24-Aug-2015  
  
Dear Dr. Niu,  
  
We have completed the review of your paper titled:  
  
Identifying minimal failure-causing schemas in the presence of multiple failures  
TOSEM-2015-0052  
authored by Niu, Xintao; Nie, Changhai; Lei, Yu (Jeff); Leung, Hareton; CHAN, ALVIN  
  
It is the decision of the TOSEM Editorial Board that we cannot accept your paper for publication in TOSEM without major revisions.  
  
I have included anonymous, verbatim reviewer comments to guide your revisions.  All three reviewers acknowledge that the paper has novelty.  However, they all raise major concerns that make the paper unacceptable in its current form. As you will see, such concerns are both about the soundness of the approach and about the empirical evaluation.  The paper therefore requires substantial rework, including, among other things:  
  
- improving the approach definition  
- reconsidering the assumptions behind the approach  
- adding further analyses (e.g. of real improvements and of statistical significance of the results)  
- revising and extending the empirical evaluation (and note that this part raised major concerns from all three reviewers)  
  
We strongly recommend and encourage you to take the reviewers’ comments in due consideration while performing your revision.  
  
If you need further information, please contact Dr. Antonia Bertolino, the TOSEM Associate Editor who was in charge of your paper. The email address is: [antonia.bertolino@isti.cnr.it](mailto:antonia.bertolino@isti.cnr.it).  
  
To expedite the re-review process, please submit your revision through Manuscript Central (<http://mc.manuscriptcentral.com/tosem>) under manuscript ID# TOSEM-2015-0052.  You may select "create a revision" for this manuscript ID which is located under "manuscripts with decisions".  In addition, please include a cover letter that specifically states how your revision addresses the concerns articulated by the referees.  
  
We look forward to receiving your revision, and thank you for your interest in TOSEM.  
  
Sincerely,  
  
  
David S. Rosenblum  
Editor-in-Chief  
ACM Transactions on Software Engineering and Methodology  
  
  
===============REVIEWS==============  
  
  
  
Recommendation: Needs Major Revision  
  
Comments:  
This work improves an existing approach, called FIC\_BS, which aims to pinpoint failure-inducing combinations of option settings (aka factor settings), such that the resulting approach can operate in the presence of multiple defects; the original approach assumes that all failures are caused by the same defect. The proposed extension is a technique, which given a failing configuration (a test case as it is referred to in the paper) and a fixed part in this configuration (i.e., options and their settings which must be remained unchanged), computes a configuration which is highly likely to pass or to reproduce the same failure. That is, the new configuration is chosen, such that it is unlikely that it reveals a different failure, which would complicate the fault characterization process. For that, a simple statistical method is developed. Furthermore, the proposed approach is evaluated using on two real and ten synthesized software systems. The results are compared 1) to those obtained from the FIC\_BS approach, after the original approach is modified in two different ways to handle multiple failures and 2) to those obtained from FDA-CIT – a feedback driven adaptive combinatorial interaction testing approach.

**Responses to Referee: 1**

**Comment 1:**

One of my major concerns is the applicability of the proposed approach in practice, which is greatly hurt by the assumptions made in the paper, including all failures are deterministic and there are no inter-option constraints.  Needless to say, these assumptions generally do not hold true in practice. While the first assumption is mentioned in the paper (Assumption 1 on page 6), the second assumption is not mentioned. This is important because in the presence of inter-option constraints, which invalidate some option setting combinations, some of the propositions and proofs given in the paper seem to become invalid and the approach needs to go through a set of nontrivial modifications. These issues need to be discussed in the paper and at the very least, the authors should elaborate on how to relax these assumptions.

Response: It is really true as the reviewer suggested that we did not clarify the assumptions that may hurt the application of our approach. Hence, we explicitly clarify the assumptions. There are three assumptions that we propose, i.e., the result of test case is deterministic, failures can be distinguished and no inter-option constraints existed. Additionally, according to the comments, we add one more section (Section 7) to explain that the impacts to our approach, as well as the measures to relax them.

**Comment 2:**

I believe what is meant by Assumption 1 in the paper is that all failures are deterministic, rather than all test results are deterministic. If so, this assumption should be restated.

Response: Here what we mean by Assumption 1 is that the test results are deterministic. Actually, these two versions of deterministic testing are both existed in the studies of Combinatorial testing. For example, Zhang [1] and Ghandehari [2] are based on the assumption that test results are deterministic, while Yilmaz [3] and Fouché [4] referred to the deterministic failure. We believed they are all reasonable. In this paper, we use the first version of testing deterministic, as it is simpler to handle (Deterministic failures may introduce the problem of test case-aware covering array [5], which is beyond of this paper). Additionally, we believe that the non-deterministic of test results is caused by the non-deterministic failures, and hence we emphasize that in the Assumption 1 and later in section 7.

[1] Zhang, Zhiqiang, and Jian Zhang. "Characterizing failure-causing parameter interactions by adaptive testing." Proceedings of the 2011 International Symposium on Software Testing and Analysis. ACM, 2011.

