The changes made in the revised manuscript are highlighted in Red in the "summary of changes" according to comments of Reviewer 3.

Reviewer: 3

Comment 1:

One of my biggest concerns with this paper is the conflicting results and claims reported in the paper. Among the existing ones the one that bothers me the most is this: Section 5.5.2 states that when it comes to reducing masking affects in the presence of a single defect, ict outperformed, sct, which in turn outperformed fda-cit (Page 19, the last two paragraphs). However, Section 5.6.2 claims that the results obtained in the presence of multiple defects "matched" the ones obtained from Section 5.5.2 (last paragraph of Section 5.6.2). However, this doesn't seem to be true!

Response: According to this comment, we have fixed these conflicting text as suggested. The details are shown in the responses to the following comments.

Comment 2:

First of all, next to last paragraph in Section 5.6.2 states that fda-cit outperformed both ict and sct in reducing masking effects. However, this conflicts with Table 27. In this table, fda-cit was clearly outperformed by sct. So, either the table or the text is wrong. Considering the conclusion stated by the authors in the last paragraph of Section 5.6.2, probably the table is wrong, i.e., fda-cit performed better. But then this contradicts (not "matches" as stated in the paper) with the conclusion obtained in Section 5.5.2. Furthermore, sct performed better than ict (according to Table 27), which also contradicts with Section 5.5.2.

Response: Yes, the original text was wrong and sorry for the misleadings we may cause. The data which are shown in Table 27 is correct, but we mistook these results (we mixed fd-cit up with sct). Hence, we updated our discussion of the results of masking effects (See the fourth paragraph from the end of Section 5.6.2, red part). Our main point is that, with respect to the reduction of masking effects when facing with software with multiple defeats, sct outperformed the remaining two approaches, and fda-cit was the second best approach, while ict was the last one.

Also as suggested, we have discussed why ict was clearly outperformed by sct at reducing masking effects under such condition (and this is the only different conclusion when compared with the results of Section 5.5). The main point is that (See the third paragraph from the end of Section 5.6.2, red part), the decreasing of the tested-t-way coverage of ict was caused by the reduction of passing test cases. This is because due to the one-bug-at-one-time strategy for handling multiple defeats, ict labeled the test cases which failed with the defeats other than the defeat under analysis as passing test cases. As a consequence, it can normally identify the MFS for the defeat, but these test cases which failed with other defeats cannot contribute to any tested-t-way coverage. Therefore, the tested-t-way coverage obtained by ict decreased.

At last, we agree with your comment that the results obtained in Section 5.6 (for multiple defeats) did not match 100% with what was obtained in Section 5.5(for single defeat). This is because there exists one difference that, with respect to the masking effects, sct outperformed the other two approaches, while ict performed the worst (Other results, i.e., the total number of generated test cases, the quality of MFS identification, and the number of test cases containing multiple MFS, matched well with the results of 5.5). Hence, we have rephrased the last paragraph of Section 5.6 (See the last paragraph of Section 5.6, red part) to be "Except for the masking effects, other results matched well with the results obtained from the experiments of a single defeat. Specifically, ict obtained the best MFS identification results and generated the least number of test cases containing multiple MFS, sct obtained the most tested-t-way coverage, and fda-cit generated the smallest number of test cases."

Comment 3:

As stated in the first paragraph above, the proposed approach claims to address three questions. However, not all of the related metrics have been used in the experiments to evaluate the proposed approach. For example, in Section 5.8 and Section 5.9, the results of t-tuples that were actually "covered" by different approaches used in the experiments, need to be reported in order to reason about how much these approaches suffered from masking effects.

Response: As suggested, we have now provided and discussed the results of t-tuples that were actually covered by different approaches in Section 5.7, 5.8 and 5.9 (See Section 5.7.2, 5.7.3, 5.8.2, 5.8.3, and Section 5.9.2, red part), so that the extent to which these approaches suffered from masking effects under these testing scenarios can be observed.

Comment 4:

"As the target of our empirical studies is to compare the ability of fault defection between our approach with traditional ones," this statement is misleading as the experiments that were carried out has nothing to do with this objective.

Response: Yes, we agree. As suggested, we have rephrased this sentence to be "As the main target of our empirical studies is to compare the ability to handle the proposed three issues between our approach with traditional ones," (See Section 5.1, 3rd paragraph, red part).

