

# Supplement to: Developing a Single Model and Test Prioritization Strategies for Event-Driven Software

Renée C Bryce, Utah State University  
 Sreedevi Sampath, University of Maryland, Baltimore County  
 Atif M Memon, University of Maryland, College Park



## 1 OVERVIEW

This supplement accompanies our paper titled, “Developing a Single Model and Test Prioritization Strategies for Event-Driven Software” [1]. We provide additional data and discussion on the APFD in both graphic and tabular forms.

## 2 DETAILED RESULTS

This section presents detailed results for our study for RQ1 using the APFD metric. For each application, we divide the prioritization criteria into two sets (static and usage event-based) and present them in two graphs to better show the trends. The graphs are shown in log-scale. The controls (*G-Best*, *Random*, and *G-Worst*) appear in each graph. We also present the APFD values for the different prioritization criteria in tabular form. Due to space constraints, the tables report the APFD after every 10% increment of the test suite execution. In each table, we highlight the *best* APFD values for each increment with a bold font, along with the results for the *G-Best* control.

**TerpCalc:** Figure 1 reports the APFD for the 10 prioritization criteria and 3 control criteria. This data is also shown in Table 10 in [1].

**TerpPaint:** Figure 2 and Table 1 show the results of our 10 prioritization criteria in relation to our controls of *Random*, *G-Best*, and *G-Worst*. In this study, *PV-LtoS* achieves the best rate of fault detection, followed by *2-way*. The *Weighted-Freq* and *MFPS* prioritization criteria are the 3<sup>rd</sup> and 4<sup>th</sup> best, with *Weighted-Freq* providing

the best APFD in the first 20% of test execution and *MFPS* providing a better APFD in the latter 80% of test execution. The *1-way* prioritization criterion is the 5<sup>th</sup> most effective, followed by *APS* and *Action-StoL*. However, *Action-StoL* has a slow start and gains competitiveness only after half of the test suite execution. The *UniqWin*, *Random*, and *Action-LtoS* are considerably less effective than the top criteria. Again, our controls show that none of the prioritization criteria perform better than the *G-Best*, nor do they perform worse than the *G-Worst*.

**TerpSpreadsheet:** Figure 3 and Table 2 show that the prioritization criteria applied to the *TerpSpreadsheet* test suite produce more variation in APFD among the prioritization criteria than our previous two results for *TerpCalc* and *TerpPaint*. The APFDs of the 10 prioritization criteria after all tests are executed range between 38.66% and 95.35%. Prioritization by *1-way* achieves the best APFD in the first 20% of the test suite and *UniqWin* produces the best APFD for the latter 80% of the test suite. Both *1-way* and *UniqWin* have APFDs that are within 1% of each other throughout the test suite execution. The second group of best prioritization criteria include *2-way* and *PV-LtoS* which are consistently within 2% APFD of each other. Prioritization by *APS* and *MFPS* have the 5<sup>th</sup> best APFD, each slightly outperforming each other at different times in the test execution. These followed in effectiveness by *Weighted-Freq*. *Random* performs worse than all of the criteria above, but three criteria are worse, including *Action-StoL*, *Action-LtoS*, and *PV-StoL*. None of the prioritization criteria are better than *G-Best*, nor are any worse than *G-Worst*.

**TerpWord:** Figure 4 and Table 3 show that the *TerpWord* test suite has the most effective rate of fault detection after all tests execute when we prioritize by *2-way*, with the exception of *MFPS* that performs better in the first 10% of test execution. The *MFPS* is the second best technique, followed by *PV-LtoS*. The *Action-LtoS*, *1-way*, *APS*, and *Weighted-Freq* are in the third tier of best prioritization criteria, all performing better than *Random*.