[2] Ghandehari, Laleh Shikh Gholamhossein, et al. "Identifying failure-inducing combinations in a combinatorial test set." Software Testing, Verification and Validation (ICST), 2012 IEEE Fifth International Conference on. IEEE, 2012.

[3] Yilmaz, Cemal, Myra B. Cohen, and Adam Porter. "Covering arrays for efficient fault characterization in complex configuration spaces." Software Engineering, IEEE Transactions on 32.1 (2006): 20-34.

[4], Sandro, Myra B. Cohen, and Adam Porter. "Incremental covering array failure characterization in large configuration spaces." Proceedings of the eighteenth international symposium on Software testing and analysis. ACM, 2009.

[5] Yilmaz, Cemal. "Test case-aware combinatorial interaction testing." Software Engineering, IEEE Transactions on 39.5 (2013): 684-706.

**Comment 3:**

Furthermore, it is important to emphasize the fact that the proposed approach is concerned with option-related failures.  It would also be interesting to evaluate and/or discuss how the proposed approach could deal with non-option-related failures.

Response: Considering the Reviewer’s suggestion, we have emphasized that our approach can only handle the option-related failures (Section 3.1, Page 7). Additionally, we discussed the problems of non-option-related failure in Section 8.5.2, as well as some measures that can alleviate this problem.

**Comment 4:**

Another concern is that a number of definitions, propositions, and proofs are given in the paper, but the relationship between the theory and the proposed approach is not clear at all. How does the proposed approach benefit from these propositions? How does the proposed approach evolve from the given theory?

Response: We are very sorry for that we didn’t make it clear of relationships between the formal modeling and proposed approach. In fact, the definitions in Section 3.1 (Definition 3.1 to Definition 3.5) are given to formalize the subjects we discussed, that is, to formally define what is the MFS. The section 3.2 are generally to reveal the relationships between the test sets and schemas. These relationships, however, is the basic foundation of how to identify the MFS, and how masking effects will affect the MFS identification. To reach this target, Proposition 3.7 and 3.8 shows that any test sets has its corresponding minimal schemas. These two propositions show the theory to obtain the MFS, as MFS is exactly the minimal schemas of failing test sets. Proposition 3.10 and 3.11 show that for two test sets that has an inclusion relationship, then their minimal schemas also have some relationships. These two propositions give the foundation of the impacts of masking effects, as the masking effects will make us incorrectly judge the outcomes of test cases, which will result in that the failing test sets we thought are include or including those real failing test sets. Based on this, we can figure out the extent to which the MFS we obtained under the masking effects will be deviated from the real MFS. Other propositions in Section 3.2, i.e., Proposition 3.6, 3.9, are the auxiliary propositions to get those four propositions. Later in Section 4, we mainly formally define the masking effects,

According to the reviewer’s comments, we have emphasized that

Then, a natural idea is that shall we remove those other failures test cases.

关系在理一下! Proposition 和 方法的关系， proposition 之间的关系。每个proposition 和 approach 加了承上启下的语句。 Response里列个表定义的关系。 Response 里加了一个，都不是多余的， 存在问题

第三个问题，理论formal model 和 提出的方法的关联? 这个方法如何从这些proposition 中获利呢？方法如何从这个theory 中演化出来呢？

**Comment 5:**

A related concern is that it is not clear at all whether the proposed approach always guarantees to find the minimal failure inducing schema (MFS) as it is defined in Definition 3.5. Note that once a failure inducing schema is found, the number test cases required to prove that this schema is an MFS grows exponentially with the number of options in the schema. It is not clear whether the proposed approach actually tries all the subsets or not. A proof is needed here!

Response:

第四个问题，其中一个相关的就是 MFS 的得出 (方法能否保证找到的就是MFS，是不是把所有的子集都看过一遍，需要证明)。 这个地方需要详细说明FiC OFOT 的原理，并给出相关理论和不足的地方

加一章 3.3 定位的原理 (FIC OFOT)，及实例

**Comment 6:**

Masking affects are not only caused by failures. Unaccounted for control dependencies, for example, can also cause masking effects, e.g., the use of –help command line option may make the sut to display a help text and exit without exercising any of the remaining option setting combinations appearing in the underlying configuration.  Therefore, the authors can consider restating Definition 4.1. It is also important to mention that Definition 3.5 is not valid in the presence of masking effects when discussing this definition; not a couple of pages later.

第五个问题Masking effects 需要重新说明，（包括unaccounted for control dependencies）

此外，masking effects 会导致 Definition 3.5 失效 (MFS，需要讨论一下)。 提一下，这是在没有mask effect 的情况下

**Comment 7:**

The proposed approach deserves a section of its own. As it is right now, the description of the approach is spread across a number of sections. For example, Section 5.2 titled “a case study using the replacement strategy” is the first section that attempts to describe how the approach works. However, as also indicated by its title, this section is supposed to introduce a feasibility study and the approach must have been introduced much more earlier. What is the input to the proposed approach? What is the exact algorithm implemented by the approach? When does the search for failure inducing combinations terminate?

