Reviewer: 2  
  
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.  
  
# Clarity of methods  
  
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.  
  
# Novelty and widespread applicability  
  
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?  
  
# Other comments  
  
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.  
  
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?  
  
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?  
  
# Specific comments:  
  
26: Was there truly 'no' effect or was it very small? Was the age-sampling ever impacted itself by the length-sampling frequency?  
  
44: I'm not sure 'require' is the right word, but certainly more is always 'better' from this perspective.  
  
48 + 65: Further -> Furthermore  
  
103: More often than what?  
  
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).  
  
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.  
  
Table 2: What is "total number of reductions"? I eventually got it but it seems like an awkward way to put it.  
  
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.  
  
I'm having a hard time matching up these steps to Fig 1. Can the numbers go on Fig 1?  
  
195: "We set the subsampling level for length frequency at numbers per haul to evaluate the AFSC length sampling design." What does this mean?  
  
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.  
  
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.  
  
214: The authors might consider referencing Appendix 2 in McAllister and Ianelli 1997. It took me a while to find it.  
  
214: Note that this is consistent only with the use of multinomial error and that it's an approximation.  
  
218: Why Oc,y,i vs. O,c,y; should all have 'i' I assume.  
  
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?  
  
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.  
  
239: Here relative input sample size is actually 'relative' making this 'relative relative' sample size. I would keep this, and change the other.  
  
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?  
  
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?  
  
281: drop very  
  
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.  
  
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.  
  
405: effectively  
  
409: drop extremely  
  
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?  
  
457: "tens of thousands of dollars" per year, per survey?  
  
169: "second (age) stage expansion processes for each bootstrap replicate" This is the first mention of 'second'? Only first-stage explicitly explained above.  
  
Steps 3 and 6 say "at pre-determined subsampling level" but we haven't been introduced to that yet making this confusing.  
  
Is there a reason for ISS instead of ESS? Either may be OK.  
  
Can the flow chart include the first- and second-stage expansion terms to match the text?  
  
Fig. 2. I believe this aggregated across species and regions but are shown within species type.  
  
Fig. 3 and elsewhere: combine regions where possible? In other cases (e.g., Fig. 6), could pick one?  
  
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.  
  
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?  
  
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.