**Hulson et al. Response to Reviews**

Overall response…

**Editorial review:**

**Review:** *I agree with both reviewers of CJFAS-2023-0164 that this manuscript provides a method for demonstrating AFSC is collecting too many length samples, but perhaps not enough age samples, in its surveys. Both reviewers noted places where the wording could be improved to facilitate understanding by readers not familiar with the AFSC surveys. I especially liked the suggestion to link the steps with the flowchart using numbers or some other approach.*

**Response:** To address the reviewers comments we have added additional detail to the methods section to further explain the expansion methods used to obtain age and length composition. We have also added text to the introduction to more precisely describe the terminology we use when discussing realized sample size and input sample size (and have taken effective sample size out, we recognize that all these related terms can get extremely confusing). We have also attempted to link Figure 1 (the flowchart) to the methods by adding numbers to denote the step described in text.**Review:** *My main concern is one raised by both reviewers, that the work is so AFSC-centric that it is difficult for readers to draw conclusions for their own situation. I agree with the reviewers that broadening the discussion to provide details that readers should consider if they want to perform a similar analysis on their survey would help. One example is to compare the results when samples are collected in 1, 2, and 5 cm bins to demonstrate how this influences the conclusions, if at all.*

**Response:** We found this to be a very helpful suggestion and in the revised analysis we have included comparison between 1, 2, and 5 cm bins. We have also included a paragraph in the discussion (Lines ##-##) that attempts to generalize these results, as it pertains to bin structure, spatial scale, expansion methods, and additional metrics.**Review:** *Another aspect of the work that deserves more attention is the assumption that the original sample is sufficient for use in the bootstrapping procedure. Some guidance to readers for what to consider when making this determination would be useful. For example, is it OK to have nearly all the lengths come from one small portion of the survey area, are there minimum sample sizes needed to conduct the analysis, and does the approach used to collect the ages (random vs stratified) matter are all questions that could be addressed in the discussion.*

**Response:** spatially spread out, total sample size, not ok to have lengths from one part of survey area, minimum sample size? Add text to methods section when introducing species**Review:** *One aspect of survey sampling that was not addressed by either reviewer is the opportunity cost of not collecting samples. By this, I mean that once the survey is done there is no possibility of increasing the sample size if a new question arises, such as spatial changes to a stock definition. I would welcome the authors’ thoughts on this aspect of data collection during surveys.*

**Response:** think about this and perhaps add to a caveats para in the discussion**Review:** *I agree with the reviewers that the title of the manuscript is a bit misleading given how little attention is paid to workforce health and efficiency. I suggest either modifying the title and abstract, or else providing some (even hypothetical) calculations demonstrating actual trade-offs at different levels of sampling. Currently, a reader could infer that the authors prefer no data collection at all to avoid workforce health issues, something I am sure they do not support.*

**Response:** change title? Get Jason to help craft that para**Review:** *The use of ISS as the only metric is a bit troubling to me, despite the excellent introduction supporting it. I wonder if some discussion about other possible metrics could be included, with some notes about why they were not selected for this work. This would help readers put the work into context of other possible ways to analyze survey data. Additionally, given the rise of self-weighting error distributions for age and length composition data in stock assessments, is ISS still a relevant metric to use when comparing length-sampling strategies? Finally, could there be important factors hidden by the use of ISS?*

**Response:** We found this to be a helpful suggestion, and in the revised analysis we have included additional metrics for consideration besides ISS, in particular the relative bias in mean length (for male and female) and the sex ratio of the length frequency data. We feel that these additional metrics provides the reader useful information to be reassured that limiting sampling of length-frequency data to 100-150 fish per haul does not have significant implications to any metric you would want to derive from length composition data (and that ISS isn’t hiding any unknown source of bias); we have noted this in the revised discussion (Lines ##-##). Albeit, the main focus of the analysis remains ISS, as sub-sampling will have impact on uncertainty to a greater extent than bias, as long as the samples are still collected randomly. On the topic of ISS relevancy, we maintain that ISS is still a relevant statistic, even given the rise of self-weighting error distributions. For example, whether using the Francis iterative reweighting method (Francis ##), the Dirichlet (Hulson et al. 2012), or the Diriclet-Multinomial (Thorson et al. ##) the age/length composition requires a sample size as a starting point to then self-weight. There are two primary points to be made about this feature of self-weighting methods. First, is that each of these self-weighting methods estimates a weighting factor that is time-invariant, where the bootstrap-simulation method used in this study provides time-dependent ISS. This then provides an indication of uncertainty in one year’s data compared to another, which is an unattainable feature in self-weighting methods. Second, the scale of the starting sample size matters. While in theory one would assume that the self-weighting methods can ‘correct’ for misspecification of scale in the starting sample size, in practice (with real-world data) this doesn’t actually occur and the stock assessment is sensitive to the starting point. We would suggest that the bootstrap-simulation ISS is a reasonable magnitude to use as a starting point as it is derived from consideration of the sources of uncertainty that are inherent to compositional data. All that said, these arguments speak into the appropriateness of one ISS compared to another (in particular, whether the bootstrap-simulation ISS would be the ‘right’ one to use) and is beyond the scope of the current paper, which is focusing on effects of sub-sampling using ISS as a metric. However, this is a current research topic that several of us are interested in pursuing further in distinct analyses from the current study.