第六个问题 提出的方法需要自己的一节。 The approach 应该比 section 5.2 更早 ?？？？？？？？？？

解决： 第五章侧重点重点是和定位算法 联合起来说。 替换策略(只是一小节) ， 然后完整的又是一小节

输入? 输出? 具体算法? 终止条件？

**Comment 8:**

The original FIC\_BS approach needs to be explained in the paper, which is not that big of a deal, so that readers are not forced to read other papers just to see what is going on.   For example, just by reading the paper it is not clear how the “fixed parts” are determined and after reading the FIC\_BS paper, one can only guess.

答：fixed part 已经在算法 和 例子 中得加以说明和强调

第七个问题 详细说明一下FIC\_BS 工作原理，解释一下 fixed parts (和之前OFOT一起说明 (第四个问题))

**Comment 9:**

The third and sixth paragraphs in Section 2 suggest that the proposed approach will use some sort of static analysis, which is not true. Therefore, these paragraphs should be rephrased to avoid misleading the reader.

确实没有用静态分析，只是一个例子表达已有问题。 重新表达? 加一句强调！

第八个问题 方法没有静态分析，需要重新组织 motivation example (第6段，第二节)

**Comment 10:**   
  
To generate “desirable” configurations, a suspiciousness score is computed for every option and failure. What is the justification for using such a scoring scheme? Doesn’t it make more sense to compute the scores for combinations of option settings rather than for individual options?

常用的方法，和 谱分析 的思路一样， 这两个一样。单个好，组合更好。

第九个问题 A suspiciousness score (对单个option) 的justification (根据) 是什么？ 难道对 combination 不是更好

**Comment 11:**

It is quite confusing that the two strategies introduced by the paper, namely “regarded as one failure” and “distinguishing failures”, are explained using the OFOT approach, which is a different fault characterization approach, than the one used in the experiments, namely FIC\_BS. How was FIC\_BS augmented with these strategies in the experiments?  For example, with the “distinguishing failures” strategy what happens if a configuration reveals a different failure?

第十个问题 策略介绍用的是OFOT， 但是实验用的是 FIC\_BS, 需要解释一下 (单独和之前第四个问题)

**Comment 12:**

In the first paragraph of Section 5.1, it is not clear what is meant by the following sentence: “The replacement must satisfy the condition that the newly generated ones will not negatively influence the original identifying process.”

重新说一下，可满足的测试用例

第11个问题 The replacement must satisfy the condition that the newly generated ones will not negatively influence the original identifying process. 是什么意思

**实验部分**

I have many concerns about the way the experiments were carried out:

**Comment 13:**

a)      The authors need to further justify the use of synthesized subject applications and elaborate more on how these subjects were actually created.

模拟程序的使用的 根据 需要更多的说明， 还有这些模拟程序是如何created 的要更多的说明

**Comment 14:**

b)      What was the MeetEndCriteria used in the experiments? Why? A sensitivity analysis for this parameter would greatly benefit the work.

MeetEndCriteria 具体是什么，为什么要用这个，对这个参数的sensitivity analysis (灵敏度分析) 很有价值，需要给出。

**Comment 15:**

c)      The configuration space models used in the experiments are relatively small. Since the authors have already used simulations, they could have also evaluated the proposed approach on larger configuration spaces.

增加几个大的，但不全测。 说明为什么

参数space 使用的相对较小。因为已经使用了模拟的，可以使方法 在更大的space 使用  
  
d)      The factors used for the real subject applications seem to be configuration options. Therefore, what were the actual test cases that were executed in the chosen configurations?

简单介绍测试用例是什么

Factors 只是option，具体的测试用例是什么? (对每条configuration，这里要考虑test-aware)  
  
e)      Please rephrase “to construct many real testing objects [for evaluations] is time-consuming”.

to construct many real testing objects [for evaluations] is time-consuming” 重新说  
  
f)      What is the degree of an MFS?

一个MFS 的degree 是什么。 参数个数，补充一下 (在试验中)，看哪个定义  
  
g)      Why was the “ignored number” metric treated differently than the rest of the evaluation metrics?  What was the issue with none of the MFS being ignored (paragraph 3 in Section 6.2.2)?

ignored number 为什么不同，(使用了全部的测试用例啊)， 为什么none of the MFS being ignored  
  
h)      In the Figures 3a-g, the points that belong to the ILP approach are almost impossible to distinguish from the rest.  
Figures 3a-g 需要使用标准化计算 , ILP的点很难区分，  
One major concern is that the experiments reported in Section 6.2 assume that for a given failure, all the test cases (configurations) revealing the failure are known a priori, which is not realistic at all. In the experiments, given a configuration space, the space is exhaustively tested, and then for each failure, every test case revealing the failure is fed to the proposed approach, and then the MFSs individually obtained for each test case are combined. Therefore, the results reported in this section are over-approximations. To be fair, the proposed approach should not work on any failures that it hasn’t discovered by itself, except for the input failure. For example, the average performance metrics obtained from failures could be reported.

