1. Are the objectives and the rationale of the study clearly stated?

Reviewer #1: The objectives and rationale for this study are clearly described.

Reviewer #2: Yes.

1. If applicable, is the application/theory/method/study reported in sufficient detail to allow for its replicability and/or reproducibility?

Reviewer #1:

Please provide additional detail in the methods section on how the age-length keys (ALKs) were constructed. Specifically, how was the bin structure selected (single or multi-centimeter bins)? Was this choice adjusted to be consistent with the range of lengths observed for each species and/or the specific bin structure that is used in the stock assessments?

**Response: In the revision we have indicated in the methods section that we are binning the length frequency (and subsequent age-length key) to 1 cm bins, that was across the range of lengths for each species and followed the bin structure used in the assessments. We also note that the framework we have constructed allows for any bin structure. In a related paper we use a similar framework (without the ageing error and growth variablity features presented in this work) to evaluate the effects of different bin structures (1 cm, 2 cm, and 5 cm). From those results we note that the bin structure did have an effect on the length composition ISS, but this was not translated to the age composition ISS, which was unaffected by the choice of bin size. We would suggest that those same results would be replicated here, where the magnitude of added uncertainty due to ageing error and growth variability would be of similar magnitudes regardless of bin structure.**

It is somewhat unclear from the methods, but I think that the “total” or combined-sex results were created by summing the male and female age compositions. Would the ‘total’ age composition be better estimated by combining males and females prior to creating the age-length key and therefore not including the bootstrapped variability of each sex separately in the total? Would this be the appropriate method if the stock assessment were intending to use the aggregate distribution? With a larger sample size, albeit possibly also with more growth variability, would the input sample sizes would be larger using this method?

**Response: We recognize that the reviewer is correct that we left it unclear as to how the combined-sex results were created and have added text to the revised methods section that describe that the combined-sex results were created by summing the male, female, and unsexed expanded population numbers at age to then compute the age composition, as the reviewer noted. The primary reason this is done is to account for the potential differences in growth between sexes (for example, for flatfish species), albeit, the reviewer makes an interesting point that the increased sample size could translate to an increased ISS. We were curious about this and ran an example for a species that doesn’t exhibit difference in growth between sexes (Gulf of Alaska Pacific Ocean perch). The following figuyre shows…**

Were the age samples in this study collected randomly in each year or were there stratified targets by length? If length-stratified in some years (which would seem best for the ALK approach), but random in other years (perhaps best for building age compositions without an ALK and having more variability when used with an ALK) would the interaction with ageing imprecision be different? If so, should the results be delineated between years with different sampling methods?

**Response: Reviewer is correct that… Run example with length-stratified vs random. indicate that preferred option would be to use the ALK, regardless of sampling methodology, so that info from length frequency, which is sampled at a much higher magnitude, is leveraged into age composition, for which ages are sampled at a much lower rate.**

1. If applicable, are statistical analyses, controls, sampling mechanism, and statistical reporting (e.g., P-values, CIs, effect sizes) appropriate and well described?

Reviewer #1:

The bootstrap approach is appropriate and the manuscript extends previous work on input sample size calculations from length and age compositions for use in fisheries stock assessment. My biggest concern is with regard to the pooling across years of the age-length key and the influence that may have on the results: A primary focus of the manuscript is the proportional contribution of growth variability to the reduction in input sample size. However, the method uses a ‘global’ ALK pooled across all years and areas. I’m assuming that assessments would normally use a year-specific key, which would not have the effects of interannual growth differences and/or strong cohorts in a single key pooled over time. If the pooled approach does not match what is done in assessments, it would seem to overestimate the reduction in input sample size due to growth. Would it be possible to add an example to the analysis comparing the results based on pooled vs annual ALKs for one of the species with a large growth effect on the input sample size (e.g. a gadid)?