**Reviewer 1:**

**Review:** *Comments to the Author  
This manuscript presents an analysis of length and age sampling for a suite of species monitored with fishery-independent trawl surveys in the North Pacific. The authors find that lengths are generally being sampled above, and ages below levels that would provide optimal effective sample sizes. The methods follow similar approaches applied elsewhere and appear appropriate for this type of data. I think that the results of this study will be of interest to others considering optimization of survey designs and/or unplanned reductions due to unique circumstances as well as for refinement of data collection targets for the specific surveys evaluated.  I suggest that some additional information on the sampling targets/properties for these surveys, as well as more detail on the specific approach used for the age-length keys would be helpful. The manuscript is well written, and I have only a few minor editorial suggestions.*

**Response:** We thank the reviewer for their helpful comments and suggestions. To address some of this reviewer’s concerns we have adding sampling information to Table # and have revised the methods section to add additional detail for how the length and age compositions are expanded (which includes formulae for the age-length key). The specific reviews regarding these topics are further addressed below.

**Review:** *The details of how the age-length keys are constructed (mentioned on lines 155-156) seem critically important to this analysis and the application of similar methods for other surveys. I think it would be worth summarizing some of the key information in this manuscript, even with the reference to Hulson et al. (2023). Specifically, do the keys for all species use 1 cm length bins regardless of the range in fish size (implied by the notation in Hulson et al. 2023)? If 2 cm bins, or whatever bin structure is used in the respective stock assessments, were used would this have implications for the calculation of effective sample size? How are sizes that are not represented in the length samples propagated in the age-length keys (an issue noted in Siskey et al. (2023)? Are age-length keys ever found to differ among strata within a survey potentially necessitating more length and age samples for sub-strata? Some of this may be beyond the scope of the manuscript to test specifically, but worth including in the discussion if deemed important.*

**Response:** We have revised the methods section and have added additional detail on how the first (length) and second (age) expansions are conducted (Lines ##-## in the revised manuscript). We also note that, while the AFSC uses a 1 cm bin structure, the formulae we use in this analysis are flexible to include other binning structures. We agree with the reviewer that it would be useful to evaluate the impact of bin structure on these results and have included analysis that alters the bin structure and compares the results of sub-sampling in these cases. We show that... We also note, that the standard operating procedure of the AFSC trawl surveys have changed over time, and currently all fish that are sub-sampled for age are also included in the length frequency samples. Thus, the mismatches noted in Siskey et al. (2023) haven’t occurred in over a decade in the historical data, and also note that these mismatches that were discussed happened extremely infrequently, on the order of less than a dozen times in the historical data. However, this was an important topic to note in the Sisky et al. (2023) paper so that surveys ensure that these data are integrated in both sources (age and length frequency). The topic of differing age-length keys among strata is certainly an important one. In the analyses we conducted we also investigated the impact of sub-sampling at the sub-region scale on length composition and it showed that the main result of 100-150 lengths per haul was consistent at that level, however, we decided not to include it in the original version of the manuscript. We have included in this version a figure that investigates sub-sampling using the Gulf of Alaska sub-regions as an example. However, investigating the optimal amount of age sampling at the strata level is beyond the scope of the current manuscript, here we show that at historical sampling levels we haven’t reached the ‘saturation’ point. We note that we are planning further investigations, similar to this work and the Siskey et al. (2023) study that investigate whether, or where, there is a saturation point for age sampling and will consider including investigations at the strata or sub-region level.

**Review:**  *Lines 269-271 and bottom panels of Figure 2: Was the age composition ISS considered at the full age sample only? Were those results tested for length frequency sampling of 100 fish/haul? If this is the recommendation for the paper, it could be helpful to report that the same relationship held at those sampling levels as well (unless I am misunderstanding something).*

**Response:** To clarify this result we have indicated that the full age specimen dataset was used while the length frequency sub-sampling level was changed in the revised results section (Lines ## - ## of the revision). To answer the reviewer’s question, yes, the full specimen age data was tested with all the length sub-sampling cases investigated.

