A Systematic Literature Review of the Test Case Prioritization Technique for Sequence of Events

Johanna Ahmad  
Software Engineering and Information System Department  
Universiti Putra Malaysia, 43 400, Serdang, Selangor, Malaysia.

Salmi Baharom  
Software Engineering and Information System Department  
Universiti Putra Malaysia, 43 400, Serdang, Selangor, Malaysia.

Abstract
Software testing is a set of activities in Software Development Life Cycle (SDLC) carried out with the intent of finding errors before deliver to the customer or top management. Generally, the software testing phase consumes 40 to 70 percent of development effort, time and costs especially for large software systems. Numerous techniques have been proposed to prioritize test cases based on certain coverage criteria with the aim to reduce time and costs of testing. This paper presents the results of a systematic literature review (SLR) of relevant primary studies as evidence of the application of test case prioritization (TCP) in the area of sequence of events. From 115 papers, the researchers identified that a total of 50 papers report complete empirical results. The review reports on different TCP used to guide the search for coverage criteria, type and size of the test cases, techniques under comparison and validity threats of previous works. One of the most significant issues in TCP is how researchers handled the same priority value issues. Most of the papers either applied random technique or did not provide any information regarding the same priority value issues. In addition, 20 percent of research emphasised the effectiveness of their technique by comparing it with the random technique. The aim of this SLR is to identify the background and limitations of previous works, serving as a starting point for guidelines on how the TCP technique can be adopted.

Keywords: test case prioritization; sequence of events; systematic review;

INTRODUCTION
Many researchers agreed that software testing is one of the most critical phases in software development in terms of costs, efforts and time. Test case prioritization (TCP) is an important technique to schedule the order of test execution and may be adopted either in regression or non-regression testing. Before 1997, people generated test cases randomly and selected the best test cases based on those randomly generated without knowing the quality of the test cases. For a large commercial system, test suites can contain thousands of test cases and can sometimes be infinite. A variety of test cases can be formed through a combination of events and input parameters. For industrial collaborators, [1] reported that they needed to run 200,000 lines of codes for seven weeks. For this reason, numerous TCP techniques have been proposed to prioritize test cases especially for large commercial systems. The first TCP is known as a hybrid technique and combines modification, minimisation, and prioritisation-based selection [2]. Based on the literature, many researchers found techniques that can reduce effort, time and costs during testing. However, there are still a number of important issues neglected by previous researchers. This paper is structured as follows; Section 2 contains details on the systematic review process. In section 3, the extracted information is analyzed to answer the research questions. Section 4 contains discussion of the results.

REVIEW METHOD
The review processes for this SLR follows SLR guidelines for software engineering per [3],[4]. Referring to the guidelines, there are three main phases involved in this SLR: planning the review, conducting a review, and reporting the review, as shown in Fig. 1.

A. Research Questions
Research questions must be defined precisely, as they are the most important part of SLR [3]. The aim of this SLR is to understand and summarize the existing evidence on test case prioritization techniques. Besides that, the researcher attempts to understand how the existing test case prioritization technique caters to the same priority values.

![Figure 1. Systematic Literature Review (SLR) phases and stages](http://www.ripublication.com)

The research questions for this SLR comprise five components. These components known as PICOC and were proposed by [4]. Table 1 shows the criteria and scope of research questions.
To achieve the aim of this study, three primary questions and seven secondary questions have been defined as follows:

**RQ1:** How has test case prioritization been adapted and applied in prioritizing test cases in the form of sequence of events?

  **RQ1.1:** What techniques are used to prioritize test case?
  **RQ1.2:** Are test cases with the same priority values handled explicitly?

**RQ2:** How are the proposed techniques evaluated?

  **RQ2.1:** Are the set of test cases for the experimental study based on industry, benchmark or case study?
  **RQ2.2:** What is the number of test cases utilized for the experiments?
  **RQ2.3:** What are the techniques under comparison?
  **RQ2.4:** What are the evaluation metrics used to measure the performance of the proposed techniques?
  **RQ2.5:** What are the main threats to the validity in the domain of test case prioritization?

