Second Response Letter to the manuscript:
"Quality assuring the quality assurance tool: A case study on applying safety-critical concepts to test framework development"

Dear Editor and Reviewer,
We thank you for this second round of professional and constructive review comments. In this letter, we provide our responses and the changes done to address the given comments.

We hope that these changes have improved the quality of the manuscript and that it is now in publishable form.

Sincerely,
Authors
September 2022

Reviewer # 1's Comments

Basic reporting, and Experimental design.
No change required.

Validity of the findings

- However, the paper still does not adequately argue in regard to the weaknesses apparent in the methodology. While most studies are far from perfect, it is important that the authors clarify the weaknesses of their study enough to allow readers to (by themselves) decide how much the conclusions can be trusted. That said, I think the changes required to reach and adequate state is relatively small.

Authors' response and action:
We thank the reviewer for this valuable comment. We have made several changes to the paper, as detailed in the answers below.

Additional Comments A

- For Your Information: I wonder whether you need to include this statement: "Applying constraints on the development process was seen as less compatible with agile development, which is why this article focuses on standards in the first group." While I understand the comment and appreciate your wish to be transparent, it might be a bit confusing to the average reader (not reading the paper in detail).

Authors' response and action:
As suggested, we removed the statement in section 2.1.

Additional Comments B

- The paper now clearly states the type of case used, but it does not argue why this case is suitable:
  "A second rationale for a single case is when the case is an extreme or a unique case, however, this does not apply in our context, rather our case is representative or typical (third rationale). The industrial partner and the products developed are typical of many embedded system manufacturers, both in the same domain of industrial communication equipment and in other similar domains. Thus, the results taken from this case are assumed to be typical of the experiences of an average embedded system manufacturer." While you can of course do a lot to present a strong positive argument, it should be sufficient/easier to simply provide a negative one here. In other words, you can, in Section 6.3, discuss the most obvious conclusions which could be different/invalid/less strong in more extreme situations. Purely as an example, embedded software developers often end up being the only expert in "their" system components or support tools/toolchains. Presumably this would decrease the value of reviews, or at least imply they should be focused on particular problems. (I could offer a few references on this and other possible talking points, but I think you have a few to choose from already.) If you provide such a small discussion, it should be enough to give a reader a feeling for the most important limits of your conclusions

Authors' response and action:
We would like to argue that the suitability of the case is typical of embedded system manufacturers. However, to acknowledge this potential limitation, we added a paragraph on generalizability in Section 6.3, including an example where one of our conclusions might not be applicable. We also added two references (second-last paragraph in Section 6.3) to back the value of industrial case studies, despite limitations on generalizability.

Additional Comments C.1 and C.2

- "…sample choice (convenience, but no strong argument for case)" -> See Comment (B), are you missing input from some important stakeholders due to who you picked for the group? (Note that the triangulation that you make use of and e.g., discuss in Section 6.3 does not help in regard to omissions in regard to sample choice.)
- "sample size (small)" -> See Comment (B), are you missing input from some important stakeholders due to how many you interviewed? (Note that the triangulation that you make use of and e.g., discuss in Section 6.3 does not help in regard to omissions in regard to sample size.)

Authors' response and action:
We agree that the sample might be perceived as limited, however, it was representative of the roles involved in developing WeOS. We updated the 2'nd paragraph of Section 6.3 with the following:

> Involving a larger number of participants would most likely have given a greater sample size and more diverse feedback. But in terms of diversity of roles participating in the focus group, only a few relevant ones were omitted, e.g. no

software architects were present but they typically have similar competences as senior developers.

Additional Comments C.3

- "not iterating" and "data saturation not being ensured" -> The triangulation is a possible argument here. However, while you refer to checking that there were no more input in the research process, it is unclear if this was the case. E.g. "The focus group ended with a summarizing event, asking the participants if there were any candidates they had expected to be presented but that were missing, and if they could share any other thoughts or ideas regarding the material that had been presented to them." (Did the participants state there was nothing more to say?)

Authors' response and action:
We clarified this by adding a new last sentence in 4.9, by extending and clarifying 5.4, and by moving a sentence from 5.3 to 5.4.

Additional Comments C.4

- "not recording audio/video from the focus group" -> You do not have to record audio/video, and you were not able to in this case. Sure. However, even Krueger states concerns when using notes and memory (it not being suitable for novices, etc.). If you handled these concerns, then state so.
I still think your analysis of the notes are a bit unclear, but if you were indeed doing this together I guess the risk you missed anything should be small.

Authors' response and action:
We have simplified the writeup of the *data collection* in Section 4.8.2, stating that we relied on multiple persons taking the notes to counter downsides of note-based data collection.

Additional Comments D

- The mentions of DoDs in Section 6.3 and 6.5 does not connect well with the rest of the text, now that you have moved much of the discussion of DoDs to an Appendix. I suggest you rephrase/reduce these parts.

Authors' response and action:
We agree and have reduced the last paragraph of section 6.3 and the last paragraph of section 6.5.