Section 6.2 所有的能揭露这个错误的测试用例都已经提前获知。 在试验中，针对每个错误， 所有的每条错误测试用例都被 fed to 这个方法，然后MFS 被combined。 这个结果因此就是over-approximations。 Approach只应该 对 自己揭露的错误进行定位。因此，平均性能是最重要的。

The discussion on page 24 states that ILP and “distinguishing failures” approaches showed the same or similar performance (i.e., similar exact number, sub number, super number, etc.), but ILP was more costly. I thought that ILP was one of the main contribution of this paper as “distinguishing failures” is a trivial extension for the existing FIC\_BS approach. The authors should elaborate more on this.

ILP代价高昂，但是结果却和 distinguish 相同。因此需要更多的说明。““distinguishing failures””只是一个简单的扩展，ILP的意义呢？

更多的masking effect 会导致结果大不一样， 在 mask 很小的时候不明显，但是多了就明显  
  
In Section 6.3, it is stated that compared to the random test generation strategy, ILP reduced the number of test cases needed by 1 to 2, on average. What is the theoretical and practical significance of this reduction? Doesn’t it suggest that the masking effects used in the experiments were easy to avoid? The approach should be evaluated more rigorously. Furthermore, I strongly believe that all the experiment sections should be rewritten, as it is quite difficult to follow what is going on and to reason about the results.

Slight improve, 但不是我们的贡献点

ILP只减少了1到2 条的意义呢？ 方法需要被更加严格的评估。 Masking effects 不是很容易被避免的吗？那么怎么会减少呢? 如果重要的话，减少更多不是更好吗？ 所有的experiment section 需要重写，因为很难follow 和 理解这些结果  
  
Section 6.4 compares the proposed approach with FDA-CIT. However, the comparisons don’t seem to be fair. First, the authors should clearly mention that FDA-CIT’s primary concern is to avoid masking effects and give every t-tuple a fair chance to be tested, not to perform fault characterization and that it can work with non-deterministic failures and in the presence of inter-option constraints. In particular, FDA-CIT uses fault characterization as a tool to reach its goals and it can work with other fault characterization approaches (as also noted by the authors).

需要强调 First, the authors should clearly mention that FDA-CIT’s primary concern is to avoid masking effects and give every t-tuple a fair chance to be tested, not to perform fault characterization

还需要强调 hat it can work with non-deterministic failures and in the presence of inter-option constraints.

还需要强调，FDA-CIT uses fault characterization as a tool to reach its goals and it can work with other fault characterization approaches (as also noted by the authors). 定位只是一个工具，可以使用其他工具

Second, in the experiments, FDA-CIT is used with t-way covering arrays, where 2 <= t <= 4. When working with failures caused by the interactions of t options, the strength of the covering arrays to be used with FDA-CIT should be at least (t+1). When a t-way covering array is used, very few instances of the failure-inducing t-tuple would appear in the array (only one instance in the worst case).  When a lower-strength covering array is used, the failure inducing t-tuple may not even appear in the array, which prevents the approach to even reveal the failure. The comparisons are not fair because although there were 6-way failures (Table XIX), the highest strength used was 4. It should have been at least 7. This can clearly be seen in Figures 5a-c, where, as t increased, FDA-CIT’s overall performance increased. Considering that the proposed approach required significantly more number of test cases than FDA-CIT (e.g., 172 vs. 626), the extra test cases required for the higher strength covering arrays may not affect the cost relationship much.

FDA-CIT 应该使用t+1的覆盖表。有6-way的错误，但是最高的维度只有4. FDA-CIT 至少应该用7维的覆盖表。 t 增加，fda-cit 能力也增加。

The paper should rigorously address the cost benefit tradeoffs. Assume that we are concerned with highly configurable systems where each factor corresponds to a configuration option. Then, what the paper refers to as a test case will correspond to a configuration. In such systems, a configuration is tested by running a battery of test cases in the configuration. For example, if we have 1000 test cases, given a configuration, we may want to execute all of these test cases in the configuration and determine the failure-inducing option setting combinations on a per test case basis. In such cases, however, the proposed approach needs to be carried separately for every test case (and even for every failure of a test case). That is, the number of configurations required by the proposed approach grows linearly with the number of test cases.  However, this is not the case for FDA-CIT as test cases share configurations in FDA-CIT. Considering that the number of test cases required by the proposed approach is already significantly higher than FDA-CIT, this may cause scalability issues in real world scenarios.

对于test case-aware covering array的情况下，他认为我的方法需要针对每个单独的测试用例 进行 定位。 因此需要的测试用例的数量呈现线性增长的趋势 随着 测试用例的个数增长。 但是FDA-CIT不是这样。 然而已经在很多测试用例的情况下，这个方法可能很难应用到实际中去。

More details should be provided about how the FDA-CIT approach was implemented. For example, the first paragraph on page 29 mentions about an “over fitting” problem. Were the classification trees computed by using n-fold cross validation to avoid over fitting as suggested by FDA-CIT? What was n? The same paragraph also suggests that every path from the root to a leaf node is considered as MFS. However, FDA-CIT only selects those paths (those MFSs) whose accuracy is above a given cutoff value. What was the cutoff value used in the experiments? How was this cutoff value selected?

how the FDA-CIT approach was implemented? 包括方法? N-fold ccorss.里的n (over fitting)， cutoff value 是多少 (只有在此之上的schema 才被认为是MFS)?  
  
