3–ST–IP5.

Lawson, T.A. 2008. Factors affecting the use of species composition data collected by observers and port samplers from purse seiners in the Western and Central Pacific Ocean. Working Paper SC4–ST–WP3.

Lawson, T.A. 2009. Selectivity bias in grab samples and other factors affecting the analysis of species composition data collected by observers on purse seiners in the Western and Central Pacific Ocean. Working Paper SC5–ST–WP–03.

Lawson, T.A. 2010. Update on the estimation of selectivity bias based on paired spill and grab samples collected by observers on purse seiners in the Western and Central Pacific Ocean. Working Paper SC6–ST–WP–02.

Lawson, T.A. 2011. Purse-Seine length frequencies corrected for selectivity bias in grab samples collected by observers. Information Paper SC7–ST–IP–02.

Lawson, T.A. 2011. Estimation of catch rates and catches of key shark species in tuna fisheries of the Western and Central Pacific Ocean using observer data. Information Paper SC7–EB–IP–02.

Pianet, R., P. Pallarés & C. Petit. 2000. New sampling and data processing strategy for estimating the composition of catches by species and sizes in the European purse seine tropical tuna fisheries. IOTC Proceedings No. 3 (2000) : 104-139. Indian Ocean Tuna Commission, Seychelles.

Appendix 2: Statement of Work for Patrick Cordue

External Independent Peer Review by the Center for Independent Experts

Review of SPC estimation of species and size composition of the western and central Pacific purse seine fishery from observer-based sampling of the catch

Scope of Work and CIE Process: The National Marine Fisheries Service's (NMFS) Office of Science and Technology coordinates and manages a contract providing external expertise through the Center for Independent Experts (CIE) to conduct independent peer reviews of NMFS scientific projects. The Statement of Work (SoW) described herein was established by the NMFS Project Contact and Contracting Officer's Representative (COR), and reviewed by CIE for compliance with their policy for providing independent expertise that can provide impartial and independent peer review without conflicts of interest. CIE reviewers are selected by the CIE Steering Committee and CIE Coordination Team to conduct the independent peer review of NMFS science in compliance the predetermined Terms of Reference (ToRs) of the peer review. Each CIE reviewer is contracted to deliver an independent peer review report to be approved by the CIE Steering Committee and the report is to be formatted with content requirements as specified in **Annex 1**. This SoW describes the work tasks and deliverables of the CIE reviewer for conducting an independent peer review of the following NMFS project. Further information on the CIE process can be obtained from www.ciereviews.org.

Project Description Estimates of species and size composition of the purse seine fishery in the western and central Pacific Ocean (WCPO) are fundamental to stock assessments of skipjack, yellowfin and bigeye tunas conducted by the Secretariat of the Pacific Community (SPC). These assessments provide the scientific basis for management decision making by the Western and Central Pacific Fisheries Commission for the world's largest tuna fishery. Estimates of purse seine catch by species and their size compositions are currently based on at-sea observer sampling protocol known as 'grab sampling', whereby 5 fish are randomly selected from each brail of each purse seine set as they are loaded onboard during observed trips, and some total catch information taken from skippers logbooks. However, it was recently recognized that grab sampling was biased because of human and other factors involved in selecting the fish for sampling and that some logsheet reporting of catches is biased towards skipjack tuna.

A series of experiments were conducted whereby grab sampling was conducted alongside another sampling method, termed 'spill sampling', in which a substantial quantity of fish was 'spilled' from selected brails during the fish loading process. These paired sampling trials allowed a bias correction to be estimated and applied to the historical grab sampling data. The corrected grab sampling data were then analyzed using statistical modeling approaches to estimate species and size composition, testing for and taking account of as appropriate potential independent variables, such as latitude, longitude, set type, year and season. The derived models are used to estimate historical catches and size composition of each species, by fleet and various spatial, temporal and operational strata.

An independent CIE review is requested to evaluate the scientific information and methodology of the SPC estimation of species and size composition of the western and central Pacific purse seine fishery. The Terms of Reference (ToRs) of the peer review are attached in **Annex 2**. The tentative agenda of the panel review meeting is attached in **Annex 3**.

