A Replication Study on Test Case Failure Prediction in the Context of Test Case Prioritization

Francis Palma
Data Science Lab,
Department of Mechanical and Industrial Engineering,
Ryerson University
Toronto, ON, Canada
francis.palma@ryerson.ca

Tamer Abdou
Data Science Lab,
Department of Mechanical and Industrial Engineering,
Ryerson University
Toronto, ON, Canada
tamer.abdou@ryerson.ca

Ayse Bener
Data Science Lab,
Department of Mechanical and Industrial Engineering,
Ryerson University
Toronto, ON, Canada
larst@affiliation.org

John Maidens
Data Science Lab,
Department of Mechanical and Industrial Engineering,
Ryerson University
Toronto, ON, Canada
maidens@ryerson.ca

Jimmy Lo
IBM Canada Lab
Toronto, ON, Canada
jimmylo@ca.ibm.com

Stella Liu
IBM Canada Lab
Toronto, ON, Canada
Stella.Liu@ibm.com

ABSTRACT

This paper provides a sample of a LaTeX document which conforms, somewhat loosely, to the formatting guidelines for ACM SIG Proceedings.[1]

KEYWORDS

ACM proceedings, LaTeX, text tagging

ACM Reference Format:
Francis Palma, Tamer Abdou, Ayse Bener, John Maidens, Jimmy Lo, and Stella Liu. 2018. A Replication Study on Test Case Failure Prediction in the Context of Test Case Prioritization. In *Proceedings of* . ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

A few examples of in-text equations in context. Notice how this equation: $\lim_{n\to\infty} x = 0$, set here in in-line math style, looks slightly different when set in display style. (See next section).

First, consider the equation, shown as an inline equation above:

$$\lim_{n\to\infty} x = 0 \tag{1}$$

Notice how it is formatted somewhat differently in the **displaymath** environment. Now, we'll enter an unnumbered equation:

$$\sum_{i=0}^{\infty} x + 1$$

[1]This is an abstract footnote

, ,

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

Table 1: Frequency of Special Characters

Non-English or Math	Frequency	Comments
Ø	1 in 1,000	For Swedish names
π	1 in 5	Common in math
\$	4 in 5	Used in business
Ψ_1^2	1 in 40,000	Unexplained usage

and follow it with another numbered equation:

$$\sum_{i=0}^{\infty} x_i = \int_0^{\pi+2} f \tag{2}$$

just to demonstrate LaTeX's able handling of numbering.

ACKNOWLEDGMENTS

The authors would like to thank Dr. Yuhua Li for providing the MATLAB code of the *BEPS* method.

The authors would also like to thank the anonymous referees for their valuable comments and helpful suggestions. The work is supported by the National Natural Science Foundation of China under Grant No.: 61273304 and Young Scientists' Support Program (http://www.nnsf.cn/youngscientists).

2 INFORMATION ABOUT THE ORIGINAL STUDY

3 INFORMATION ABOUT THE REPLICATION

3.1 Study Design

3.1.1 Data Collection and Processing.

3.1.2 Variable Selection.

3.1.3 Analysis Method. We apply the Wilcoxon rank sum and Kruskal-Wallis tests [1] to compare two distributions of the results to see if one outperforms the other using a 95% confidence level (*i.e.*,

Table 2: Some Typical Commands

Command	A Number	Comments
`\author`	100	Author
`\table`	300	For tables
`\table*`	400	For wider tables

p-value<0.05). For any comparison exhibiting a statistically significant difference, we further compute the Vargha and DelaneyâĂŹs Â12 statistics to quantify the importance of the difference.

3.1.4 Replication Package. All the data used in this replication study are publicly available on github[2]. In particular, we provide: (i) the original dataset, (ii) the predicted rank data based on the original models, (iii) the predicted rank data based on the replicated and modified models, and (iv) the R sources and all the generated figures.

4 COMPARISON OF REPLICATION RESULTS TO ORIGINAL RESULTS

Table 3: Wilcoxon Rank Sum Test for Replicated Traditional Metrics-based and Modified Traditional Metrics-based Models.