**Review:** *The manuscript notes (lines 99-101) that it is important to consider optimizing the samples per haul as well as the number of hauls with samples. It would be helpful to provide additional information on this, perhaps as part of Table 1, where the average number of hauls sampled for age and length as well as the number of total fish captured, and total hauls made could be added. This would allow an easier evaluation of how much subsampling has occurred and how many additional hauls could potentially be sampled, per the recommendations on lines 404-407.*

**Response:** We agree with the reviewer that this would be useful information and have added the average number of hauls sampled for age and length frequency, and the total number of hauls in which the species was captured to Table S1 in the supplementary material section (referenced in Lines ##-## of the revision). We added this to the supplementary materials section rather than the main text in order to keep the number of tables and figures low.  
  
**Review:** *Lines 53-69: Suggest also referencing Thorson et al. (2023) in this paragraph.*

**Response:** We have referenced Thorson et al (2023) in this paragraph and have also revised this paragraph to indicate that ‘realized sample size’ is the sample size that accounts for additional uncertainty in composition data, and ‘input sample size’ is the sample size used in the multinomial likelihood when fitting composition data. We recognize that the terms ‘effective sample size’, ‘realized sampled size’, and ‘input sample size’ can get confusing and used interchangeably, however, we have attempted to make our use of terminology consistent within the revision as well as consistent with current use of terminology (agreeing much with terminology used in Thorson et al. 2023). This reference is on Line ##, and the revised paragraph starts on Line ## in the revised manuscript.

**Review:** *Lines 86-88 and line 452: Suggest referencing Link et al. (2008) here.*

**Response:** We have referenced Link et al. (2008) in this sentence (Line ## of the revised manuscript).

**Review:** *Lines 94-97: This manuscript does not evaluate the costs of otolith reading. I suggest omitting this topic, as there could be subsampling of otoliths to be prepared and read that would achieve the same cost reduction in that process as achieved for reducing the collection of otoliths in the field.*

**Response:** Our intention of including this sentence in the manuscript was to provide context for the entirety of the collection-to-assessment process when considering age composition data and all the associated costs. We agree with the reviewer that we did not evaluate the costs of otolith reading and that cost reduction can be achieved in a variety of ways, however, evaluating costs was not our intent when including this sentence. We agree with the suggestion to omit this sentence, as upon further reflection it doesn’t add to paragraph, and have omitted this sentence in the revised manuscript.

**Review:** *Line 121: consider rephrasing since the analysis considers decreasing not increasing the number of length and age samples.*

**Response:** We have rephrased this question in the introduction to reflect that we are investigating decreasing sampling intensities (and appreciate that ‘increasing’ could cause confusion here, Lines ## - ## in the revision).

**Review:** *Line 130-131: Do you mean species-specific exceptions for ‘maximum’ length subsample size? I understand the protocol to be collecting up to the target sample size depending on the catch.*

**Response:** Currently on the AFSC surveys there is a ‘minimum’ length subsample size, that is, field scientists are instructed to collect ‘at least’ a certain amount of length samples for a certain species. In this study, we flip that coin to help provide guidance to these field scientists on what the ‘maximum’ needs to be, as it has been the case (and as we show) that many more fish were sampled for sex-specific lengths than needed to be. We recognize, however, that this phrase could be confusing given the motivation of our study and have omitted it from the revision for clarity.

**Review:** *Lines 139-140: For sampling protocol #2 are the lengths for fish sampled for otoliths also included with the length samples when analyzed (do the 2-20 fish count toward the 100 for both sampling and input sample size calculations)?*

**Response:** As we note in a previous response, yes, for at least two decades the lengths of fish sampled for otoliths are included in the length frequency samples in AFSC surveys.

**Review:** *Lines 234-237: It is worth noting here that the ISS is analogous to the term “effective sample size” as used in the introduction.*

**Response:** As noted in a response above, we have reworded the introduction to distinguish between ‘realized sample size’ and ‘input sample size’, and have removed ‘effective sample size’ from the revision. We have not noted this in the revision, but we note here that we agree with the terminology used in Thorson et al. (2023) and reserve the use of ‘effective sample size’ to denote the degree of agreement between a modeled age composition and an observed age composition as a model output diagnostic.

**Review:** *Lines 294-296 and Table 2: I think the text and caption are backwards here and should be 150 to 100 rather than 100 to 150, since the reductions are smaller to larger.*

**Response:** The reviewer is correct here that these were inadvertently switched in the original version. We have corrected this in the revision in both the text (Line ##) and in the caption of Table 2.

