# Response to comments of Dave

***Dave[3]***

Response: We have mentioned partition testing in the abstract.

***Dave[4]***

Response: We have added more recent references about ART.

***Dave[5]***

Response: A fault may result in a failure in case there is no protection. Here, we would not distinguish faults and failures. Instead, we focus on the occurrence of a fault, namely the output of an input does not match the expected one. Thus, we believe that “fault” is okay.

***Dave[6]***

Response: We hereby confirm that your understanding is precise.

***Dave[7]***

Response: Our experimental results showing DRT outperform RT in terms of fault detection efficiency.

***Dave[8]***

Response: The comparison between DRT and RT or RPT focuses on fault detection efficiency in terms of the number of test cases and time required for detecting a given number of faults. That is, to detect the same number of faults, DRT uses fewer test cases than RT or RPT, and requires less time than RT or RPT.

***Dave[9]***

Response: The plural of formula has been changed to formulas throughout the paper.

***Dave[10]***

Response: We confirm that “Formulas 1 to 4” is correct. By the way, we have corrected the overall time complexity for DRT, and it should be O(mn).

***Dave[11]***

Response: To invoke a web service, one first needs to know its input parameters through analysis of the input message in its WSDL, and generate test data, and finally wrap generated test data in a SOAP message.

***Dave[12]***

Response: We intend keep this list to support the subsequent argument that SOA testing is more challenging than traditional testing. After a second thought, we have rephrased most points. We try to connect these points with the necessity of DRT.

1) *Lack of access to service implementation*: Normally service owners will not make source code of their web services accessible. In this context, service users only have access to the service interface defined in a WSDL file, which means that white-box testing approaches are not possible.

2) *Incomplete documentation or specification*: A service provider may normally offer an incomplete or inaccurate description of a service’s functional and non-functional behavior. This makes it difficult to decide whether a test pass or not, especially some behaviour assumptions or restrictions behind implementation are missing.

3) *Lack of control*: Unlike traditional software testing where testers can control the execution of software under test, there is no chance to intervene of execution of web service under test since a web service is often deployed in a remote service container.

4) *Side effects caused by testing*: A large number of tests may introduce extra communication load, and hence affect the performance of the Web service under test. This indicates the number of tests should be reduced as much as possible.

***Dave[13]***

Response: We have changed the description of this point.

***Dave[14]***

Response: We have added several recent references.

***Dave[15]***

Response: We have added some recent references.

***Dave[16]***

Response: One way to determine the importance of a partition is based on the number of detected faults within it in the testing history.

***Dave[17]***

Response: We have added some recent references.

***Dave[18]***

Response: Correct.

***Dave[19]***

Response: Ideally, a partition should be homogeneous, that is, if one input is fault-revealing, all others inputs in the same partition will be fault-revealing, too. This explained why partition testing is widely adopted in practice to save the testing effort. Accordingly, we changed “effectiveness” to “practicability”. Is this change okay?

***Dave[20]***

Response: We have added a note about the parameter \varepsilon.

***Dave[21]***

Response: We have revised the references to formulas in a consistent way.

***Dave[22]***

Response: Yes, we have the same assumption with you.

***Dave[23]***

Response: We have checked the full-stops after all formulas. Just keep as they are.

***Dave[24 ]***

Response: OK.

***Dave[25]***

Response: We confirm that the changes you have made are correct.

***Dave[26]***

Response: After checking some relevant literature, no special notation for an interval. So, just keep as it is.

***Dave[27]***

Response: Please be noted that partitioning cannot be automated, human inputs are necessary. As illustrated in the experiments discussed later, we can employ Decision Table to partitioning, where decision rules have to be specified with human effort. For your information, we are still improving the prototype, and an updated version will be ready very soon.

***Dave[28]***

Response: Both effectiveness and efficiency are covered. The subsequence empirical study evaluated the effectiveness and efficiency of RT, RPT, and DRT in terms of F-, F2-, and T-measure. These measurement (including F-, F2-, and T-measure) also reflect the effectiveness of RT, RPT, and DRT. For instance, to detect 5 faults, 10 test cases were used for DRT, 15 test cases were used for RPT, and 20 test cases were used for RT, which means that after executing 10 test cases, DRT detected 5 faults, and the number of faults detected by RT and RPT is less than 5. Therefore, we can think that DRT has a higher effectiveness.

***Dave[29-30]***

Response: We have checked the formatting styles for Tables. We also changed column headings from “plan” to “option”, while keep PLAN A/B/C as table titles.

***Dave[31]***

Response: We have moved the “$” sighs to the left in Table 6.

***Dave[32]***

Response: OK. I have add some references for different measures.

***Dave[33]***

Response: We have explained why these measures were adopted in our experiment.

***Dave[34]***

Response: We have rewritten this paragraph. Please check it.

***Dave[35]***

Response: We have deleted the last line of the previous paragraph.

***Dave[36]***

Response: We have rephrased this paragraph.

***Dave[37]***

Response: We decided to use “scheme”. Accordingly, we have changed all “schema” to “scheme” in all figures.

***Dave[38]***

Response: We confirm that your understanding is precise.

***Dave[39]***

Response: We have followed the practical guidelines proposed by Andrea Arcuri in [38].

***Dave[40]***

Response: Usually, the thread to external validity is related to subject programs and the associated faults.

***Dave[41]***

Response: We have elaborated a bit more about the distinct faults.