Projects	Treatment Groups	p-value	VD.A
Closure Compiler	Traditional2 Traditional	0.7167	0.4847091 (negligible)
Commons Lang	Traditional2 Traditional	0.6197	0.47375 (negligible)
Commons Math	Traditional2 Traditional	0.9552	0.5037858 (negligible)
JFreeChart	Traditional2 Traditional	0.6271	0.4583333 (negligible)
Joda Time	Traditional2 Traditional	1	0.5008 (negligible)

Table 4: Wilcoxon Rank Sum Test for Replicated Similarity Metrics-based and Replicated Traditional Metrics-based Models.

Projects	Treatment Groups	p-value	VD.A
Closure Compiler	Similarity Traditional	0.3468	0.5395568 (negligible)
Commons Lang	Similarity Traditional	0.8283	0.4884722 (negligible)
Commons Math	Similarity Traditional	0.8121	0.4848567 (negligible)
JFreeChart	Similarity Traditional	0.469	0.4383681 (negligible)
Joda Time	Similarity Traditional	0.7468	0.4728 (negligible)

Table 5: Wilcoxon Rank Sum Test for Replicated Similarity Metrics-based and Modified Similarity Metrics-based Models.

Projects	Treatment Groups	p-value	VD.A
Closure Compiler	Similarity2 Similarity	0.803	0.4894737 (negligible)
Commons Lang	Similarity2 Similarity	0.8967	0.4930556 (negligible)
Commons Math	Similarity2 Similarity	0.8934	0.5086533 (negligible)
JFreeChart	Similarity2 Similarity	0.8931	0.4878472 (negligible)
Joda Time	Similarity2 Similarity	0.9221	0.5088 (negligible)

[2]https://github.com/franpalma/pred-rep/

Table 6: Wilcoxon Rank Sum Test for Modified Similarity Metrics-based and Replicated Similarity Metrics-based Models.

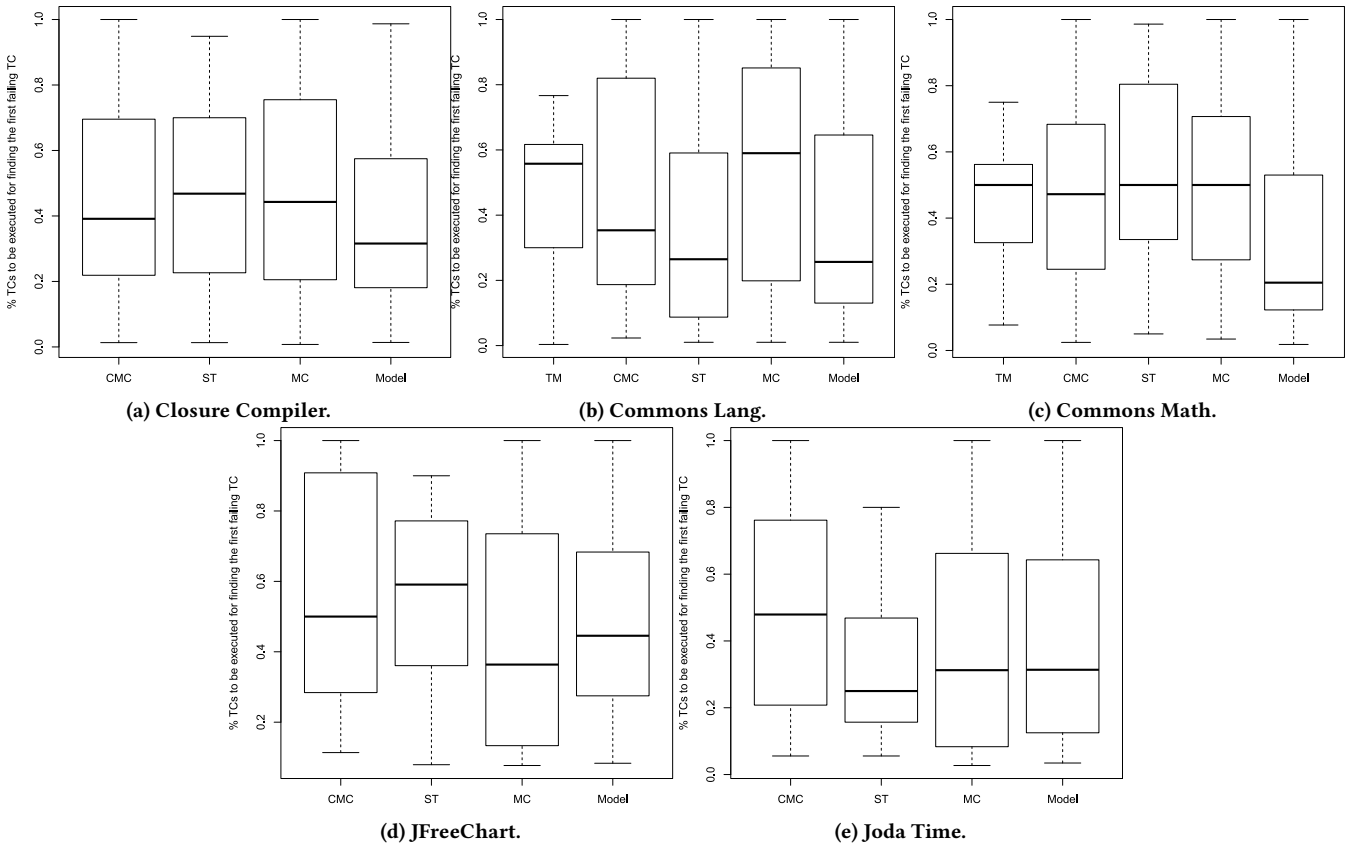
Projects	Treatment Groups	p-value	VD.A
Closure Compiler	Similarity2 Traditional	0.506	0.5279778 (negligible)
Commons Lang	Similarity2 Traditional	0.7389	0.4823611 (negligible)
Commons Math	Similarity2 Traditional	0.9277	0.4940508 (negligible)
JFreeChart	Similarity2 Traditional	0.2871	0.4097222 (small)
Joda Time	Similarity2 Traditional	0.822	0.4808 (negligible)