It is good that the authors published the subject applications and the configurations used in the experiments online. However, it is not quite clear how to reproduce the experiments, e.g., how to configure these subject applications and how to run the test cases. Furthermore, the synthesized subject applications and the respective test cases are missing.

如何运行这些实验。 人工的application 少了  
  
  
Minor issues:  
  
+ Combinatorial interaction testing is often abbreviated as CIT, not as CT.

Page2 : Another term for CT is Combinatorial Interaction Testing, which is abbreviated as CIT. In this paper, they are uniformly cited as Combinatorial testing (CT)

+ Second paragraph on page 14, the second “the left part” should be “the right part”

has been fixed

+ Fifth paragraph on page 14, “an scenario” -> “a scenario”

has been fixed

+ Second paragraph in Section 6.1.2, the sentence that starts with “In fact,” is incomplete.

这个句子不完整 In fact, there are about 768 out of 4608 test cases (16.7%) in hsqldb with 2rc8 version. ?？？？？？？？ 重新说一下

At last, special thanks to you for your good comments.  
  
Additional Questions:  
Review's recommendation for paper type: Full length technical paper  
  
Does this paper present innovative ideas or material?: Yes  
  
In what ways does this paper advance the field?: This work improves an existing approach, called FIC\_BS, which aims to pinpoint failure-inducing combinations of option settings (aka factor settings), such that the resulting approach can operate in the presence of multiple defects; the original approach assumes that all failures are caused by the same defect. The proposed extension is a technique, which given a failing configuration (a test case as it is referred to in the paper) and a fixed part in this configuration (i.e., options and their settings which must be remained unchanged), computes a configuration which is highly likely to pass or to reproduce the same failure. That is, the new configuration is chosen, such that it is unlikely that it reveals a different failure, which would complicate the fault characterization process. For that, a simple statistical method is developed. Furthermore, the proposed approach is evaluated using on two real and ten synthesized software systems. The results are compared 1) to those obtained from the FIC\_BS approach, after the original approach is modified in two different ways to handle multiple failures and 2) to those obtained from FDA-CIT – a feedback driven adaptive combinatorial interaction testing approach.  
  
Is the information in the paper sound, factual, and accurate?: No  
  
If not, please explain why.: One concern is that a number of definitions, propositions, and proofs are given in the paper, but the relationship between the theory and the proposed approach is not clear at all. How does the proposed approach benefit from these propositions? How does the proposed approach evolve from the given theory?  
  
A related concern is that it is not clear at all whether the proposed approach always guarantees to find the minimal failure inducing schema (MFS) as it is defined in Definition 3.5. Note that once a failure inducing schema is found, the number test cases required to prove that this schema is an MFS grows exponentially with the number of options in the schema. It is not clear whether the proposed approach actually tries all the subsets or not. A proof is needed here!  
  
Rate the paper on its contribution to the body of knowledge in software engineering (none=1, very important=5): 3  
  
What are the major contributions of the paper?: 1) A set of definitions, propositions and proofs concerned with fault characterization in the presence of masking effects. However, it is not clear how this theory supports the proposed approach.  
  
2) An approach for pinpointing failure inducing options in the presence of masking effects, however the assumptions made in the paper greatly hurt the practicality of the approach.  
  
3) A series of experiments comparing the proposed approach to extended versions of FIC\_BIS and FDA-CIT. However, there are many concerns about the way these experiments were carried out.  
  
Rate how well the ideas are presented (very difficult to understand=1, very easy to understand=5): 3  
  
Rate the overall quality of the writing (very poor=1, excellent=5): 3  
  
Does this paper cite and use appropriate references?: Yes  
  
If not, what important references are missing?:  
  
Should anything be deleted from or condensed in the paper?: No  
  
If so, please explain.:  
  
Is the treatment of the subject complete?: No  
  
If not, What important details / ideas/ analyses are missing?: There are some concerns about the way the proposed approach is evaluated and the relationship between the theory and the proposed approach.  
  
Please help ACM create a more efficient time-to-publication process: Using your best judgment, what amount of copy editing do you think this paper needs?: Moderate  
  
Most ACM journal papers are researcher-oriented. Is this paper of potential interest to developers and engineers?: Maybe  
  
  
Referee: 2  
  
Recommendation: Needs Major Revision  
  
Comments:  
This paper addresses the problem of identifying the combination  
of parameters which leads to a failure detected by a combinatorial test suite.  
It gives a theory of the "minimal failure causing schema" (MFS), i.e. the smallest  
set of parameters which causes the failure. Then it illustrates how FII approaches,  
described in other papers, can identify these MFS.  
Then starts the original part of the paper which investigates the effect  
of multiple failures on FII algorithms. The paper shows that these additional  
failures may have a masking effect on the failure under study, and lead to a  
non-optimal identification of the MFS. The authors propose a solution to this  
problem based on the replacement of masked test cases.  
Finally, the paper includes four empirical studies that help appreciate  
the impact of masking effects, that compare the proposed approach to existing  
ones without replacement, that show that the proposed replacement strategy is better  
than random replacement, and  finally compare the approach to another  
approach handling masking effects. These empirical studies show that the  
proposed approach is superior than the other ones, but at the cost of  
evaluating additional test cases.  
  