- R. Bryce is with the Department of Computer Science, Utah State University, Logan, UT 84322.  
E-mail: Renee.Bryce@usu.edu
- S. Sampath is with the Department of Information Systems, University of Maryland, Baltimore County, Baltimore, MD 21250.  
E-mail: atif@cs.umd.edu
- A M Memon is with the Department of Computer Science, University of Maryland, College Park, MD 20742.  
E-mail: atif@cs.umd.edu

*UniqWin* is comparable to *Random* as both maintain an APFD within 2% of each other. *Action-StoL* has worse APFD than *Random*. None of the prioritization criteria perform better than *G-Best*, nor do any perform worse than *G-Worst*.

**CPM:** Figure 5 and Table 4 show the results for CPM. Prioritization by *1-way* and *2-way* produce the best APFD. *1-way* has the best APFD during the first 10% of test suite execution and *2-way* has the best APFD for the latter 90% of the test suite. *PV-StoL* produces the worst APFD and *Action-StoL* has the second worst APFD, both producing results worse than *Random*. *MFPS*, *APS*, *PV-LtoS*, *Weighted-Freq*, and *Action-LtoS* fall in between the best and worst prioritization criteria. None of our prioritization criteria have a better APFD than *G-Best*, nor do any have a worse APFD than *G-Worst*.

**Masplas:** Figure 6 and Table 5 show the results for Masplas. The FDD (Fault Detection Density) of this test suite is quite low, so we observe a more dramatic difference in the APFD that results from the different prioritization criteria. In addition, a close look at the fault matrix reveals that a small number of test cases find the majority of faults.

The *Weighted-Freq* approach is by far the best prioritization technique and chooses the first 15 tests such that they rapidly identify all of the faults. All of the faults are found before even 10% of the test suite is executed. *Action-LtoS* has the second best APFD, followed by *2-way*, *APS*, *1-way*, and *PV-LtoS*. All five of these criteria produce an APFD within 2% of the best technique throughout the entire test selection process. Three prioritization criteria perform slightly worse than *Random*, including *UniqWin*, and *Action-StoL*. *PV-StoL* produces the worst APFD and is only slightly better than *G-Worst*. This is our first result in which a prioritization criteria, namely *Weighted-Freq*, produces an APFD that is better than the APFD from the greedy *G-Best* algorithm. Again, none of our criteria is worse than *G-Worst*.

**Book:** Figure 7 (corresponding to Table 11 in [1]) presents the results from Book.

Next, we present the results of evaluating the effectiveness of the prioritization criteria using our second metric, i.e., number of test cases executed to detect all the faults, for all our subject applications.

In [1], Table 13 shows the number of tests that are executed to locate 100% of the faults for the 7 applications, the 10 prioritization criteria, and the controls. We report the results by grouping the applications that have the same best prioritization techniques.

**TerpCalc and MASPLAS:** In [1], Table 13 shows the number of tests executed before 100% fault detection. The prioritization technique of *Weighted-Freq* finds 100% of the faults soonest after 217 test cases for TerpCalc and after 15 test cases for MASPLAS. For both applications, this is faster than the *G-Best*. As described earlier in section 5.5 of [1], this is due to the greedy implementation of the *G-Best* algorithm that selects one test at a time where each "next test" is selected to cover the

maximum number of faults in relation to the previously selected tests. Aside from *Weighted-Freq*, we observe that the APFD results for *G-Best* are better than any of the other prioritization criteria for both applications

**TerpPaint and Book:** In [1], Table 13 shows that *Action-StoL* finds 100% of the faults sooner than the other prioritization criteria, using 151 test cases for TerpPaint and 65 for Book. Prioritization by *PV-StoL* is the worst case, using 299 test cases to find all faults for TerpPaint. Prioritization by *APS* and *Random* are the worst for Book, using 124 test cases to find all faults. The other criteria fall between these two extremes. We observe room for improvement in comparison to the *G-Best* for the best case, but even our worst prioritization technique is still slightly better than or equal to the *G-Worst*.