**RQ3:** Is there any test case prioritization technique conducted in the area of sequence of events?

### B. Data Sources

Ten online databases have been chosen as data sources: which are ACM Digital Library, Emerald, Elsevier, Google Scholar, IEEExplore Digital Library, ScienceDirect, Scopus, SpringerLink, Taylor & Francais and Wiley. The selection of online databases was based on the list available online database subscribed by the University Putra Malaysia’s Library under the “Computer Science” subject category.

### C. Search Strategy

Since the search process distinguishes systematic literature reviews from the traditional review, it is necessary to follow the search strategy. Trial searches with combinations of terms derived from the research questions need to be done before finalizing the search string. The following steps have been taken to formulate the search query for this SLR:

- Derivation of major terms from the research question.
- Identifying synonyms of major terms.
- Identifying keywords in relevant papers or books.
- Using Boolean OR for the alternatives synonyms or variants of each keyword.
- Using Boolean AND to link major terms.

The researcher attempted a number of search strings before arrive the conclusion that “software testing” as an expression is not suitable as a term, because many papers do not use these two words together even if related to software testing [5]. Nevertheless, the researcher would find too many unrelated papers if using the term "testing" alone. Based on that reason and number of trials, the researcher decided to use the term software and test linked together both terms with a Boolean AND. Using the terms "software" and "test", the researcher will find almost all papers related to software testing, and expression of "test case prioritization" as second major term [5]. The resulting search string is described as follows:

```
("Software" AND "Test") OR "Test Case Prioritization") OR
("Test Case Prioritization" AND "Sequence Based") OR
("Test Case Prioritization" AND "Search Based") OR ("Test Case Prioritization" AND "Event Based") OR ("Test Case Prioritization" AND "State Based")
```

The researcher performed a manual check of the results of the search string and realised that some of the online databases needed advanced settings such as IEEExplore and Scopus. The researcher needs to add alternative words and expressions to the search string.

### D. Study Selection

In this research, study selection involves selection of the online databases and identification of the search string, and applies inclusion and exclusion as identified in the early stage. The aim at this stage is to ensure the comprehensive of the selection of papers for this research.

### E. Inclusion and Exclusion Criteria

Inclusion and exclusion criteria for this SLR is based on research question [3]. Since this SLR is focused on the test case prioritization technique, in order to select only relevant papers, it is necessary to define inclusion and exclusion criteria. The inclusion criteria are as follows:

- All papers must be published in English.
- All papers must be published from 1 January 2005 to 25 November 2015.
- All papers must be focused on test case generation and test case prioritization.

Each of the papers is filtered by exclusion criteria before being accepted for the next stage. The exclusion criteria for this SLR were as follows:

- Papers not published in English.
- Duplicate study areas.
- Papers that only contains opinion pieces, viewpoints, progress research or incomplete results.
- Theses.
- Papers with less than three pages.
- Papers that do not report any empirical study in their paper.

<table>
<thead>
<tr>
<th>Criteria</th>
<th>Scope</th>
</tr>
</thead>
<tbody>
<tr>
<td>Population</td>
<td>Sequence Based, Event Based, Search Based, State Based</td>
</tr>
<tr>
<td>Intervention</td>
<td>Test case prioritization technique</td>
</tr>
<tr>
<td>Comparison</td>
<td>NA</td>
</tr>
<tr>
<td>Outcomes</td>
<td>Structure of TCP technique applied in Sequence Based, Event Based, Search Based, State Based</td>
</tr>
<tr>
<td>Context</td>
<td>Review(s) of any empirical studies of the test case prioritization</td>
</tr>
</tbody>
</table>

The researcher performed a manual check of the results of the search string and realised that some of the online databases needed advanced settings such as IEEExplore and Scopus. The researcher needs to add alternative words and expressions to the search string.

### D. Study Selection

In this research, study selection involves selection of the online databases and identification of the search string, and applies inclusion and exclusion as identified in the early stage. The aim at this stage is to ensure the comprehensive of the selection of papers for this research.

### E. Inclusion and Exclusion Criteria

Inclusion and exclusion criteria for this SLR is based on research question [3]. Since this SLR is focused on the test case prioritization technique, in order to select only relevant papers, it is necessary to define inclusion and exclusion criteria. The inclusion criteria are as follows:

- All papers must be published in English.
- All papers must be published from 1 January 2005 to 25 November 2015.
- All papers must be focused on test case generation and test case prioritization.