***Dave[42]***

Response: As mentioned above, we have followed the guideline of statistical experiments in the field of software engineering proposed by Andrea Arcuri and Lionel Briand [38]. That is, when analysing or comparing randomized algorithms, there is a common rule of thumb using at least n = 30 observations, which is also adopted in other fields, such as medicine and behavioural science, to make statistical results significant.

***Dave[43]***

Response: We have used underlying and bold to indicate the signification. [Table 17-19 not changed with this]

***Dave[44]***

Response: In the previous version, there were some mistakes in Section 3.2. We have re-run the experiments and updated the experimental results accordingly. [NEED CONFIRMATION]

***Dave[45]***

Response: This should be okay. We use scenarios to mean different cases or settings.

***Dave[46]***

Response: After a second thought, we decided to change “h” to \varepsilon^{\*}, which is better than “h”, indicating a special value of \varepsilon.

***Dave[47-48]***

Response: Accordingly, we have changed “h1” and “h2” to “\varepsilon^{\*}\_1” and “\varepsilon^{\*}\_2”, respectively. We have also added a note on the axes and values Figure 6. [Figure 6(b) is correct? NEED CONFIRMATION]

***Dave[49]***

Response: Correct.

***Dave[50]***

Response: We have added scheme 1/2 in all subfigures of Figures 7~9.

***Dave[51]***

Response: All the changes you made are good.

***Dave[52]***

Response: We have added two recent references, which was published in 2014 and 2015, respectively. [Reference 40 is correct? NEED CONFIRMATION]

Response: I have re-checked this reference, and found that only a little of content is related to test web services. Therefore, I think we should delete this reference. Besides, I have added another paper which was published in the ACM Computing Surveys.

***Dave[53]***

Response: As suggested, we have added a reference on specification-based testing of Web services.

***Dave[54]***

Response: The previous one (Majdi et al.) is not a high-level journal paper, thus has been removed. We have added an alternative and more details on the newly added reference are provided. [Newly added reference on BPEL is not related to testing of Web services. Consider the replacement.]

***Dave[55]***

Response: We have re-descripted this reference and added some details as suggested [Description on the new reference is not understandable. NEED REPHASE!]

***Dave[56-57]***

Response: Added accordingly.

***Dave[58]***

Response: We added some details on the proposed MT framework for Web services.

***Dave[59]***

Response: We have added two references as suggested.

***Dave[60]***

Response: We have added a reference as suggested.

***Dave[61]***

Response: Yes, we have listed more algorithms such as FSCS-ART and FSCS-ART-with-Forgetting techniques. And we have added 5 reference as suggested. [Not just add some references. List more algorithms? RRT? Linear-order ART.]

***Dave[62]***

Response: No, DRT and ART hold completely different ideas. And I have changed my previous sentence to “Following the idea of software cybernetics, AT was proposed, and ….” [This is not correct!]

***Dave[63]***

Response: Many researches has shown that AT performed better than RT and RPT in terms of T-measure and increasing the number of faults, which means that AT has higher efficiency and effectiveness than RT and RPT. [Not answer the question: Outperform in what sense? Failure detection? F-measure? E-/P-measure?]

***Dave[64]***

Response: We have added some details on O-DRT.

***Dave[65]***

Response: During the testing process, the number of execution detecting a fault less than the number of execution which does not detect a fault. Based on above observation, Lv et al think that the parameter \varepsilon in Formulas 1 and 2 should be bigger than the parameter \dalta in Formulas 3 and 4, that is, In [16], there are two parameters in DRT. Moreover, Lv et al. used a theoretical analysis to identify the relationships between \varepsilon and \delta., which can be described as follow:

1 / (\theta\_x) – 1 < \varepsilon / \delta < 1 / (\theta\_y) – 1,

where \theta\_x and \theta\_y are partitions failure rates, and \theta\_x = max{\theta\_i| i =1,2,…,m} (assume that there are m partitions), \theta\_y = max{\theta\_i| i =1,2,…,m, i \ne x}.

From the above formula, if the value of a parameter is available, then the value of the other parameter can be calculated from that formula.

The difference between our work and their work is reflected in two points. On the one hand, there are only one DRT parameter in our study. On the other hand, our previous work [1] found that the DRT performance can be influenced by the number of partitions and the parameter \varepsilon used for adjusting selection probabilities of partitions. Therefore, we identified the relationships between parameter and the number of partitions.

[NEED REPHASE!]

***Dave[66]***

Response: Your understanding is not correct. Because our theoretical analysis identifies the relationships between number of partitions and DRT parameter and does not concern failure rates. By the way, in our study, DRT only has one parameter. Hence, the word “DRT parameters” should be “DRT parameter”. [Our theoretical analysis does not concern failure rates at all! ]

***Dave[67]***

Response: DRT obtains better performance than both RT and RPT according to F-, F2-, and T-measure.

***Dave[68]***

Response: We replaced “outstanding” with “better”.

***Dave[69]***

Response: Just keep as it is.

***Dave[70]***

Response: Just keep as it is.

***Dave[71]***

Response: Guan Wang is a master student when he worked on this paper. In China, we do not distinguish master in Science and master in Engineering.

***Dave[72]***

Response: Just keep as it is.

***Dave[73]***

Response: We copy Prof Cai’s bibliography from his co-authored papers and paste it to here.