**TerpSpreadsheet and CPM:** In [1], Table 13 shows that *2-way* find 100% of all faults in the fewest test cases or TerpSpreadsheet and CPM. The second best prioritization techniques for these applications are also static: *PV-StoL* is the second best technique for TerpSpreadsheet, while *Action-LtoS* is the second best for CPM. None of our prioritization criteria for TerpSpreadsheet find 100% of all faults in fewer test cases than *G-Best*, however all 10 criteria find 100% of faults in fewer tests than *Random* and *G-Worst*. The *2-way* and *PV-LtoS* techniques use fewer test cases than *G-Best* for CPM, but the other 8 prioritization techniques are significantly worse. The *G-Worst* ordering for both Spreadsheet and CPM uses the entire test suite to find 100% of the faults.

**TerpWord:** In [1], Table 13 shows that *1-way* is better than the other criteria in finding 100% of the faults, using 94 out of 300 test cases. We consider *1-way* as quite effective as this technique results in finding 100% of the faults before the greedy *G-Best* algorithm. *Action-StoL* is the third best, using 100 test cases. The other 7 prioritization criteria find all of the faults in fewer than the *Random* and *G-Worst* with the exception that *APS* uses the same number of tests as *Random* and *MFPS* uses 10 additional tests beyond *Random*.

## REFERENCES

- [1] R. Bryce, S. Sampath, and A. M. Memon, "Developing a Single Model and Test Prioritization Strategies for Event-Driven Software," *IEEE Trans. Softw. Eng.*, to appear.

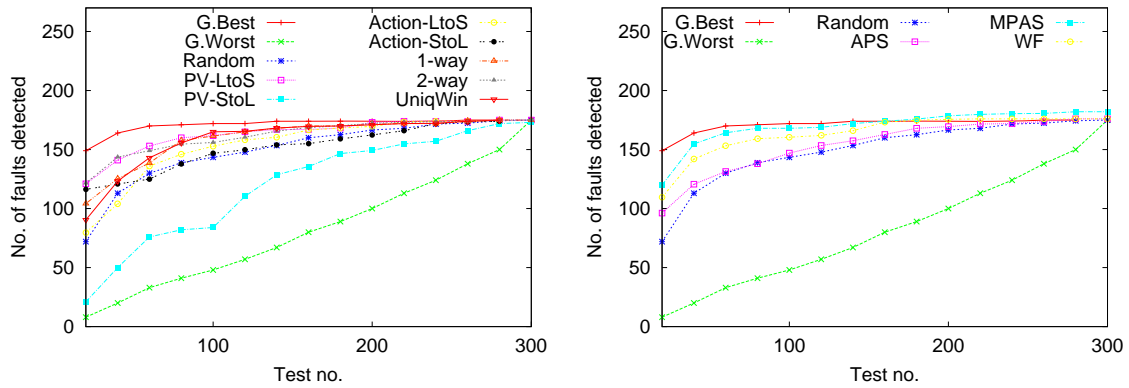


Fig. 1: Terpcalc: rate of fault detection using prioritized test orderings.

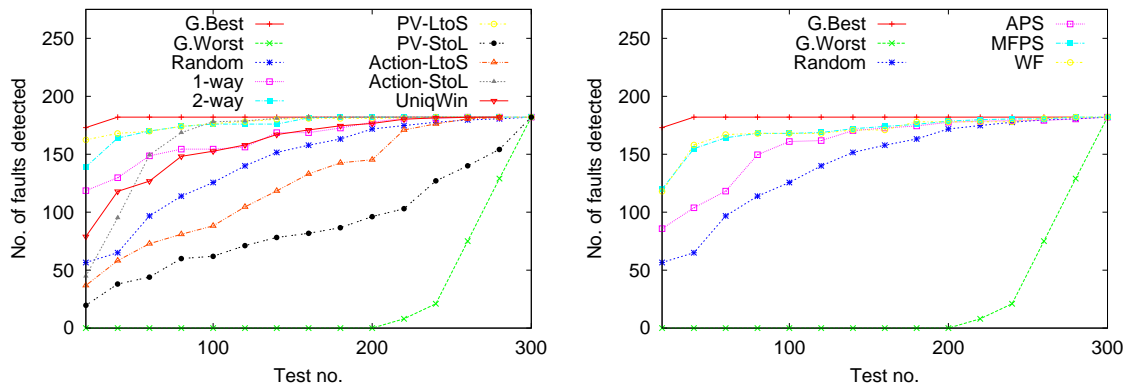


Fig. 2: Terppaint: rate of fault detection using prioritized test orderings.

