Critique of Coverage Correlated with Test Suite Effectiveness Paper for CS5371 Soft Test for Mobile & Emb Sys

Jeremy Solmonson
School of Security Engineering
University of Colorado at Colorado Springs
Colorado Springs, CO 80922
Email: jsolmons@uccs.edu

***Abstract*—This is paper is a critique of "Coverage Is Not Strongly Correlated with Test Suite Effectiveness." In accordance with homework 1 requirements the critique will be based on: suggestion for acceptance, summary of paper, evaluation of paper, positive points, negatives points, and potential future work.**

I. Suggestion for Acceptance

I would accept this paper with revisions. The author needs to explain the decision process for threats to validity in their experiment and further elaborate on why the chosen techniques were used (aka. Kendall t)

II. Summary of Paper

The current standard for ensuring quality within the testing process is to verify a high degree of code coverage is examined. This is based on the assumption that testing more code will lead to a higher degree of fault removals and increased code quality. This paper provides evidence showing that is not the case. Instead, increasing the number of test method, or ignoring the test size, within a test suite is more effective. As a result, test implementers and test policy writers should focus on alternative strategies for testing instead of code coverage.

III. Evaluation

The idea of determining the most effective method for testing, or eliminating methods that are ineffective for testing, is desirable for the testing community. Testing policies and testing standards are driving by effective practices. The authors are on an interesting path for their work.

The main problem is ensuring the authors conclusions are accurate. This goes against the status quo that a higher degree of code coverage ensures a higher degree of code quality. As a result, this critique will primarily focus around the methods of testing and measuring.

Due to the limited number test programs and the implementation of random generated tests, the sample size is insufficient to correlate the results. The experiment doesn't seem to have been run multiple times to reduce the degree of error. Further, the results were most mixed based on the selected program. As a result, another threat to validity, that the authors could control, is to reduce the outliers by increasing sample size or execute multiple random test suites.

The authors' outlined testing methods of related research, and provided their own testing method, but never compared or contrasted their testing method with the methods used by others. Where did they differ? Where where they similar? I think the purpose was to demonstrate their similarities, but I can only infer rather than know the author's intent. Recommend a brief section on compare/contrast.

The author's choose to correlate the results with the Kendall t. There are reasons the author's choose this method but little explanation was given. Elaborate on why this was the best method to correlate the data.

While it is necessary to clarify terminology, this was provided half way into page 3 when the terms were already used multiple times. Either remove the sub-section or provide sooner in paper.

Is the hypothetical graph really necessary? Wasted space?

IV. Positive Points

+ Interesting topic with valuable insights. Provides evidence that contradicts the status quo of current testing methods.

+ Provides a summary of relevant research and elegantly injects additional information from "real-world" programs.

+ Goes above the current standard of testing by using large programs and very large test suites.

V. Negative Points

- Multiple threats against validity - while the author addressed the potential problems with the research, few solutions were provided to remove these threats in future research.

VI. Potential Future Work

One area that needs to be explored to validate the authors research is a larger sample of test programs. While the authors selected these programs for the unique attributes (open source, high degree of KLOC, 1000+ unit tests) the five programs do not provide a comprehensive sample size. One test (HSQLDB) performing as an outliers accounts for 20% of their accuracy rate. Additional tests should be performed to better understand the efficiency rate of the authors contributions.