Each of the papers is filtered by exclusion criteria before being accepted for the next stage. The exclusion criteria for this SLR were as follows:

- Papers not published in English.
- Duplicate study areas.
- Papers that only contains opinion pieces, viewpoints, progress research or incomplete results.
- Theses.
- Papers with less than three pages.
- Papers that do not report any empirical study in their paper.
F. Data Extraction and Quality Assessment

The quality assessment for this study covered quantitative and qualitative studies since there is no restriction in terms of experimental design. A quality assessment study checklist is used to ensure the data extraction process meets the quality criteria [6]. The quality assessment checklist has been designed following guidelines proposed by [3]. Table II shows a list of general questions to measure the quality of selected studies. Three scales are coded for the quality assessment checklist and given a score; Yes = 1; Partially = 0.5; No = 0. Based on the item checklist, each article ranged from 0 (very poor) to 5 (very good).

<table>
<thead>
<tr>
<th>No</th>
<th>Item</th>
<th>Answer</th>
</tr>
</thead>
<tbody>
<tr>
<td>SQ 1</td>
<td>Are the aims and objectives of the research clearly stated?</td>
<td>Yes/No</td>
</tr>
<tr>
<td>SQ 2</td>
<td>Is the research design clearly specified?</td>
<td>Yes/No/Partially</td>
</tr>
<tr>
<td>SQ 3</td>
<td>Have the researcher(s) properly carried out the process of data collection?</td>
<td>Yes/No/Partially</td>
</tr>
<tr>
<td>SQ 4</td>
<td>Have the researcher(s) given enough data to support their results and conclusions?</td>
<td>Yes/No/Partially</td>
</tr>
<tr>
<td>SQ 5</td>
<td>Is there involved comparison of other technique in the experiment?</td>
<td>Yes/No</td>
</tr>
</tbody>
</table>

TABLE II. QUALITY ASSESSMENT CHECKLIST

FINDINGS

From the first stage of the search process, the researcher found 2,294 papers using the defined search terms. After screening titles and abstracts, only 115 were potentially relevant. Each of the papers was filtered for referring inclusion and exclusion criteria before being accepted for the synthesis of evidence. At this stage, irrelevant studies and duplicate studies were eliminated. The researcher read the full articles if the titles and abstracts were not sufficient to categorize the paper relevant or not to the research area. Finally, 50 studies were selected for providing answers to the formulated research questions. Fig. 2 shows results of the search and selection papers process.

A. Quality of Factors

Table III indicates the quality assessment scores for final identified papers consisting of 50 studies. 6 studies (12 percent) are fair, 9 studies (18 percent) are good and 35 studies (70 percent) are very good quality. None of the paper rated as poor quality. Therefore, all selected papers were included in the next phase for further analysis.

<table>
<thead>
<tr>
<th>Quality Scale</th>
<th>Very poor (&gt;=1)</th>
<th>Poor (&gt;=2)</th>
<th>Fair (&gt;=3)</th>
<th>Good (&gt;=4)</th>
<th>Very Good (&gt;=5)</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of studies</td>
<td>0</td>
<td>0</td>
<td>6</td>
<td>9</td>
<td>35</td>
<td>50</td>
</tr>
<tr>
<td>Percentage (%)</td>
<td>0</td>
<td>0</td>
<td>12</td>
<td>18</td>
<td>70</td>
<td>100</td>
</tr>
</tbody>
</table>

DISCUSSIONS

This section presents and discusses the results related to the research questions. A detailed description of the findings will be presented in this section with the research questions and its sub-questions.

RQ1: How has test case prioritization been adapted and applied in prioritizing test cases in the form of sequence of events?

The purpose of this research question is to investigate and understand test case prioritization in the current situation. Researchers have shown increased interest in test case prioritization techniques, since it is very important to reduce time and cost during the testing phase. However, only a few research works have used the test case prioritization techniques for populations such as sequence based, event based, search based and state based. The RQ1 is divided into two sub-questions.

RQ1.1: What are techniques to prioritize test cases?