	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
1-way	68.56	77.49	79.89	80.54	84.24	85.25	86.41	86.83	86.92	86.92
2-way	83.87	90.05	92.19	92.56	93.39	94.18	94.18	94.18	94.18	94.18
PV-LtoS	<b>90.62</b>	<b>91.43</b>	<b>93.17</b>	<b>94.67</b>	<b>95.54</b>	<b>95.54</b>	<b>95.54</b>	<b>95.69</b>	<b>95.69</b>	<b>95.69</b>
PV-StoL	14.19	22.46	29.82	33.13	35.47	37.45	40.28	43.32	44.48	45.54
Action-LtoS	24.48	37.01	42	48.8	54.99	59.37	62.22	64.85	65.31	65.35
Action-StoL	34.42	73.74	84.96	85.76	86.64	86.75	86.75	86.75	86.75	86.8
UniqWin	57.25	65.98	75.79	78.61	82.04	83.29	83.97	84.43	84.48	84.48
APS	51.41	61.36	79.14	79.78	82.94	83.34	84.26	84.38	84.43	84.48
MFPS	79.35	86.97	88.25	88.59	90.08	90.68	91.27	91.37	91.48	91.82
Weighted-Freq	79.83	87.14	87.57	87.57	88.53	90.24	90.24	90.24	90.56	90.56
Random	31.64	48.84	59.52	65.61	69.8	72.31	74.32	74.89	75.13	75.18
G-Best	<b>95.18</b>	<b>98.60</b>	<b>98.60</b>	<b>98.60</b>	<b>98.60</b>	<b>98.60</b>	<b>98.60</b>	<b>98.60</b>	<b>98.60</b>	<b>98.60</b>
G-Worst	0.17	0.17	0.17	0.17	0.17	0.17	0.51	2.88	9.17	11.17

TABLE 1: APFD for Terppaint (each increment of 10% of the test suite is 30 tests).

	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
1-way	<b>92.43</b>	<b>92.43</b>	92.43	92.43	92.43	94.19	94.19	94.19	94.37	94.37
2-way	77.81	93.18	94.79	94.96	94.96	94.96	94.96	94.96	94.96	95
PV-LtoS	78.63	91.78	91.78	94.18	94.18	94.18	94.18	94.18	94.18	94.2
PV-StoL	1.09	2.35	4.07	7.02	8.77	13.38	35.93	36.62	38.66	38.66
Action-LtoS	6.36	33.18	50.1	66.26	72.07	73.12	73.12	73.12	73.27	73.3
Action-StoL	13.5	43.34	65.09	66.82	66.82	70.62	72.01	73.6	73.73	73.73
UniqWin	91.48	91.48	<b>93.19</b>	<b>93.37</b>	<b>94.15</b>	<b>95.06</b>	<b>95.15</b>	<b>95.15</b>	<b>95.35</b>	<b>95.35</b>
APS	69.77	82.18	84.47	84.47	90.14	90.14	90.14	90.14	90.14	90.14
MFPS	73.48	82.12	83.19	85.16	88.06	88.96	89.46	89.68	89.68	89.68
Weighted-Freq	58.14	81.92	81.92	83.53	84.86	89.19	89.19	89.19	89.19	89.19
Random	34.88	64.9	72.13	78.89	79.16	79.41	80.37	80.5	80.55	80.68
G-Best	<b>99.45</b>	<b>99.45</b>	<b>99.45</b>	<b>99.45</b>	<b>99.45</b>	<b>99.45</b>	<b>99.45</b>	<b>99.45</b>	<b>99.45</b>	<b>99.45</b>
G-Worst	0.17	0.17	0.17	0.17	0.17	0.17	0.17	0.17	3.58	6.23

TABLE 2: APFD for Spreadsheet (each increment of 10% of the test suite is 30 tests).