Comment 5:

"One observation from this table is that the number of test cases generated by our approach was smaller than that of the sct approach." This statement contradicts with the results reported in Table 10. Besides the gcc results, as far as I can see there is at least one other comparison where sct outperformed ict (t=2 on Hsqldv).

Response: Yes, we agree this sentence is not strict. It is true that sct only outperformed ict at t=2

on Hsqldb, but ict outperformed sct at t = 3 and 4 on Hsqldb. Ict also outperformed sct at all the degrees (t =2, 3, and 4) on Tomcat and Jflex, and at t=3 and 4 on Tcas. In summary, there are 10 out of 15 cases in which ict needed fewer test cases, while sct only outperformed ict at 5 out of 15 cases. Besides, the extent to which ict outperformed sct is not trivial. Specifically, for the cases in which ict needed fewer test cases than those of sct, ict reduced about 22.2 test cases on average. On the contrary, sct only reduced about 4.7 test cases on average when sct needed fewer test cases.

According to your comment, we have rephrased the sentence to make it more rigorous -- "One observation from this table is that, in most cases, the number of test cases generated by our approach was smaller than that of the sct approach." (See Section 5.2.2, 2nd paragraph, red part)

Comment 6:

The number of t-tuples that are covered in passing test cases or that are marked as MFS, reported as tested t-way coverage in the paper. However, tested t-way coverage has a different meaning. So, this needs to be fixed.

Response: We agree that the original definition of tested-t-way coverage is a little bit different from what we mentioned in this paper. Specifically, the original definition (See the study[1], Page 4, Definition 8) is "Given a configuration space model and a value of t, the tested t-way interaction criterion is satisfied, iff, for each t-pair 1) the t-tuple was present in at least one configuration in which the test case t passed, 2) the t-tuple is designated as a faulty interaction, or 3) the t-tuple was present in at least one configuration in which the test case failed with a non-option-related cause." In our experiments, all the failures are option-related, and the remaining two conditions coincide with the formula with which we used to compute the tested-t-way tuples. Hence, the computation of the tested-t-way coverage in our experiments satisfied the original definition. As suggested, we have emphasized this point to avoid the confusion (See Section 5.3, 1st paragraph, Red part).

[1] C. Yilmaz, E. Dumlu, M. Cohen, and A. Porter, "Reducing masking effects in combinatorial interaction testing: A feedback driven adaptive approach," Software Engineering, IEEE Transactions on, vol. 40, no. 1, pp. 43–66, Jan 2014.

Comment 7:

How were the synthetic models created? Justifications are needed. Furthermore, it would be great if the authors could actually publish all these models to ensure the repeatability of the experiments.

Response: As suggested, we have published all these synthetic models, as well as the corresponding information of MFS, in the Appendix (See Appendix B), so that these results obtained from these subjects can be reproduced directly. These models are created by considering how these approaches performed under different conditions. More specifically, for the synthetic models used in Section 5.7, we need to consider how approaches performed under

a different number of parameters (8 to 100) instead of some specific number of parameters obtained from previous experiments. Additionally, we also consider the different number of MFS (1 to 90) that contained in these synthetic models instead of only 1 or 2. In Section 5.8, these models are created by considering the various probabilities of triggering MFS (1% to 98%) instead of 100% or only some specific probabilities. We also considered the number of times (1 to 47) that unsafe values are introduced each time the MFS identification proceeded in the experiments.

Comment 8:

The Threats to Validity section needs to be updated. For example, it vaguely addresses external treats regarding the use of the actual subject applications. What about the ones concerned with the synthetic models used in the experiments.

Response: As suggested, we have updated the section on Threats to Validity. Specifically, we have rephrased the external treats regarding the use of actual subject applications and discussed the use of the synthetic software (See Section 5.10, 1st paragraph, red part).

At last, we appreciate your valuable and helpful comments. The revision according to your comments improved the quality of this paper.

Reviewer: 1

Public Comments (these will be made available to the author)
I'm satisfied with the authors' responses to my comments and happy to accept this version for publication.

At last, thanks for your satisfaction of our revision according to your helpful comments.

Reviewer: 2

Public Comments (these will be made available to the author)
The authors did a good job with this revision and I am happy to recommend the acceptance of the paper.

At last, we appreciate your valuable comments and satisfaction of our work.