**Response: A global ALK was not used, but the age-length pairs for estimating growth variablity was. we ran an example for PCod to show the differences between pooled and annual growth data… many assessments pool age-length data to estimate growth**

Reviewer #2:

Additional details about the estimated relationships between nominal and input sample sizes should be added as noted in the review.

**Response: we added a table…**

1. Could the manuscript benefit from additional tables or figures, or from improving or removing (some of the) existing ones?

Reviewer #1:

I think it would be helpful to include a figure/panels showing results similar to the upper panels of figure 5, but comparing with the age range in the observed data for each species and also the number of length bins with data for the ALK for that species. Perhaps these have more of a clear effect than the CVs?

**Response: Hmm, these might be interesting**

Figure 2 (and other boxplots): What is shown by these boxplots - median, 25th, 75th and range?

**Response: add to caption**

Figure 4, upper panel: These axes should not extend below, as that space has no meaning. Also, the ellipses need a description in the caption: why do they extend far below the points (are they forced to cover 0,0)?, do/should they have 95% coverage of the points?

**Response:** **easy fix using coord\_cartesian(xlim = c(0, Inf), ylim = c(0, Inf)), I suppose the reason for the below 0,0 is that they are not computed on a log scale?**

Figure 5 lower panel: These results seem inconsistent with figure 3 as the whiskers all extend above a value of 1.0 (but not in Figure 3). Is it possible that the bootstrapped values and not the annual harmonic means are being plotted here?

**Response: Need to think on this one a bit - certainly showing different things (lumped vs split) - also might want to explain what the error bar is in Fig 5.**

Reviewer #2:

Yes, it would be useful to provide a table with summary statistics for the results shown in Figure 4.

**Response:**

1. If applicable, are the interpretation of results and study conclusions supported by the data?

Please provide suggestions (if needed) to the author(s) on how to improve, tone down, or expand the study interpretations/conclusions. Please number each suggestion so that the author(s) can more easily respond.

Reviewer #1: Yes. But see comments on the use of the pooled ALK.

Reviewer #2: Yes.

1. Have the authors clearly emphasized the strengths of their study/theory/methods/argument?

Reviewer #1: Yes

Reviewer #2: Yes

1. Have the authors clearly stated the limitations of their study/theory/methods/argument?

Reviewer #1:

Per the above comments on the ALK approach, the relative effect of growth on input sample size may be overestimated; however an example comparison for one species with large variability due to growth could address this.

**Response:**

Reviewer #2:

No, they need to provide more context about alternative approaches to accounting for ageing error and growth variability in stock assessments.

**Response:**

1. Does the manuscript structure, flow or writing need improving (e.g., the addition of subheadings, shortening of text, reorganization of sections, or moving details from one section to another)?

Reviewer #1: The manuscript is well written and organized.

Reviewer #2: The flow and structure are fine.

1. Could the manuscript benefit from language editing?

Reviewer #1: No

Reviewer #2: No

Reviewer #1: Minor editorial corrections:

Lines 153-155: Which years were included in this analysis?

**Response: corrected**

Line 339: “Bias”

**Response: corrected**

Line 380: Instead of “half-way”, perhaps consider “partially” since the relative magnitude of the effects from ageing error in the assessment and in the ISS bootstrap have not been quantified.

**Response: corrected**

Line 400: “as example”

**Response: corrected**

Line 403: “has largest” “and smallest”

**Response: corrected**

Reviewer #2: Big picture comments

The manuscript provides a valuable estimate of adjustments that should be made to account for input sample sizes (ISS) in statistical catch-at-age models in which length compositions are converted to age compositions using an age-length key. This approach is used at the Alaska Fisheries Science center where the authors work and in many other places. The manuscript is technically sound and surely valuable for improving the assessments at AFSC. However, it has some major gaps when it comes to providing context for this approach.