Requirements for CIE Reviewers: Two CIE reviewers shall conduct an impartial and independent peer review in accordance with the SoW and ToRs herein. CIE reviewer team shall have the combined expertise and current working experience in stock assessment, fisheries and mathematical statistics, and statistical modeling. Familiarity with the pelagic tuna fisheries and/or catch sampling protocols is desirable. Each CIE reviewer's duties shall not exceed a maximum of 14 days to complete all work tasks of the peer review described herein.

Location of Peer Review: Each CIE reviewer shall conduct an independent peer review during the panel review meeting scheduled in Noumea, Caledonia during October 22-25, 2012.

Statement of Tasks: Each CIE reviewer shall complete the following tasks in accordance with the SoW, ToRs, and Schedule of Milestones and Deliverables herein.

Prior to the Peer Review: Upon completion of the CIE reviewer selection by the CIE Steering Committee, the CIE shall provide the CIE reviewer information (full name, title, affiliation, country, address, email) to the COR, who forwards this information to the NMFS Project Contact no later than the date specified in the Schedule of Milestones and Deliverables. The CIE is responsible for providing the SoW and ToRs to the CIE reviewers. The NMFS Project Contact is responsible for providing the CIE reviewers with the background documents, reports, foreign national security clearance, and other information concerning pertinent meeting arrangements. The NMFS Project Contact is also responsible for providing the Chair a copy of the SoW in advance of the panel review meeting. Any changes to the SoW or ToRs must be made through the COR prior to the commencement of the peer review.

Foreign National Security Clearance: When CIE reviewers participate during a panel review meeting at a US government facility, the NMFS Project Contact is responsible for obtaining the Foreign National Security Clearance approval for CIE reviewers who are non-US citizens. For this reason, the CIE reviewers shall provide requested information (e.g., first and last name, contact information, gender, birth date, passport number, country of passport, travel dates, country of citizenship, country of current residence, and home country) to the NMFS Project Contact for the purpose of their security clearance, and this information shall be submitted at least 30 days before the peer review in accordance with the NOAA Deemed Export Technology Control Program NAO 207-12 regulations available at the Deemed Exports NAO website:
http://deemedexports.noaa.gov/
http://deemedexports.noaa.gov/compliance_access_control_procedures/noaa-foreign-national-registration-system.html

Pre-review Background Documents: At least two weeks before the peer review, the NMFS Project Contact will send (by electronic mail or make available at an FTP site) to the CIE reviewers the necessary background information and reports for the peer review. In the case where the documents need to be mailed, the NMFS Project Contact will consult with the CIE Lead Coordinator on where to send documents. CIE reviewers are responsible only for the pre-review documents that are delivered to the reviewer in accordance to the SoW scheduled deadlines specified herein. The CIE reviewers shall read all documents in preparation for the peer review.

Panel Review Meeting: Each CIE reviewer shall conduct the independent peer review in accordance with the SoW and ToRs, and shall not serve in any other role unless specified herein. **CIE reviewer shall not be required to participate in a consensus review.** Each CIE reviewer shall actively participate in a professional and respectful manner as a member of the meeting review panel, and their peer review tasks shall be focused on the ToRs as specified herein. The

NMFS Project Contact is responsible for any facility arrangements (e.g., conference room for panel review meetings or teleconference arrangements). The NMFS Project Contact is responsible for ensuring that the Chair understands the contractual role of the CIE reviewers as specified herein. The CIE Lead Coordinator can contact the Project Contact to confirm any peer review arrangements, including the meeting facility arrangements. **Modifications to the SoW and ToRs can not be made during the peer review, and any modification to the SoW or ToRs prior to the peer review must be approved by the COR and CIE Lead Coordinator.**

Contract Deliverables - Independent CIE Peer Review Reports: Each CIE reviewer shall complete an independent peer review report in accordance with the SoW. Each CIE reviewer shall complete the independent peer review according to required format and content as described in Annex 1. Each CIE reviewer shall complete the independent peer review addressing each ToR as described in Annex 2.

Specific Tasks for CIE Reviewers: The following chronological list of tasks shall be completed by each CIE reviewer in a timely manner as specified in the **Schedule of Milestones and Deliverables**.