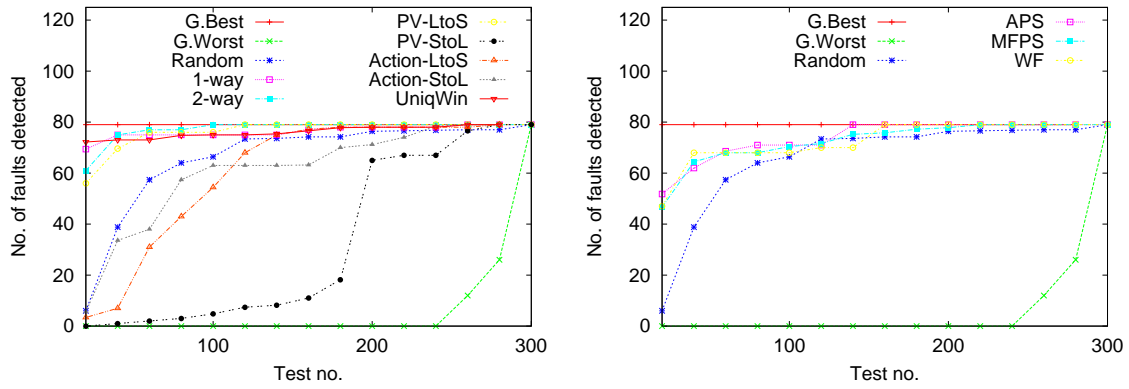


Fig. 3: Terpspreadsheet: rate of fault detection using prioritized test orderings.

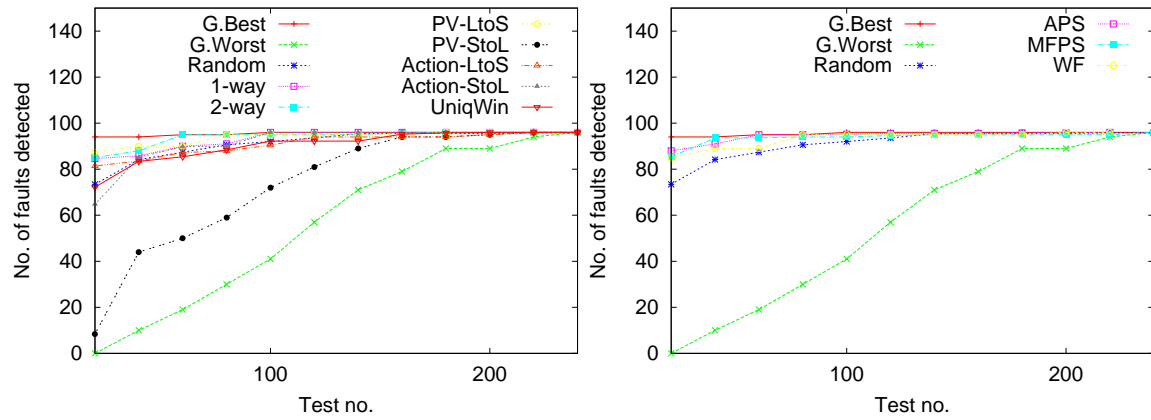


Fig. 4: TerpWord: Rate of fault detection using prioritized test orderings.