Table 7: Wilcoxon Rank Sum Test for Replicated Similarity Metrics-based, Modified Similarity Metrics-based, and Replicated Traditional Metrics-based Models.

Projects	Treatment Groups	p-value
Closure Compiler	Similarity2 Similarity Traditional	0.963
Commons Lang	Similarity2 Similarity Traditional	0.942
Commons Math	Similarity2 Similarity Traditional	0.9714
JFreeChart	Similarity2 Similarity Traditional	0.5647
Joda Time	Similarity2 Similarity Traditional	0.9423

REFERENCES

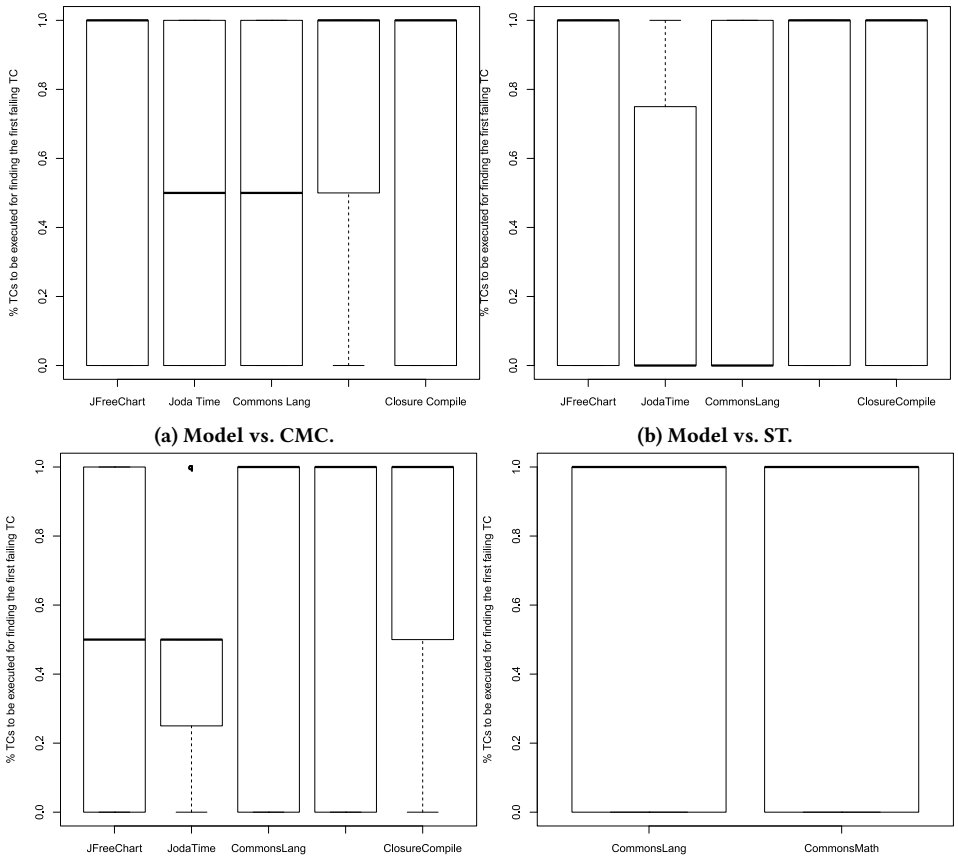
[1] David J Sheskin. 2003. *Handbook of Parametric and Non-parametric Statistical Procedures.* crc Press.