1) Conduct necessary pre-review preparations, including the review of background material and reports provided by the NMFS Project Contact in advance of the peer review.
2) Participate in the panel review meeting in Noumea, Caledonia during 22-25 October 2012.
3) In Noumea, Caledonia during 22-25 October 2012 as specified herein, and conduct an independent peer review in accordance with the ToRs (**Annex 2**).
4) No later than **November 9, 2012,** each CIE reviewer shall submit an independent peer review report addressed to the "Center for Independent Experts," and sent to Mr. Manoj Shivlani, CIE Lead Coordinator, via email to shivlanim@bellsouth.net, and CIE Regional Coordinator, via email to Dr. David Die ddie@rsmas.miami.edu. Each CIE report shall be written using the format and content requirements specified in Annex 1, and address each ToR in **Annex 2**.

Schedule of Milestones and Deliverables: CIE shall complete the tasks and deliverables described in this SoW in accordance with the following schedule.

17 September 2012	CIE sends reviewer contact information to the COR, who then sends this to the NMFS Project Contact
8 October 2012	NMFS Project Contact sends the CIE Reviewers the pre-review documents
22-25 October 2012	Each reviewer participates and conducts an independent peer review during the panel review meeting
9 November 2012	CIE reviewers submit draft CIE independent peer review reports to the CIE Lead Coordinator and CIE Regional Coordinator
23 November 2012	CIE submits CIE independent peer review reports to the COR

30 November 2012	The COR distributes the final CIE reports to the NMFS Project Contact and regional Center Director

Modifications to the Statement of Work: This 'Time and Materials' task order may require an update or modification due to possible changes to the terms of reference or schedule of milestones resulting from the fishery management decision process of the NOAA Leadership, Fishery Management Council, and Council's SSC advisory committee. A request to modify this SoW must be approved by the Contracting Officer at least 15 working days prior to making any permanent changes. The Contracting Officer will notify the COR within 10 working days after receipt of all required information of the decision on changes. The COR can approve changes to the milestone dates, list of pre-review documents, and ToRs within the SoW as long as the role and ability of the CIE reviewers to complete the deliverable in accordance with the SoW is not adversely impacted. The SoW and ToRs shall not be changed once the peer review has begun.

Acceptance of Deliverables: Upon review and acceptance of the CIE independent peer review reports by the CIE Lead Coordinator, Regional Coordinator, and Steering Committee, these reports shall be sent to the COR for final approval as contract deliverables based on compliance with the SoW and ToRs. As specified in the Schedule of Milestones and Deliverables, the CIE shall send via e-mail the contract deliverables (CIE independent peer review reports) to the COR (William Michaels, via William.Michaels@noaa.gov).

Applicable Performance Standards: The contract is successfully completed when the COR provides final approval of the contract deliverables. The acceptance of the contract deliverables shall be based on three performance standards:
(1) The CIE report shall completed with the format and content in accordance with **Annex 1**,
(2) The CIE report shall address each ToR as specified in **Annex 2**,
(3) The CIE reports shall be delivered in a timely manner as specified in the schedule of milestones and deliverables.

Distribution of Approved Deliverables: Upon acceptance by the COR, the CIE Lead Coordinator shall send via e-mail the final CIE reports in *.PDF format to the COR. The COR will distribute the CIE reports to the NMFS Project Contact and Center Director.

Support Personnel:

William Michaels, Program Manager, COR
NMFS Office of Science and Technology
1315 East West Hwy, SSMC3, F/ST4, Silver Spring, MD 20910
William.Michaels@noaa.gov Phone: 301-427-8155

Manoj Shivlani, CIE Lead Coordinator
Northern Taiga Ventures, Inc.
10600 SW 131st Court, Miami, FL 33186
shivlanim@bellsouth.net Phone: 305-383-4229

Roger W. Peretti, Executive Vice President
Northern Taiga Ventures, Inc. (NTVI)
22375 Broderick Drive, Suite 215, Sterling, VA 20166
RPerretti@ntvifederal.com Phone: 571-223-7717

Key Personnel:

NMFS Project Contact:

Keith Bigelow
NOAA National Marine Fisheries Service
Pacific Islands Fisheries Science Center
2570 Dole Street, Honolulu, HI 96822
Email: keith.bigelow@noaa.gov Phone: 808-983-5388

Annex 1: Format and Contents of CIE Independent Peer Review Report

1. The CIE independent report shall be prefaced with an Executive Summary providing a concise summary of the findings and recommendations.