**Review:** *Line 350: ‘greater’*

**Response:** ‘Great’ has been changed to ‘greater’ (Line ## of the revision).

**Review:** *Lines 367-381: I’ not convinced it is very complicated to have species specific sampling rates for a subset of species where it may be warranted by real differences in ISS for a given rate of sampling and where the value of that information may be much higher than other species (e.g., walleye pollock).*

**Response:** Between this paragraph and the following paragraph our intent was to show that our results were consistent across species, and in terms of uncertainty that would be subsequently applied within stock assessments, there is no need to have sampling levels be species-specific. That said, we agree with the reviewer that species-specific sampling rates for important species would not be overly complicated (and is currently the case in the AFSC trawl surveys when considering the minimum length sample size discussed above). We have taken this statement out of the revision, as it is not needed to make the point that our results were consistent across species and sampling rates do not need to be species-specific.

**Review:** *Lines 398-400: This seems speculative and was not tested by this study. I suggest omitting this statement.*

**Response:** We have removed this sentence from the revision.

**Review:** *Lines 401-407: With no evidence that the sampling asymptote is being approached, perhaps the recommendation should be to retain ‘at least’ historical levels of sampling and consider additional research (e.g., sampling one or more species’ ages at a higher rate for a short time) into whether collection of additional samples would improve information quality via ISS?*

**Response:** We agree with the reviewer on this point and have added ‘at a minimum’ before ‘historical levels’ (Line ## of the revision) and a sentence recommending further simulation and field research to evaluate the implications of increasing age sampling (Lines ##-## of the revision).

**Review:** *Lines 452-465: The discussion of age reading cost doesn’t seem necessary for this manuscript since it wasn’t investigated specifically.*

**Response:** reframe this para towards workforce health and efficiency

**Review:** *Lines 463-465: See general comment above on reporting the number of hauls sampled and made; without this information it is impossible to know how much change could be made by adjusting the number of fish per haul and number of hauls.*

**Response:** As noted above, we have included the average number of hauls sampled for age and length and the average number of hauls in which the species was caught to Table # in order for readers to be able to evaluate how much adjusting could be possible.

**Reviewer 2:**  
**Review:** *Comments to the Author  
This manuscript explores the impact of reductions in AFSC groundfish survey length and age sampling on effective composition sample size using a bootstrap simulation approach. The main result is that the AFSC groundfish survey program is collecting more lengths than necessary to nearly saturate effective sample size, that reducing the number of lengths sampled all the way to 50 per haul has (almost?) no effect on ages from subsampled fish that have been ‘expanded’ with an age-length key, and that effective sample size for ages has not saturated at current age sampling levels.  
  
The topic of adapting to reduced survey capacity is timely, the results seem important for the AFSC, and despite the complexity in the sampling and expansion (some of which I struggled with), the overall approach is a simple one that other jurisdictions could use as well. My biggest concern is with clarity of the methods, the sampling protocols, and the expansion protocols. My next biggest concern is about applicability beyond the included surveys. I expand on these below.*

**Response:** We appreciate the reviewer’s suggestions and concerns. To address this reviewers concerns we have added more detail in the methods section on how the length and age compositions are expanded and we have included additional analyses that evaluate alternative bin structures in order to make these results more broadly applicable.  
  
**Review:** *I spent an inordinate amount of time trying to understand what was done with the sampling and the two expansion levels. I'm sure this is obvious to those used to the AFSC protocols, but if I am any indication, the typical CJFAS reader not from AFSC/NOAA may struggle to figure out what was done. Because of the confusion this created, it made it very hard initially to follow the results and interpret them. I think I have now mostly wrapped my head around what was done, but I think expanding and clarifying the writing of the methods and being explicit about defining terms would go a long way. Many of my specific comments below reference my initial confusion.  
  
Two major places to focus are the method 'stages' and the flow chart. The method steps do not match up easily with the flow chart making it hard to figure out what was done. Second-stage expansion is never defined or explained beyond the word 'second (age)' and another section that I think is defining second-stage expansion without ever mentioning it. How the age-length key factors in, as it is featured in the flow chart, is not explained in the core of the methods.  
  
This is obvious in retrospect after much thought and reading of NOAA reports, but it's never explicitly stated that the age-length key is only applied to fish with age samples (I presume?), vs. converting all lengths to ages or expanding ages similarly to how lengths are. The text had me that confused. I presume this second-stage expansion is done primarily because of any length-stratified age sampling that either happens or has happened historically.*

**Response:** We have largely rewritten the methods section that introduces the length and age composition expansion, including additional details and formulae to help readers understand the application that we used in this study (Lines ##-## in the revised version). We have also attempted to streamline out terminology, by removing ‘effective sample size’ from the revision and instead use ‘realized sample size’ and ‘input sample size’ and define these terms in the introduction (Lines ##-## in the revision). We appreciate the reviewer’s suggestion to add the steps to the flow chart and have added this change to the revision in Figure 1. We agree that the way we used the ‘age-length key’ terminology was confusing, and in the revised methods section have provided more detail on how the age-length key is constructed (which it now should be clear that the age-length key is applied only to fish with age samples). We have as well removed it from the flow chart figure, as it is contained within the ‘estimating population numbers-at-age’ step. The reviewer is correct that the second stage expansion has been used due to length-stratified sampling, but it is also applied to randomly sampled ages as well as a method to leverage expanded abundance-at-length information (which is much more highly sampled) to age information.  
  