	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
<i>1-way</i>	68.56	77.49	79.89	80.54	84.24	85.25	86.41	86.83	86.92	86.92
<i>2-way</i>	83.87	90.05	92.19	92.56	93.39	94.18	94.18	94.18	94.18	94.18
<i>PV-LtoS</i>	<b>90.62</b>	<b>91.43</b>	<b>93.17</b>	<b>94.67</b>	<b>95.54</b>	<b>95.54</b>	<b>95.54</b>	<b>95.69</b>	<b>95.69</b>	<b>95.69</b>
<i>PV-StoL</i>	14.19	22.46	29.82	33.13	35.47	37.45	40.28	43.32	44.48	45.54
<i>Action-LtoS</i>	24.48	37.01	42	48.8	54.99	59.37	62.22	64.85	65.31	65.35
<i>Action-StoL</i>	34.42	73.74	84.96	85.76	86.64	86.75	86.75	86.75	86.75	86.8
<i>UniqWin</i>	57.25	65.98	75.79	78.61	82.04	83.29	83.97	84.43	84.48	84.48
<i>APS</i>	51.41	61.36	79.14	79.78	82.94	83.34	84.26	84.38	84.43	84.48
<i>MFPS</i>	79.35	86.97	88.25	88.59	90.08	90.68	91.27	91.37	91.48	91.82
<i>Weighted-Freq</i>	79.83	87.14	87.57	87.57	88.53	90.24	90.24	90.24	90.56	90.56
<i>Random</i>	31.64	48.84	59.52	65.61	69.8	72.31	74.32	74.89	75.13	75.18
<i>G-Best</i>	<b>95.18</b>	<b>98.60</b>	<b>98.60</b>	<b>98.60</b>	<b>98.60</b>	<b>98.60</b>	<b>98.60</b>	<b>98.60</b>	<b>98.60</b>	<b>98.60</b>
<i>G-Worst</i>	0.17	0.17	0.17	0.17	0.17	0.17	0.51	2.88	9.17	11.17

TABLE 3: APFD for Word (each increment of 10% of the test suite is 25 tests).

	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
<i>1-way</i>	<b>83.79</b>	87.78	91.54	94.79	94.79	94.79	94.79	94.79	94.99	94.99
<i>2-way</i>	83.72	<b>90.8</b>	<b>91.72</b>	<b>95.64</b>	<b>95.64</b>	<b>95.64</b>	<b>95.64</b>	<b>95.64</b>	<b>95.64</b>	<b>95.64</b>
<i>PV-LtoS</i>	83.53	88.77	88.77	92.71	92.71	94.26	94.26	94.26	94.26	94.26
<i>PV-StoL</i>	16.38	25.6	26.44	28.76	30.33	34.64	39.15	39.58	42.18	43.09
<i>Action-LtoS</i>	<b>82.62</b>	<b>88.05</b>	<b>90.36</b>	<b>90.81</b>	<b>93.05</b>	<b>94.2</b>	<b>94.2</b>	<b>94.2</b>	<b>94.2</b>	<b>94.2</b>
<i>Action-StoL</i>	19.01	28.19	30.74	31.81	35.82	38.68	41.48	44.53	45.9	46.51
<i>UniqWin</i>	31.4	36.47	49.13	55.66	60.49	63.12	64.9	66.85	66.98	67.28
<i>APS</i>	75.72	87.05	90.21	90.67	90.8	91.43	91.77	91.99	92.26	92.28
<i>MFPS</i>	65.6	66.8	68.21	68.61	69.62	76.17	77.37	79.48	79.48	79.48
<i>Weighted-Freq</i>	80.68	85.9	85.9	85.9	87.28	87.28	88.09	89.43	89.43	89.54
<i>Random</i>	48.63	57.55	64.51	69.19	73.03	75.37	77.37	78.24	78.45	78.49
<i>G-Best</i>	95.27	97.37	97.37	97.37	97.37	97.92	97.92	97.92	97.92	97.92
<i>G-Worst</i>	9.14	11.18	12.97	26.96	27.6	30.99	30.99	30.99	33.15	34.64

TABLE 4: APFD for CPM (each increment of 10% of the test suite is 89 tests).