2. The main body of the reviewer report shall consist of a Background, Description of the Individual Reviewer's Role in the Review Activities, Summary of Findings for each ToR in which the weaknesses and strengths are described, and Conclusions and Recommendations in accordance with the ToRs.

   a. Reviewers should describe in their own words the review activities completed during the panel review meeting, including providing a brief summary of findings, of the science, conclusions, and recommendations.

   b. Reviewers should discuss their independent views on each ToR even if these were consistent with those of other panelists, and especially where there were divergent views.

   c. Reviewers shall provide a critique of the NMFS review process, including suggestions for improvements of both process and products.

   d. The CIE independent report shall be a stand-alone document for others to understand the weaknesses and strengths of the science reviewed.

3. The reviewer report shall include the following appendices:

   Appendix 1: Bibliography of materials provided for review
   Appendix 2: A copy of the CIE Statement of Work
   Appendix 3: Panel Membership or other pertinent information from the panel review meeting.

Annex 2: Terms of Reference for Peer Review of the

Review of SPC estimation of species and size composition of the western and central Pacific purse seine fishery from observer-based sampling of the catch

1. Evaluate and provide recommendations on the statistical methods used to estimate species and size composition (skipjack, yellowfin and bigeye tuna) in the purse seine fishery, with particular attention to the following issues:

   a. The need for, and approaches to, simultaneous estimation of catches for all three species

   b. The need for, and approaches to, simultaneous estimation of the size and species composition.

   c. Approaches to take into account factors that influence the size and species composition, e.g., season, location, set type, and vessel flag.

   d. Approaches for interpolation (e.g., statistical versus substitution) where sample data are low or absent, especially for the estimation of historical catches.

   e. Approaches to characterize uncertainty in estimates of the catch by species;

   f. The ability to provide reliable estimates of catch by species at different levels, e.g. the set, the trip, vessel flag, assessment region.

   g. Approaches to analyze the paired spill and grab sample trials.

2. Based on the findings of (1) above provide recommendations for:

   a. Protocols for the sampling of species and size composition by scientific observers aboard purse seine vessels; and/or

   b. Recommendations for future experimental work that would lead to the determination of new sampling protocols.

Annex 3: Tentative Agenda

Review of SPC estimation of species and size composition of the western and central Pacific purse seine fishery from observer-based sampling of the catch

Secretariat of the Pacific Community
Promenade Roger Laroque
Noumea, New Caledonia
Phone: +687 262000
22-25 October 2012

1. Overview of observer sampling protocols including the paired spill and grab sample trials

2. Overview of logsheet data collection and known issues

3. Overview of historical and recent approaches to the estimation of species and size composition of purse seine catches.

4. Interactive examination of the issues outlined in the TOR. SPC staff will be available to produce data summaries and run statistical analyses to support this examination.

5. Reviewers prepare draft reports,

6. Presentation of preliminary review findings by the reviewers, with discussion and feedback from SPC staff.

Appendix 3: Panel membership and meeting agenda

The review panel consisted of two CIE appointed reviewers:

Mr. Patrick Cordue, Fisheries Consultant, New Zealand
Dr. Joseph Powers, Associate Professor, Department of Oceanography and Coastal Sciences Louisiana State University.

The meeting was chaired by Dr. Harley on Monday and then was run informally on Tuesday, Wednesday, and Thursday (with Dr. Hampton chairing when he was present). The agenda followed was:

Monday 22 October 2012:

Introduction and welcome by Dr. Harley.

Presentations:

Background and overview	Dr. Williams
Current sampling methods and results	Dr. Lawson
Set-based binomial modelling	Dr. Hoyle
Mis-reporting on logsheets	Dr. Williams

Tuesday 23 October 2012

Morning:	Reviewers formulate and document requests
Afternoon:	Reviewers begin formulating/drafting preliminary findings. SPC staff work on requests.