**Review:** *The bootstrap approach to composition effective sample size is not new, the impact on age effective samples themselves is not new, and obviously reducing length samples will reduce length sample effective samples—it's just a matter of what the curve looks like. I do think there's a niche this paper fills in terms of the specific length sample, age subsample, age-length key, 2-stage expansion approach and whether reducing the length samples impacts the ages (which I thought was better motivated in the discussion of Siskey et al. 2023 than in this manuscript). A main conclusion is that the AFSC is probably collecting far more lengths than it needs. This is certainly important for the AFSC. Can the authors make this applicable to a broader audience and comment on the degree to which this is likely happening in other agencies around the world? How often is this approach used elsewhere? Are there lessons for other jurisdictions to learn?*

**Response:** look into other places? Try to tie in more generality and add a para to the discussion on this. To our knowledge, there have been no papers that have looked into, ‘how much is enough’ in terms of length frequency sampling.  
  
**Review:** *Given the title and framing of the paper around workforce health and efficiency, are there any numbers or literature that could be brought in on this side of the trade-off? The manuscript does cite some literature on otolith aging costs, but the main effort examined here, and the main result, is about reducing length samples.*

**Response:** change around how we talk about workforce health  
  
**Review:** *How often in the second row of Fig 2 were the number of fish aged affected or were there gaps in the age-length key? Presumably this would be affected somewhat by which of the above protocols is used and what number is sampled by species/region, which isn't clear. Could this information be included somewhere?*

**Response:** The following table shows the instances in which an age-length pair was dropped from the age-length key because there were no corresponding lengths in the length frequency data from the base data (without any resampling or sub-sampling):

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Year | Species | Region | Length (cm) | Age | % of total |
| 2014 | arrowtooth flounder | AI | 85 | 27 | 0.53% |
| 2014 | Pacific cod | AI | 111 | 9 | 0.38% |
| 2014 | Pacific cod | AI | 113 | 13 | 0.38% |
| 2018 | northern rock sole | EBS | 47 | 22 | 0.32% |
| 2019 | Alaska plaice | EBS | 59 | 27 | 0.32% |
| 2021 | arrowtooth flounder | EBS | 77 | 15 | 0.25% |
| 2017 | northern rockfish | GOA | 17 | 3 | 0.39% |
| 2019 | Pacific cod | GOA | 96 | 8 | 0.40% |
| 2019 | Pacific cod | GOA | 97 | 10 | 0.40% |
| 2021 | Pacific cod | GOA | 96 | 7 | 0.39% |

Two primary considerations to note here: (1) the age-length pairs that were dropped for each of these species represents a fraction of a percent of the total age-length pair sample size that is used to construct the age-length key (far right column), and (2) these age-length pairs are at the extremes of the size range for each of these species (either at the large or small extreme). The following table shows the average number of age-length pairs that were dropped due to length frequency subsampling:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Length sub-sampling level | Region | Avg # of species | Avg ages dropped | Avg % of total |
| 50 | AI | 6 | 1.2 | 0.47% |
| 50 | EBS | 7 | 1.1 | 0.32% |
| 50 | GOA | 7 | 1.2 | 0.47% |
| 100 | AI | 6 | 1.2 | 0.46% |
| 100 | EBS | 7 | 1.1 | 0.31% |
| 100 | GOA | 7 | 1.2 | 0.47% |
| 150 | AI | 6 | 1.2 | 0.46% |
| 150 | EBS | 7 | 1.1 | 0.31% |
| 150 | GOA | 7 | 1.3 | 0.48% |
| 200 | AI | 6 | 1.2 | 0.45% |
| 200 | EBS | 7 | 1.1 | 0.30% |
| 200 | GOA | 7 | 1.2 | 0.46% |
| 250 | AI | 6 | 1.2 | 0.45% |
| 250 | EBS | 7 | 1.1 | 0.30% |
| 250 | GOA | 7 | 1.2 | 0.48% |

Again, we note that the average number of age-length pairs dropped was small, and the percent of the total age-length sample size in these instances was a fraction of a percent. In the revision we have not adjusted the methodology to account for this, primarily because of the negligible impact it would have on the results, but also because we wanted to stay consistent with how the age-length key is computed by the survey group at AFSC. However, we note that it would be relatively trivial to account for these missing age-length pairs by including the length that is not reflected in the length frequency data to that data prior to the first-stage expansion. To address this concern we have some text highlighting the small amount of instances in which this occurred in the revised version of the manuscript (Lines ##-##).  
  