First, I believe the manuscript does not adequately compare their approach to the alternative in which length compositions are input directly into the model and fit to expected compositions calculated from a growth curve and estimated variability in length at age each. That alternative approach allows ageing error (as well as potentially process error in growth and selectivity) to be incorporated directly into the model, and was used by Thorson and Haltuch (2019) as well as in the assessments of all the stocks explored by Stewart and Hamel (2014). The paragraph starting on line 368 suggests that incorporation of ageing error in the calculation of expected age compositions addresses only half of the problem, but when length compositions are used directly, there is no need for incorporation of ageing error or growth variability in the calculation of input sample size.

**Response: need to explain the difference between the modeled and observed quantities, that both contain these sources of error, and that the data weighting applied to the fit between the two should also include these sources of error. when fitting caal would still need an iss that has growth and ageing error accounted for. but, more work to be done on how to develop iss for caal data.**

Second, the manuscript does not address other data weighting approaches, such as the iterative approach developed by Francis (2011) or those that include estimated parameters promoted by Thorson (numerous, including 2023 paper co-authored by the lead author of this manuscript). It would be useful to see how the down-weighting calculated via a model-based data-weighting method compared to the adjustments proposed in the manuscript.

**Response: beyond scope of current ms, but a study dealing with implcations of starting values for iss would be useful. hve shown that misspecifying these values can cause bias, with the implication that not accounting for these sources of error could cause bias. add para to discussion on this**

Minor issues

Line 54: “obtained from hauls within a trip that is targeting a specific species” Perhaps this is the norm in Alaska, but in much of the world it’s more common to sample trips, not hauls, and most fisheries do not target a single species.

**Response: corrected - adjusted text to reflect that either hauls or trips may be sampled and that multiple species may be targeted**

Line 66: “The statistical weight assigned to annual composition data…” The statistical weight is often a combination of one of the methods listed and some adjustment applied within the model according to some data-weighting scheme.

**Response: Guess we should clarify that it is the input weight? not all models run iterative reweighting.**

Line 81: the proposed naming convention is valuable and welcome

**Response: corrected**

Line 97: “duet o” to “due to”

Line 108: “a number of studies” is supported only by a single reference: “Liao et al. (2013).

Line 212: “Growth variability was incorporated by resampling from all lengths associated with a given age and sex.” This is confounded with ageing error. Even if there were zero variability in growth, this calculation would show a range of lengths associated with a given age due to ageing error except for ages beyond the point where growth had reached an asymptote.

**Response: totally, how to address this…**

Line 303: “A positive relationship is observed, by species group, between the number of age samples taken per haul and the age composition ISS per sampled haul (top panel of Figure 4).” It would be useful to provide a table with summary statistics for the results shown in Figure 4: the average input sample size per nominal age sample and/or the average ISS per haul. This would make it easier to compare to the results of Stewart and Hamel (2014) and to adjust nominal sample sizes without repeating the bootstrap procedure.

**Response: add a table, but note that the realtionship is extremely variable and it wouldn’t be as simple as adjusting nominal sample size and that hte bs should be run for each new year, doesn’t need to be run for the whole time series**

Line 416: Change “…the package we built…” to “…the R package we built…”. The DESCRIPTION file in the R package should be updated so that the citation could be filled in so that the information at <https://benwilliams-noaa.github.io/surveyISS/authors.html#citation> is more complete (that same information is provided by the R command citation(“iss”) if the package is installed.

**Response: corrected**

Line 575: “Number ages” to “Number of ages”

**Response: corrected**

Line 580: “plts” to “plots”

**Response: corrected**

Line 580: “elipses” to “ellipses”

**Response: corrected**

Also, some explanation of how the ellipses were calculated would be useful. Adding a linear regression may also be useful.

**Response: maybe just switch to linear for all**

From the editor (Andre Punt) The reviewers raise several very relevant and important points. Reviewer #2 requests consideration of some broader issues. You should either adopt these or provide reasons why this is not appropriate. I am likely to request a 2nd review of the MS by reviewer #2