(a) Closure Compiler. **(b) Commons Lang.** **(c) Commons Math.**

(d) JFreeChart. **(e) Joda Time.**

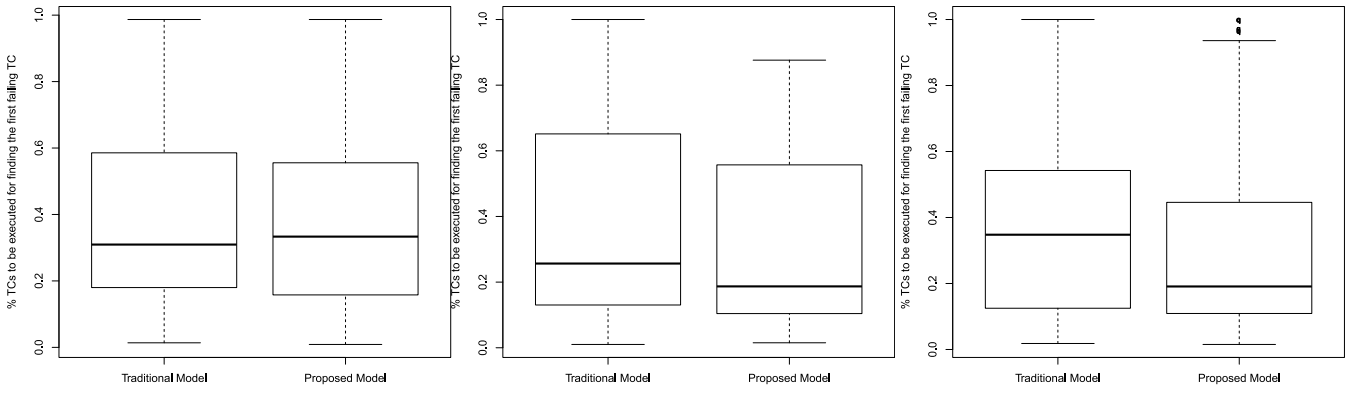
Figure 1: The Ranks of the First Failing Test Cases using TM, CMC, MC, ST, and the Replicated Regression Model for each Version of the Project.



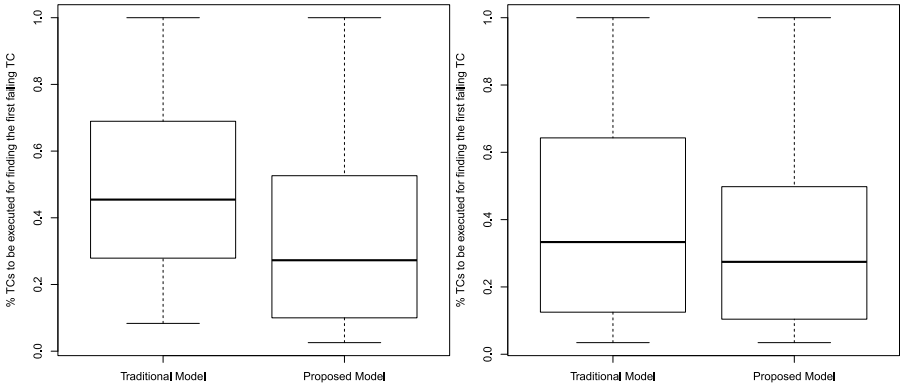
(a) Model vs. CMC. **(b) Model vs. ST.**

(c) Model vs. MC. **(d) Model vs. TM.**

Figure 2: The bBxplots of the Effect Size Measures for Finding the First Fault using CMC, MC, ST, TM and the Regression Model for each Version of the Project.