Wednesday 24 October 2012

Morning:	Reviewers continue work on preliminary findings. SPC staff work on requests.
Afternoon:	Presentation and discussion of results from reviewers requests. Discussion with SPC staff who had at-sea experience with observer sampling. Additional follow-up requests.

Thursday 25 October 2012

Morning:	Reviewers continue work on preliminary findings. SPC staff work on requests.
Afternoon:	Presentation and discussion of results from additional requests. Presentation of reviewers preliminary findings and discussion.

Appendix 4: Reviewer requests and selected responses

Two requests were made for additional presentations by the SPC scientists. The first request was formulated on the morning of 23 October 2012 by both reviewers in consultation. I made an additional request on the evening of 24 October 2012 (dated 25 October 2012).

The SPC scientists fulfilled our requests, where possible, in a timely manner. The initial request for statistics with regard to grab-sampling data assumed that 5 fish were sampled per brail. Dr. Lawson advised that this was not always the case and that results would have to be interpreted with this in mind. The second lot of requests were made with the knowledge that there was some variation in the number of fish sampled per brail. For the second request, number 1, was not able to be responded to as the positions of the sampled brails were not recorded on the spill-sample database (despite being on the spill-sample form). Results for request 2, number 2, were presented on 25 October 2012 and they are reproduced here to show the peculiar local modes in the number of brails per set (Figure 2.1; which is still to be explored) and the remarkable variation in the average number of fish sampled per brail (Figure 2.2). It is also noteworthy that the brail-fullness code for an associated set is almost never equal to 1 (which denotes a full brail; Figure 2.3).

Annex 1: First request

CIE Tuna purse seine catch-sampling review

P. Cordue
J. Powers

23 October 2012

Request for statistics and plots

1. For 2004-2010 when observer coverage was about 10% or better:

Histograms of total catch per set, by year, split associated, unassociated.

For each of the ten main 5 degree x 5 degree squares, by year, associated sets only:

- Histograms of total catch per set (with mean, sd)
- Histograms of proportion of skipjack per set (from observer data, with mean, sd)
- Total number of sets
- Proportion of sets sampled

2. For grab sampling data with correct ordering of samples across brails (all years combined):

Histogram of number of brails per set, split by associated, unassociated
Histogram of average brail weight per set, split by associated, unassociated
Histogram of x/y per set where:

(a) x = average weight of fish in first brail; y = average weight of fish in last brail
(b) x = average weight of fish in first two brails; y = average weight of fish in last two brails
(c) x = average weight of fish in first three brails; y = average weight of fish in last three brails

Annex 2: Second request

CIE Tuna purse seine catch-sampling review

P. Cordue

25 October 2012

Request for statistics and plots (in order of priority)

1. For all spill-sampling data:

a. Compare the brail-fullness codes of spill-sampled brails with the brail-fullness codes of all brails. The question is whether the selection of brails for spill sampling looks like it is random/uniform with regard to brail-fullness.

Suggestion: a bar plot showing for each fullness-code the proportion for spill-sampled brails and the proportion for all brails (aggregated across all spill-sampled sets)

b. Look at whether the selection of brails for spill sampling is random/uniform with regard to brail order.

Suggestion:

Categorize sets according to the number of brails: 2, 3, …, 9, and 10 or more. Present a bar plot for each category: with less than 10 brails show the proportion of sets for which each brail position was sampled (e.g., when 2 brails: #1 proportion and #2 proportion); for the 10 or more brail category, show the proportion of sets for which each brail position was the first brail sampled. Also show the number of sets for each category.

2. For all grab sampling data (split by associated, unassociated)

Histogram of number of brails per set
Histogram of average number of fish sampled per brail per set
Histogram of brail fullness-codes aggregated across all sets

3. Some continued exploration of the large increase in 2010, in some 5 degree squares, of observed associated-sets with a very low proportion of skipjack. Is it genuine or some problem with misidentification of species or a change in definition of associated/unassociated sets or mis-recording of the type of set?

Annex 3: Response to second request, number 2.

These graphs were supplied by Dr. Lawson.

