The study by Li et al. presents a closed simulation analysis to test the effectiveness of including random effects on numbers-at-age transitions. I have found the manuscript to be well written and of interest to practitioners of fisheries stock assessment. I have a few questions and issues raised below but would recommend the publication of this work in CJFAS after minor revisions.

What do “G” and “D” stand for in OFG and OFD (lines 271 and 272)? These indices also appear in Figure 7. Why not add a Table with a list of the indices, since the appear in Table 7, and how the Score of each index is calculated (is a high value good or bad?). Figure 7 uses “Index” in its legend, but the text talks about “Performance Metrics”, be consistent throughout the article. Having a Table would also simplify the caption of Figure 7.

In Section 2.2, no mention is made of how movement is handled in the operating models, and I realize that section 2.2.3 is exclusively about movement dynamics. But why not also add movement to the last sentence of 2.2?

Minor editorial note: Page 3, line 74 “may not varying” should be “may not be varying”

How many parameters are estimated by each estimation model, and was there any exploration of comparing these models, and how they would stand to a model selection process based on AIC values? In Section 2.3, when NAA random effects are included, how many more parameters are estimated and how does it impact the complexity of model fitting? While I understand that models of different complexities are to be compared against simulated data, it would be useful to document the number of parameters that are estimated in the different models, and how that compares with the number of available simulated observations. This information could perhaps be added to Table S1?

In Section 2.4, the authors state that the assessments use the most recent 20 years of data available. Why not use all the data available to that point instead? What is the purpose of not using data beyond this 20 year time window?

The measures chosen to compare the different estimation models and the MSE results are appropriate and well defined.

Lines 274 to 277: what is the purpose of the weighting described here?

In the Results section, the first two paragraphs use “For simplicity”. Could the authors explain what exactly is made simpler? The endless ability to produce results when using a closed simulation framework means that a large number of those results could be presented in figures, so decisions are made to select the key findings. The selection of Figures 3 to 7 followed a thought process that the authors went through, and it would be great to add a few sentences about this. Figure 1 describes 2 OM versions. Figure 2 presents the 5 different estimation models with and without random effects. It thus follows that figures 3 and 4 have 10-1=9 different entries on the x axis. The 3 columns used to compare the fishing exploitation regimes considered are shared by Figures 3 to 7. Is the order of the x axis based on model complexity increasing from left to right? If so, state that in the figure caption.

Why not combine figures 5 and 6 into a single figure? They share the same x axis and could be grouped together. The figure captions for both of these figures are insufficient. They do not describe the groupings used for columns and rows. In figure 5, what are the 3 different rows? I deduce that it corresponds to region 1, region 2 and global, but this information should appear in the caption.

Wouldn’t Figure 5 be easier to interpret if the same y axis limits were shared by the different panels? That way the level of catch variation could be compared for the different estimation models and also between the different groupings presented.

Figure 7 brings it all together and has the 3 different fishing scenarios as columns, the 10 estimation models as rows, but now the colors are associated with the performance indices. The order of the indices in the legend on the right of the figure should be the same as in the panels, they are currently in opposite order. I believe that this will improve the readability of this figure. There is some shading added to separate the NAA and noNAA estimation models, please state that in the captions. Same comments for Supplementary figures S14, S15, S33, S34 and S35.

The word “interestingly” appears too many times in the Results section. State what is interesting about the results, and use other words to portray your findings. Or were the results surprising and unexpected? What were the expectations of the authors about the results, and how were those found to be different from the simulation results obtained?

Lines 566 to 567: “slightly different” is too vague. I don’t understand that first sentence, “slightly different” from what?

Line 578, the proposed methods would be an “intermediate” estimation approach when movement dynamics are not accounted for. What are the other approaches to consider? Does “intermediate” mean something between a survey index and a fully spatial state-space model. What is done before and after this intermediate step?

There are many instances of missing capital letters in the article titles in the References. In BibTeX the capital letters would require curly braces around them, otherwise they will be rendered as lowercase. For example, Berger et al. (2012) should be “Accounting for spatial population structure at scales relevant to life history improves stock assessment: The case for Lake Erie walleye *Sander vitreus*”.

If the recommendations of the paper are followed (NAA random effects should be a default starting point in state-space stock assessments), the inclusion of random effects on numbers-at-age transitions should be included as a default setting in WHAM, and provided as an alternative in setting up model runs.

Overall I am very favorable with the publication of this manuscript. The number of figures in the Appendix is perhaps excessive, but I also do not have strong opinions about what should and should not be included.

One question I had was about WHAM and its governance. Currently, both WHAM and its MSE tools exist in personal git repositories of authors (/lichengxue and /timjmiller). The work leading to the development of these tools was in part under the auspice of NOAA. As such, is there a plan to establish custodianship of these tools? This is especially important to consider since the tools are useful to many other scientific institutions outside of the United States. Will these tools outlive the careers of its primary developers, and what risk is involved for an agency to base its stock assessment work in these tools? WHAM and whamMSE are publicly available from the main developer’s personal repository on GitHub, and a clear NOAA Disclaimer appears there, but I still wanted to raise the question.

How closely related are WHAM and whamMSE? They currently exist as two separate packages maintained by two developers. Is the integration of MSE tools for WHAM part of the development milestones, or was it strictly developed for the current article? This relates to my earlier question about governance and WHAM development. Will the whamMSE functionalities be developed in parallel with WHAM?

I really appreciate the authors work towards implementing useful tools that can assist other stock assessment scientists in their work.

A more thorough discussion is warranted to explore trade-offs between implementing models of various levels of complexity, and the difficulty to do so by practitioners. If the tools presented can only be developed and implemented by a handful of practitioners, their endorsement and use in the fisheries science community might be limited.

Sorry for the delay in submitting this review and thank you for the opportunity to critique this interesting piece of work.