(a) Closure Compiler. **(b) Commons Lang.** **(c) Commons Math.**



(d) JFreeChart. **(e) Joda Time.**

Figure 3: The Ranks of the First Failing Test Cases using TM, CMC, MC, ST, and the Replicated Regression Model for each Version of the Project.

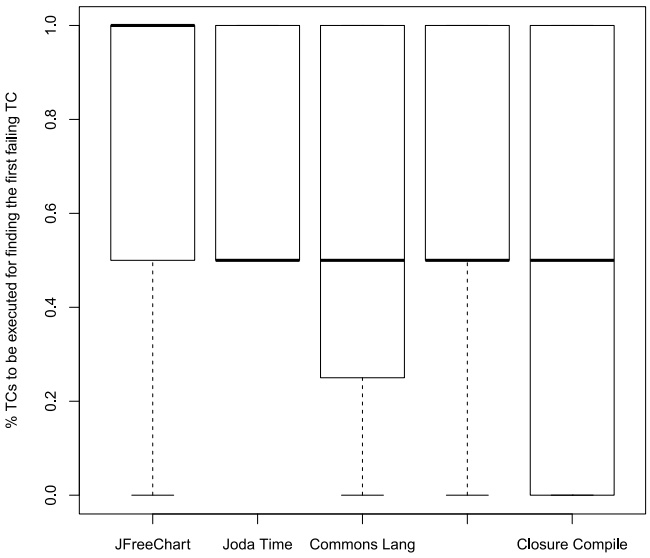
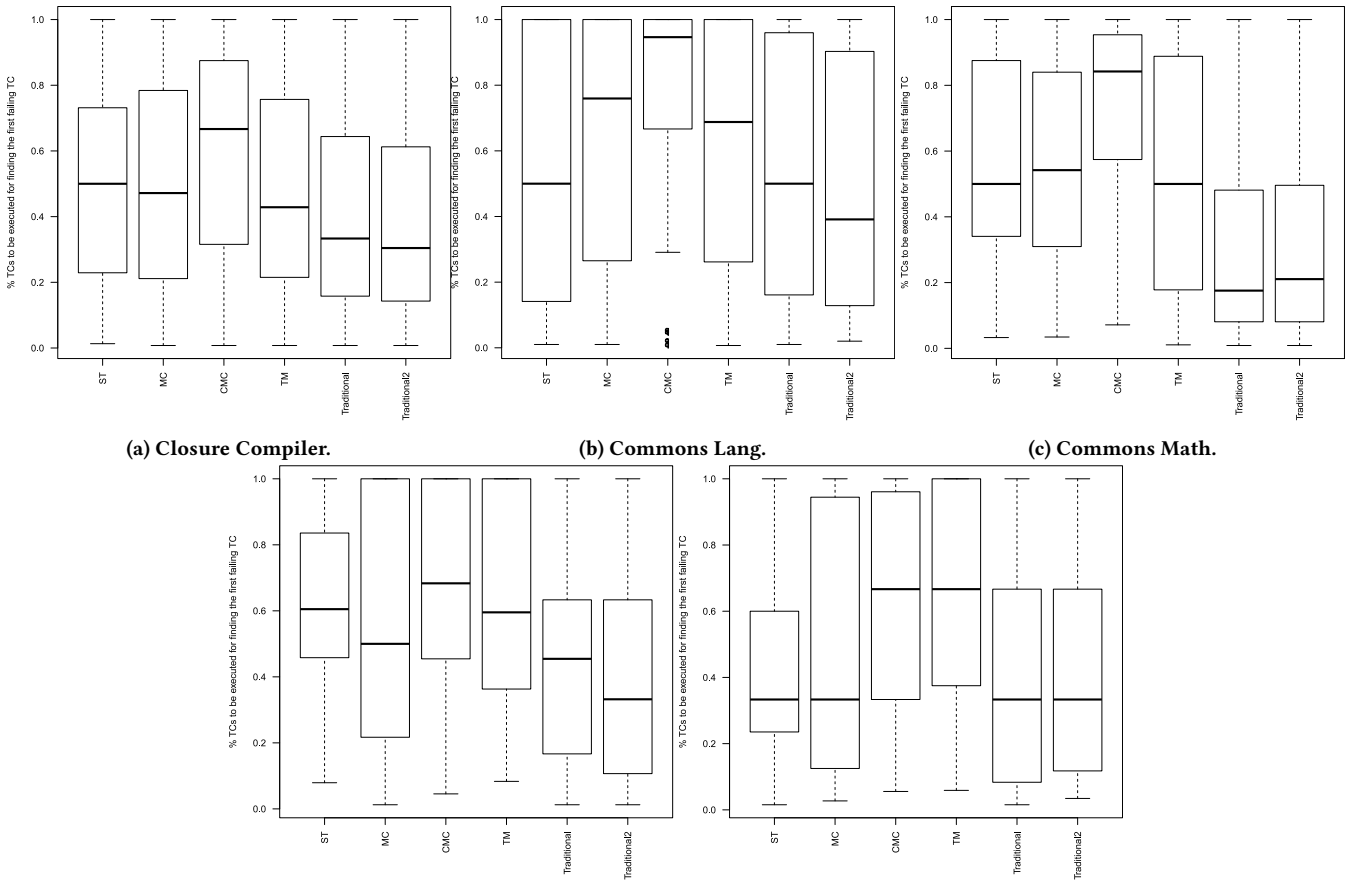