From the SLR studies, a majority of the papers (18 percent) implemented code coverage to prioritize test cases. Some of the papers combined techniques such as combination of code coverage, requirement coverage, and execution cost. Researchers have stated that they combined more than one technique to maximize the number of discovered faults [7]. Previous research has ruled out the execution cost. There could be reasons why some researchers applied more than one technique [5]. From the literature, some researchers agreed that using more than one coverage criteria can break ties and solve same priority value issues [8]. Based on the SLR evaluations, 20 percent combined more than one technique. Table IV depicts the most popular and utilized techniques based on SLR evaluations.
In recent years, there has been increasing interest in test case prioritization. Some of the research can prioritize multiple test suites and test cases. Unfortunately, only 1 study reported that they handled same priority value cases. "Break ties" and "same weight value" term referred to same priority value issues. The study described in detail how their practical weight prioritization factors can solve same priority value issues. Four factors identified as the best factors to maximize the number of discovered faults: cost factors, time factors, defect factors and complex factors. Nevertheless, the subject program is not in the form of sequence of events. Most studies (98 percent) either applied random technique or did not provide any information regarding same priority value issues. Random methods can be categorized as poor weight algorithm and can caused wrong prioritization and bias issues [19]. This evidence means that there is a need to improve the ability of test case prioritization technique especially for the test cases in the form of sequence of events.

RQ2: How are the proposed techniques being evaluated?

In accessing the effectiveness and efficiency of existing technique, the reporting of the empirical study is very important. To avoid missing any information regarding the empirical study for each of the papers, RQ2 was further divided into five sub-questions as follows:

RQ2.1: Are the set of test cases for the experimental study based on industry, benchmark or case study?

Type of test cases used for the experiment is an important factor to validate the effectiveness of the proposed approach. In this SLR, the researchers defined three categories for type of test cases, industry-based, benchmark or case study. Most of the papers that used a case study agreed to apply real software system or industry for their future work [7],[19],[29],[30]. Of 50 papers, 25 reported that they used a benchmark as their test cases. 19 papers reported their test cases from a case study. The remaining 5 papers reported they applied real software system test cases for their experiment. Only one paper reported that they combined two industry projects and five case studies to determine that the proposed prioritization technique positively increased numbers of faults detection [20].

RQ2.2: Number of test cases utilized for the experiments.

**TABLE IV. MOST POPULAR AND UTILIZED TECHNIQUES**

<table>
<thead>
<tr>
<th>Techniques</th>
<th>Authors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Code Coverage</td>
<td>[9], [7], [10], [11], [12], [13], [14], [15], [16], [17]</td>
</tr>
<tr>
<td>Requirement Coverage</td>
<td>[18], [19], [20], [21], [22]</td>
</tr>
<tr>
<td>Fault Coverage</td>
<td>[23], [24], [25]</td>
</tr>
<tr>
<td>Interaction Coverage</td>
<td>[26], [27], [15], [28]</td>
</tr>
</tbody>
</table>

RQ1.2: Are test cases with same priority values handled explicitly?

Test case prioritization is considered an emerging research area. There are already wide sets of data and case studies that have been developed. Some of them have been published publicly for the use of others. With regards to the size of data sets, only 34 percent coming for large data sets in this review. The large data sets have up to 300,000 lines of codes and 200,000 test cases. 18 percent comes from medium data sets with up to 14,437 lines of codes and 5,000 test cases. A majority of the data sets at 48 percent come from small data sets. Small data sets have less than 1,000 lines of codes and fewer than 100 test cases. Most of the papers found it easy to test and manage if they applied small test sizes in the experiment. Small test sizes, means less effort, time, and cost for the experiment [19].

RQ2.3: Techniques under comparison.

Many techniques have been developed by numerous researchers since 1997. In order to validate the effectiveness of their TCP technique, 20 percent of researches compared their technique with random technique. The experimental results shows that their TCP technique is more effective than a random technique. However, existing TCP technique failed to resolves some issues which are considered as limitations to their research. The limitations will be used as a basis to improve the TCP technique. In summary, Table V depicts major limitations of existing TCP techniques.