The paper is clearly written and addresses various aspects of the problem  
(identification of failure causes in the presence of masking effects).  
It includes formal modelling, good illustrative examples, a proposal  
for solution and various empirical studies. There is substantial work  
in the paper and real contributions.

**Responses to Referee: 2**

The major problem of the paper is the treatment of randomness in the empirical  
studies. Three of the four empirical studies involve some level of randomness  
(random replacement for studies 2 and 3, generation of 2-4 way coverage).  
The authors compute the average of 30 tries, but don't report on notions  
like confidence interval. Although I am not a statistician, I am convinced  
that there is not sufficient analysis to provide statistical evidence.  
Also, in study 4, you are comparing average values; so you should probably do ANOVA  
analysis or something similar.

试验中随机的评估。 文章中随机的生成覆盖表30次， 还有 random replacement。 但是没有给出致信区间。 （studies 2 和 3） study 4 比较平均，应该给出 ANOVA 给出，  
  
Detailed remarks  
  
Page 2 "We can get five two-way suspicious interactions..." five should be six!

Has been fixed

Page 6 : definition 3.3, you should not use k in v\_k\_1 or v\_k\_t because k is  
intended to represent the number of parameters that influence the SUT.

Using v\_b\_1 and v\_k\_t

Page 6: Definition 3.5  
The definition should be clarified. It states that "all test cases ... trigger ... failure F".  
This is not the case with the example of Table II where only 4 test cases trigger Ex1.  
In most cases, the failure is not triggered by ALL test cases. So you should make clear  
which is the set where all test cases trigger a failure.

If any test case which contains the schema c, and trigger a particular failure F

page 9: Proposition 3.11  
Could you give a proof of this proposition?  
Probably I don't understand the term "antithesis" correctly  
but as far as I understand it, the antithesis is the exact opposite of the thesis.  
So if both thesis and antithesis are true, everything is true!

给个证明

Proposition 的问题，不能两者都相反。

page 11: typo in section 4.3 "and the other test cases (t2,t3) failed" should be "(t2,t4)".

Has been fixed  
  
page 12, first line "t1 and t3 should..." should be "t1 and t2 should..."

Has been fixed  
  
page 13 section 5.1, 4th paragraph, 2nd line "as it may not always BE possible..."

Has been fixed

page 14 line 7 "the left part" should be "the right part"

Same as third minor question as CYlimaz. And has been fixed.  
  
page 16 section 5.2 "fixed part needed to be testED in each iteration"

Has been fixed  
  
page 17, Figure 2  
The suspiciousness matrix is related to e2 and e3 and not e1 and e2

Has been fixed

page 17, Figure 2  
I don't understand the choice of t4'. Why don't you choose "00001221"  
which has a better suspiciousness score?

需要解释fic\_bs ， fic\_bs 需要每次生成额外测试用例的时候保持一部分。  
  
page 18, table XVI  
what is the "bug pairs", how did you choose these pairs, why are they not  
the same in the versions of the same program? (were they fixed with the new version?).

需要解释bug pairs

page 19, section 6.1.2 and table XVIII.  
how many tests of the "failures" count do contain the MFS?  
From your text, it seems that all failures did contain the MFS, and  
that none of the tests did contain the second fault alone.  
It also seems that there are only two faults in each version.

所有的failure是由MFS引起的 。 没有单独触发第二个错误的测试用例， 每个version都有两个错误。 (有的，是由单独引发第二个故障的测试用例存在，不然，我们无法观测，masking 的部分就是两个fault 同时存在的时候。我需要把这个表再细化如下)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Software | Version | All tests | Failure | | Mask 多少条 |
|  |  |  | #1() 多少条。 | #2 多少条 | （）本该#2但是却全被#1 覆盖了 |

是的，这个  
  
page 20 section 6.2 "we observed that the number of parameters..."  
On which case studies did you make these observations? the ones of table XVIII?  
Probably not the case studies of table XVIII as the largest one has 15 parameters (and not 30).  
So you should make more precise from which observations you deduced the number  
of parameters.

软件的 大小， Range from 10 到 30， 但是没有这么大的软件， 是的，需要重新添加新的大的软件。

方案一：此外，我们还参考了 。。。。。。。 文章 文章的subject 的 syncate ()， 应该有大的

方案二： 不提这个  
  
page 22: last two lines  
testing object 2 has no best strategy  
testing object 3 has distinguishing failures as best strategy  
So the numbers do not correspond between this paragraph and figure 3.