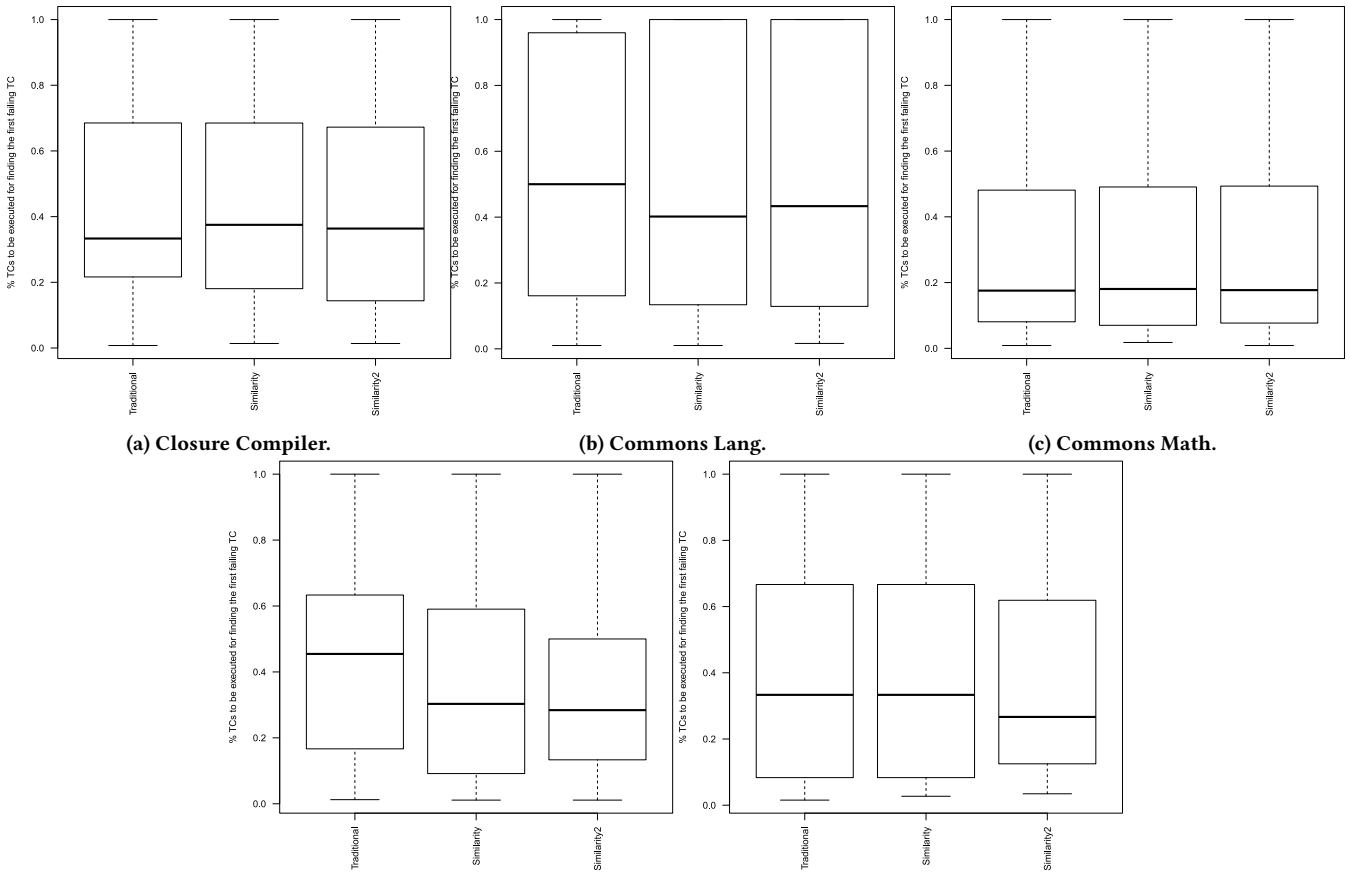
Figure 4: The Boxplots of the Effect Size Measures for Finding the First Fault using the Two Prediction Models for each Version of each Project.