**Review:** *Stewart and Hamel (2014) used 25,000 bootstrap replicates. Can the authors show, even if just for an example species-region, that 500 is sufficient to stabilize the results?*

**Response:** In the revised supplementary material section (Figure S1) we show an example of length composition ISS across 1,000 iterations.  
  
**Review:** *Line 26: Was there truly 'no' effect or was it very small? Was the age-sampling ever impacted itself by the length-sampling frequency?*

**Response:** add relative iss to figure 2 on bottom. Maybe change ‘no’ to ‘minimal’  
  
**Review:** *Line 44: I'm not sure 'require' is the right word, but certainly more is always 'better' from this perspective.*

**Response:** We have reworded this sentence to reflect that based on theory, one would expect the precision to increase as the number of samples is increased (Line ## of the revision).

**Review:** *Lines 48 + 65: Further -> Furthermore*

**Response:** We have changed ‘Further’ to ‘Furthermore’ here (Line ## of the revision).  
  
**Review:** *Line 103: More often than what?*

**Response:** ‘More often’ has been replaced with ‘has been’ (Line ## of the revision).  
  
**Review:** *Line 155: This is the last that age length keys get mentioned and given the passive voice in this sentence it's not clear if it is something that the authors did as part of the methods (except for appearing in the flowchart).*

**Response:** As mentioned above, we have added additional detail to the methods section about how the age-length key is constructed. Our intention is that these additions will help clarify how we did this analysis.

**Review:** (Note that this review was further down in the list of comments from Reviewer 2, we have moved it here to be in order with the line numbers)*Line 169: "second (age) stage expansion processes for each bootstrap replicate" This is the first mention of 'second'? Only first-stage explicitly explained above.*

**Response:** We note in a previous response that we have added text to the revised methods section that describes both the first-stage (length) and second-stage (age) expansion process and have included formulae describing these methods.  
  
**Review:** *Line 171: "Bootstrap resampling was performed either with replacement (wr) or without replacement (wor) depending on the needs of a particular protocol." Can you help the reader understand why one or the other (wr vs. wor) in steps 1-7? After extensive re-reading and thinking I think I follow why (one is a bootstrap resampling and one emulates an actual subsample and so a fish is a fish), but this was not obvious at first.*

**Response:** We have added a sentence following this in the revision that attempts to clarify when each type of resampling was conducted (Lines ##-##), and the reviewer is correct that one is bootstrap resampling where another is used to emulate actual sampling, for example, when sub-sampling occurred in the analysis.  
  
**Review:** *Table 2: What is "total number of reductions"? I eventually got it but it seems like an awkward way to put it.*

**Response:** We have changed ‘Average total number of reductions in length frequency samples’ to ‘Average reductions in total length frequency samples per survey’ in the revised version to try and make the caption for Table 2 less awkward.  
  
**Review:** *Step 3 and 6: "From the resampled lengths in step 2, subset the haul-specific length samples (wor) at the pre-determined subsampling level." I was confused here. I later realized this is total or sex-specific, but this wasn't clear yet.*

**Review:** (Note that this review was further down the list of comments that were received from Reviewer 2, we have placed it here as it is related to the previous review and respond to both of these in the following)*Steps 3 and 6 say "at pre-determined subsampling level" but we haven't been introduced to that yet making this confusing.*

**Response:** We have added text to step 3 (Line ## of the revision), as well as define ‘sub-sampling at a pre-determined level’ when explaining when with/without replacement was used (in addressing a previous review, Lines ##-## in the revision) in order to help clarify.

**Review:** *Steps 3 and 6 say "at pre-determined subsampling level" but we haven't been introduced to that yet making this confusing.*

**Response:** clarify  
  
**Review:** *I'm having a hard time matching up these steps to Fig 1. Can the numbers go on Fig 1?*

**Response:** We felt this was a great suggestion and have added the step numbers from text to Figure 1 in the revision.  
  
**Review:** *Line 195: "We set the subsampling level for length frequency at numbers per haul to evaluate the AFSC length sampling design." What does this mean?*

**Response:** In the revision we have added a sentence prior to this (Lines ##-##) to provide clarity and distinction between how we evaluated sub-sampling for lengths (hauls level) and ages (total survey level). We also added text to the end of this sentence to provide rationale as to why we were evaluating length sampling at the haul level (Lines ##-##).  
  