为什么 ? 是的

对的，是我粗心了，重新描述 test object Fig 3. A 应该是ob 1 ILP, ob 2 no. ob 3, distinguish. Ob 11 regarded as one good.  
  
---  
  
Additional Questions:  
Review's recommendation for paper type: Full length technical paper  
  
Does this paper present innovative ideas or material?: Yes  
  
In what ways does this paper advance the field?: This paper addresses the problem of identifying the combination  
of parameters which leads to a failure detected by a combinatorial test suite.  
It gives a theory of the "minimal failure causing schema" (MFS), i.e. the smallest  
set of parameters which causes the failure. Then it illustrates how FII approaches,  
described in other papers, can identify these MFS.  
Then starts the original part of the paper which investigates the effect  
of multiple failures on FII algorithms. The paper shows that these additional  
failures may have a masking effect on the failure under study, and lead to a  
non-optimal identification of the MFS. The authors propose a solution to this  
problem based on the replacement of masked test cases.  
Finally, the paper includes four empirical studies that help appreciate  
the impact of masking effects, that compare the proposed approach to existing  
ones without replacement, that show that the proposed replacement strategy is better  
than random replacement, and  finally compare the approach to another  
approach handling masking effects. These empirical studies show that the  
proposed approach is superior than the other ones, but at the cost of  
evaluating additional test cases.  
  
Is the information in the paper sound, factual, and accurate?: Yes  
  
If not, please explain why.:  
  
Rate the paper on its contribution to the body of knowledge in software engineering (none=1, very important=5): 3  
  
What are the major contributions of the paper?: The major contributions are the proposal for a replacement strategy and the empirical studies.  
  
Rate how well the ideas are presented (very difficult to understand=1, very easy to understand=5): 5  
  
Rate the overall quality of the writing (very poor=1, excellent=5): 4  
  
Does this paper cite and use appropriate references?: Yes  
  
If not, what important references are missing?:  
  
Should anything be deleted from or condensed in the paper?: No  
  
If so, please explain.:  
  
Is the treatment of the subject complete?: No  
  
If not, What important details / ideas/ analyses are missing?: The major problem of the paper is the treatment of randomness in the empirical  
studies. Three of the four empirical studies involve some level of randomness  
(random replacement for studies 2 and 3, generation of 2-4 way coverage).  
The authors compute the average of 30 tries, but don't report on notions  
like confidence interval. Although I am not a statistician, I am convinced  
that there is not sufficient analysis to provide statistical evidence.  
Also, in study 4, the authors are comparing average values; so they should probably do ANOVA  
analysis or something similar.  
  
Please help ACM create a more efficient time-to-publication process: Using your best judgment, what amount of copy editing do you think this paper needs?: Moderate  
  
Most ACM journal papers are researcher-oriented. Is this paper of potential interest to developers and engineers?: Yes  
  
  
Referee: 3  
  
Recommendation: Needs Major Revision  
  
Comments:  
Summary:  
This paper proposes a technique to identify the failure inducing input interactions  
in the presence of multiple faults. The challenge lies in that these faults may cause  
multiple failures that mask each other. As such, an input including the failure inducing  
interaction may not trigger the expected failure because it is masked by another failure.  
Their idea is that for a failure under study, replace the failing test cases that induce  
other failures with test cases that either pass or induce the same failure (under study).  
This is done by fixing part of the input while altering the rest. The mutation is guided  
by so-called the suspiciousness score. An input element value is suspicious regarding  
a failure if the value frequently occurs in the executions with the same failure. The  
technique would pick an alternation scheme that causes the minimal suspiciousness scores  
regarding failures other than the one under study. The technique is evaluated on 5 versions  
of 2 real-world programs and 10 small synthesized programs. The results show that the  
technique has decent improvement over the state of the art based on feedback driven  
combinatorial testing. But the improvement over simpler strategies is not as substantial  
as expected.  
  
Pros:  
+ Failure inducing interaction identification in the presence of multiple faults is an  
 realistic and interesting challenge.  
  
+ The proposed solution is reasonable although the idea is bit straightforward, which may  
 be a good thing.  
  
+ The technique is evaluated on two real-world applications.  
  
+ The comparison with the state of the art (based on feedback driven combinatorial testing)  
  seems promising  
  
Cons:  
- The improvement over simpler strategies is not substantial. For example, using the suspiciousness  
  score only reduces 1-2 test cases on average compared to using a random selection approach.  
  Given the overhead of computing and maintaining the suspiciousness scores, random selection may be  
  a more cost-effective solution. The benefits of using mutation and ILP do not seem to go beyond  
  the very simple distinguish-failure strategy much, which simply does not consider the failing cases that do  
  not belong to the same fault. The two have almost the same performance for 4/5 metrics in Fig. 3.  
  Even for the metric that the proposed method shows benefits, I wonder if one can easily improve  
  the performance of the distinguish-failure strategy by having more test cases.