(a) Closure Compiler. **(b) Commons Lang.** **(c) Commons Math.**

(d) JFreeChart. **(e) Joda Time.**

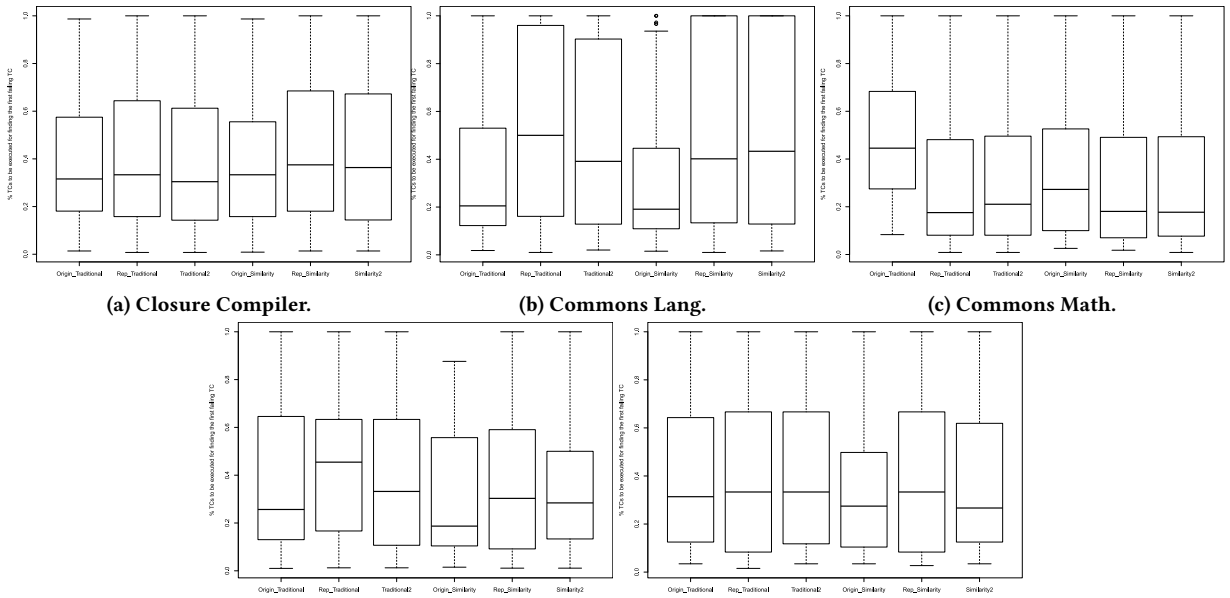
Figure 5: The Comparison among the Ranks of the First Failing Test Cases using TM Model, CMC Model, MC Model, ST Model, the Replicated Traditional Metrics-based Regression Model, and the Modified Traditional Metrics-based Regression Model for each Version of the Project.



(a) Closure Compiler. (b) Commons Lang. (c) Commons Math.

(d) JFreeChart. (e) Joda Time.

Figure 6: The Comparison among the Ranks of the First Failing Test Cases using Traditional Metrics Model, Similarity Metrics Model, and Modified Similarity Metric Model for each Version of the Project.



(a) Closure Compiler. **(b) Commons Lang.** **(c) Commons Math.**

(d) JFreeChart. **(e) Joda Time.**

Figure 7: The Comparison of Ranks of the First Failing Test Cases using Original Traditional Metrics-based Model, Replicated Traditional Metrics-based Model, Modified Traditional Metrics-based Model, Original Similarity Metrics-based Model, Replicated Similarity Metrics-based Model, Modified Similarity Metrics-based Model for each Version of the Project.