**Review:** *Line 197: "Additionally, to subsample ages, we reduced the proportion of the total number of ages sampled in step 6 to evaluate the consequences of reductions in overall age sampling,..."  
They are already subsampled, right? This is to test the effect of reduced subsampling? This was fuzzy at first to me.*

**Response:** The reviewer is correct, they are already subsampled and our analysis was to test further subsampling. We added text to the end of this sentence in the revision to provide reasoning as to why we investigated ages at the total survey level (as a main consideration is the effort expended by an age reading laboratory, Lines ##-##).  
  
**Review:** *Line 217: Effective sample size? 'Relative' sounds like relative to something and made me think it was a ratio at first, but it's a sample size count that comes out of this equation. Stewart and Hamel use 'realized sample size' and the referenced NOAA report does too.*

**Response:** This was a typo in the original version and should have been ‘realized’, we have changed ‘relative’ to ‘realized’ here and where appropriate in the revised version (Lines ## and ## of the revision for this particular suggestion).  
  
**Review:** *Line 214: The authors might consider referencing Appendix 2 in McAllister and Ianelli 1997. It took me a while to find it.*

**Response:** We have referenced Appendix 2 of McAllister and Ianelli (1997) here (Line ## of the revision).  
  
**Review:** *Line 214: Note that this is consistent only with the use of multinomial error and that it's an approximation.*

**Response:** While the reviewer is correct that it is consistent with multinomial error, it is also consistent with other error structures (i.e., Dirichelt (Hulson et al. 2012) or Dirichlet-Multinomial (Thorson et al. 2017)). To address this review we have changed ‘evaluate the level of intra-haul correlation’ to ‘approximate the level of intra-haul correlation’ and have noted that this approximation can be used in these error structures when fitting compositional data in stock assessments in the revision (Lines ##-##).  
  
**Review:** *Line 218: Why Oc,y,i vs. O,c,y; should all have 'i' I assume.*

**Response:** This was a typo in the original version, these terms should have all been Oc,y, as they are the ‘observer’ proportion (or, the ‘original’ proportion computed from the non-resampled data), this has been corrected in the revision (Line ##).  
  
**Review:** *R package: The readme has “The goal of swo is to ...” and the DESCRIPTION Description is still  “What the package does (one paragraph).” I would encourage the authors to also archive the code somewhere (e.g., Zenodo).  
Perhaps this is beyond what is reasonable at this stage, but it would be very helpful if there were any part of this package that could be run with different input data or at least without an AFSC username and password. If some functions could be used for other data, can that be described in a vignette, help files, or similar?*

**Response:** We have provided example data and have developed the description…  
  
**Review:** *Line 227: "we computed a relative sample size that indicated the change in uncertainty caused by sub-sampling length frequency and age specimen data." This is worded in a way that reinforced to me at first that it's a ratio in Eq. 1. However, it's a sample size number. This doesn't indicate the change in uncertainty itself --- it is an effective number that when lower reflects greater uncertainty compared to when it's higher.*

**Response:** Related to a response to a review above, ‘relative’ here was a typo in the original version and should have been ‘realized’. We have changed ‘relative’ to ‘realized’ here and in other occurrences in the revised manuscript to keep terminology consistent (Line ## for this particular review).  
  
**Review:** *Line 239: Here relative input sample size is actually 'relative' making this 'relative relative' sample size. I would keep this, and change the other.*

**Response:** Similar to the previous response, we agree with the reviewer and have changed ‘relative’ to ‘realized’ sample size where appropriate in the revised version. However, at this occurrence we are speaking about a ‘relative’ statistic that compares the realized sample size from the full dataset to the realized sample size for a sub-sampled dataset.  
  
**Review:** *Do the authors have an idea what's causing these sex differences in ISS for flatfish? Is it because of sex-specific M's or clustering?*

**Response:** There could be a number of causes for the sex differences in ISS flatfish. The lower ISS for flatfish indicates that there is more intra-haul correlation for males than females, or that the length composition from haul-to-haul is more variable for males than females. A number of processes could cause this, as the reviewer notes, one could be sex-specific M differences, it could also be due to availability differences to the survey between sexes (i.e., the males are less available to the survey than females). We have added an explanation along these lines in the revised results section to note this result (Lines ##-##).  
  
**Review:** *Line 269: "Across the length frequency haul level sub-sampling cases evaluated, the magnitude of age composition ISS for all stocks within each region was unaffected by length sub-sampling (bottom panels of Figure 2)."  
I was confused by this result for a while, and I didn't feel like the text did enough to explain this result. As I eventually understood it, the age-sampling protocols (e.g., 2-20 fish per trawl) is not affected by the length sampling because even the lowest level examined was 50 fish per trawl leaving more than 20 fish. But it isn't a given (quote from Siskey et al. 2023) "This creates mismatches between the length bins for expanded length compositions and those associated with aged fish, resulting in “lost” ages when the age–length key is applied to the expanded length compositions in design-based estimators." Can the authors expand on this within this manuscript and better explain the findings?*

**Response:** We note that Reviewer 1 also had a comment here that was very similar. In the revision we have added text to this section (Lines ##-##) to more adequately explain these results. We have also added text to the discussion to more thoroughly describe the implications of these results (per a following review below). In terms of the effect on age samples by reducing lengths, above we included a table with statistics describing how many ages were lost with subsampling, showing that it was a negligible amount of data lost that would have no impact on the main results and conclusions of this analysis.  
  