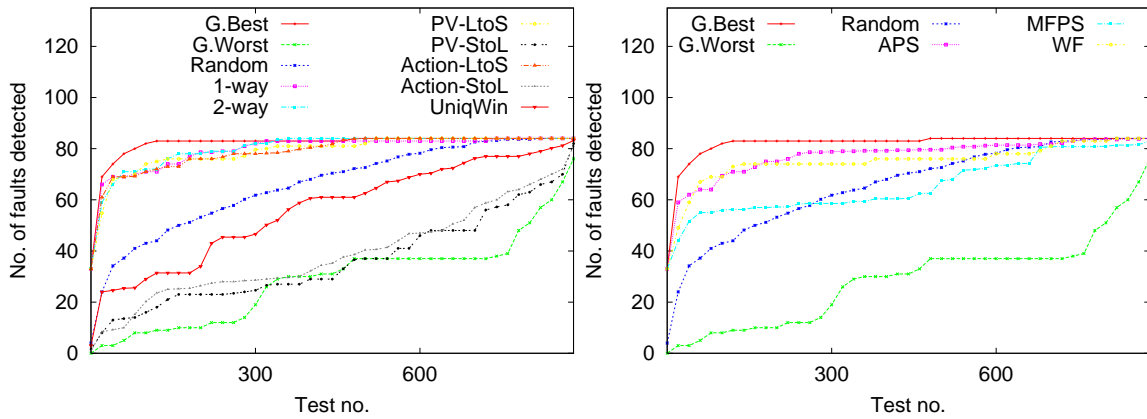


Fig. 5: CPM: rate of fault detection using prioritized test orderings

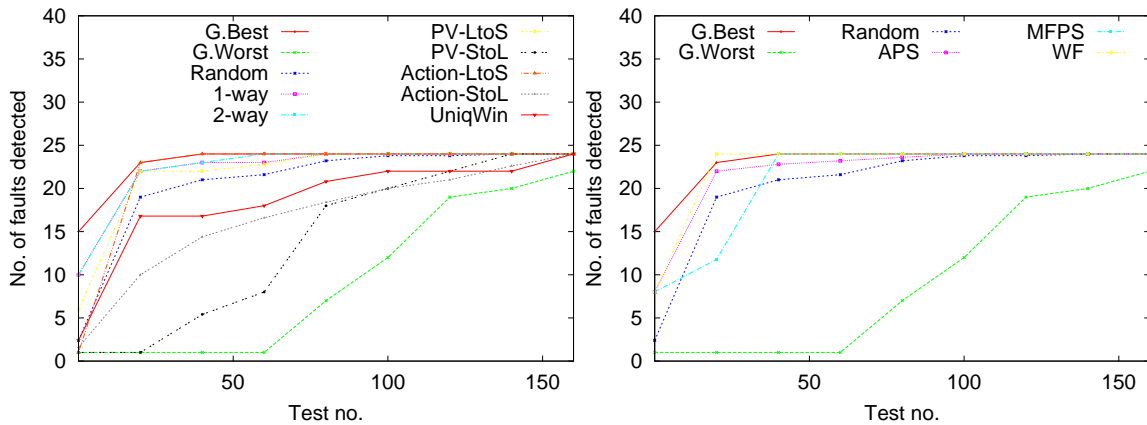


Fig. 6: MASPLAS: Rate of fault detection

	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
<i>1-way</i>	89.6	93.04	93.04	95.56	95.56	95.56	95.56	95.56	95.56	95.56
<i>2-way</i>	90.98	90.98	94.28	97.06	97.06	97.06	97.06	97.06	97.06	97.06
<i>PV-LtoS</i>	86.05	89.74	89.74	93.38	94.84	94.84	94.84	94.84	94.84	94.84
<i>PV-StoL</i>	4.44	4.44	26.61	30.08	50.16	53.91	57	58.1	58.85	58.85
<i>Action-LtoS</i>	94.35	97.87	97.87	97.87	97.87	97.87	97.87	97.87	97.87	97.87
<i>Action-StoL</i>	40.76	51.79	55.69	62.43	67.25	69.23	70.25	70.49	72.48	72.48
<i>UniqWin</i>	68.65	68.65	68.65	75.53	80.38	81.17	81.17	81.17	81.77	82.02
<i>APS</i>	90.6	92.71	93.34	94.44	95.8	96.21	96.21	96.21	96.21	96.21
<i>MFPS</i>	44.63	91.8	91.8	91.8	91.8	91.8	91.8	91.8	91.8	91.8
<i>Weighted-Freq</i>	<b>98.4</b>	<b>98.4</b>	<b>98.4</b>	<b>98.4</b>	<b>98.4</b>	<b>98.4</b>	<b>98.4</b>	<b>98.4</b>	<b>98.4</b>	<b>98.4</b>