**TABLE V. MAJOR TCP TECHNIQUES LIMITATIONS**

<table>
<thead>
<tr>
<th>Limitations</th>
<th>Authors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Failed to prioritize multiple suites</td>
<td>[23], [21]</td>
</tr>
<tr>
<td>Failed to prioritize test case with same priority value</td>
<td>[7], [23], [29], [31], [32], [33], [30], [34], [12], [35], [13], [26], [20], [36], [8], [37], [38], [24], [39], [40], [41], [42], [25], [17], [43], [44], [45], [46], [47], [28], [48], [49], [50], [51], [52], [53]</td>
</tr>
<tr>
<td>Ignore the practical weight factors</td>
<td>[23], [33]</td>
</tr>
<tr>
<td>Test cases from small data sets</td>
<td>[7], [29], [18], [9], [54], [55], [27], [28], [21], [40], [56], [16], [57], [45], [47], [50], [22], [51]</td>
</tr>
</tbody>
</table>

**TABLE VI. TECHNIQUES UNDER COMPARISON**

<table>
<thead>
<tr>
<th>Techniques</th>
<th>Authors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Code Coverage</td>
<td>[9], [7], [10], [11], [12], [13], [14], [15], [16], [17]</td>
</tr>
<tr>
<td>Requirement Coverage</td>
<td>[18], [19], [20], [21], [22]</td>
</tr>
<tr>
<td>Fault Coverage</td>
<td>[23], [24], [25]</td>
</tr>
<tr>
<td>Interaction Coverage</td>
<td>[26], [27], [15], [28]</td>
</tr>
</tbody>
</table>

RQ2.4: What is the evaluation performance of the proposed techniques?

Numerous evaluation metrics have been proposed in the selected primary studies that employ test case prioritization technique. The most frequently employed evaluation metric is Average of the Percentage of Faults Detected (APFD) with 58 percent. Average Percentage Statement Coverage (APSC) and Average Percentage of Faults Detected per Cost (APFDC), both were counted for 8 percent. Furthermore, 6 percent were counted for the usage of Normalize Percentage of Faults Detected (NAPFD). The remaining evaluation metrics reached as 2 percent, as shown in Table VI. In terms of the number of evaluation metric techniques employed, 12 primary studies used more than one evaluation metric.
RQ2.5: What are the main threats to the validity in the domain of test case prioritization?

In order to answer this research question, the researcher looked at the validity in terms of the effectiveness of the measurement [5]. The most frequently observed threats were found when researchers modified the codes and performed code inspection and testing to reduce failure bugs [36]. Besides that, most of the papers stated that they applied APFD only as a metric; therefore, they may get other results if other metrics are used, as APFD is not the only metric available [8], [36], [33], [48]. Table VII shows the most frequently threats researchers cater in their research. Out of 50 papers, 20 papers do not mention how they catered to threats in their experiments.

### TABLE VII. VALIDITY THREATS

<table>
<thead>
<tr>
<th>Threats</th>
<th>Number of papers</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Internal</td>
<td>29</td>
<td>58%</td>
</tr>
<tr>
<td>External</td>
<td>26</td>
<td>52%</td>
</tr>
<tr>
<td>Construct</td>
<td>23</td>
<td>46%</td>
</tr>
<tr>
<td>Conclusion</td>
<td>2</td>
<td>4%</td>
</tr>
<tr>
<td>NA</td>
<td>20</td>
<td>40%</td>
</tr>
</tbody>
</table>

RQ3: Is there any test case prioritization technique conducted in the area of sequence of events?

There has been a lot of research done into TCP since its proposal in 1997 [19]. Based on the SLR conducted, only 36 percent applied sequence of event as their subject programs in their TCP technique. However, none of the papers catered to the same priority value issue in their experiments. As mentioned in RQ 1.2, various TCP techniques applied random techniques to solve same priority value issues, even though most of the researchers agree that the random technique is ineffective.

### CONCLUSIONS

The aim of this SLR was to identify the status of research on TCP in the area of sequence of events. Solving the same priority value issue is one of the concerns of the research into this SLR. The analysis in this SLR is based on the constructed research question. A total of 50 primary studies were selected and analyzed. Results of this SLR show that many papers have still not reported their empirical results completely. However, the results have shown that TCP research continues to increase annually as the TCP technique is increasingly important [14]. The essence of this SLR may be used to identify areas for possible improvement in TCP technique for sequences of events.

### ACKNOWLEDGMENT

The authors appreciate the efforts of the community of Software Engineering Universiti Putra Malaysia (UPM) for supporting this research.

### REFERENCES