**Review:** *Line 281: drop very*

**Response:** ‘Very’ has been removed in the revision.  
  
**Review:** *Line 297: "While age composition ISS was unaffected by the length frequency sub-sampling level, the age composition ISS did markedly decrease as the proportion of total specimen age data decreased "  
I eventually got it but "proportion of total specimen age data" seemed awkward.*

**Response:** We agree that ‘proportion of total specimen age data’ is awkward and have changed this to ‘magnitude of specimen age data’ for clarity  
  
**Review:** *Line 351: "We also show that the age composition ISS was unaffected by reductions in length sampling effort." Stated as is, this is obvious and uninteresting. Needs to be more nuanced to the age-length key approach here.*

**Response:** change this  
  
**Review:** *Line 405: effectively*

**Response:** ‘Effective’ has been changed to ‘effectively’ in the revision (Line ##).  
  
**Review:** *Line 409: drop extremely*

**Response:** ‘Extremely’ has been removed in the revision.  
  
**Review:** *Line 418: "a reduction in age sampling effort for flatfish and rockfish has a relatively larger impact on the reduction in age composition ISS compared to gadids." Can the authors comment on why this might be?*

**Response:** We have added reasoning to the discussion section of the revised manuscript to try and explain this result (Lines ##-##). Overall, we suggest that the intra-haul correlation, or patchy spatial distribution, will have an effect, as well as the number of ages in the age composition that are observed (i.e., rockfish and flatfish are longer lived than gadids).  
  
**Review:** *Line 457: "tens of thousands of dollars" per year, per survey?*

**Response:** Need to work on this paragraph

**Review:** *Is there a reason for ISS instead of ESS? Either may be OK.*

**Response:** We have noted above in our response to Reviewer 1 that we have revised the manuscript so that we are consistent with our terminology both within our manuscript and with the current state of literature. We suggest that ‘input sample size’ (ISS) is a distinct term from ‘effective sample size’ ESS and propose that ISS be used when describing the sample size used to weight composition data within the stock assessment model and ESS be used to describe the degree of agreement, or fit, between observed and assessment model estimated composition data as an output diagnostic.  
  
**Review:** *Can the flow chart include the first- and second-stage expansion terms to match the text?*

**Response:** As noted in previous responses, we have included the step numbers in Figure 1 to be consistent with the text in the methods section, and have revised the methods section to add clarity surrounding the specific expansion methods used in our analysis.  
  
**Review:** *Fig. 2. I believe this aggregated across species and regions but are shown within species type.*

**Response:** We agree with the reviewer that this figure caption was confusing in the original version. We have revised this figure caption to indicate, as the reviewer commented, that the boxplots are shown by species type and are aggregated across species and region.  
  
**Review:** *Fig. 3 and elsewhere: combine regions where possible? In other cases (e.g., Fig. 6), could pick one?*

**Response:** In the revision we have followed the reviewer’s suggestion and have shown the main results by combining across regions in the Figures. For the additional analyses we conducted (length bins and sub-region impacts) we selected the GOA survey to use as an example.  
  
**Review:** *The proportional representation of reduced input sample size (e.g., Fig. 3) seems like a more powerful way to illustrate potential impacts and yet I can see the value in showing the actual effect of sample sizes themselves as well. In some cases, could the two be combined particularly if the regions are combined or an illustrative one is included? E.g., one row of Fig. 4 and one row of Fig. 5.*

**Response:** In general, in the revision we have attempted to streamline the results shown in the figures by aggregating across regions. We have also followed this suggestion, for example, in the revised Figure 6 in which both the ISS and relative ISS are shown in the same plot.

**Review:** *One of the more important results is likely the lower half of Fig. 2 not showing an impact. To highlight that there is 'no' (or very little?) impact here, could it be shown as a proportion where it's easier to see differences?*

**Response:** We agree with the reviewer’s suggestion and in the revised Figure 3 have included the relative ISS for age composition across the length-frequency sub-sampling levels to drive this point home.  
  
**Review:** *Fig. 6: If the regions are combined or an example region is chosen, this would give more space to make the aspect ratio 1 to 1, making it easier to interpret given the important finding is that the relationship is not 1 to 1.*

**Response:** We again agree with the reviewer’s suggestion and have aggregated across regions in this figure (revised Figure 7).