<i>Random</i>	76.33	80.51	85.57	87.59	89.91	90.69	90.69	90.91	90.91	90.91
<i>G-Best</i>	99.5	100	100	100	100	100	100	100	100	100
<i>G-Worst</i>	87.15	87.15	87.15	89.22	89.81	90.85	92.06	92.18	92.18	92.28

TABLE 5: APFD for Masplas (each increment of 10% of the test suite is 17 tests).

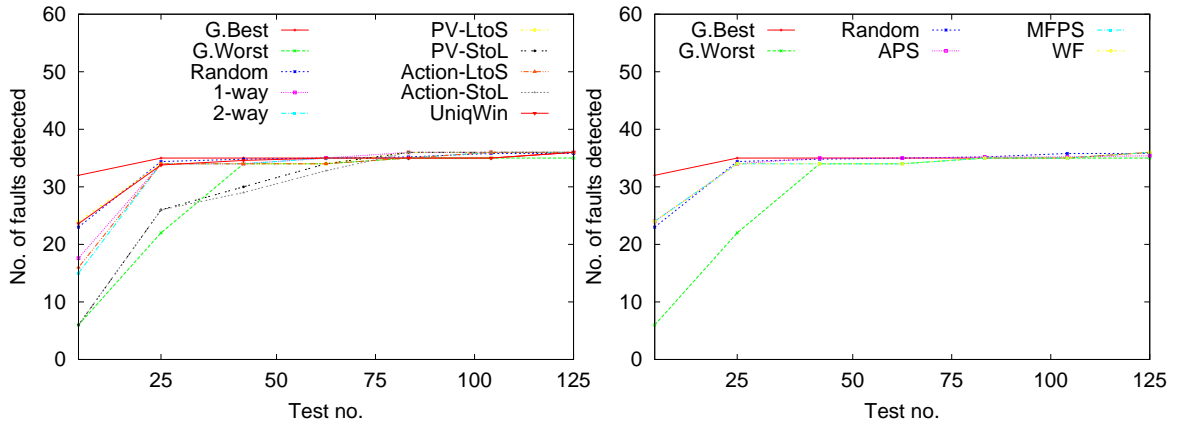


Fig. 7: Book: Rate of fault detection

Prio. Tech.	Calc (300)	Paint (300)	SSheet (300)	Word (250)	Book (125)	CPM (890)	Masplas (169)
<i>1-way</i>	272	253	258	<b>94</b>	72	740	68
<i>2-way</i>	283	162	<b>91</b>	168	83	<b>346</b>	58
<i>PV-LtoS</i>	274	226	114	170	76	518	72
<i>PV-StoL</i>	294	299	262	210	68	890	139
<i>Action-LtoS</i>	294	284	266	218	96	492	30
<i>Action-StoL</i>	297	<b>151</b>	257	100	<b>65</b>	890	148
<i>UniqWin</i>	287	267	253	183	111	889	160
<i>APS</i>	271	298	135	214	124	867	87
<i>MFPS</i>	254	267	217	224	118	692	31
<i>Weighted-Freq</i>	<b>217</b>	267	155	195	115	804	<b>15</b>
<i>Random</i>	292	288	286	214	124	814	125
<i>G-Best</i>	246	34	21	100	19	480	27
<i>G-Worst</i>	300	300	300	232	124	890	164

TABLE 6: Number of tests for 100% fault detection.