# 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: Yes, you have a correct understanding.

***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: Yes, we should. I have added a short sentence to illustrate that \varepsilon is the adjustment of the selection probabilities of partitions.

***Dave[21]***

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

***Dave[22]***

Response: Yes, you can assume that. From the sentence “Letting \theta\_M denote the maximum failure rate, and s\_M denote **partitions** with that failure rate”, we can observe that more than one partition may have the maximum failure rate.

***Dave[23]***

Response: No, we think it is better to keep them.

***Dave[24 ]***

Response: OK.

***Dave[25]***

Response: Yes, all the changes you have made are correct.

***Dave[26]***

Response: Yes, it is. There is no a special notation denoting an interval according to the standard of ISO 80000-2:2009 which introduces the mathematical signs and symbols to be used in the natural sciences and technology. Therefore, I think that any capital letter is ok.

***Dave[27]***

Response: Sorry, I made a mistake. Our tool does generate test cases, but cannot automatically generate partitions. There are some problems with the current tool, and I will make a new one as soon as possible.

***Dave[28]***

Response: Yes, we conducted an empirical study to directly evaluate the efficiency of RT, RPT, and DRT in terms of F-, F2-, and T-measure. Meanwhile, the values of 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: OK. I have changed the column headings in each table, and checked the IEEEtran\_HOWTODO file, and finally changed all tables according to the guidelines. However, TABLE 3 is so big that I added an order, as follow: \resizebox{\textwidth}{12mm}{…}. Is that order okay?

***Dave[31]***

Response: Yes, I have put the “$” sighs on the left side in Table 6.

***Dave[32]***

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

***Dave[33]***

Response: In many related researches, the F-measure and T-measure are used for evaluating the fault-detection efficiency and effectiveness of the DRT testing technique. Therefore, we also employ the F-measure and T-measure as metrics to evaluate the fault-detection efficiency and effectiveness of different techniques.

***Dave[34]***

Response: No, we don’t. I have rewritten this paragraph, and described the decision table through a more formal method.

***Dave[35]***

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

***Dave[36]***

Response: Yes, I have rephrased this paragraph.

***Dave[37]***

Response: It should be scheme.

***Dave[38]***

Response: Yes. I hereby confirm that your understanding is precise.

***Dave[39]***

Response: I follow the practical guide of Andrea Arcuri in [38].

***Dave[40]***

Response: Usually, the thread to external validity is concerned with object programs and the faults. Based on this, we can defend this threat in terms of objects and faults. I have added some new arguments to defend this thread.

***Dave[41]***

Response: Yes, we can. I have added a sentence to describe the distinct faults.

***Dave[42]***

Response: Following the guidelines (Section 7 in [38]) of Andrea Arcuri and Lionel Briand, to analyse and compare randomized algorithms, there is a common rule of thumb using at least n = 30 observations, which is adopted in many fields of science (e.g. medicine and behavioural science) to make statistical results significant.

***Dave[43]***

Response: OK, we have employed underlying and bold to indicate signification.

***Dave[44]***

Response: In the previous version, there are some errors in Section3.2. After correcting the errors, I ran the experiment again. Therefore, the data is different from the previous data.

***Dave[45]***

Response: Sorry, I do not know, but in our previous work (Adaptive Partition Testing), we used the “scenarios” to express the same meaning.

***Dave[46]***

Response: After rethinking, I think \varepsilon^{\*} is a better symbol, indicating the special value of \varepsilon.

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

Response: I have changed “h1” and “h2” to “\varepsilon^{\*}\_1” and “\varepsilon^{\*}\_2”, respectively, and added a short sentence to explain the axes and values Figure 6.

***Dave[49]***

Response: Yes, you captured the intended meaning.

***Dave[50]***

Response: I’m sorry. I have added these labels.

***Dave[51]***

Response: Yes, all the changes you made are reasonable.

***Dave[52]***

Response: I referenced two new papers, one of them was published in 2014, and the other was published in 2016.

***Dave[53]***

Response: I have referenced one new paper.

***Dave[54]***

Response: Yes, we have something more to say. Because this paper was published in a not good journal, I have replaced it with another paper.

***Dave[55]***

Response: I have already made a more description of this paper.

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

Response: I have referenced some papers.

***Dave[58]***

Response: Yes, we do. And I have added some sentences to describe this method.

***Dave[59]***

Response: OK. I have referenced two papers.

***Dave[60]***

Response: OK. I have referenced one paper.

***Dave[61]***

Response: Yes, I have referenced five papers.

***Dave[62]***

Response: AT is based on the same observation as ART.

***Dave[63]***

Response: AT has been shown to perform better than RT and RPT in fault detection.

***Dave[64]***

Response: Yes, we can, and I have changed this sentence to another one which is more detailed. Li et al. proposed O-DRT, which has an objective function and a pre-defined parameter f that is criterion. During the testing process, the value of the objective function is calculated. If this value is greater than f, then test profile will be adjusted to a theoretically optimal one.

***Dave[65]***

Response: Yes, I can. During the test 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 in Formulas 1 and 2 should be bigger than the parameter in Formulas 3 and 4, that is, In [16], there are two parameters in DRT. However, there is just one parameter in our study. In [16], Lv et al. studied the relationship of parameters in DRT, and gave an equation. If the value of a parameter is available, then the value of the other parameter can be calculated from that equation.

***Dave[66]***

Response: Yes, you have a correct understanding.

***Dave[67]***

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

***Dave[68]***

Response: I replaced this word with “better”.

***Dave[69]***

Response: Prof. Sun

***Dave[70]***

Response: I used to have the same opinion as you. However, Beijing university of Aeronautics and Astronautics changed its English name to Beihang university in 2002.

***Dave[71]***

Response:

***Dave[72]***

Response: same as 70

***Dave[73]***

Response: I changed the first sentence.