多加一些 错误， 多加一些 error 的形式， 可以导致 mask 非常多，因此就会有很好的结果

- The evaluation can be improved. It is currently evaluated on only 2 programs with 5 versions, each version  
  having two bugs. I wonder why not collect more bugs for each version and have more programs (and versions)?  
  I would rather see the author’s trade the current space used in the over-detailed experiment set-up for more  
  programs and bugs. The synthetic programs are not that useful.

多加一些 真实的软件， 有更多的错误， 可以和 张志强 simens 的 (和上一个的解答 形成一致)  
  
- It is unclear what is the termination condition of this failure inducing interaction identification process. The paper  
  seems to indicate that it terminates when the MFS is computed. But from my point of view, the computation  
  of MFS is determined by the test suite you have. This seems to make it a chicken-and-egg problem.

重新说一下，终止条件 不能是终止条件， 需要 具体 给出 条件。

- Part of the technique hinges on properly classifying failures, which is a hard challenge in general.  
  The authors should discuss how they achieve this.

错误分类， 是非常复杂的， 我只是对每个错误的 exception 信息 进行分类， 如果不是这样的话 ，我们可能需要对 执行路径 和 程序行为来进行 统计分类 方法， 这在这篇文章之外  
  
- I don't understand why ILP is needed. A simple linear search algorithm shall do the work. Please explain.

测试用例全是整数的

ILP 只是其中的一种解法  
  
- The paper still contains a lot of grammatical problems. It has to go through very rigorous proof-reading.  
  
  
Specifics:  
  
Abstract: "theory lack"=> "theory lacks"

Has been fixed

page 3: "newly regenerated" => "newly generated"

Has been fixed

page 3: what is a factor?

page 3: "suffered multiple failures" => "encountered multiple failures"

Has been fixed

page 3: "import masking effects" => "induce masking effects"  
Has been fixed

Table II, there seems to be a soundness issue here. For ex2, it is possible that the programs just fails with  
<7,2,4,5> and <11,2,4,5> but not <5,2,4,5>. But know the technique seems to think that <-,2,4,5> would fail  
the program. How can your technique address this in general?

page 7: "L: The number of failures ...". You should distinguish faults and failures (through out the paper).  
They have separate meaning in software engineering but the paper simply uses failures in all cases.

failures 和 fault 的定义

page 7: "T(c)<=T\_Fm", should it not be "T(c) >=T\_Fm"?

不是得，T（c）是单个的failing testcases Tfm可以有多个  
  
pages 8-9: The proofs are not that useful. They are quite obvious.

？？

page 9: "impacts of masking..." => "impact of masking", the same problem occurs a few times.

page 10: "one failure-fail"=> "one-fault failure"

注意这个

page 11: "significantly impact on" => "has significant impact on", the same occurs a few times.

page 11: "We offer..." => "Consider"

Has been fixed

page 11: "The pass of..." => "The passings"

Has been fixed

page 12: "In other word",=> "In other words", the same happens a few times

page 14: "between test case" => "between test cases"

Has been fixed

page 14: "triggers other failure"=> "triggers other failures"

page 14: what do you mean by "the maximal possible failure"?

page 14: "the corresponding failure", do you mean "all other failures" here?

page 16: "number of attempts is"

page 19: "SUT have" => "SUTs have"

The observations in 6.1 are not new. So you are not gaining much out of this experiment

.  
page 20: "can happened"

Can be in the form Has been fixed

page 24: what do you mean by "One issue is the redundancy...", please rephrase.

page 26: "produce the generated..."

句子不好， 改成 guarantee that the generated test cases should cover all the ?

page 32: "of constrains"

Has been fixed

page 32: "covering array with considering" => "covering array while considering"

Has been fixed

page 32: "this constraint" => "these constraints"

Has been fixed

page 32: "First, the work that aims..", not a sentence.  
  
  
  
  
  
Additional Questions:  
Review's recommendation for paper type: Full length technical paper  
  
Does this paper present innovative ideas or material?: Yes  
  
In what ways does this paper advance the field?: This paper proposes a technique to identify the failure inducing input interactions in the presence of multiple faults.  
  
Is the information in the paper sound, factual, and accurate?: Yes  
  
If not, please explain why.:  
  
Rate the paper on its contribution to the body of knowledge in software engineering (none=1, very important=5): 3  
  
What are the major contributions of the paper?: Please see the review below.  
  
Rate how well the ideas are presented (very difficult to understand=1, very easy to understand=5): 4  
  
Rate the overall quality of the writing (very poor=1, excellent=5): 3  
  
Does this paper cite and use appropriate references?: Yes  
  
If not, what important references are missing?:  
  
Should anything be deleted from or condensed in the paper?: Yes  
  
If so, please explain.: The proofs in pages 8-9 can be removed. The experiment set-up can be shortened.  
  
Is the treatment of the subject complete?: No  
  
If not, What important details / ideas/ analyses are missing?: The experiment can be improved.  
  
Please help ACM create a more efficient time-to-publication process: Using your best judgment, what amount of copy editing do you think this paper needs?: Moderate  
  
Most ACM journal papers are researcher-oriented. Is this paper of potential interest to developers and engineers?: Maybe  
  
===================================