Figure 2.1 Number of brails per set

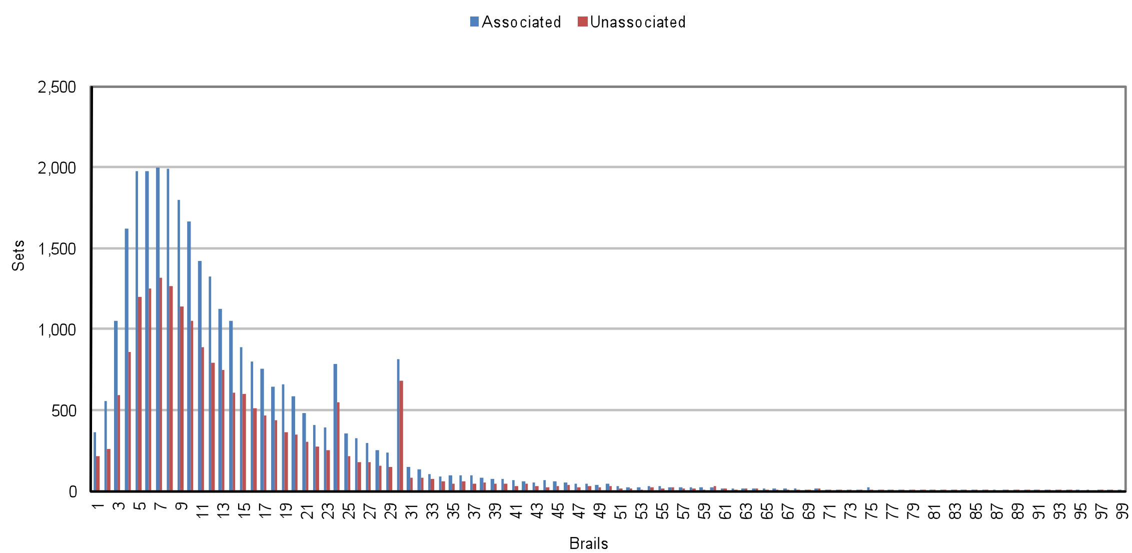

Figure 2.2 Average number of fish sampled per brail per set

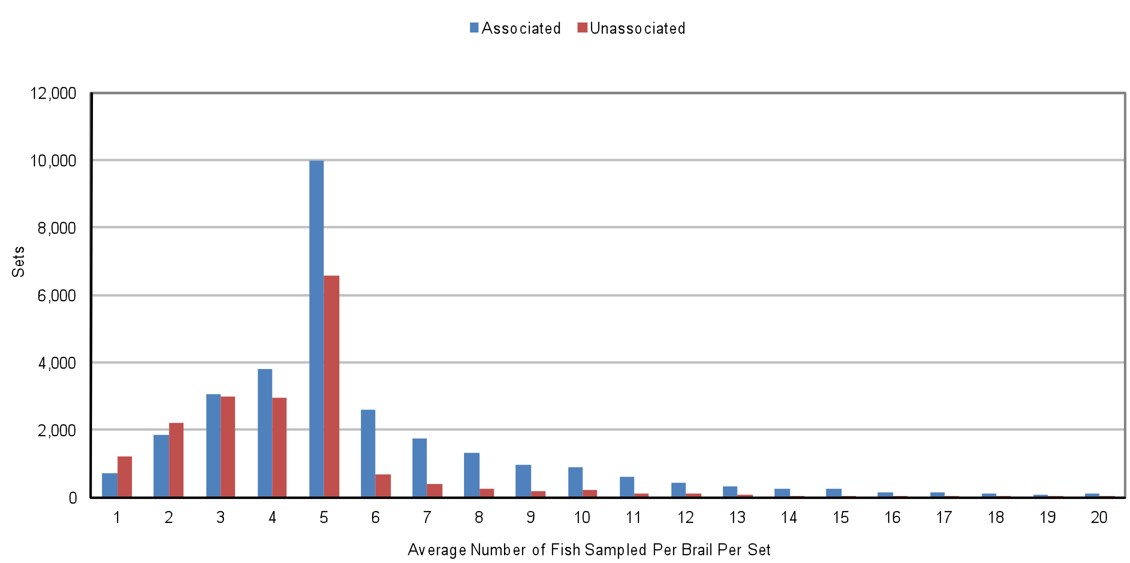

Figure 2.3 Percentage of brails by fullness. Associated: N = 36,342. Unassociated: N = 31,242.